MORTALITY FORECASTING WITH A SPATIALLY PENALIZED SMOOTHED VAR MODEL*

BY

LE CHANG AND YANLIN SHI

ABSTRACT

This paper investigates a high-dimensional vector-autoregressive (VAR) model in mortality modeling and forecasting. We propose an extension of the sparse VAR (SVAR) model fitted on the log-mortality improvements, which we name "spatially penalized smoothed VAR" (SSVAR). By adaptively penalizing the coefficients based on the distances between ages, SSVAR not only allows a flexible data-driven sparsity structure of the coefficient matrix but simultaneously ensures interpretable coefficients including cohort effects. Moreover, by incorporating the smoothness penalties, divergence in forecast mortality rates of neighboring ages is largely reduced, compared with the existing SVAR model. A novel estimation approach that uses the accelerated proximal gradient algorithm is proposed to solve SSVAR efficiently. Similarly, we propose estimating the precision matrix of the residuals using a spatially penalized graphical Lasso to further study the dependency structure of the residuals. Using the UK and France population data, we demonstrate that the SSVAR model consistently outperforms the famous Lee–Carter, Hyndman–Ullah, and two VAR-type models in forecasting accuracy. Finally, we discuss the extension of the SSVAR model to multi-population mortality forecasting with an illustrative example that demonstrates its superiority in forecasting over existing approaches.

KEYWORDS

Mortality forecasting, high-dimensional vector-autoregressive model, cohort effect, accelerated proximal gradient, multi-population modeling.

JEL codes: C18, C32, C52, C53.

*This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Astin Bulletin 51(1), 161-189. doi:10.1017/asb.2020.39 © 2020 by Astin Bulletin. All rights reserved.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

1. INTRODUCTION

Mortality modeling and forecasting have become standard assessment tools to measure mortality and longevity risks. Among the 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 and reviews of the LC model have been extensively discussed in the existing literature – for example, see Booth *et al.* (2006), Renshaw and Haberman (2006), Barrieu *et al.* (2012), and Hunt and Blake (2014). Among them, using the functional principal component analysis, the Hyndman–Ullah (HU) model proposed by Hyndman and Ullah (2007) demonstrates outstanding forecasting performance. In contrast, another important stream focuses on applying the high-dimensional vector-autoregressive (VAR) models with meaningful and effective dimensionality reduction techniques (see, e.g., Yang and Wang, 2013; Zhou *et al.*, 2014; Li and Lu, 2017), 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 of all VAR-type models for mortality modeling and forecasting. In life insurance practice, the issue of $p >> n$ is widely acknowledged. That is, the number of variables/parameters (e.g., age) is usually greater than the sample size (e.g., year), so that there are insufficient data to fit a full VAR-type model. Hence, the first concern is an appropriate dimension-reduction approach. A recent paper of Li and Lu (2017) addressed this by focusing on the cohort effects[1] and adopting a sparse VAR (SVAR) model with restrictive Granger causality matrix (spatial-temporal VAR (STAR) model, such that only mortality rates for neighboring ages could interact. In contrast, Guibert *et al.* (2019) adopted a pure data-driven approach and employed the SVAR model with an elastic-net (ENET) penalty estimation method. Second, VAR-type models need to be stationary to produce meaningful estimates and forecasts. Considering that the age-wide mortality rates are nonstationary time series, Li and Lu (2017) considered a coherent structure to impose a powerful constraint in the Granger causality matrix and impose smoothness penalties. Hence, the forecast mortality rates will be smoothed and those of neighboring age groups will not diverge. In contrast, Guibert *et al.* (2019) resolved this issue by working on the time-differenced log-mortality rates (mortality improvements/evolutions), which are inherently stationary.

Despite the effectiveness of both approaches, some new concerns remain unsolved. STAR considers the sparsity of Granger causality matrix in a relatively ad hoc manner, in which only lagged mortality rates of ages $x-2$, $x-1$, and x can affect contemporary rates of age x. However, assuming only significant effects between neighboring cohorts can be too restrictive because information from other surrounding age groups can further contribute to the mortality forecasting in many situations. In contrast, SVAR employs a data-driven approach, which allows a more flexible coefficient

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

(Granger causality) matrix than that considered in STAR. However, the price of such flexibility is the inevitable divergence in forecast mortality rates and the questionable interpretation of coefficients. For example, unlike STAR, the SVAR model does not impose any smooth constraints. Consequently, forecast rates are very likely to diverge in SVAR, even for those of neighboring ages. 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 45 years of age were significantly influenced by the lagged evolution of 95 years of age, which is counterintuitive. As a result, the desirable cohort effects could be overwhelmed and not explainable in the SVAR model.

To address those concerns, we propose a novel spatially penalized smoothed VAR model (SSVAR), which may nest STAR and SVAR model as special cases. Similar to Guibert *et al.* (2019), we focus on mortality improvements to resolve the stationary issue. The coefficient matrix is also determined by a data-driven approach. Distinct from the SVAR model, we employ smoothness penalties, as considered in Li and Lu (2017), to produce smooth age-dependent parameters. More importantly, we use the adaptive weights determined by the spatial distances between age groups to penalize coefficients differently in the sparsity penalty. By including the adaptive weights, the coefficients for mortality improvements of remote age groups are more likely to shrink to zero. This approach employs the data-driven method to estimate the coefficient matrix, yet retains the benefits of the STAR model to consider meaningful cohort effects. That is, our model can appropriately balance the data-driven approach in SVAR and the demand of cohort effects in STAR. Moreover, smoothness penalties largely reduce the divergence in forecast mortality rates of neighboring age groups. Finally, our SSVAR model can be straightforwardly extended to model multi-population cases, as with STAR and SVAR.

We provide empirical evidence of male and female mortality data for the UK and France, sourced from Human Mortality Database (2019). Using the crude rates ranging from 1950 to 2016, we systematically compare the forecasting performance of the LC, HU, SVAR, STAR, and SSVAR models on the age groups from 0 to 100 years. As measured by the root of mean squared forecast error (RMSFE), the SSVAR model consistently beats other competing models up to the 16-step-ahead forecasting horizon for all four datasets. We then conduct long-term forecasting analysis to compare the life expectancies calculated by different models up to 2050. Finally, when the UK and France males are jointly modeled, SSVAR still outperforms those multi-population extensions of LC, HU, SVAR, and STAR models in most cases. The same conclusions also hold for the multi-population forecasting of females.

The major contributions of this paper are twofold. First, the proposed SSVAR model significantly complements the recent studies of Li and Lu (2017) and Guibert *et al.* (2019). Our data-driven approach is more flexible than the

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

STAR model and successfully retains its advantages in avoiding divergence in forecast and realizing interpretable coefficients. Second, we provide technical details of minimizing the objective function of SSVAR, which incorporates the nonstandard smoothing ($L2$) penalty and the sparsity ($L1$) penalty with adaptive weights. To solve SSVAR efficiently, we further derive a novel estimation approach that uses the accelerated proximal gradient (APG) algorithm, which is compatible for other VAR-type models. In relation to this, we investigate the precision matrix (inverse of the variance–covariance matrix) of the residuals to capture the contemporary dependency between mortality improvements for different ages. A spatially penalized graphical Lasso is proposed to estimate this sparse precision matrix. Additionally, for the proposed SSVAR model, we demonstrate its compatibility with multi-population mortality forecasting, and superiority in forecasting accuracy when compared with LC, HU, and existing VAR-type models.

The rest of this paper is organized as follows. In Section 2, we propose the SSVAR model and its estimation algorithm with technical details and further discuss the dependency structure of the residuals. In Section 3, we conduct empirical studies with the single population case, while, in Section 4, we discuss the multi-population extension. Finally, Section 5 concludes this paper.

2. METHODS

2.1. The vector autoregression model

Throughout this study, we are interested in the log-mortality improvement rates $y_{i,t}$ such that:

$$y_{i,t} = \ln(m_{i,t}) - \ln(m_{i,t-1}), \tag{2.1}$$

where $m_{i,t}$ is the crude mortality rate at age i in year t. $t = 1, \ldots, T$. $i = 1, \ldots, N$. T and N define the number of years and number of ages under consideration, respectively. In general, the log-mortality rates are considered not stationary because they exhibit a clear decreasing trend over time. To remove the linear trend in $\ln(m_{i,t})$, we model the series $y_{i,t}$, the mortality improvement rates instead.

To capture both the temporal dependence and the cohort effect of the mortality improvements, many authors (Lazar and Denuit, 2009; Salhi and Loisel, 2017; Li and Lu, 2017; Guibert *et al.*, 2019) have suggested VAR-structured models. For the sake of interpretability, we consider a lag-1 VAR model as in Li and Lu (2017), abbreviated as VAR(1), which only regresses the log-mortality improvements at time t on those measured at time $t-1$.[2] VAR with more lags can be easily extended based on VAR(1). If we further define $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{N,t})'$, an $N \times 1$ vector of the log-mortality improvements at time t, a VAR(1) model can be written as:

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \tag{2.2}$$

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

where $\boldsymbol{\alpha}$ is an N-dimensional intercept vector, $\boldsymbol{B}$ is an $N \times N$ Granger causality matrix that evaluates the long-run co-movement of the mortality improvements, and $\boldsymbol{\epsilon}_t$ is an $N \times 1$ vector of error terms. In our study, we assume a Gaussian distribution of error, such that $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ defines the variance–covariance matrix of the error.

One great benefit of using VAR-type models in mortality modeling is that the Granger causality matrix, $\boldsymbol{B}$, can provide direct interpretations of the period and cohort effects. For example, as discussed in Li and Lu (2017) and Guibert *et al.* (2019), the diagonals of $\boldsymbol{B}$ represent the period effect that captures the relationship between the log-mortality improvements in the current and previous years. For the cohort effects, we focus on those between surrounding cohorts (i.e., age groups that are spatially close to each other), which are measured by entries around the diagonals of $\boldsymbol{B}$. This nests the effect investigated in Li and Lu (2017), which strictly refers to the cohort effects of individuals born in the same and neighboring years, as measured by sub-diagonals of $\boldsymbol{B}$.

However, when the dimension of age N is large or even moderate, the number of parameters in a VAR model is enormous, leading to very unstable estimations and problematic interpretations of the cohort effects. If the target is to forecast the mortality for ages from 0 to 100 (e.g., $N = 101$), the number of parameters in the Granger causality matrix $\boldsymbol{B}$ reaches 101^2, which is much larger than the sample size that we can generally obtain, and thus leads to an infeasible ordinary least squares estimation. Instead of allowing the relationships among all ages to exist, SVAR that shrinks many entries in $\boldsymbol{B}$ to zero can be considered. Over the last decade, SVAR has been widely discussed in the statistics literature (Song and Bickel, 2011; Fan *et al.*, 2011; Basu *et al.*, 2015; Davis *et al.*, 2016). One most commonly used SVAR model is achieved by adding a Lasso penalty ($L1$) to force some of the coefficients in the Granger causality matrix $\boldsymbol{B}$ to be exactly zero. Other penalties, such as ENET (Zou and Hastie, 2005) which combines ridge and Lasso penalties, have also been considered in the context of VAR models (Basu *et al.*, 2015).

2.2. Spatially penalized smoothed VAR

To consider the period and cohort effects in mortality modeling, Li and Lu (2017) proposed a STAR model by imposing an ad hoc sparse structure on the Granger causality matrix. More specifically, only mortality rates of ages right next to each other can interact, which can be too restrictive. Additionally, this assumption can be unrealistic for many populations, where the cohort effects do not exist for some of the cohorts. In contrast, Guibert *et al.* (2019) used a SVAR model with the ENET penalty to model the log-mortality improvements with a pure data-driven approach. In other words, zero entries in coefficient matrix $\boldsymbol{B}$ are completely determined by the data. However, the resulting coefficient matrix can sometimes be unexplainable. For instance, the empirical analysis in Guibert *et al.* (2019) indicated that the mortality improvements for remote ages could still influence each other (e.g., age 45 years and age 95 years).

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

In addition, the original SVAR model did not account for the smoothness penalty, which could result in divergence in forecast rates. Detailed definition and discussions on the non-divergence and how this is asymptotically ensured by imposing the smoothing penalties can be found in Appendix A. As demonstrated in our empirical analysis, using the smoothing penalties, together with the imposed adaptive weights, also help to reduce unexplainable results of SVAR, even with limited sample sizes.

To address those issues, we propose a smoothed version of SVAR with adaptive weights, namely, SSVAR, to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{B}$ in model (2.2), such that:

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{B}} \frac{1}{2}\sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1} \right\|_2^2 + \lambda \sum_{i,j}^{N} w_{ij}|\beta_{ij}| + \eta_1 \left\| \boldsymbol{\alpha}^T \boldsymbol{H} \right\|_2^2 + \eta_2 \left\| \boldsymbol{B}^T \boldsymbol{H} \right\|_F^2, \tag{2.3}$$

where $\|\cdot\|_2$ and $\|\cdot\|_F$ define $L2$ norm for a vector and Frobenius norm for a matrix, respectively; β_{ij} denotes the ijth entry of the Granger causality matrix $\boldsymbol{B}$; λ is the tuning parameter that aims to shrink some of β_{ij} to zero and thus controls the sparsity of $\boldsymbol{B}$; and η_1 and η_2 are the tuning parameters that control the smoothness of the intercept $\boldsymbol{\alpha}$ and the Granger causality matrix $\boldsymbol{B}$, respectively.

One of the major merits of our proposed model is that the adaptive weights, w_{ij}, $i,j = 1,\ldots,N$, are added to penalize β_{ij} differently in the $L1$ penalty term. The choice of weights w_{ij} can be specified by the users. In the mortality modeling context, it is widely believed that the mortality improvements of age i are more likely to propagate to those of ages close to i, rather than the ages far away from i. That is, for larger $|i-j|$, β_{ij} should be more penalized and thus more likely to shrink toward zero. Therefore, in our study, we consider w_{ij} an exponential function[3] of distance between two ages, i and j, such that:

$$w_{ij} = \exp\left\{\frac{|i-j|}{\theta}\right\}, \tag{2.4}$$

where $|i-j|$ measures absolute distance between two ages, i and j, and θ is another tuning parameter that controls the sensitivity of shrinkage to the distance between i and j. In contrast to Li and Lu (2017) and Guibert *et al.* (2019), our method only penalizes β_{ij} with a larger $|i-j|$ more strongly, rather than imposing a predetermined or purely data-driven sparsity structure of $\boldsymbol{B}$. In other words, some β_{ij} with moderately small $|i-j|$ are still likely to be nonzero, so that important effects between surrounding cohorts can be discovered to improve the forecasting accuracy. For those with large $|i-j|$, the β_{ij} are very prone to shrink to zero to avoid the uninterpretable estimates.

Another important feature of our model is that the $L2$ type of penalties for $\boldsymbol{\alpha}$ and $\boldsymbol{B}$ are used to provide smooth curves of these age-dependent parameters. This aims to effectively reduce the divergence in forecast rates of neighboring

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

ages, as considered in Li and Lu (2017). The $N \times N$ penalty matrix $\boldsymbol{H}$ in (2.3) is defined as:

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \ddots & \vdots & \vdots \\ 0 & -1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -1 & 0 \end{bmatrix}. \tag{2.5}$$

That is, we penalize the squared sum of the differences between adjacent entries in rows of $\boldsymbol{\alpha}^T$ and $\boldsymbol{B}^T$, so that abnormal changes of these parameters across ages can be smoothed out. In addition, the smoothing parameters η_1 and η_2 can be set differently to allow different amounts of smoothness for the intercept $\boldsymbol{\alpha}$ and the Granger causality matrix $\boldsymbol{B}$. Larger values of η_1 and η_2 produce smoother curves of $\boldsymbol{\alpha}$ and $\boldsymbol{B}$.

It is worth noting that the structure of the Granger causality matrix in our proposed model can be viewed as a generalization of those in STAR and SVAR and thus may achieve the best of both worlds. When $w_{ij} \equiv 1, \forall i, j = 1, \ldots, N$, $\eta_1 = 0$, and $\boldsymbol{H} = \text{diag}(N)$, our model reduces to SVAR using the ENET penalty, which combines the $L1$ and $L2$ penalty of $\boldsymbol{B}$. When $w_{ij} \equiv 1, \forall i, j = 1, \ldots, N$, and $\eta_1 = \eta_2 = 0$, our model simply reduces to SVAR using the Lasso penalty only. When $w_{ij} = 0$ if $i = j, j+1, j+2$, and $w_{ij} = \infty$ otherwise, our model forces β_{ij} to be exactly zero for $i \neq j, j+1, j+2$ such that:

$$\boldsymbol{B} = \begin{bmatrix} \beta_{11} & 0 & 0 & \cdots & \cdots \\ \beta_{21} & \beta_{22} & 0 & \cdots & \cdots \\ \beta_{31} & \beta_{32} & \beta_{33} & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \cdots & 0 & \beta_{N(N-2)} & \beta_{N(N-1)} & \beta_{NN} \end{bmatrix},$$

which coincides with the sparsity structure of the Granger causality matrix in Li and Lu (2017). Under this setting of w_{ij}, our model approximates to STAR in Li and Lu (2017), while the difference is that we model the log-mortality improvement rates, rather than the log-mortality rates themselves in STAR. Therefore, by choosing the adaptive weights w_{ij} defined in (2.4), our proposed method, SSVAR, considers the common belief implied by the cohort effects that mortality improvements of closer ages are more likely to be related. Its sparsity structure of the coefficient matrix (e.g., definition of closeness) is still determined by the data, rather than being ad hoc.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

2.3. Algorithms for numerical optimization

Adding the adaptive weights and smoothing penalties poses a great challenge in estimating the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{B}$, as in (2.3). Standard Lasso/ENET algorithm (glmnet) for SVAR or penalized least squares for STAR cannot be directly applied to solve our proposed SSVAR. In this section, we discuss the technical details of our algorithm to estimate the intercept vector $\boldsymbol{\alpha}$ and the Granger causality matrix $\boldsymbol{B}$ for SSVAR. Recall the objective function of SSVAR:

$$Q(\boldsymbol{\alpha}, \boldsymbol{B}) = \frac{1}{2}\sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1}\|_2^2 + \lambda \sum_{i,j}^{N} w_{ij}|\beta_{ij}| + \eta_1 \|\boldsymbol{\alpha}^T \boldsymbol{H}\|_2^2 + \eta_2 \|\boldsymbol{B}^T \boldsymbol{H}\|_F^2. \tag{2.6}$$

Unlike the traditional SVAR with a Lasso/ENET penalty (e.g., $w_{ij} \equiv 1, i, j = 1, \ldots, N$ and $\eta_1 = \eta_2 = 0$), the intercept $\boldsymbol{\alpha}$ in (2.6) cannot simply be removed by replacing the data with their centered versions because $\boldsymbol{\alpha}$ also appears in the penalty term. However, because the penalty term of $\boldsymbol{\alpha}$ is quadratic, we can still obtain a closed-form solution for $\boldsymbol{\alpha}$ if $\boldsymbol{B}$ is given. To minimize $Q(\boldsymbol{\alpha}, \boldsymbol{B})$ with respect to $\boldsymbol{\alpha}$, we consider the partial derivative:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} Q(\boldsymbol{\alpha}, \boldsymbol{B}) = T\boldsymbol{\alpha} - \sum_{t=1}^{T} \left(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}\right) + 2\eta_1 \boldsymbol{H}\boldsymbol{H}^T \boldsymbol{\alpha}. \tag{2.7}$$

Setting $\frac{\partial}{\partial \boldsymbol{\alpha}} Q(\boldsymbol{\alpha}, \boldsymbol{B}) = 0$ yields

$$\boldsymbol{\alpha} = \left(T \cdot \boldsymbol{I}_N + 2\eta_1 \boldsymbol{H}\boldsymbol{H}^T\right)^{-1} \sum_{t=1}^{T} \left(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}\right), \tag{2.8}$$

where $\boldsymbol{I}_N$ defines an $N \times N$ identity matrix.

It is worth mentioning that minimizing $Q(\boldsymbol{\alpha}, \boldsymbol{B})$ with respect to $\boldsymbol{B}$ is a non-smooth optimization problem because the $L1$ penalty of $\boldsymbol{B}$ is imposed. In this study, we apply an APG algorithm (Daubechies *et al.*, 2004; Beck and Teboulle, 2009) to estimate $\boldsymbol{B}$.

Consider the following minimization problem:

$$\underset{x}{\operatorname{argmin}}\{F(x) := f(x) + g(x)\}, \tag{2.9}$$

where $f : R^n \to R$ is a differentiable convex function and $g : R^n \to (-\infty, \infty]$ is a proper, lower semi-continuous convex function. Minimizing $Q(\boldsymbol{\alpha}, \boldsymbol{B})$ with respect to $\boldsymbol{B}$ is exactly the problem in (2.9), with:

$$f(\boldsymbol{B}) = \frac{1}{2}\sum_{t=1}^{T} \left\|\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1}\right\|_2^2 + \eta_2 \left\|\boldsymbol{B}^T \boldsymbol{H}\right\|_F^2 \quad \text{and} \quad g(\boldsymbol{B}) = \lambda \sum_{i,j}^{N} w_{ij}|\beta_{ij}|, \tag{2.10}$$

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

such that the differentiable and non-differentiable parts are separated. To apply the APG method, the gradient of $f(\boldsymbol{B})$ is required, which is given by:

$$\nabla f(\boldsymbol{B}) = -\sum_{t=1}^{T} \left(\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1}\right) \boldsymbol{y}_{t-1}^T + 2\eta_2 \boldsymbol{B}^T \boldsymbol{H}\boldsymbol{H}^T, \tag{2.11}$$

It can then be shown that the proximal operator of function $g(\boldsymbol{B})$ with a step size s is exactly the soft thresholding function used by Donoho and Johnstone (1994):

$$\text{prox}_{sg}(\boldsymbol{B}) = S_{s\lambda_{ij}}(\boldsymbol{B}) = \begin{cases} \beta_{ij} - s\lambda_{ij} & \text{if } \beta_{ij} > s\lambda_{ij} \\ 0 & \text{if } |\beta_{ij}| \le s\lambda_{ij} \\ \beta_{ij} + s\lambda_{ij} & \text{if } \beta_{ij} < -s\lambda_{ij} \end{cases}, \quad i,j = 1,\ldots,N, \tag{2.12}$$

where $\lambda_{ij} = \lambda w_{ij}$, which controls the amount of shrinkage for β_{ij}, $i,j = 1,\ldots,N$. Therefore, according to Beck and Teboulle (2009), the APG method to solve $\boldsymbol{B}$ in our model leads to the following iterative method:

$$\begin{aligned} \widetilde{\boldsymbol{B}}_k &= \boldsymbol{B}_k + \frac{k-1}{k+2}(\boldsymbol{B}_k - \boldsymbol{B}_{k-1}) \\ \boldsymbol{B}_{k+1} &= S_{s_k\lambda_{ij}}\left(\widetilde{\boldsymbol{B}}_k - s_k \nabla f(\boldsymbol{B}_k)\right), \end{aligned} \tag{2.13}$$

where $\boldsymbol{B}_k$ denotes the iterated $\boldsymbol{B}$ in the kth iteration step; s_k is a suitably chosen kth step size, such that $0 < s_k < 1/L$, where L is the Lipschitz constant of ∇f; and $\nabla f(\boldsymbol{B}_k)$ is the gradient of $f(\boldsymbol{B})$ evaluated at $\boldsymbol{B}_k$ and $\boldsymbol{\alpha}$, as given in (2.8). When solving the standard Lasso problem, the APG method is also known as the fast iterative soft thresholding algorithm. Based on Beck and Teboulle (2009), let $\boldsymbol{B}^*$ be any minimizer of $F(\boldsymbol{B})$, then:

$$0 \le F(\boldsymbol{B}_k) - F(\boldsymbol{B}^*) \le \frac{\|\boldsymbol{B}_0 - \boldsymbol{B}^*\|_F^2}{s(k+1)^2} \quad \text{for} \quad k \ge 1. \tag{2.14}$$

where $\boldsymbol{B}_0$ is the starting value of the iteration for $\boldsymbol{B}$. That is, the APG method achieves a global rate of convergence of $O(1/k^2)$, which is much faster than other classical gradient methods. Details of the proof of the convergence rate can be found in Beck and Teboulle (2009).

Once the iteration in (2.13) converges, the final iteration $\boldsymbol{B}_k$ is served as the estimator $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\alpha}}$ is obtained by assigning $\boldsymbol{B} = \widehat{\boldsymbol{B}}$ in (2.8). In the empirical application to the UK and France mortality data, our proposed algorithm always converges within 100 iterations, which is a common choice of the maximum number of iteration in the APG method.

In addition, because the traditional SVAR or other constrained VAR models can be viewed as special cases of our model, as mentioned previously, our

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

proposed numerical algorithm for SSVAR can also be used to solve other VAR-type models very efficiently. R programming codes for this algorithm are available upon request.

2.4. Tuning parameter selection

For a given pair of tuning parameters θ,[4] λ, η_1, and η_2, our proposed model can be efficiently solved by the algorithm introduced in Section 2.3. It is critical to select appropriate tuning parameters λ, η_1, and η_2, as they control the sparsity and smoothness of the age-dependent parameters, respectively. One typical method to select tuning parameters is to optimize cross-validated forecasting accuracy. Following Song and Bickel (2011), we use an h-step-ahead RMSFE to measure the out-of-sample forecast accuracy, which is computed as:

$$\mathrm{RMSFE}_h(\lambda, \eta_1, \eta_2) = \sqrt{\frac{1}{N(T_1 - T_0 - h)} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-h} \left(\widehat{y}_{i,t+h}^{(\lambda,\eta_1,\eta_2)} - y_{i,t+h} \right)^2}, \quad (2.15)$$

where $[T_0, T_1]$ is the evaluation period, which is given by the last third of the data in our study; $\widehat{y}_{i,t+h}^{(\lambda,\eta_1,\eta_2)}$ denotes the h-step-ahead forecast for ith age based on the observations up to time t; and $y_{i,t+h}$ defines the observed log-mortality improvements at time $t+h$ for ith age. We then choose a grid of values of λ, η_1, and η_2 and search the pair of $(\lambda, \eta_1, \eta_2)$ that minimizes the RMSFE_h.

2.5. Dependency structure of the residuals

Alongside estimating $\boldsymbol{B}$, analyzing the dependency structure of the residuals is also very important in this study. Such error structures can capture the contemporaneous interactions between the mortality improvements for different ages. Many studies have investigated the dependence among residuals in mortality modeling. Biffis and Millossovich (2006), Debón *et al.* (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., $N = 101$). Therefore, Li and Lu (2017) proposed a constrained spatial autoregressive model to model the residual for ith age $\epsilon_{i,t}$ by its neighboring terms $\epsilon_{i-1,t}$ and $\epsilon_{i+1,t}$. This specification significantly reduces the number of parameters from $\frac{N(N+1)}{2}$ to $2(N-1)$. However, similar to estimating $\boldsymbol{B}$, only assuming the dependency between residuals from neighboring ages can be too restrictive.

In the same spirit of SSVAR, we introduce a more data-driven approach that deals with the curse of dimensionality concurrently to study the dependency structure of residuals. More specifically, we propose adding adaptive

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

weights calculated based on the age distance $|i-j|$ to the graphical Lasso (Friedman *et al.*, 2008), namely, the spatial penalized graphical Lasso[5]:

$$\widehat{\boldsymbol{K}} = \arg\min_{\boldsymbol{K}} \log\det(\boldsymbol{K}) - \mathrm{tr}(\boldsymbol{S}\boldsymbol{K}) - \rho \sum_{i,j}^{N} w_{ij}|K_{ij}|, \tag{2.16}$$

where $\boldsymbol{K}$ defines the precision matrix (the inverse of the covariance matrix), such that $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$; det and tr represent the determinant and trace of a matrix, respectively; $\boldsymbol{S}$ defines the sample covariance of the residuals, which can be calculated as $\boldsymbol{S} = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y}_t - \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{B}}\boldsymbol{y}_{t-1})(\boldsymbol{y}_t - \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{B}}\boldsymbol{y}_{t-1})^T$ in our model; ρ is a tuning parameter that controls the sparsity of $\boldsymbol{K}$; w_{ij} is the adaptive weights, as defined in (2.4); and K_{ij} defines the (i,j)th entry of the precision matrix $\boldsymbol{K}$.

In contrast to the existing literature, we focus on estimating the precision matrix K, rather than the covariance matrix $\boldsymbol{\Sigma}$, because the precision matrix can be standardized to encode the partial correlation or partial contemporaneous correlation (PCC) of two variables. More specifically, PCC is defined as a measure of the strength and direction of a causal relationship between two variables given that all other variables are known, which can be computed through the precision matrix:

$$\mathrm{PCC}(\epsilon_{i,t}, \epsilon_{j,t}) = -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}. \tag{2.17}$$

Therefore, a zero K_{ij} means that the residual for age i is conditionally independent with the one for age j contemporaneously. As noted in (2.16), the adaptive weights, w_{ij}, are added to penalize K_{ij} differently, such that the residuals for two ages that are more distant should be more likely to be independent.

3. Empirical application

In this paper, we use mortality data for the UK and France obtained from the Human Mortality Database (2019). Following Booth *et al.* (2006), we choose an opportune range of data, from 1950 to 2016, to have a reliable and complete dataset.[6] The crude male and female mortality rates are studied separately.

To illustrate the effectiveness of our proposed model, we fit the data from 1950 to 2000 and perform 16-step-ahead forecasting, as considered in Li and Lu (2017). The results of the LC, HU, SVAR, STAR, and SSVAR are compared. The out-of-sample RMSFE is the criterion for forecasting performance comparison.

3.1. Results of forecasting performance

To comprehensively compare the forecasting performance across all models, we consider both RMSFEs over age groups and time horizons only and an

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

TABLE 1

SUMMARY OF RMSFE OVER AGE GROUPS FOR THE FORECAST MORTALITY DATA. THE BOLD NUMBER IS THE MINIMUM RMSFE AMONG DIFFERENT METHODS.

	Male					Female				
Models	All	Mean	*SD*	Q_1	Q_3	All	Mean	*SD*	Q_1	Q_3
Panel A: UK										
LC	0.1953	0.1743*	0.0884	0.1062	0.2387	0.1696	0.1486*	0.0822	0.0745	0.2034
HU	0.1998	0.1722*	0.1019	0.0897	0.2199	0.1477	0.1277*	0.0747	0.0534	0.1772
SVAR	0.1719	0.1498*	0.0847	0.0829	0.2145	0.1303	0.1091	0.0715	0.0552	0.1352
STAR	0.1714	0.1499*	0.0836	0.0749	0.2159	0.1311	0.1120	**0.0685**	0.0613	0.1433
SSVAR	**0.1570**	**0.1381**	**0.0751**	**0.0699**	**0.1933**	**0.1285**	**0.1062**	0.0726	**0.0490**	**0.1299**
Panel B: France										
LC	0.2501	0.2008*	0.1499	0.0927	0.2813	0.1866	0.1558*	0.1032	0.0744	0.1882
HU	0.2396	0.1933*	0.1423	0.0800	0.2819	0.1752	0.1483*	0.0937	0.0572	0.2106
SVAR	0.1800	0.1564*	0.0897	0.0857	0.2037	0.1613	0.1339*	0.0905	0.0604	0.1977
STAR	0.1902	0.1649*	0.0953	0.0863	0.2060	0.1635	0.1342*	0.0939	0.0630	0.1687
SSVAR	**0.1702**	**0.1475**	**0.0852**	**0.0739**	**0.1888**	**0.1471**	**0.1232**	**0.0808**	**0.0562**	**0.1681**

overall measure, as follows:

$$RMSFE_x = \sqrt{\frac{1}{16}\sum_{h=1}^{16}(\ln m_{x,T+h} - \ln \hat{m}_{x,T+h})^2}$$

$$RMSFE_h = \sqrt{\frac{1}{101}\sum_{x=0}^{100}(\ln m_{x,T+h} - \ln \hat{m}_{x,T+h})^2} \tag{3.1}$$

$$RMSFE_{all,h} = \sqrt{\frac{1}{101 \times h}\sum_{i=1}^{h}\sum_{x=0}^{100}(\ln m_{x,T+i} - \ln \hat{m}_{x,T+i})^2}$$

$RMSFE_x$ ($RMSFE_h$) is the RMSFE averaged over all 16 forecasting steps (101 age groups) for age group x (time horizon h). $RMSFE_{all,h}$ is the overall measure considering both dimensions up to step h. We report UK and France results for all the five models in Tables 1 and 2, as well as in Figures 1 and 2.

Figure 1(a) displays the $RMSFE_x$. For UK males, except for a few groups among 25 to 30, the VAR-type models produce smaller RMSFE than LC and HU for almost all cases. In particular, the SVAR model is purely data-driven and non-smooth, which could lead to large errors in older ages (e.g., age 100 years). The differences between the results of STAR and SSVAR are not quite visually observable. Despite this, as summarized in Panel A of Table 1, SSVAR produces the smallest mean $RMSFE_x$ across all age groups (0.1381), which is 20% smaller than that of LC (0.1743) and HU (0.1722) and around 10% smaller than those of SVAR (0.1498) and STAR (0.1499). This is consistent

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

TABLE 2

RMSFEs OVER TIME HORIZONS OF THE FORECAST MORTALITY DATA. THE BOLD NUMBER IS THE MINIMUM RMSFE AMONG DIFFERENT METHODS.

	Male					Female				
Steps	LC	HU	SVAR	STAR	SSVAR	LC	HU	SVAR	STAR	SSVAR
Panel A: UK										
2001	0.1224	0.0782	0.1058	0.0790	**0.0760**	0.1258	**0.0949**	0.1174	0.1052	0.1093
2002	0.1398	0.0888	0.0868	0.0908	**0.0848**	0.1291	0.0938	0.0875	0.0904	**0.0860**
2003	0.1337	0.0975	0.0870	0.0892	**0.0798**	0.1437	0.1065	0.1140	**0.1035**	0.1052
2004	0.1354	0.1243	0.1099	0.1060	**0.0985**	0.1405	0.0970	0.0969	0.0973	**0.0911**
2005	0.1405	0.1477	0.1189	0.1171	**0.1085**	0.1302	**0.0971**	0.1039	0.1136	0.1034
2006	0.1678	0.1382	0.1303	0.1209	**0.1148**	0.1568	0.1350	0.1325	0.1296	**0.1205**
2007	0.1844	0.1451	0.1356	0.1396	**0.1253**	0.1436	0.1188	0.1085	0.1120	**0.1016**
2008	0.1944	0.1667	0.1504	0.1438	**0.1356**	0.1769	0.1578	0.1323	**0.1257**	0.1333
2009	0.1851	0.1894	0.1665	0.1655	**0.1516**	0.1765	0.1556	0.1347	**0.1315**	0.1365
2010	0.2058	0.2199	0.1866	0.1859	**0.1709**	0.1749	0.1573	**0.1342**	0.1375	0.1353
2011	0.2322	0.2565	0.2221	0.2214	**0.2043**	0.1989	0.1730	0.1477	0.1528	**0.1420**
2012	0.2348	0.2696	0.2264	0.2303	**0.2126**	0.1781	0.1739	0.1391	0.1488	**0.1327**
2013	0.2408	0.2726	0.2326	0.2391	**0.2146**	0.1788	0.1679	0.1322	0.1392	**0.1312**
2014	0.2502	0.2777	0.2337	0.2314	**0.2130**	0.2032	0.1796	**0.1543**	0.1656	0.1616
2015	0.2414	0.2661	0.2151	0.2180	**0.1979**	0.2025	0.1897	**0.1573**	0.1594	0.1581
2016	0.2346	0.2557	0.2097	0.2139	**0.1913**	0.2160	0.1956	0.1633	**0.1545**	0.1702
Panel B: France										
2001	0.1060	0.0524	0.0567	**0.0522**	0.0528	0.1305	**0.1152**	0.1312	0.1254	0.1233
2002	0.1242	**0.0784**	0.0823	0.0800	0.0787	0.1020	**0.0864**	0.0948	0.1095	0.1037
2003	0.1516	0.1122	0.0887	0.0889	**0.0853**	0.1553	**0.1207**	0.1351	0.1373	0.1290
2004	0.1896	0.1708	0.1263	0.1327	**0.1200**	0.1505	0.1170	0.1190	0.1250	**0.1099**
2005	0.1841	0.1800	0.1202	0.1262	**0.1130**	0.1684	0.1447	0.1475	0.1594	**0.1413**
2006	0.2172	0.2207	0.1420	0.1490	**0.1330**	0.1714	0.1389	0.1380	0.1346	**0.1208**
2007	0.2353	0.2450	0.1635	0.1716	**0.1543**	0.1706	0.1462	0.1479	0.1498	**0.1278**
2008	0.2444	0.2548	0.1757	0.1873	**0.1656**	0.1861	0.1589	0.1572	0.1589	**0.1435**
2009	0.2375	0.2447	0.1662	0.1765	**0.1568**	0.1765	0.1591	0.1583	0.1590	**0.1450**
2010	0.2512	0.2555	0.1829	0.1967	**0.1731**	0.2020	0.1851	0.1701	0.1637	**0.1443**
2011	0.2599	0.2573	0.1857	0.2005	**0.1722**	0.2027	0.1958	0.1675	0.1743	**0.1472**
2012	0.3048	0.2945	0.2225	0.2359	**0.2107**	0.2033	0.2009	0.1682	0.1714	**0.1544**
2013	0.3171	0.3006	0.2388	0.2511	**0.2265**	0.2314	0.2381	0.2091	0.2130	**0.1937**
2014	0.3424	0.3217	0.2596	0.2750	**0.2457**	0.2345	0.2365	0.2001	0.2018	**0.1780**
2015	0.3161	0.2918	0.2290	0.2392	**0.2154**	0.2093	0.2192	0.1803	0.1743	**0.1613**
2016	0.3479	0.3238	0.2660	0.2830	**0.2536**	0.2308	0.2375	0.2105	0.2129	**0.1929**

with the result of the overall measure $RMSFE_{all,16}$, which is reported in the "All" column of Table 1. Additionally, the standard deviation (*SD*) suggests that the variation of $RMSFE_x$ of SSVAR is also the smallest among all models. The Q_1 and Q_3 measures further support that the SSVAR model consistently outperforms other counterparties. As for the UK females, the first noticeable

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

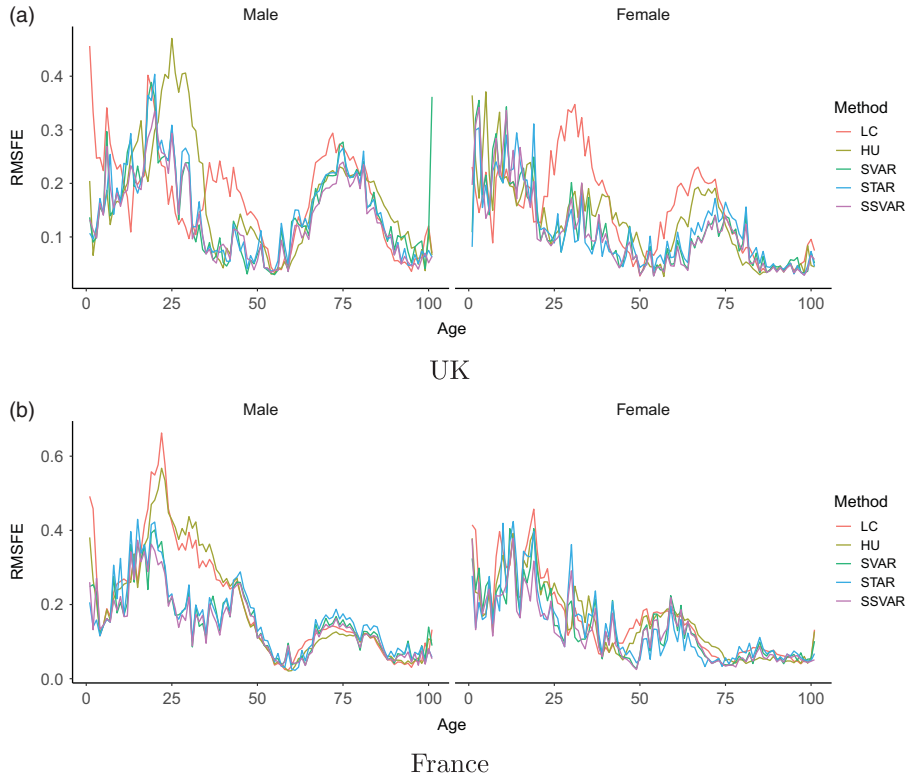


FIGURE 1: RMSFEs over forecasting age groups.

difference is the large $RMSFE_x$ generated by LC for the 20–75 years of age groups. SSVAR still leads to the best performance, as measured in Panel A of Table 1, except for the standard deviation.

The outcomes for France are largely consistent with the observations for the UK. In Figure 1(b), for males, LC and HU lead to larger $RMSFE_x$ for ages 20–40 years than the VAR-type models. For females, the produced $RMSFE_x$ by the five models are roughly close to each other. As measured in Panel B of Table 1, SSVAR results in the best RMSFE according to all statistics for both sexes. For instance, the $RMSFE_{all,16}$ of SSVAR fitted for males (0.1702) is around 30% smaller than that of LC (0.2501) and HU (0.2396) and 10% smaller than those of SVAR (0.1800) and STAR (0.1902). To test the statistical significance, we perform the DM test proposed in Diebold and Mariano (2002) on the forecast rates between SSVAR and each of the other four models. For its pairwise nature, the DM test is performed with the null (alternative) hypothesis such that the forecast rates of SSVAR are as accurate as (more accurate than) those of the tested model. As indicated by * in Table 1, except in the case of UK females, SSVAR model consistently outperforms the rest at the 5% significance level.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

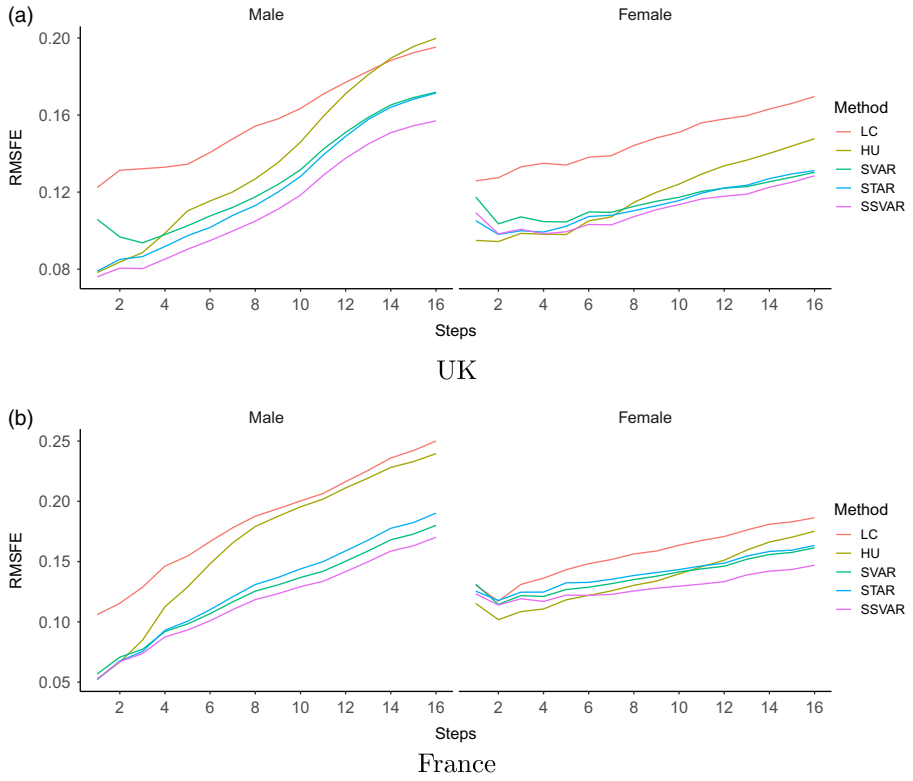


FIGURE 2: RMSFEs over forecasting steps.

Figure 2 plots the $RMSFE_{all,h}$ for h ranging from 1 (2001) to 16 (2016). Distinct differences can be observed in all cases. For UK males, the curve of SSVAR is consistently below those of all other competing models. This is confirmed with the results observed in Panel A of Table 2, where SSVAR leads to the smallest $RMSFE_h$ in all steps. The outcomes for UK females are largely similar, except that $RMSFE_{all,h}$ of SSVAR and STAR are similar in the first few steps. $RMSFE_{all,h}$ of SSVAR begins to outperform the rest from the fourth horizon (2004) and remains the best up to the last horizon (2016). As shown in Table 2, the individual $RMSFE_h$ of SSVAR is the smallest in 7 of the 16 steps, which is still the best performing result. Despite this, the advantage of SSVAR over the rest competing model is comparatively narrowed, contrasting the results of UK female to those of UK male. One possible explanation is that cohort effects could be different for male and female mortality (see Renshaw and Haberman, 2006 for the case of UK). To further improve the performance of SSVAR, we might adjust θ in (2.4) and thus allow for population-dependent adaptive weights. Further exploration on this matter remains for future works.

The conclusions on the France data are again largely consistent. Figure 2(b) suggests that the $RMSFE_{all,h}$ of SSVAR starts to beat the others from the third

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

and the seventh step for males and females, respectively. For $RMSFE_x$ reported in Panel B of Table 2, except for the first two (four) steps in males (females), SSVAR produces the best result in all other scenarios. Contrasting the results of the UK and France, in Figure 2, it is interesting to note the difference between the STAR and SVAR models. For the UK, the STAR model almost consistently outperforms the SVAR model throughout the steps, whereas the opposite holds for the France data. This implies that the cohort effect for neighboring ages is more important in the mortality improvements for the UK than for France. As discussed in Section 2, our proposed SSVAR model is capable of modeling the cohort effects with a more flexible and data-driven structure. More specifically, SSVAR adopts a less restricted coefficient matrix $\boldsymbol{B}$ than does STAR to allow for more complicated cohort structures. Compared with SVAR, SSVAR uses the spatial information by employing the adaptive weights on the age dimension. This ensures that the parameters of $\boldsymbol{B}$ are not "overly" data-driven to overwhelm the explainable cohort structures. It is also the main reason that our model can consistently beat other VAR-type competitors.

3.2. SVAR versus SSVAR

To understand the superiority of our proposed SSVAR model, we first discuss the fitted intercepts $\boldsymbol{\alpha}$. For a VAR-type model without time trends, the effects of auto-regressors will die out quickly, which is known as geometric decay. Thus, the intercepts will dominate in mid- to long-term forecasts (described as "mean-reversion" in the literature). Thus, appropriately modeling the intercept is a crucial issue for forecasting mortality rates accurately.

We compare the fitted $\boldsymbol{\alpha}$ of SVAR and SSVAR in Figure 3. As discussed in Section 2, the SVAR model considered in Guibert *et al.* (2019) is a purely data-driven framework. Hence, the intercept is not enforced to 0 as β_{ij} or penalized to impose smoothness among ages. Given crude mortality rates, this will lead to unavoidable roughness and irregular patterns on the age dimension. In Figure 3(a), random variations among ages are observed for both UK males and females, as fitted by SVAR. For instance, the intercept of males suddenly increases from −0.01 for the age of 98 years to 0.02 for the age of 99 years, and then drops quickly to −0.05 for the age of 100 years. In addition, there are a few positive fitted intercepts for ages 25–50 years. Without further adjustment, using this result may introduce uninterpretable forecasts, such as increasing mortality rates for young age groups in the long run. The results of SVAR for the France data also demonstrate similar patterns, where roughness is shown across ages and abrupt changes are produced at very old ages (e.g., 100 years of age).

As for SSVAR, the intercepts are penalized appropriately, with an independent smoothness parameter (η_1) from that for β_{ij} (η_2). The effectiveness of our approach can be observed in Figure 3. In all four cases, the fitted $\boldsymbol{\alpha}$ is much more smoothed out than that of SVAR. There are also no irregular

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

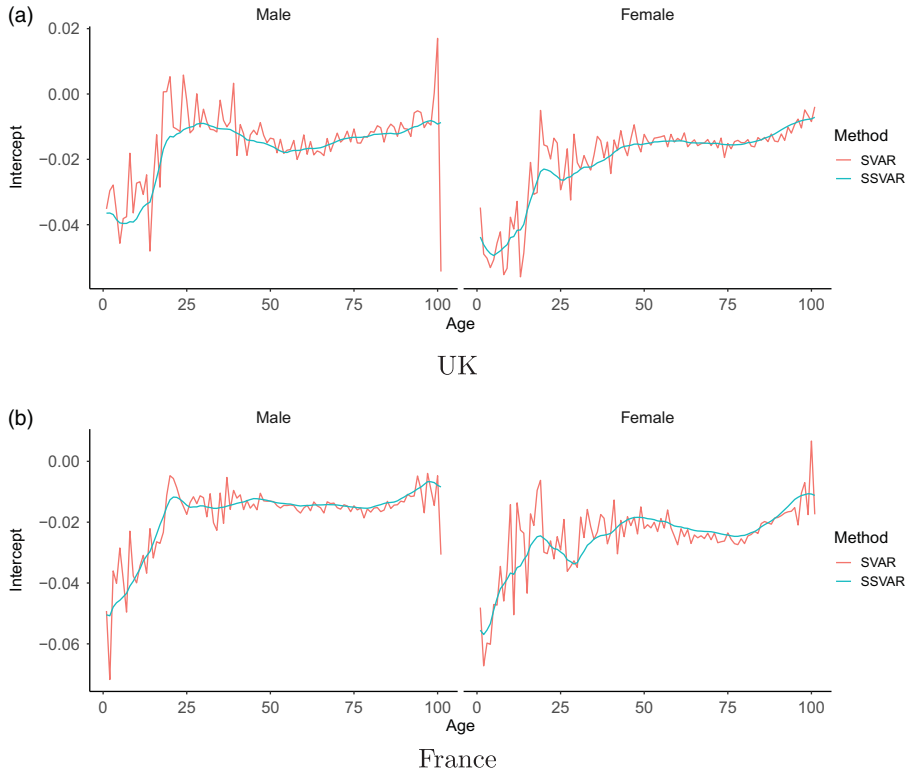


FIGURE 3: Fitted intercepts.

patterns or abrupt changes among all age groups. In particular, none of the fitted $\boldsymbol{\alpha}$ are above 0, which indicates that mortality rates will continue to decrease in the long run. Thus, in contrast with the results of the SVAR model, the smoothed intercepts of SSVAR will produce more meaningful interpretation of mortality improvements. This is essential in actuarial and demographic practice. For example, smoothness ensures that life insurance premiums will not inconsistently change when policyholders grow older.

Another important difference between our proposed SSVAR and SVAR is the adoption of adaptive weights. A purely data-driven framework in SVAR may lead to unexplainable fitted coefficients. For instance, mortality improvements of very young ages may be associated with those of very old ages. Using the w_{ij} as described in (2.4), SSVAR ensures that nonzero β_{ij} is more likely to occur for local age groups (j close to i). When j is largely different from i, β_{ij} will tend to be penalized to 0, which is more consistent with actuarial and demographic practice. The fitted β_{ij} of the SVAR and SSVAR models are plotted in Figures 4 and 5, respectively.

In Figure 4, it is clear that the sparsity of fitted β_{ij} for the SVAR model is of a more random nature. For instance, nonzero $\beta_{i,100}$ are estimated for UK

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

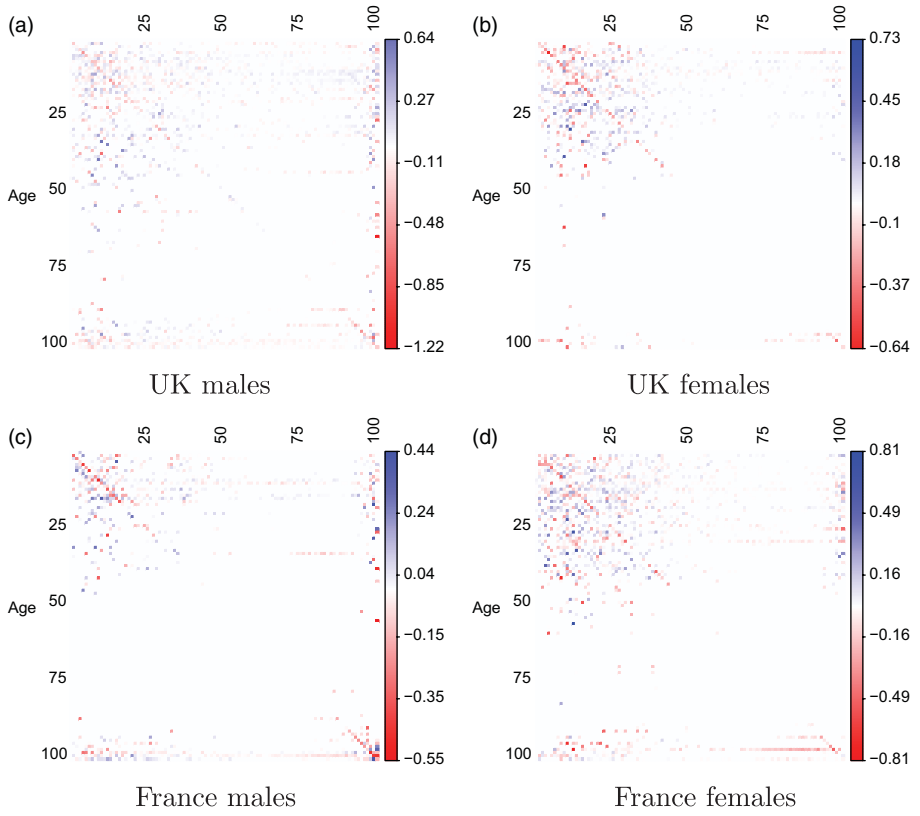


FIGURE 4: Fitted coefficients of the SVAR model.

males for many ages i, from young (as low as 0) to old. This outcome is highly likely a result of noise among the data and not quite interpretable, since mortality improvements of young ages are not likely to be affected by those of old ages. Further, the cohort effects may be "contaminated" by the data-driven yet uninterpretable results. That is, the effects of the neighboring ages may be offset by those of irrelevant groups. This may explain the superiority of STAR over SVAR for the UK data, where cohort effects for neighboring ages are strong.

In Figure 5, nonzero fitted β_{ij} produced by SSVAR model concentrate on the diagonals. This is as expected, where only spatially close age groups can have a meaningful effect on the mortality improvements. Compared with the results of Figure 4, the randomness is largely removed in all cases, and the level of sparsity is increased. It is interesting to note that the "spread" of nonzero β_{ij} is wider for females than for males for both the UK and France in the young age groups. This suggests that there are potentially more groups contributing to the cohort effects for females than for males. This is also evidence that the ad hoc STAR structure may not generally suit for all empirical data.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

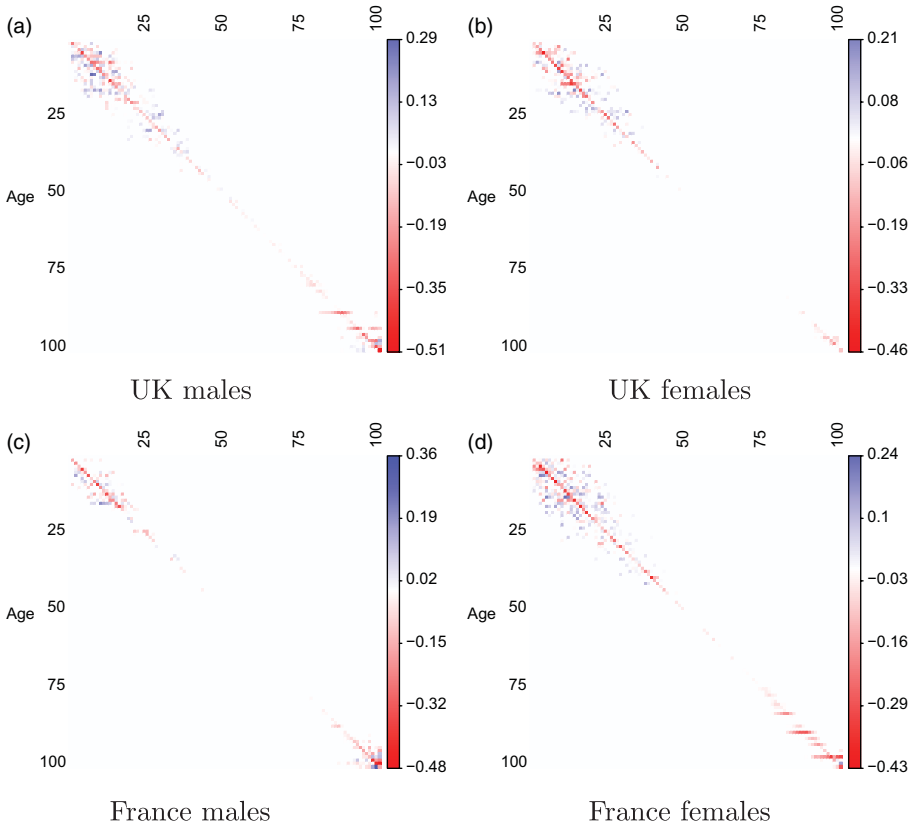


FIGURE 5: Fitted coefficients of the SSVAR model.

Finally, we compare the residuals produced by the SVAR and SSVAR models, using the method discussed in Section 2. As a result of the influence of the adaptive weights, we expect that the correlations of residuals in SSVAR may exhibit similar structures to those of the fitted β_{ij}. This is evidenced in Figures 6 and 7. When the SVAR model is employed, the correlations of residuals between many young age groups and the 100-year-olds are far from 0, especially for males. Again, this is counterintuitive, as we do not expect the 100-year-old group to significantly contribute to the contemporaneous error structures of young age groups. In contrast, nonzero correlations are only observed for residuals of nearby ages (on the diagonals) when the SSVAR model is used. This provides further evidence to support the effectiveness of our proposed spatially penalized approach. It is worth noting that some negative and thus counterintuitive correlations are observed in both Figures 6 and 7. Indeed, those negative correlations are inevitable random noises of a high-dimensional modeling framework with limited sample sizes, like the mortality data. Compared to the SVAR model, the number of those negative correlations are much more reduced for SSVAR. In particular, negative correlations are of

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

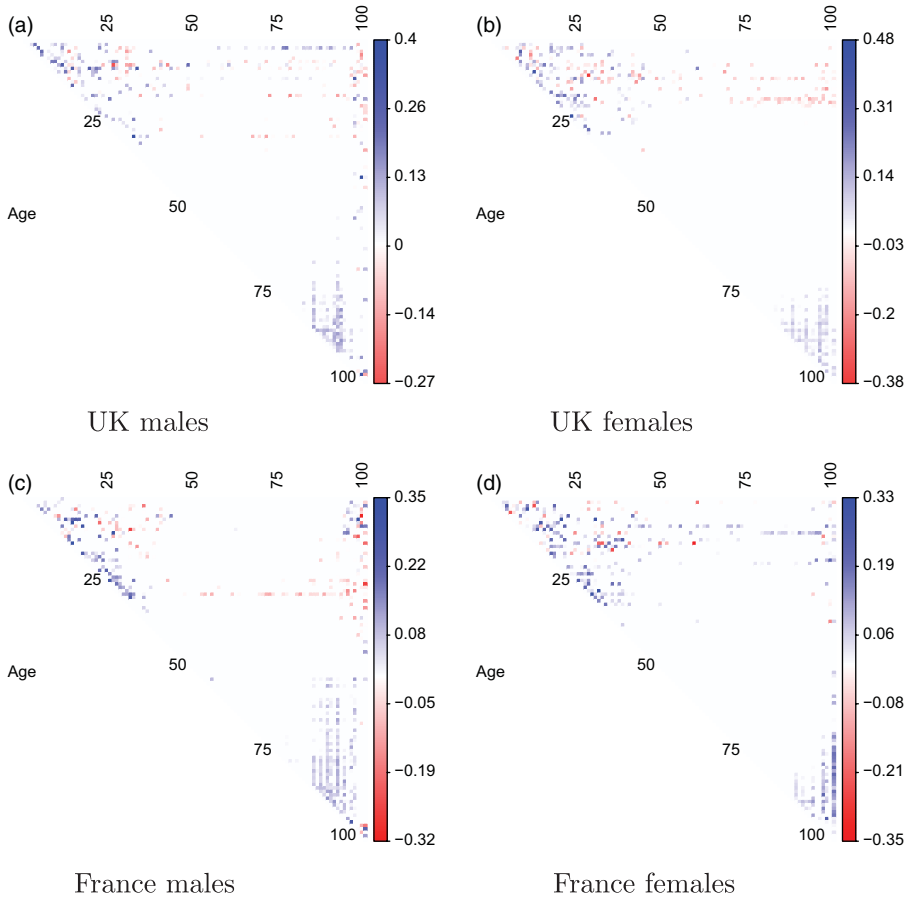


FIGURE 6: Correlations of residuals fitted by the SVAR model.

a pure random nature in Figure 7, rather than being clustered as in Figure 6. With a larger sample size, we believe there should be a smaller number of such negative correlations. Using the imposed adaptive weights and smoothing penalties, we expect that with an increasing sample size, the number of negative correlations of our SSVAR model will be reduced faster than that of the SVAR model.

Thus, the error structure of the SSVAR model is of a more interpretable structure than that of the SVAR model. This may improve the accuracy of relevant inferences, such as the prediction intervals (PIs).

3.3. Long-term application: Life expectancies

We now examine the long-term application of the four models. The complete datasets are employed (from 1950 to 2016) and the mortality rates are forecast

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

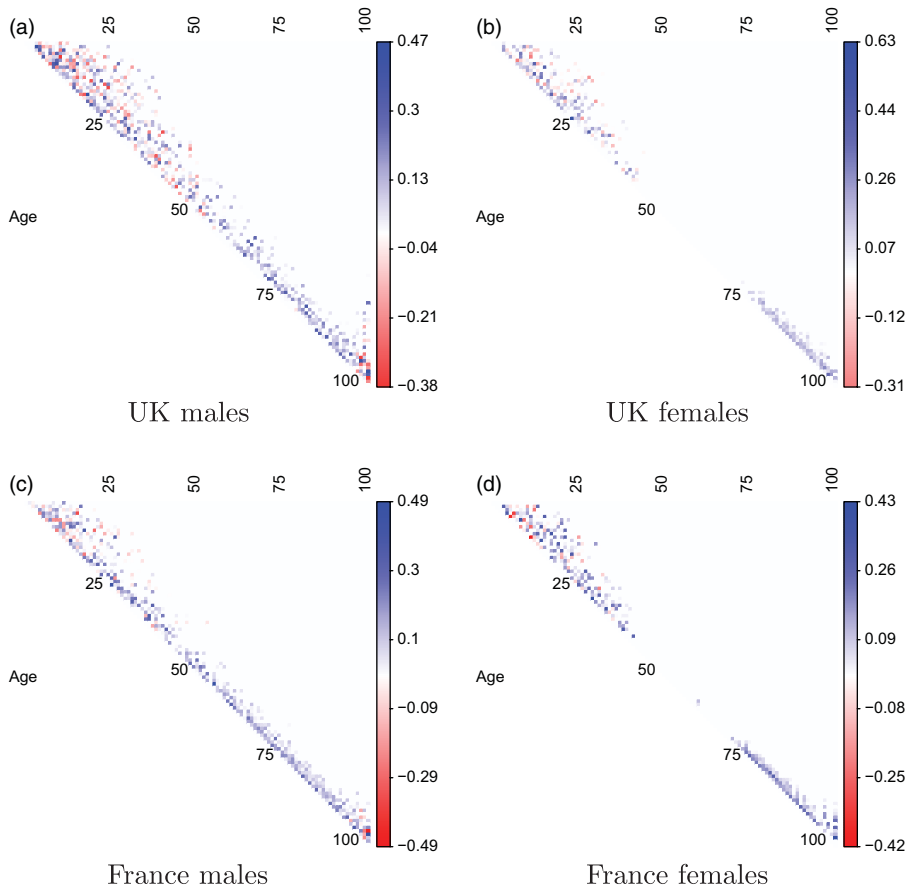


FIGURE 7: Correlations of residuals fitted by the SSVAR model.

to 2050 for both the UK and France males and females. Among all information generated from mortality rates, life expectancies are widely studied in demographic research and investigated in actuarial practice. We then study the produced life expectancies at the age of 0 years (e_0) in Figure 8.

Figure 8 reports both mean/point forecasts (solid line) and 95% PIs (dashed line) of forecasts of e_0 up to 2050. The PIs of all models are calculated using the 2.5th and 97.5th percentiles of the simulated samples with 2000 replicates. In all cases, despite the similarities of mean forecasts, the PIs of the LC model are relatively wider than those of the rest. For the UK male point life expectancies, e_0 generated by LC is higher than that of the VAR models in 2017. However, the increment in e_0 of LC is smaller than the others. From around 2035, the LC model produces the shortest e_0. In 2050, the results of STAR (84.41), HU (84.40), and SSVAR (84.38) are very close to each other and slightly higher

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

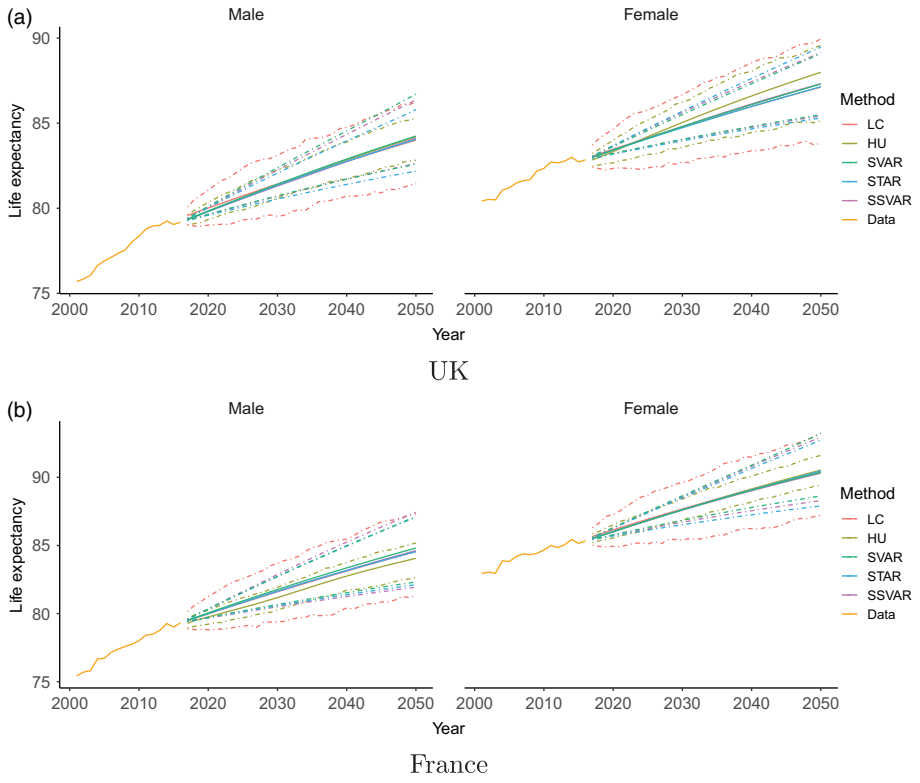


FIGURE 8: Actual and forecast life expectancies: 2001–2050.

than those of SVAR (84.06) and LC (84.0). Apart from the results of HU, the differences of the point forecast UK female e_0 are even smaller and range from 87.12 (LC) to 87.48 (SSVAR) in 2050. In terms of the PIs, despite some minor differences, the widths of all three VAR-type models are close to each other and slightly narrower (wider) than those of HU for males (females), with similar increments throughout times.

The outcomes for France are largely consistent with our observations for the UK. In Figure 8(b), we still find that point forecasts of e_0 do not largely diverge from model to model, except that those of HU are uniformly the lowest for males. The PIs of LC are still wider than those of the VAR-type models. Besides, there are only two noticeable differences. First, the point forecasts of e_0 produced by the SSVAR model are the longest in most years, for both males and females. Second, the point forecasts of e_0 generated by LC are no longer the shortest. For females, despite the similarity, SVAR leads to the smallest life expectancies in most scenarios. In 2050, the point forecasts range from 84.23 (HU) to 84.69 (SSVAR) for males and vary between 90.10 (SVAR) and 90.38 (SSVAR) for females.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

4. AN EXTENSION TO MULTI-POPULATION MODELING

Accompanied with global improvements in public health, medical advances, transportation, and technology, it is widely recognized that the mortality rates between 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 take account of the common patterns in a large group of populations (see Li and Lee, 2005; Cairns *et al.*, 2011, and Li and Hardy, 2011). Li and Lu (2017) and Guibert *et al.* (2019) also extended their VAR-type models to simultaneously analyze several populations. In this section, we explore the extension of our proposed SSVAR to multi-population mortality modeling. For simplicity, we only describe the extension for a two-population case, while situations with more than two populations can be easily generalized.

We denote $y_{1,i,t}$ and $y_{2,i,t}$ as the log of mortality improvements for the first and second population at age i in year t, respectively. Further, we merge these two different N-dimensional vectors into a $2N$-dimensional vector $\boldsymbol{y}_t = (y_{1,1,t}, \ldots, y_{1,N,t}, y_{2,1,t}, \ldots, y_{2,N,t})'$. Then, the VAR model for a single population as in (2.2) is fitted to this combined data, and our proposed SSVAR as in (2.3) can be implemented to estimate a $2N$-dimensional intercept vector $\boldsymbol{\alpha}$ and a $2N \times 2N$ Granger causality matrix $\boldsymbol{B}$. However, because the dimension of the parameters increases, we need to reconstruct the adaptive weights w_{ij} and the smoothing penalty matrix $\boldsymbol{H}$.

It is worth noting that the $L1$ penalty for $\boldsymbol{B}$ as in (2.3) can also be written as:

$$\lambda \sum_{i,j}^{N} w_{ij}|\beta_{ij}| = \|\boldsymbol{P} \circ \boldsymbol{B}\|_1, \tag{4.1}$$

where the penalty matrix $\boldsymbol{P} = \lambda \boldsymbol{W}$, $\boldsymbol{W}$ is an $N \times N$ weight matrix with the ijth element being w_{ij}, $\circ$ is the Hadamard product (entrywise product), and $\|\cdot\|_1$ represents the $L1$ norm or the absolute sum of all elements in a matrix. In a two-population case, we can extend the $2N \times 2N$ penalty matrix $\boldsymbol{P}^{(2)}$ as in the following form:

$$\boldsymbol{P}^{(2)} = \begin{bmatrix} \lambda_1 \boldsymbol{W} & \lambda_2 \boldsymbol{W} \\ \lambda_1 \boldsymbol{W} & \lambda_2 \boldsymbol{W} \end{bmatrix},$$

where we allow different tuning parameters for the first and second population, considering that different populations may exhibit different levels of sparsity in $\boldsymbol{B}$. We maintain the structure of w_{ij} as in (2.4) because we believe that the log-mortality improvements for close ages are more correlated, both within and between populations. Similarly, to produce smoothed age-dependent parameters, we extend the $2N \times 2N$ smoothing penalty matrix $\boldsymbol{H}^{(2)}$ as:

$$\boldsymbol{H}^{(2)} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{H} \\ \boldsymbol{H} & \boldsymbol{H} \end{bmatrix}. \tag{4.2}$$

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

TABLE 3

MULTI-COUNTRY COMPARISON: $RMSFE_{all,16}$. THE BOLD NUMBER IS THE MINIMUM RMSFE AMONG DIFFERENT METHODS.

Country	Sex	LL	HU	SVAR	STAR	SSVAR
UK	Male	0.1676	0.1853*	0.1827*	0.1712*	**0.1617**
UK	Female	0.1374*	0.1523*	0.1302	0.1308	**0.1276**
France	Male	0.1874*	0.2180*	0.1749*	0.1903*	**0.1693**
France	Female	0.1600*	**0.1492**	0.1635*	0.1637*	0.1526

Finally, SSVAR for a two-population model can be achieved by solving:

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{B}} \frac{1}{2}\sum_{t=1}^{T} \left\|\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1}\right\|_2^2 + \left\|\boldsymbol{P}^{(2)} \circ \boldsymbol{B}\right\|_1 + \eta_1 \left\|\boldsymbol{\alpha}^T \boldsymbol{H}^{(2)}\right\|_2^2 + \eta_2 \left\|\boldsymbol{B}^T \boldsymbol{H}^{(2)}\right\|_F^2. \quad (4.3)$$

For illustration purposes, we fit the males of the UK and France together using multi-population extension of LC, namely the Li–Lee (LL) model (Li and Lee, 2005), the extension of HU, namely the coherent HU (CHU) model (Hyndman *et al.*, 2013), and the three VAR models.[7] Similar to Section 3, we also consider the sample period of 1951–2000 and forecast mortality rates for 2001–2016. The case of females for both countries is further investigated following the same approach. The resulting $RMSFE_{all,16}$ are reported in Table 3. Clearly, consistent with our observations in Section 3, the SSVAR model leads to the best overall forecasting performance in almost all scenarios. The DM test also suggests that SSVAR significantly (at 5% level, as indicated by *) outperforms other competing models for three out of four populations. The results of $RMSFE_x$ and $RMSFE_h$ are also robust to those of the single-country analysis, which are available upon request.

5. CONCLUDING REMARKS

In this paper, we propose an effective spatially penalized smoothed sparse VAR model (SSVAR) to model and forecast mortality improvements. Three key results can be drawn from our study. First, the SSVAR model estimated by our proposed numerical algorithm is effective in mortality forecasting. As measured by the RMSFE, our SSVAR model almost uniformly outperforms the LC, HU, STAR, and SVAR models. This conclusion consistently holds for both males and females of the UK and France when the sample of 1950–2000 is fitted and forecasts of 2001–2016 are produced. Second, the SSVAR model has a more general framework that retains the advantages of both the STAR and SVAR models. More specifically, SSVAR adopts the data-driven approach

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

as in SVAR, which is more flexible than the STAR model. Meanwhile, its estimated Granger causality matrix is more interpretable than that of SVAR. In addition, the smoothness penalty is effective in reducing divergence of forecast rates, as the estimated intercepts among neighboring ages of SSVAR are much less random than those of SVAR. Finally, in contrast to the standard factor models, such as LC, the extension of SSVAR to multi-population mortality modeling is straightforward. Our proposed SSVAR model still outperforms the Li–Lee, CHU, STAR, and SSVAR under most of those scenarios. Our results demonstrate the effectiveness and usefulness of SSVAR in mortality forecasting.

As observed in Section 2.5, the precision matrix $\boldsymbol{K}$ is produced after parameters $\boldsymbol{\alpha}$ and $\boldsymbol{B}$ are estimated using SSVAR via a penalized least squares type of objective function. Future research can be conducted to simultaneously estimate $\boldsymbol{\alpha}$, $\boldsymbol{B}$, and $\boldsymbol{K}$ using a penalized likelihood approach to achieve higher efficiency of the estimators. In addition, extensions may be developed to consider the potential heteroskedasticity of mortality rates. Instead of working with mortality improvements, the SSVAR model may also be extended to model mortality rates directly with appropriate stationarity constraints.

Acknowledgment

The authors would like to thank the Australian National University and Macquarie University for research support. We particularly thank the Editor (Montserrat Guillen) and two anonymous referees for providing valuable and insightful comments on earlier drafts. The usual disclaimer applies.

Notes

1. Note that the cohort effects studied in this paper essentially refer to those between more generalized surrounding cohorts (i.e., age groups spatially close to each other), rather than the effect within the same cohort only. As discussed below, the cohort effect we consider nests that investigated in Li and Lu (2017).

2. As suggested in Li and Lu (2017), the coefficients of the VAR(1) can be easily explained by the cohort effects. For instance, the impacts of $y_{i-1,t-1}$ and $y_{i-2,t-1}$ on $y_{i,t}$ are cohort effects of the same and the younger cohorts. From a pure data-driven perspective, although more lags could lead to higher forecasting accuracy, the resulting coefficients are much less interpretable. To balance the interpretability and the merits of the pure data-driven approach, we only focus on the lag 1 case in this paper. Our preliminary analysis using (partial) autoregressive functions also supports the adequacy of the VAR(1) specification. The proposed model, however, is directly extensible to contain more lags.

3. Other spatial similarity functions, such as the Matérn family (Cressie and Wikle, 2015, p. 126), may also be applied but they involve extra tuning parameters, which is more computationally intensive for mortality forecasting and is not considered in this paper.

4. In our study, we consider a fixed $\theta = 0.9$ to reduce the computation cost. In addition, we find that the results are not very sensitive to the choice of θ if θ does not vary much.

5. The objective function below comes from the penalized log-likelihood of the multivariate normal distribution. Details of this penalized log-likelihood without weights added can be found in Friedman *et al.* (2008).

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

6. We also modeled mortality data of other countries, including Ukraine and Russia, which have experienced structural changes. In those cases, the three VAR-type models demonstrate fairly similar performances. The detailed results are available upon request.

7. It is worth noting that as studied in Villegas *et al*. (2017), the selection of a benchmark for the multi-population factor models needs a more thorough analysis to consider explainable demographic properties. For illustrative purpose and to be consistent with Section 3, we only consider the multi-population extensions of the previously investigated single-population models. A more detailed analysis remains for the future works.

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.

BASU, S. and MICHAILIDIS, G. (2015) Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, **43**(4), 1535–1567.

BECK, A. and TEBOULLE, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**(1), 183–202.

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

BOOTH, H., HYNDMAN, R., TICKLE, L. and 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. and KHALAF-ALLAH, M. (2011) Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, **41**, 29–59. ISSN 1783-1350.

CRESSIE, N. and WIKLE, C.K. (2015) *Statistics for Spatio-Temporal Data*. New Jersey: John Wiley & Sons.

DAUBECHIES, I., DEFRISE, M. and DE MOL, C. (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, **57**(11), 1413–1457.

DAVIS, R.A., ZANG, P. and ZHENG, T. (2016) Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, **25**(4), 1077–1096.

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.

DIEBOLD, F.X. and MARIANO, R.S. (2002) Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **20**(1), 134–144.

DONOHO, D.L. and JOHNSTONE, J.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

FAN, J., LV, J. and QI, L. (2011) Sparse high-dimensional models in economics. *Annual Review of Economics*, **3**(1), 291–317.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

GUIBERT, Q., LOPEZ, O. and PIETTE, P. (2019) Forecasting mortality rate improvements with a high-dimensional var. *Insurance: Mathematics and Economics*, **88**, 255–272.

HUMAN MORTALITY DATABASE. (2019) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). URL http://www.mortality.org.

HUNT, A. and BLAKE, D. (2014) A general procedure for constructing mortality models. *North American Actuarial Journal*, **18**(1), 116–138.

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.

HYNDMAN, R.J. and ULLAH, M.S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, **51**(10), 4942–4956.

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.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

LEE, R.D. and CARTER, L.R. (1992) Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.
LI, H. and LU, Y. (2017) Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin: The Journal of the IAA*, **47**(2), 563–600.
LI, H., LU, Y. and LYU, P. (2018) Coherent mortality forecasting for less developed countries. SSRN Scholarly Paper ID 3209392. URL http://dx.doi.org/10.2139/ssrn.3209392.
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.
RENSHAW, A.E. 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. (2017) Basis risk modelling: A cointegration-based approach. *Statistics*, **51**(1), 205–221.
SONG, S. and BICKEL, P.J. (2011) Large vector auto regressions. arXiv preprint arXiv:1106.3915.
VILLEGAS, A.M., HABERMAN, S., KAISHEV, V.K. and MILLOSSOVICH, P. (2017) A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA*, **47**(3), 631–679.
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.
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.
ZOU, H. and HASTIE, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, **67**(2), 301–320.

LE CHANG (Corresponding author)
Research School of Finance, Actuarial Studies, and Statistics
Australian National University
Canberra, ACT, Australia
E-Mail: le.chang@anu.edu.au

YANLIN SHI
Department of Actuarial Studies and Business Analytics
Macquarie University
Sydney, NSW, Australia
E-Mail: yanlin.shi@mq.edu.au

APPENDIX A

A.1. Technical details on the non-divergence

In this paper, we define the non-divergence (also known as the coherence in the existing literature, e.g., see Li and Lee, 2005 and Li and Lu, 2017) as follows.

Definition 1. *Non-divergence means that for the h-step-ahead forecasts,* $|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}| = O_p(1)$*, for all* $i, j \in \{1, 2, ..., N\}$*. That is, when* $h \to \infty$*,* $|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}|$ *will not diverge to infinity.*

We will use the following result to discuss the non-divergence property of the SSVAR and SVAR model.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

Lemma 1. *Assume that $y_{i,t} = \ln m_{i,t} - \ln m_{i,t-1}$ follows a stationary VAR(1) process as described in (2.2), for any two ages $i \neq j$ and $i, j \in \{1, 2, ..., N\}$, we have that:*

$$|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}| = O_p(1) + h|\hat{y}_{i,T+h} - \hat{y}_{j,T+h}|,$$

Proof. Note that $\ln \hat{m}_{i,T+h} = E(\ln m_{i,T+h}|\Omega_T)$, where Ω_T is the information set up to time T. Using the given conditions, $\hat{y}_{i,T+h} = E(y_{i,T+h}|\Omega_T) = [\sum_{s=0}^{h-1} \hat{\boldsymbol{B}}^s \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{B}}^h \boldsymbol{y}_T]_i = [(\boldsymbol{I}_N - \hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{B}}^h(\boldsymbol{y}_T - \hat{\boldsymbol{\alpha}})]_i$ and $\hat{y}_{i,T+k} - \hat{y}_{i,T+h} = [(\hat{\boldsymbol{B}}^k - \hat{\boldsymbol{B}}^h)(\boldsymbol{y}_T - \hat{\boldsymbol{\alpha}})]_i$. Therefore, we have that:

$$\begin{aligned}\sum_{k=1}^{h} \hat{y}_{i,T+k} &= \sum_{k=1}^{h} (\hat{y}_{i,T+k} - \hat{y}_{i,T+h} + \hat{y}_{i,T+h}) = \sum_{k=1}^{h} \left[\left(\hat{\boldsymbol{B}}^k - \hat{\boldsymbol{B}}^h\right)(\boldsymbol{y}_T - \hat{\boldsymbol{\alpha}})\right]_i + h\hat{y}_{i,T+h} \\ &= \left[\left(\sum_{k=1}^{h} \hat{\boldsymbol{B}}^k - h\hat{\boldsymbol{B}}^h\right)(\boldsymbol{y}_T - \hat{\boldsymbol{\alpha}})\right]_i + h\hat{y}_{i,T+h} = O_p(1) + h\hat{y}_{i,T+h},\end{aligned}$$

when $h \to \infty$, since $\boldsymbol{y}_t$ is stationary. Further,

$$\ln \hat{m}_{i,T+h} = \ln m_{i,T} + \ln \hat{m}_{i,T+1} - \ln m_{i,T} + \ln \hat{m}_{i,T+2} - \ln \hat{m}_{i,T+1} + ... = \ln m_{i,T} + \sum_{k=1}^{h} \hat{y}_{i,T+k}.$$

Hence,

$$\begin{aligned}|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}| &= |\ln m_{i,T} - \ln m_{j,T} + \sum_{k=1}^{h} (\hat{y}_{i,T+k} - \hat{y}_{j,T+k})| \\ &= O_p(1) + h|\hat{y}_{i,T+h} - \hat{y}_{j,T+h}|.\end{aligned}$$

□

Theorem 1. *Given the same assumptions as in Lemma 1, and suppose the conditions listed below also hold*

(i) h and T go to infinity at the same rate;
(ii) $\boldsymbol{H}$ is defined as in (2.5),
(iii) $\eta_1 \|\boldsymbol{\alpha}^T \boldsymbol{H}\|_2^2$ and $\eta_2 \|\boldsymbol{B}^T \boldsymbol{H}\|_F^2$ are both $O_p(T)$; and
(iv) Smoothing penalties η_1 and η_2 go large with T^3,

the SSVAR model fitted as in (2.3) has non-divergent out-of-sample forecasts when $T \to \infty$.

Proof. Note that in (2.3), $\frac{1}{2}\sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{B}\boldsymbol{y}_{t-1}\|_2^2$ is $O_p(T)$. Thus, for the penalty terms to be effective, both $\eta_1 \|\boldsymbol{\alpha}^T \boldsymbol{H}\|_2^2$ and $\eta_2 \|\boldsymbol{B}^T \boldsymbol{H}\|_F^2$ should also be $O_p(T)$, as assumed in (iii). Further with Assumption (iv), $\|\boldsymbol{\alpha}^T \boldsymbol{H}\|_2^2$ is then $O_p(1/T^2)$. For a fitted SSVAR model, with Assumption (ii), this is equivalent to $\sum_{i=1}^{N} (\hat{\alpha}_{i+1} - \hat{\alpha}_i)^2 = O_p(1/T^2)$. Since N is finite, we can further infer that $|\hat{\alpha}_{i+1} - \hat{\alpha}_i| = O_p(1/T)$, where $\hat{\alpha}_i$ is the ith element of $\hat{\boldsymbol{\alpha}}$. For the same reason, it can be shown that $|\hat{\beta}_{i+1,j} - \hat{\beta}_{i,j}| = O_p(1/T)$, and $\hat{\beta}_{i,j}$ is the ith row and jth column of $\hat{\boldsymbol{B}}$. Both $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}$ are defined as in (2.2).

For a finite N, we also notice that:

$$|\hat{y}_{i+1,T+h} - \hat{y}_{i,T+h}| = |\hat{\alpha}_{i+1} - \hat{\alpha}_i + \sum_{k=1}^{N} (\hat{\beta}_{i+1,k} - \hat{\beta}_{i,k})\hat{y}_{k,T+h-1}| = O_p(1/T).$$

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press

Further, use Lemma 1 and the Assumption (i) such that h and T go to infinity at the same rate, it can be shown that:

$$|\ln \hat{m}_{i+1,T+h} - \ln \hat{m}_{i,T+h}| = O_p(1) + h|\hat{y}_{i,T+h} - \hat{y}_{j,T+h}| = O_p(1).$$

This result can then be straightforwardly extended to any two ages within $1, 2, ..., N$, since $\ln \hat{m}_{i+r,T+h} - \ln \hat{m}_{i,T+h} = \ln \hat{m}_{i+r,T+h} - \ln \hat{m}_{i+r-1,T+h} + \ln \hat{m}_{i+r-1,T+h} - \ln \hat{m}_{i+r-2,T+h} + ... + \ln \hat{m}_{i+1,T+h} - \ln \hat{m}_{i,T+h}$, which completes the proof. □

Clearly, without the smoothing penalties, $|\hat{y}_{i,T+h} - \hat{y}_{j,T+h}|$ of the SVAR model is $O_p(1)$, such that $|\ln \hat{m}_{i+1,T+h} - \ln \hat{m}_{i,T+h}| = O_p(h)$, and thus the out-of-sample forecasts are divergent.

Following the same logic and principles, the above results can be directly extended to the multi-population scenarios as discussed in Section 4 of the paper. In particular, if the $\boldsymbol{H}^{(2)}$ matrix is defined as in (4.2), the forecasts will not diverge between the two populations. If $\boldsymbol{H}^{(2)} = diag(\boldsymbol{H}, \boldsymbol{H})$, clearly the non-divergence will only exist for the ages of the individual population only but not across those of the two populations. This can be useful to allow potentially divergent forecasts for less developed countries (Li *et al.*, 2018).

APPENDIX B

B.1. Computational packages

We use the software R to perform the computations of all models. LC, HU, and CHU are estimated using the demography package. LL is fitted by the MortalityForecast package. SVAR, STAR, and SSVAR are computed with codes written by the authors. The forecasts of LC and HU models are based on the ARIMA model, which are the default option in the demography package.

https://doi.org/10.1017/asb.2020.39 Published online by Cambridge University Press