EXTENDING THE LEE–CARTER MODEL WITH VARIATIONAL AUTOENCODER: A FUSION OF NEURAL NETWORK AND BAYESIAN APPROACH

BY

AKIHIRO MIYATA AND NAOKI MATSUYAMA

ABSTRACT

In this study, we propose a nonlinear Bayesian extension of the Lee–Carter (LC) model using a single-stage procedure with a dimensionality reduction neural network (NN). LC is originally estimated using a two-stage procedure: dimensionality reduction of data by singular value decomposition followed by a time series model fitting. To address the limitations of LC, which are attributed to the two-stage estimation and insufficient model fitness to data, single-stage procedures using the Bayesian state-space (BSS) approaches and extensions of flexibility in modeling by NNs have been proposed. As a fusion of these two approaches, we propose a NN extension of LC with a variational autoencoder that performs the variational Bayesian estimation of a state-space model and dimensionality reduction by autoencoding. Despite being a NN model that performs single-stage estimation of parameters, our model has excellent interpretability and the ability to forecast with confidence intervals, as with the BSS models, without using Markov chain Monte Carlo methods.

KEYWORDS

Lee–Carter model, state-space model, variational autoencoder, variational Bayesian inference.

1. INTRODUCTION

Longevity risk management and economic valuation of insurance and pension liabilities, which have recently received considerable attention, require statistical mortality models that provide stable long-term predictions even for single populations with limited data. The Lee–Carter (LC) model (Lee-Carter, 1992)

Astin Bulletin 52(3), 789–812. doi:10.1017/asb.2022.15

is a basic statistical mortality model with desirable properties, wherein the log mortality is determined by the sum of a fixed age impact and a product of time index and age sensitivity.

The model fitting procedure of LC originally comprises two stages: singular value decomposition (SVD)-based dimensionality reduction of the log-mortality data and fitting of a time series model to the obtained time components. Various extensions of LC have been proposed in the literature, for example, a Poisson regression version of LC by Brouhns *et al*. (2002), a cohort extension of LC by Renshaw and Haberman (2006), and a two-factor period effect model by Cairns *et al*. (2006). Cairns *et al*. (2009) provides a quantitative comparison of eight stochastic mortality models including LC.

Although LC achieves good interpretability and ease of estimation using SVD and drifted random walk (RW), it has a limitation in its accuracy primarily because of two issues: incoherency between parameters because of the two-stage estimation and insufficient fitting to the nonlinearity of data.

To address the first issue, various single-stage estimations of LC in Bayesian settings, which are implemented by Markov chain Monte Carlo (MCMC) methods, have been proposed. Czado *et al*. (2005) considers a single-stage Bayesian estimation of LC in a Poisson regression form and that in the state-space form was considered by Pedroza (2006), Kogure and Kurachi (2010), Cairns *et al*. (2011), and Fung *et al*. (2017). Notable extensions of the Bayesian LC model are proposed by Kogure and Kurachi (2010) in a risk-neutral form and Cairns *et al*. (2011) in a multi-population form. Moreover, Fung *et al*. (2017) introduces a general Bayesian state-space (BSS) modeling framework of the extensions of LC, which allows stochastic volatility in the period effect.

To address the second issue, many nonlinear extensions of LC using neural network (NN) methods have been proposed. NNs are typically defined by a network structure comprising multiple layers of neurons and an activation function that outputs a transformation of the weighted input to each neuron. As Cybenko (1989) demonstrates that any compactly supported continuous function can be uniformly approximated by a two-layer NN with a continuous sigmoidal activation function, NNs have a universal approximation capability for any function in a broad function class. The most basic NNs are feedforward NNs (FNNs) that have no cyclic connections, whereas NNs that have cyclic connections are called recurrent NNs (RNNs). NNs are also classified in supervised and unsupervised. Furthermore, convolutional NN (CNN), a sparse connected FNN to learn the neighborhood effect of the data, and RNN are often used for learning sequential data.

Richman and Wüthrich (2021) proposes a NN-based generalization of LC using a fully connected network (FCN) with multiple hidden layers, which is called deep FCN. Perla *et al*. (2021) considers many supervised NN extensions of LC and shows the superiority of one-dimensional (1D) CNN over deep FCN and long short-term memory (LSTM), a RNN suitable for learning long sequential data. Wang *et al*. (2021) proposes mortality forecasts using two-dimensional (2D) CNN to capture the neighborhood effect of the mortality

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

data. Schnürch and Korn (2021) also considers 2D CNN and achieves mortality forecasts with confidence intervals. These NN approaches calibrate their models with huge training data such as the data of all countries of the Human Mortality Database (HMD; http://www.mortality.org). The huge training data will contribute to the stability and accuracy of predictions as a countermeasure to the seed robustness and overlearning problems common in NN approaches. While these NN approaches achieve single-stage estimation with relatively high prediction accuracy, they lose interpretability of the model and are not necessarily suitable for relatively small training data (e.g., single population data).

On the other hand, Hainaut (2018) proposes a replacement of SVD with an unsupervised NN for dimensionality reduction, NN-analyzer which is more commonly known as an autoencoder (AE), and Nigri *et al*. (2019) proposes an application of LSTM to time components obtained from SVD, both of which have the interpretability of the model, but with the limitation of two-stage estimation. Thus, the existing NN approaches for the mortality prediction are subject to the problem of either a two-stage estimation or loss of interpretability.

To achieve a single-stage estimation of the parameters without losing interpretability, we introduce a variational AE (VAE) proposed by Kingma and Welling (2013) to the mortality prediction; our method implies a fusion of NN and Bayesian approach. VAE, one of the representative generative NNs (i.e., NNs for generating new data by learning features of training data), performs the variational Bayesian estimation of the state-space model and the AE-based dimensionality reduction simultaneously.

The rest of this study is organized as follows. Section 2 discusses how the existing approaches extend the original LC model and the limitations. Section 3 presents an overview of the VAE approach. We propose a model in a generalized state-space form in Section 4 and discuss how to apply the VAE algorithm to the inference of the proposed model in Section 5. In Section 6, we apply our model to the data from the HMD and present numerical results including the calibration procedure of the model, performance comparison with LC, parameter comparison with LC to show the interpretability of the model, forecasts with confidence intervals, and remarks on the effects of changing the number of neurons in the model. Finally, Section 7 concludes this study.

2. Extensions of LC

LC defines the log-mortality rate at age x in calendar year t as follows:

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t, \tag{2.1}$$

where α_x is the average log mortality at age x measured over the observation period, and the bilinear term $\beta_x \kappa_t$ can be interpreted as the product of

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

age-specific sensitivity factor and year-specific mortality improvement factor. The age-specific factor β_x and the year-specific factor κ_t are obtained by the SVD of the log-mortality data net of the age-specific averages $\{\alpha_x\}$, where the following model identification constraints are required:

$$\begin{cases} \sum_x \beta_x = 1 \\ \sum_t \kappa_t = 0. \end{cases} \tag{2.2}$$

The prediction of future mortality in LC can be performed via a two-stage procedure: obtaining $\{\kappa_t\}$ by SVD and then fitting a time series model to $\{\kappa_t\}$. It is common to apply a drifted RW model determined by:

$$\kappa_t = \kappa_{t-1} + \mu + \epsilon_t, \tag{2.3}$$

where ϵ_t follows an *iid* (i.e., independent and identically distributed) standard normal distribution.

The interpretability of LC has given rise to various extensions, for example, Renshaw and Haberman (2006) (RH) makes a cohort extension by adding a new term $\beta'_x \kappa'_{t-x}$ on the right-hand side of Equation (2.1).

However, LC is known to have limitations in estimation accuracy, primarily because of incoherency among variables caused by the two-stage estimation and the limitation of nonlinear representation capability of the bilinear form.

For the first issue, single-stage estimations of LC in Bayesian settings have been proposed. A direct translation of LC into a state-space form in Pedroza (2006) is given by:

$$\begin{aligned} &\text{Observation equation:} \quad x_t = \alpha + \kappa_t \beta + \varepsilon_t, \varepsilon_t \sim iid\ N\left(0, \sigma_\varepsilon^2 I\right), \\ &\text{State equation:} \quad \kappa_t = \kappa_{t-1} + \mu + \epsilon_t, \epsilon_t \sim iid\ N\left(0, \sigma_\epsilon^2\right), \end{aligned} \tag{2.4}$$

where $x_t = \left(\log m_{t,0}, \log m_{t,1}, \ldots, \log m_{t,n}\right)^T$; $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_n)^T$; $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$; n denotes the maximum age observed. Here, MCMC is required to obtain posterior joint distributions for the parameters α, β, σ_ε, μ, and σ_ϵ. Because the MCMC is computationally intensive, the number of age categories to be estimated is often limited in many previous studies. Although many Bayesian extensions of LC have been proposed, they follow the bilinear form as in LC or RH, which results in limited nonlinear representation capabilities.

For the second issue, the extension of the nonlinear representation capabilities, NN approaches have been recently proposed; our study is in this context. The reason behind the application of NNs to the nonlinear extension of models is the universal approximation capability of NNs that is demonstrated by Cybenko (1989). NNs generally comprise an input layer, hidden layers, an output layer, neurons in each layer, links between the neurons in different layers, and activation functions. In an FNN, the d_{i+1}-dimensional vector y^{i+1} representing the output value of neurons in the i+1-th layer is determined by the d_i-dimensional output vector y^i in the previous layer, the activation function

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

ϕ, the weights $W^i \in \mathbb{R}_{d_{i+1} \times d_i}$, and the bias $b^i \in \mathbb{R}_{d_i+1}$ as follows:

$$y^{i+1} = \phi\left(W^i y^i + b^i\right). \tag{2.5}$$

NNs are trained to obtain weights W^i that minimize loss functions under a given network structure and activation function, typically using a gradient descent method (GDM).

Nigri *et al*. (2019) proposes a two-stage NN extension of LC in which the LSTM was applied to the extrapolation of time series components obtained by SVD. Perla *et al*. (2021) considers NN extensions of LC, which could perform single-stage estimations and demonstrates that 1D CNN outperformed FCN proposed by Richman and Wüthrich (2021) and LSTM. CNN, proposed by LeCun *et al*. (1990), is generally known as an effective method for 2D images and 3D spatial data; however, it has been recently used for 1D time series data. CNN replaces the product term $W^i y^i$ in Equation (2.5) with the convolution, as described below; it is often performed in three stages. In the first stage, the convolution with shared weights is performed; in the second stage, the value obtained by the convolution is nonlinearly transformed by an activation function; finally, a pooling function is used to output the value. The 1D CNN uses the convolution filter $W^{i,j} \in \mathbb{R}_{d \times m} (j = 1, \ldots, J)$ instead of the weight W^i in Equation (2.5), where $m \in \mathbb{N}$ denotes the kernel size of the filter and J denotes the number of filters. Then, the output data of layer i, $y^i \in \mathbb{R}_{d \times T}$, is transformed as follows:

$$y_{j,k}^{i+1} = \phi\left(\sum_{s=1}^{m}\sum_{l=1}^{d} W_{l,s}^{i,j} y_{l,k+s-1}^{i} + b^{i,j}\right);$$

$$k = 1, \ldots, T + 1 - m; y^{i+1} \in \mathbb{R}_{J \times (T+1-m)}. \tag{2.6}$$

After the above transformation, the pooling function is employed. Generally, the pooling function returns the maximum, minimum, or average value within each window region of the input data. For an h-dimensional CNN input, $y^i \in \mathbb{R}_{d \times T}$ and $y^{i+1} \in \mathbb{R}_{J \times (T+1-m)}$ are changed to $y^i \in \mathbb{R}_{h \times d \times T}$ and $y^{i+1} \in \mathbb{R}_{h \times J \times (T+1-m)}$. Wang *et al*. (2021) proposes mortality forecasts using two-dimensional (2D) CNN to capture the neighborhood effect of the mortality data. Schnürch and Korn (2021) also considers 2D CNN and achieves mortality forecasts with confidence intervals. The confidence intervals are not based on the endogenous randomness as in the BSS models, but on the exogenous randomness driven by the random seed for the NN, which corresponds to the model uncertainty.

Hainaut (2018) proposes a two-stage estimation of LC with a NN-based dimensionality reduction as an alternative to SVD. The NN algorithm called NN-analyzer in Hainaut (2018) can be classified as AE. Generally, AE is a NN that uses a low-dimensional hidden layer and learns such that the input data are close to the output of the AE reconstructed via the hidden layer and

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

can extract nonlinear features and characterize multidimensional data in low-dimensional components. The first part of AE, which outputs low-dimensional features from the original data, is called an encoder f^{enc}, and the second part, which reconstructs data from low-dimensional features, is called a decoder f^{dec}.

For the input data $X(t)$ at time t given by a vector of age-specific log-mortality rates net of the age-specific averages, the reconstructed data $\widehat{X}(t)$ by f^{dec} and the d-dimensional latent factor $\kappa_t^{nn} = \left(\kappa_t^{nn,1}, \ldots, \kappa_t^{nn,d}\right)$, the NN-analyzer is described as follows:

$$\begin{cases} \kappa_t^{nn} := f^{enc}(X(t)); \\ \widehat{X}(t) := f^{dec}\left(\kappa_t^{nn}\right). \end{cases} \tag{2.7}$$

The loss function is given by the squared error between $\widehat{X}(t)$ and $X(t)$, and the parameters of f^{enc} and f^{dec} are estimated using a GDM to minimize the loss function.

Moreover, the NN analyzer has the following features:

- It has a symmetric network structure with three hidden layers using hyperbolic tangent sigmoidal and identity function as activation functions for both f^{enc} and f^{dec}.
- The input and output data are evenly divided in subgroups, and each subgroup is exclusively connected to a specific neuron in the input and output layers, resulting in a sparsely connected AE (SAE), which is expected to prevent overlearning.
- A genetic algorithm is used to identify appropriate initial values for the GDM.

Finally, using the latent factors, the mortality model is expressed as follows:

$$\log m_t = \alpha + f^{dec}\left(\kappa_t^{nn}\right), \tag{2.8}$$

where m_t and the average mortality rate α are vectors of values for each age. The decoder term in Equation (2.8) gives a nonlinear generalization of the bilinear term of LC; however, the prediction requires extrapolating the latent factor κ_t^{nn} by a time series model, resulting in a two-stage estimation. We introduce VAE to perform the single-stage estimation of the AE-based extension of LC. For more theoretical background on NNs and AEs, we refer to Wüthrich and Merz (2022).

3. VARIATIONAL AUTOENCODER (VAE)

VAE, proposed by Kingma and Welling (2013), is a type of deep generative model with an AE structure that performs unsupervised learning

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

with dimensionality reduction to a latent space. VAE assumes a probability distribution for the latent space, unlike the conventional AE, and can be implemented without using MCMC techniques. The aim of VAE is to identify $p_\theta(x)$ that denotes the generative distribution of the data x with generative parameter θ; in the process, it can acquire the dimensionally reduced latent representation of data as a probability distribution. Assume that, for $T \in \mathbb{N}$, there exists a set of latent variables $Z = \{z_t\}_{t=1}^T$ that generates a sample dataset $X = \{x_t\}_{t=1}^T$ with probability $p_\theta(x_t|z_t)$ for each t, and z_t follows the prior distribution $p_\theta(z_t)$, where $p_\theta(z_t)$ and $p_\theta(x_t|\, z_t)$ follow probability distributions differentiable with respect to the parameters to be identified.

Because it is generally difficult to directly estimate the multidimensional posterior distribution $p_\theta(Z|X)$, a variational approximation with the parameter φ, $q_\varphi(Z|X)$, is used; the approximation is often implemented in the form of a mean field approximation via a factor decomposable distribution as follows:

$$q_\varphi(Z|X) = \prod_{t=1}^{T} q_\varphi(z_t|x_t)\,. \tag{3.1}$$

From the AE perspective, the approximate distribution $q_\varphi(z_t|x_t)$ and generative distribution $p_\theta(x_t|\, z_t)$ are considered probabilistic encoder and decoder, respectively. The generative parameter θ and variational parameter φ are learned as network parameters in VAE to maximize the evidence lower bound (ELBO) given by Equation (3.2). The meanings of the ELBO become clear by rewriting Equation (3.2) with the Kullback–Leibler (KL) divergence represented by $D_{KL}[||]$, which gives asymmetric distances between distributions, as shown in Equations (3.3) and (3.4). We also refer to Section 11.6.3 of Wüthrich and Merz (2022) for more details:

$$\text{ELBO} = \int q_\varphi(Z|X) \log \frac{p_\theta(Z, X)}{q_\varphi(Z|X)} dZ. \tag{3.2}$$

$$\text{ELBO} = \log p_\theta(X) - D_{KL}\big[q_\varphi(Z|X)||p_\theta(Z\,|X)\big]\,,$$

$$\text{where } D_{KL}\big[q_\varphi(Z|X)||p_\theta(Z\,|X)\big] = \int q_\varphi(Z|X) \log \frac{q_\varphi(Z|X)}{p_\theta(Z|X)} dZ. \tag{3.3}$$

$$\text{ELBO} = \int q_\varphi(Z|X) \log p_\theta(X|Z)\, dZ - D_{KL}\big[q_\varphi(Z|X)||p_\theta(Z)\big]\,,$$

$$\text{where } D_{KL}\big[q_\varphi(Z|X)||p_\theta(Z)\big] = \int q_\varphi(Z|X) \log \frac{q_\varphi(Z|X)}{p_\theta(Z)} dZ. \tag{3.4}$$

Equation (3.3) shows that the ELBO maximization implies maximizing the log-likelihood of the data with penalizing the approximation error given by $D_{KL}\big[q_\varphi(Z|X)|\,|p_\theta(Z|X)\big]$. From the perspective of AE, Equation (3.4) shows that the integral term of the ELBO indicates an expected value of the negative

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

reconstruction error of the data by the decoder $p_\theta(X|Z)$ because $\log p_\theta(X|Z)$ is closer to 0 for higher reconstruction probability and takes a greater negative value for lower reconstruction probability; the second (KL) term of the ELBO acts as a regularizer that penalizes the KL distance between the approximate posterior distribution $q_\varphi(Z|X)$ and the prior distribution $p_\theta(Z)$. Thus, the ELBO maximization simultaneously performs the data reconstruction error minimization, which is essential for AEs, and regularization. While the distributions are usually selected such that the KL term of Equation (3.4) can be analytically calculated, the first term is difficult to analytically calculate. Thus, a sampling approximation by z_t following $q_\varphi(z_t|x_t)$ and a reparameterization technique (called reparameterization trick) for z_t are required to optimize parameters using the GDM. The reparameterization is determined by a differentiable bivariate function of the data point x_t and an auxiliary noise variable ϵ_t as follows:

$$z_t = g_\varphi(x_t, \epsilon_t)\,. \tag{3.5}$$

In particular, assuming that the *iid* noise ϵ_t follows standard normal distribution, $g_\varphi(x_t, \epsilon_t)$ can be expressed as follows:

$$\begin{aligned} g_\varphi(x_t, \epsilon_t) &= \mu_t + \sigma_t \epsilon_t; \\ (\mu_t, \sigma_t) &= f_\varphi^{enc}(x_t)\,, \end{aligned} \tag{3.6}$$

where the parameters are given by a vector-valued function $f_\varphi^{enc}(x_t)$ in the encoder.

For $L \in \mathbb{N}$ denoting the number of samples from $g_\varphi(x_t, \epsilon_t)$, the sampling approximation of the first term is given by:

$$\int q_\varphi(z_t|x_t) \log p_\theta(x_t|\, z_t)\, dz_t \approx \frac{1}{L} \sum_{l=1}^{L} \log p_\theta\left(x_t|z_{l,t}\right). \tag{3.7}$$

4. The model

We propose a nonlinear extension of LC, written as a state-space model with a latent variable that follows a drifted RW process. For the data $\{x_t\}_{t=1}^T$ comprising vectors of age-specific log-mortality rates for each observation year t and latent variables $\{z_t\}_{t=1}^T$ for all observation years, their joint probability and graphical representation (Figure 1) are as follows:

$$p_{\theta,\xi}(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T) = \prod_{t=1}^{T} p_\theta(x_t|z_t) \prod_{t=1}^{T-1} p_\xi(z_{t+1}|z_t)\, p_\xi(z_1)\,. \tag{4.1}$$

Although a drifted RW is assumed for $p_\xi(z_{t+1}|z_t)$ as in the original LC model, we propose a generalization by replacing $\alpha + \kappa_t \beta$ in Equation (2.4)

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

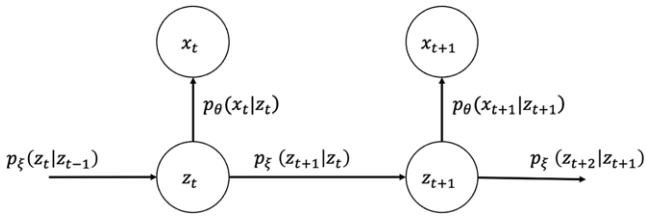

FIGURE 1. Graphical model representation of state-space LC.

with a nonlinear vector-valued function denoted by $f_\theta(z_t)$, where the function belongs to a broad nonlinear function class that can be obtained by a NN.

Using $x_t = \log m_t = (\log m_{t,0}, \log m_{t,1}, \ldots, \log m_{t,n})$, our model is given by:

$$\begin{aligned} \log m_t &= f_\theta(z) + \varepsilon_t, \\ z_t &= \mu_\xi + z_{t-1} + \sigma_\xi \eta_t, t > 1 \\ z_1 &= z_0 + \sigma_\xi \eta_1 \end{aligned} \tag{4.2}$$

where ε_t and η_t are *iid* noises that $\varepsilon_t \sim N(0, \sum_\theta)$; $\sum_\theta = diag(\sigma_{\theta 0}^2, \sigma_{\theta 1}^2, \ldots, \sigma_{\theta n}^2)$; $\eta_t \sim N(0,1)$.

Note that $(\sigma_{\theta 0}^2, \sigma_{\theta 1}^2, \ldots, \sigma_{\theta n}^2)$, (μ_ξ, σ_ξ) and the initial value z_0 can be obtained as learning parameters of a NN.

Assuming that $p_\xi(z_1)$ follows a normal distribution, the log-likelihood of the generative distribution to be maximized is given as follows:

$$\log p_\theta(x_t|z_t) = -\log\sqrt{(2\pi)^{n+1}\left|\sum_\theta\right|} - \frac{1}{2}(x_t - f_\theta(z_t))^{\mathrm{T}} \sum_\theta^{-1} (x_t - f_\theta(z_t)), \tag{4.3}$$

where $\left|\sum_\theta\right| = \sigma_{\theta 0}^2 \sigma_{\theta 1}^2 \ldots \sigma_{\theta n}^2$.

5. VAE FOR THE MODEL

In this section, a VAE is used to estimate the parameters of the state-space model, and the state-space model specification chosen will determine the expression for the loss function used when fitting the VAE to data.

5.1. The loss function

To apply the GDM for the variational inference, it is necessary to derive the loss function given by the sign-reversed ELBO of the model. The joint posterior distribution $p_\theta(z_1, z_2, \ldots, z_T|x_1, x_2, \ldots, x_T)$ with the generative parameter θ is approximated by $q_\varphi(z_1, z_2, \ldots, z_T|x_1, x_2, \ldots, x_T)$ with variational parameter φ where the approximation distribution q_φ is assumed to be factorizable as expressed in Equation (3.1). The structure of the VAE used for the variational

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

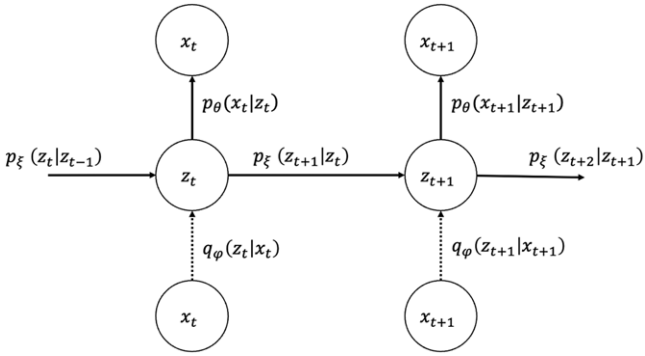

FIGURE 2. Graphical model representation of VAE for state-space LC (dotted line: approximation).

inference of the proposed model is represented in a graphical model (Figure 2) where the vertical network at each time t has an AE structure with encoder $q_\varphi(z_t|x_t)$ and decoder $p_\theta(x_t|\, z_t)$.

The ELBO to be maximized for the proposed model is given by:

$$\begin{aligned}\text{ELBO} &= \int \dots \int q_\varphi(z_1, \dots, z_T|x_1, \dots, x_T) \\ &\qquad \log\frac{p_{\theta,\xi}(x_1, x_2, \dots, x_T, z_1, z_2, \dots, z_T)}{q_\varphi(z_1, \dots, z_T|x_1, \dots, x_T)} dz_1 \dots dz_T \\ &= \sum_{t=1}^{T} \int q_\varphi(z_t|x_t) \log p_\theta(x_t|z_t)\, dz_t \\ &\quad - D_{kL}\big[q_\varphi(z_1|x_1)||\, p_\xi(z_1)\big] \\ &\quad - \sum_{t=1}^{T-1} \int q_\varphi(z_t|x_t)\, D_{kL}\big[q_\varphi(z_{t+1}|x_{t+1})\,||\, p_\xi(z_{t+1}|z_t)\big]\, dz_t, \end{aligned} \tag{5.1}$$

where the KL terms can be calculated analytically from the assumptions, given by Equation (3.6) and (4.2), as follows:

$$D_{KL}\big[q_\varphi(z_1|x_1)||p_\xi(z_1)\big] = \frac{(\mu_1 - z_0)^2}{2\sigma_\xi^2} + \frac{\sigma_1^2}{2\sigma_\xi^2} - \log\frac{\sigma_1}{\sigma_\xi} - \frac{1}{2},$$

$$D_{KL}\big[q_\varphi(z_{t+1}|x_{t+1})\,||p_\xi(z_{t+1}|z_t)\big] = \frac{\big(\mu_{t+1} - (\mu_\xi + z_t)\big)^2}{2\sigma_\xi^2} + \frac{\sigma_{t+1}^2}{2\sigma_\xi^2} - \log\frac{\sigma_t}{\sigma_\xi} - \frac{1}{2},\ t \geq 1.$$

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

Replacing z_t in Equation (5.1) with the sampling approximations $\{z_{l,t}\}$ that follows the reparametrized distribution $g_\varphi(x_t, \epsilon_t)$ given by Equation (3.6), the ELBO is approximated by:

$$\text{ELBO} \approx \frac{1}{L}\sum_{l=1}^{L}\left(\sum_{t=1}^{T}\log p_\theta\left(x_t|z_{l,t}\right) - D_{kL}\left[q_\varphi(z_1|x_1)\,||p_\xi(z_1)\right]\right.$$
$$\left. - \sum_{t=1}^{T-1} D_{KL}\left[q_\varphi\left(z_{l,t+1}|x_{t+1}\right)||p_\xi\left(z_{l,t+1}|z_{l,t}\right)\right]\right). \tag{5.2}$$

Finally, the loss function (i.e., sign-reversed ELBO) is given by:

$$-\frac{1}{L}\sum_{l=1}^{L}\left(\sum_{t=1}^{T}\left(\log p_\theta\left(x_t|z_{l,t}\right) - \frac{\left(\mu_t - \left(\mu_\xi + z_{l,t-1}\right)\right)^2}{2\sigma_\xi^2} - \frac{\sigma_t^2}{2\sigma_\xi^2} + \log\frac{\sigma_t}{\sigma_\xi} + \frac{1}{2}\right)\right), \tag{5.3}$$

where $z_{l,0} = z_0 - \mu_\xi$, and $\log p_\theta\left(x_t|z_{l,t}\right)$ is obtained by replacing z_t in Equation (4.3) with $z_{l,t}$. Note that $\left(\sigma_{\theta_0}^2, \sigma_{\theta_1}^2, \ldots, \sigma_{\theta_n}^2\right)$, (μ_ξ, σ_ξ), z_0 and (μ_t, σ_t) given by Equation (3.6) are learning parameters in the VAE network. Details of the derivation of the loss function are given in Appendix.

The mortality predictions can be obtained by decoding the projected state random variable $z_{t,}$ in the form of mean values or confidence intervals.

5.2. The network configuration

The proposed VAE has an FCN architecture comprising three hidden layers between the input and output layers; it has the following specifications:

- Although the numbers of neurons in the first and third hidden layers are variable, the number of neurons in the latent layer (second hidden layer) is fixed at 1 to be consistent with the dimension of the state random variable z_t.
- Both the encoder and decoder use hyperbolic tangent sigmoidal function and identity function as activation functions.
- The encoder (first hidden layer) includes the reparameterization unit given in Equation (3.6).
- A fixed age-impact vector corresponding to α in Equation (2.4) is used as a learning parameter that is deducted before encoding and added after decoding, whose initial value is given by a vector of the age-specific averages of the observed log-mortality rates.

The architecture is illustrated in Figure 3. The model is coded from scratch in Python without using optimization packages, and the codes for the most

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

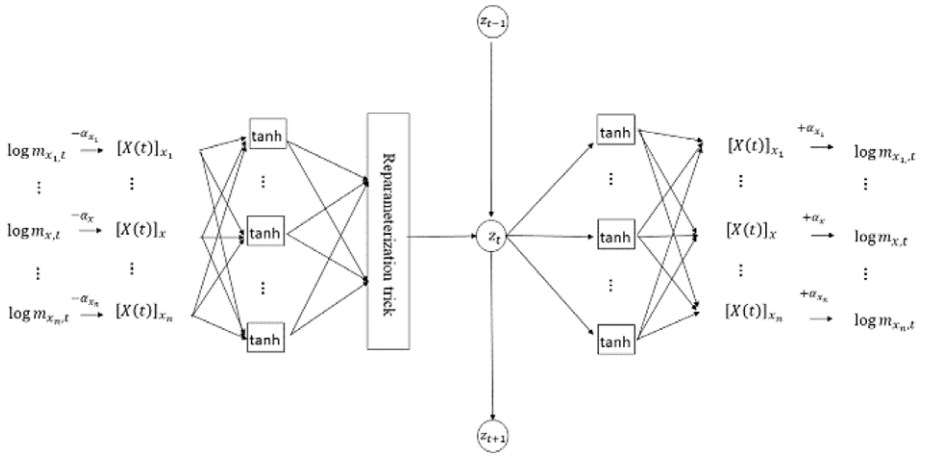

FIGURE 3. Network architecture of VAE.

characteristic part of the model are presented in the online Supplementary Material.

6. NUMERICAL APPLICATION

This section shows the results of applying the proposed model to the data from the HMD. The HMD data used here are the central mortality rates for ages 0–99 years, mixed gender, from 1961 to 2018. In this study, the observation period begins in 1961 to coincide with the introduction of the universal health insurance in Japan. For the accuracy evaluation, the data are divided into training data from 1961 to 2000 and test data from 2001 to 2018.

6.1. Calibration procedures

The hyperparameters to be determined for our model are the learning rate, the number of learning epochs, and the number of neurons in the first and third hidden layers. The number of neurons in the second hidden layer is fixed at 1 to be consistent with the dimension of the state variable of the model. The hyperparameters are selected to minimize the median of 10 minimum squared errors (MSEs), which correspond to 10 random seeds, over the validation data that consist of the latter 5 years (i.e., from 1996 to 2000) of the original training data; the selection universe of the hyperparameters contains the numbers of epochs from 5000 to 50,000 and the numbers of neurons in the three hidden layers from 10–1–10 to 50–1–50, where the configurations are limited to the symmetric type (i.e., the same number of neurons in the first and third hidden layers) common in AEs. Note that cross-validations are not available for time

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

TABLE 1

VALIDATED HYPERPARAMETERS FOR SIX COUNTRIES.

	Japan	US	Spain	Swiss	Canada	Denmark
Neurons	20-1-20	50-1-50	50-1-50	50-1-50	50-1-50	50-1-50
Epochs	25,000	10,000	45,000	30,000	15,000	45,000
Learning rate	0.00001	0.00001	0.00002	0.00003	0.00001	0.00002

TABLE 2

MSE COMPARISON BETWEEN VAE AND LC OVER TEST DATA FOR SIX COUNTRIES.

	Japan	US	Spain	Swiss	Canada	Denmark
LC	3.2695555	1.437645	11.57248	13.82407	2.050331	15.04826
VAE	1.4990569	1.380066	9.330155	9.814298	1.894726	14.39155

series models including our model. Since a larger number of epochs reduces the effect of the number of samples for the Monte Carlo integration of the loss function, the number L in Equation (5.3) is fixed at 10. The validated hyperparameters for six countries, including Japan and the United States (US), are given as follows.

The model, which is coded from scratch in Python without using optimization packages, requires a relatively large number of epochs as shown in Table 1, but the total run-time per country is about 10 min in the Google Colaboratory (https://colab.research.google.com) due to its shallow network structure. Although not used in this study, normalizing the input data to being centered with unit variance can reduce the number of epochs.

6.2. Performance comparison with LC

Using the hyperparameters given in Table 1, prediction performances are estimated in the same way over the test data. Table 2 summarizes the comparison of the prediction accuracy between LC (SVD+RW) and the VAE over the test data for six countries, where the accuracy measure is the same MSE as for the validation of the hyperparameters.

The observation period for the training data begins in 1961, when the universal health insurance was introduced in Japan, but longer observation periods are likely to improve forecasting accuracy in other countries. This model fits both in the context of the literature on BSS mortality models, where prediction accuracy is not often discussed, and in the context of the literature on NN models for mortality prediction, where prediction accuracy is the focus. As a NN model, this model focuses more on achieving features not found in

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

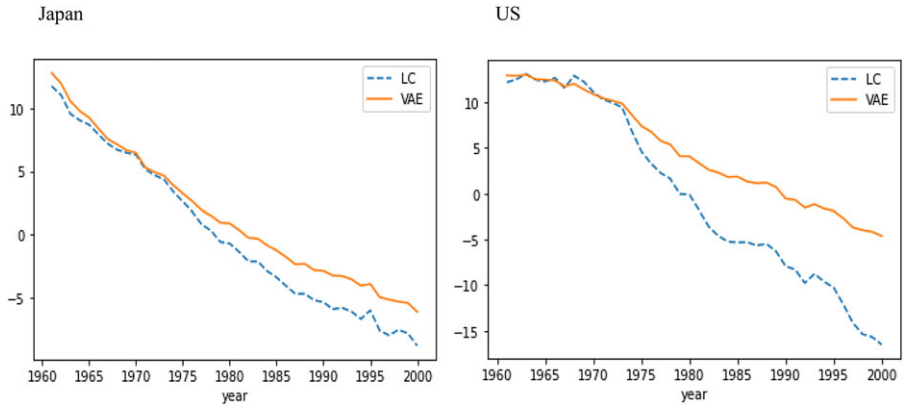

FIGURE 4. VAE's μ_t and LC's κ_t over training data for Japan (left) and US (right).

previous NN models than on prediction accuracy, as shown in the following sections. In the following sections, we will focus our analysis on Japan and US.

6.3. Interpretability of the model

In this section, the interpretability, one of the key features of the model, is demonstrated by the comparison of the components of the model with all parameters of LC over the training data for Japan and US, using the hyperparameters given in Table 1.

Figure 4 gives the comparison of LC's year-specific factor κ_t with μ_t, denoting the mean of the latent factor of the VAE generated by one random seed, over the training data for Japan and US. The descending curves of μ_t can be interpreted as indicating medical progress as well as the LC's year-specific factor κ_t. The slope of μ_t is more gradual than that of κ_t, for both Japan and US, indicating that excessive mortality reductions in the long-term predictions are less likely to occur than in LC.

Figure 5 compares LC's age sensitivity factor β and the decoder's sensitivity to μ_t, given by $\frac{d}{d\mu_t} f_\theta(\mu_t)$ for one random seed, for all ages in 1970, 1980, 1990, and 2000, over the training data for Japan and US. The number of humps in the decoder's sensitivity curves is roughly similar to that of LC's age sensitivity factor β, for both Japan and US, capturing country-specific characteristics.

Figure 6 shows that the learning parameter α of the VAE, generated by one random seed, is consistent with the fixed age factor α of LC, given by the averaged mortality, over the training data for Japan and US.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

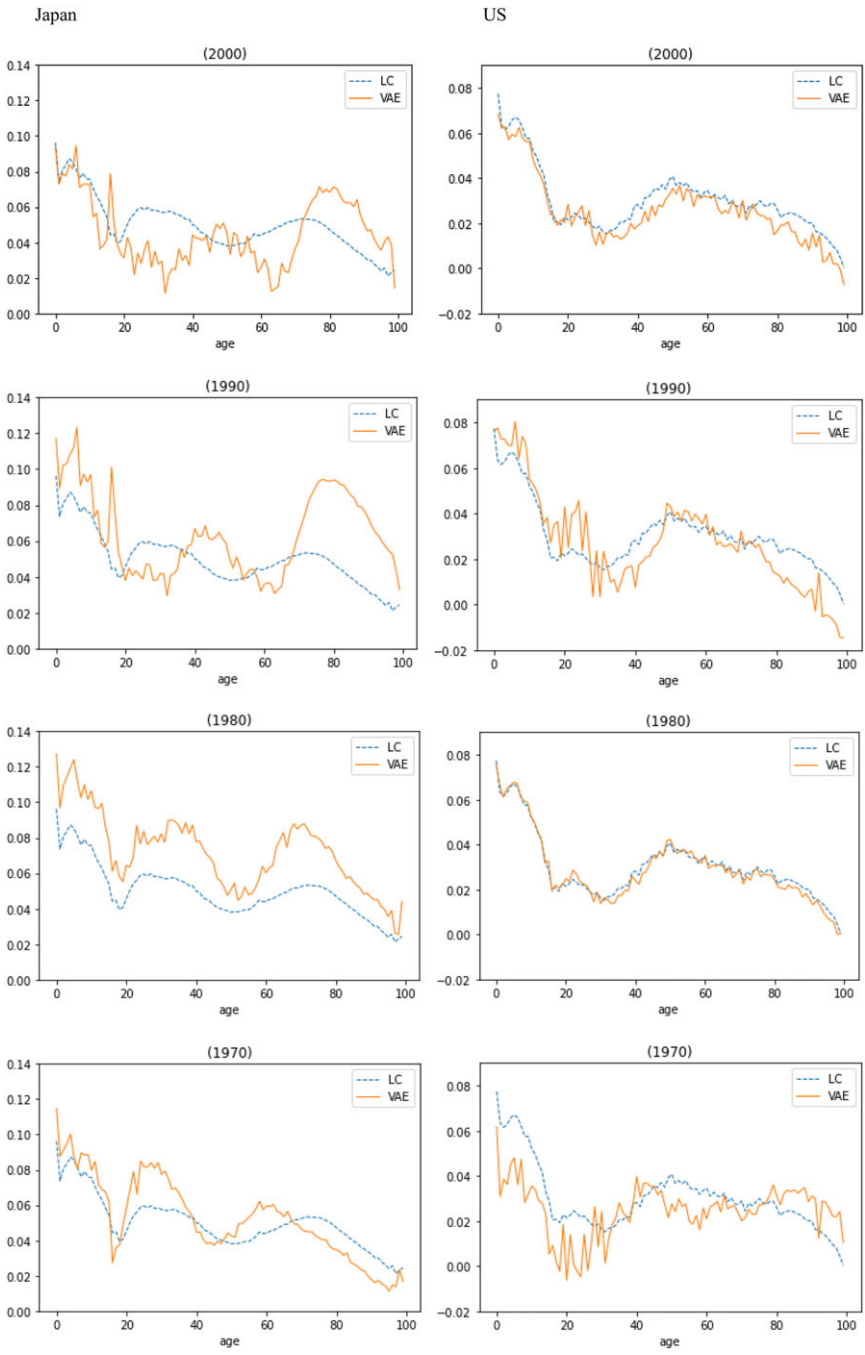

FIGURE 5. Decoder's sensitivity to μ_t and LC's β over training data for Japan (left) and US (right).

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

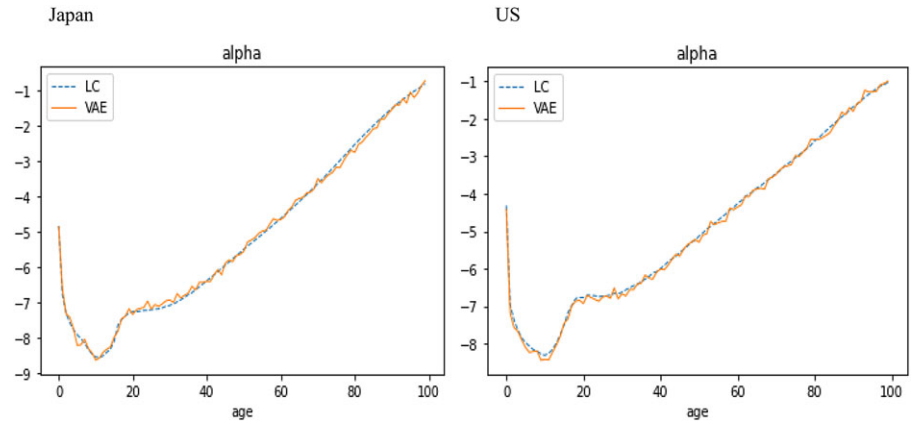

FIGURE 6. VAE's α and LC's α over training data for Japan (left) and US (right).

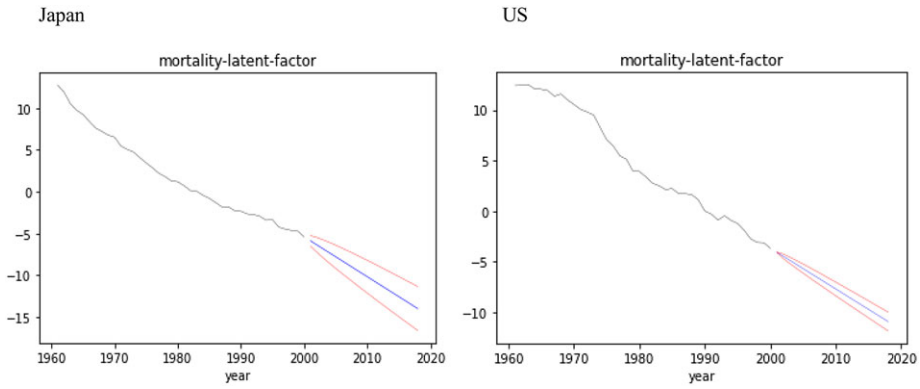

FIGURE 7. Forecasts with confidence intervals for latent factor z_t over test data for Japan (left) and US (right).

6.4. Forecasts with confidence intervals

Existing NN models for the mortality forecasts with confidence intervals are solely based on the exogenously given randomness such as random seed for NNs or added randomness for the time series models in two-stage estimations. In this section, we show that our model has an ability to forecast mortality with confidence intervals based on the endogenous randomness, as with BSS formulations, using the hyperparameters given in Table 1.

Figure 7 shows the forecasts with 97.5% confidence intervals for the latent factor z_t over the test data for Japan and US, appending μ_t over the training data. The natural connection between the mean μ_t of the latent variable encoded over the training data and the mean of the state model z_t estimated over the test data shows the effect of the variational approximation.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

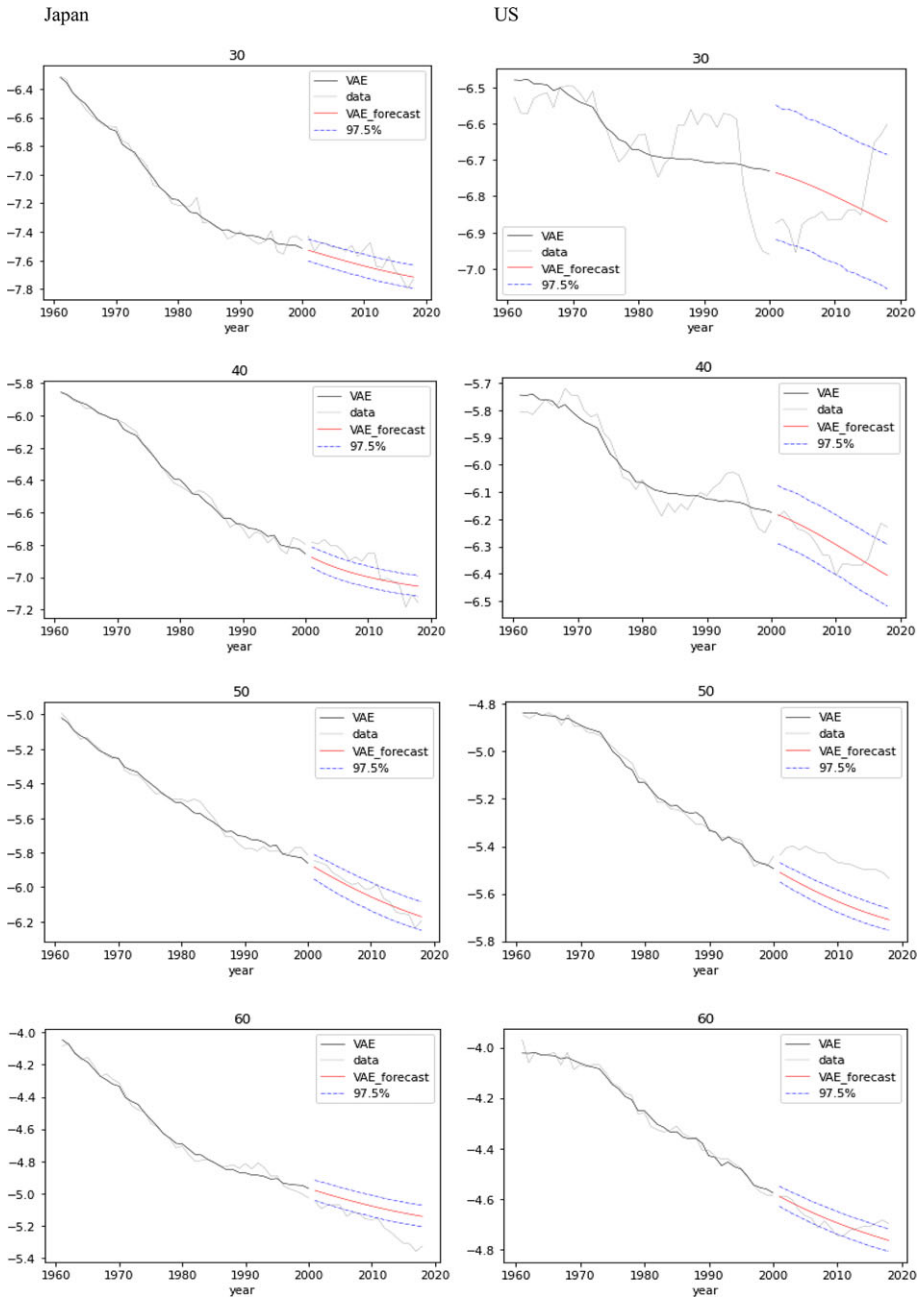

FIGURE 8. Forecasts with confidence intervals for mortality by age (from 30 to 80 years) over test data for Japan (left) and US (right).

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

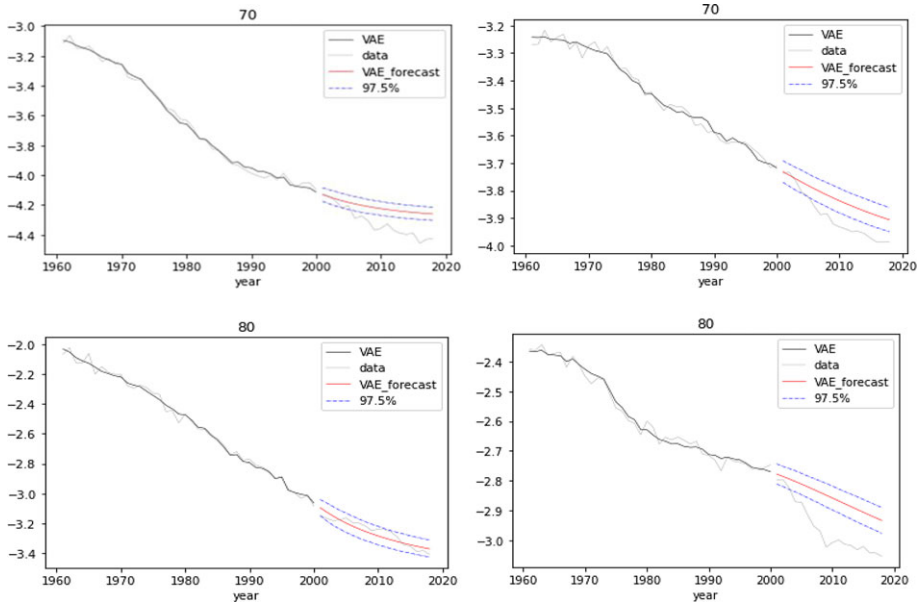

FIGURE 8. Continued.

Figure 8 gives mortality forecasts with 97.5% confidence intervals at ages 30, 40, 50, 60, 70, and 80 years over the test data for Japan and US, appending the actual mortality and the reconstructed mortality over the training data.

6.5. Remarks on changing number of neurons

The prediction accuracy of the VAE model is generally improved by increasing the number of neurons in the first and third hidden layers. Table 3 summarizes the comparison of the prediction accuracy by MSE of LC and VAE models (from 10–1–10 to 50–1–50) over the test data for Japan and US, using a fixed number of epochs per country. The results demonstrate that the number of neurons in a VAE that can outperform LC varies among countries and that if the number of neurons is extremely large, the prediction accuracy decreases because of overlearning.

We also consider the smoothness of mortality curves in ultralong-term predictions desirable in the economic valuation of insurance/pension liabilities and longevity risk management.

Figure 9 shows the 50-year predictions of Japanese mortality by VAE (from 10–1–10 to 50–1–50), where the dark-colored curves are the predictions. For the 50-year projection in Japan, all data (i.e., from 1961 to 2018) are used for the training data.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

TABLE 3

MSE COMPARISON OF LC AND VAE (FROM 10-1-10 TO 50-1-50) OVER TEST DATA FOR JAPAN AND US.

Model	Japan	US
SVD+RW(LC)	3.2695555	1.4376452
VAE(10-1-10)	**2.5506942**	1.7537529
VAE(20-1-20)	**1.4990569**	1.4427476
VAE(30-1-30)	**1.3890170**	**1.3534000**
VAE(40-1-40)	**1.3699607**	**1.3851189**
VAE(50-1-50)	**1.4554506**	**1.4325700**

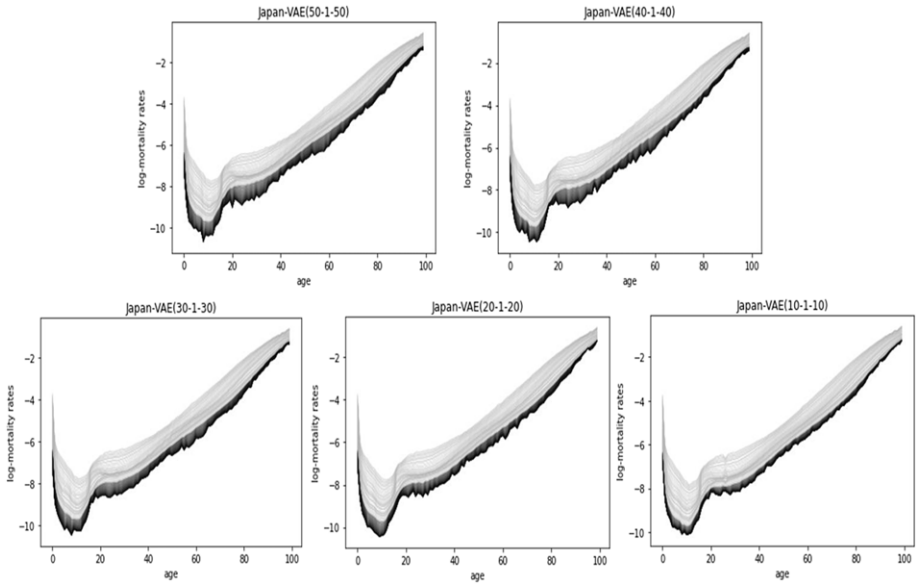

FIGURE 9. Fifty-year predictions by VAE (from 10–1–10 to 50–1–50), Japan.

Improving the prediction accuracy of VAE trades-off with the smoothness of the prediction curves, but introducing an asymmetric configuration into the hidden layers of VAE can improve the smoothness of prediction while maintaining high prediction accuracy. It is effective to increase and decrease the numbers of neurons in the first and third hidden layers, respectively.

For example, VAE (50–1–3) yields relatively smooth prediction curves as shown in Figure 10, and the prediction accuracy (MSE:3.0416) remains better than LC (MSE:3.2696) on the same data as for Table 3.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

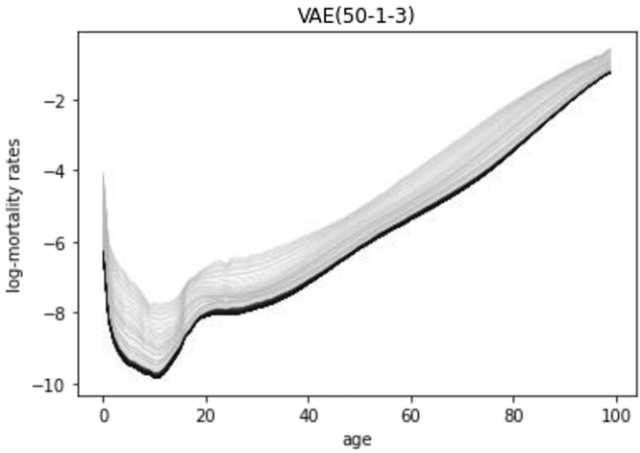

FIGURE 10. Fifty-year prediction by VAE(50-1-3), Japan.

7. CONCLUSIONS

This paper proposes a NN-based generalization of LC using VAE that performs mortality forecasts with confidence intervals based on the endogenous randomness (i.e., not by seed randomness for NNs), as with the BSS models, in a single-stage procedure without losing interpretability of the model. Our model fills a gap in the literature of NN extensions of LC, since previous NN models either enable single-step estimation of parameters but lose interpretability of the model, or retain interpretability but estimate parameters in two steps. The model also can yield relatively smooth mortality curves in long-term predictions due to the dimensionality reduction capability of the VAE.

However, our model has the limitations that it employs a 1D RW with *iid* residuals for the latent state model, and thus the number of neurons in the second hidden layer is limited to one; the limitations are intended to avoid sampling approximations of multiple integrals that reduce estimation efficiency and often require MCMC. Dimensional extensions of the latent state model (i.e., multiple neurons in the second hidden layer) and the introduction of the non-*iid* residuals to the latent state model are future work; they can improve representation capabilities of the model and may allow its extension to cohort and multiple population models. However, if one has multiple neurons in the second hidden layer and may have to deal with an identifiability issue, see Example 7.28 in Wüthrich and Merz (2022).

ACKNOWLEDGMENTS

The authors thank the anonymous referees for valuable comments which helped to improve the paper substantially.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://doi.org/10.1017/asb.2022.15.

REFERENCES

BROUHNS, N., DENUIT, M. and VERMUNT, J.K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**(3), 373–393.

CAIRNS, A.J., BLAKE, D. and DOWD, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *North American Actuarial Journal of Risk and Insurance*, **73**(4), 687–718.

CAIRNS, A.J., BLAKE, D., DOWD, K., COUGHLAN, G.D., EPSTEIN, D., ONG, A. and BALEVICH, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 1–53.

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**(1), 25–59.

CYBENKO, G. (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.

CZADO, C., DELWARDE, A. and DENUIT, M. (2005) Bayesian Poisson log–bilinear mortality projections. *Insurance: Mathematics and Economics*, **36**(3), 260–284.

FUNG, M.C., PETERS, G.W. and SHEVCHENKO, P.V. (2017) A unified approach to mortality modelling using state-space framework: Characterisation, identification, estimation and forecasting. *Annals of Actuarial Science*, **11**(2), 343–389.

HAINAUT, D. (2018) A neural-network analyzer for mortality forecast. *ASTIN Bulletin*, **48**(2), 481–508.

KINGMA, D.P. and WELLING, M. (2013) Auto-encoding variational Bayes. arXiv:1312.6114.

KOGURE, A. and KURACHI, Y. (2010) A Bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics*, **46**(1), 162–172.

LEE, R.D. and CARTER, L. (1992) Modeling and forecasting US mortality. *Journal of the American statistical association*, **87**(419), 659–671.

LECUN, Y., BOSER, B.E., DENKER, J.S., HENDERSON, D., HOWARD, R.E., HUBBARD, W.E. and JACKEL, L.D. (1990) Handwritten digit recognition with a back-propagation network. *Neural Information Processing Systems 2* (NIPS 1989), pp. 396–404.

NIGRI, A., LEVANTESI, S., MARINO, M., SCOGNAMIGLIO, S. and PERLA, F. (2019) A deep learning integrated Lee–Carter model. *Risks*, **7**(1), 33.

PEDROZA, C. (2006) A Bayesian forecasting model: Predicting U.S. male mortality. *Biostatistics*, **7**, 530–550.

PERLA, F., RICHMAN, R., SCOGNAMIGLIO, S. and WÜTHRICH, M.V. (2021) Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, **7**, 572–598.

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.

RICHMAN, R. and WÜTHRICH, M.V. (2021) A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, **15**(2), 346–366.

SCHNÜRCH, S., and KORN, R. (2021) Point and interval forecast of death rates using neural networks. *ASTIN Bulletin*, **52**(1), 333–360.

WANG, C.W., ZHANG, J. and ZHU, W. (2021) Neighbouring prediction for mortality. *ASTIN Bulletin*, **51**(3), 689–718.

WÜTHRICH, M.V. and MERZ, M. (2022) Statistical foundations of actuarial learning and its applications. Available at SSRN id=3822407.

AKIHIRO MIYATA
Graduate Student at AMS
Meiji University
Tokyo, Japan

NAOKI MATSUYAMA (CORRESPONDING AUTHOR)
Graduate School of Advanced Mathematical Sciences (AMS)
Meiji University
Tokyo, Japan
E-mail: ma2yama@meiji.ac.jp

APPENDIX

Derivation of the loss function

The joint distribution $p_{\theta,\xi}(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T)$ following Equation (4.2) satisfies

$$\log p_{\theta,\xi}(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T) = \sum_{t=1}^{T} \log p_\theta(x_t|z_t) + \sum_{t=1}^{T-1} \log p_\xi(z_{t+1}|z_t) + \log p_\xi(z_1)\,.$$

Using the above equation and the mean field approximation given by Equation (3.1), the ELBO of the model can be rewritten as follows:

$$\begin{aligned}
\mathrm{ELBO} &= \int \ldots \int q_\varphi(z_1, \ldots, z_T|x_1, \ldots, x_T) \\
&\quad \log \frac{p_{\theta,\xi}(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T)}{q_\varphi(z_1, \ldots, z_T|x_1, \ldots, x_T)} dz_1 \ldots dz_T \\
&= \int \ldots \int q_\varphi(z_1|x_1) \ldots q_\varphi(z_T|x_T) \left(\sum_{t=1}^{T} \log p_\theta(x_t|z_t) \right. \\
&\quad \left. + \sum_{t=1}^{T-1} \log p_\xi(z_{t+1}|z_t) + \log p_\xi(z_1) - \sum_{t=1}^{T} \log q_\varphi(z_t|x_t) \right) dz_1 \ldots dz_T \\
&= \int \ldots \int q_\varphi(z_1|x_1) \ldots q_\varphi(z_T|x_T) \left(\sum_{t=1}^{T} \log p_\theta(x_t|z_t) \right. \\
&\quad \left. - \log \frac{q_\varphi(z_1|x_1)}{p_\xi(z_1)} - \sum_{t=1}^{T-1} \log \frac{q_\varphi(z_{t+1}|x_{t+1})}{p_\xi(z_{t+1}|z_t)} \right) dz_1 \ldots dz_T
\end{aligned}$$

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

$$= \sum_{t=1}^{T} \int q_\varphi(z_t|x_t) \log p_\theta(x_t|z_t)\, dz_t - D_{KL}\big[q_\varphi(z_1|x_1)||p_\xi(z_1)\big]$$

$$- \sum_{t=1}^{T-1} \int q_\varphi(z_t|x_t)\, D_{KL}\big[q_\varphi(z_{t+1}|x_{t+1})||p_\xi(z_{t+1}|z_t)\big]\, dz_t, \tag{A1}$$

where, from the assumptions given by Equation (3.6) and (4.2), the components of the KL terms follow the normal distributions as follows:

$$q_\varphi(z_1|x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(z_1-\mu_1)^2}{2\sigma_1^2}},$$

$$p_\xi(z_1) = \frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-\frac{(z_1-z_0)^2}{2\sigma_\xi^2}},$$

$$q_\varphi(z_{t+1}|x_{t+1}) = \frac{1}{\sqrt{2\pi}\sigma_{t+1}} e^{-\frac{(z_{t+1}-\mu_{t+1})^2}{2\sigma_{t+1}^2}},$$

$$p_\xi(z_{t+1}|z_t) = \frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-\frac{(z_{t+1}-(\mu_\xi+z_t))^2}{2\sigma_\xi^2}}, \quad t \geq 1.$$

Then the KL terms in Equation (A1) can be calculated analytically as follows:

$$D_{KL}\big[q_\varphi(z_1|x_1)||p_\xi(z_1)\big] = \int q_\varphi(z_1|x_1) \log \frac{q_\varphi(z_1|x_1)}{p_\xi(z_1)} dz_t$$

$$= \frac{(\mu_1 - z_0)^2}{2\sigma_\xi^2} + \frac{\sigma_1^2}{2\sigma_\xi^2} - \log\frac{\sigma_1}{\sigma_\xi} - \frac{1}{2}, \tag{A2}$$

$$D_{KL}\big[q_\varphi(z_{t+1}|x_{t+1})||p_\xi(z_{t+1}|z_t)\big]$$

$$= \int q_\varphi(z_{t+1}|x_{t+1}) \log \frac{q_\varphi(z_{t+1}|x_{t+1})}{p_\xi(z_{t+1}|z_t)} dz_{t+1}$$

$$= \frac{\big(\mu_{t+1} - (\mu_\xi + z_t)\big)^2}{2\sigma_\xi^2} + \frac{\sigma_{t+1}^2}{2\sigma_\xi^2} - \log\frac{\sigma_t}{\sigma_\xi} - \frac{1}{2}, \quad t \geq 1. \tag{A3}$$

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press

Substituting (A2) and (A3) into Equation (A1) gives

$$\begin{aligned}\text{ELBO} = \sum_{t=2}^{T} \int q_\varphi(z_t|x_t) \Bigg(& \log p_\theta(x_t|z_t) - \frac{\left(\mu_t - \left(\mu_\xi + z_{t-1}\right)\right)^2}{2\sigma_\xi^2} \\ & - \frac{\sigma_t^2}{2\sigma_\xi^2} + \log \frac{\sigma_t}{\sigma_\xi} + \frac{1}{2} \Bigg) dz_t \\ & + \int q_\varphi(z_1|x_1) \left(\log p_\theta(x_1|z_1) - \frac{(\mu_1 - z_0)^2}{2\sigma_\xi^2} - \frac{\sigma_1^2}{2\sigma_\xi^2} + \log \frac{\sigma_1}{\sigma_\xi} + \frac{1}{2} \right) dz_1.\end{aligned}$$

Using the sampling values $\{z_{l,t}\}$ from $g_\varphi(x_t, \varepsilon_t)$, the sampling approximation of the ELBO is given by:

$$\begin{aligned}\text{ELBO} \approx \frac{1}{L} \sum_{l=1}^{L} \Bigg(\sum_{t=1}^{T} \Bigg(& \log p_\theta\left(x_t|z_{l,t}\right) - \frac{\left(\mu_t - \left(\mu_\xi + z_{l,t-1}\right)\right)^2}{2\sigma_\xi^2} \\ & - \frac{\sigma_t^2}{2\sigma_\xi^2} + \log \frac{\sigma_t}{\sigma_\xi} + \frac{1}{2} \Bigg) \Bigg),\end{aligned}$$

where $z_{l,0} = z_0 - \mu_\xi$.

Finally, the loss function of the model is given by the sign reversal of the approximation of the ELBO.

https://doi.org/10.1017/asb.2022.15 Published online by Cambridge University Press