Annals of Actuarial Science (2021), **15**, pp. 519–548
doi:10.1017/S1748499520000275

PAPER

Mortality forecasting using a Lexis-based state-space model

Patrik Andersson [1]* and Mathias Lindholm[2]

[1]Department of Statistics, Uppsala University, Uppsala, Sweden and [2]Department of Mathematics, Stockholm University, Stockholm, Sweden
*Corresponding author. E-mail: patrik.andersson@statistics.uu.se

(Received 29 November 2019; revised 11 May 2020; accepted 30 July 2020; first published online 11 September 2020)

Abstract

A new method of forecasting mortality is introduced. The method is based on the continuous-time dynamics of the Lexis diagram, which given weak assumptions implies that the death count data are Poisson distributed. The underlying mortality rates are modelled with a hidden Markov model (HMM) which enables a fully likelihood-based inference. Likelihood inference is done by particle filter methods, which avoids approximating assumptions and also suggests natural model validation measures. The proposed model class contains as special cases many previous models with the important difference that the HMM methods make it possible to estimate the model efficiently. Another difference is that the population and latent variable variability can be explicitly modelled and estimated. Numerical examples show that the model performs well and that inefficient estimation methods can severely affect forecasts.

Keywords: Non-linear non-Gaussian state-space models; Exponential family PCA; Stochastic approximation EM; Particle filter; Mortality forecasting; Hidden Markov model

1. Introduction

Understanding and forecasting mortality is an important part of demographic research and policy making, due to its connection to, for example, pensions, taxation and public health. A closely related area of application is within actuarial science and, in particular, life insurance. A first step in understanding mortality patterns is to construct a model describing observed death counts or mortality rates, or "force of mortality", across age groups ("period mortality") or within cohorts (indexed with respect to time of birth).

One of the earliest contributions to the area of mortality forecasting is the so-called "Gompertz law of mortality" (Gompertz, 1825), see also the survey Pitacco (2018) and the references therein for more on other mortality laws. A more recent important contribution to the area is the Lee–Carter model (Lee & Carter, 1992), where a log-linear multivariate Gaussian model is assumed for the mortality rates, across age groups and calendar time. The model is a factor model, where the factor loadings are given by the first component in a principal component decomposition, thus inducing dependencies across age groups as well as reducing the dimension of the problem. Concerning forecasting, the model assumes that the calendar time effect is governed by a one-dimensional Gaussian random walk with drift. Consequently, the Lee–Carter model treats the mortality rates as a stochastic process. An alternative, and very natural, interpretation of the Lee–Carter model is as a Gaussian hidden Markov model (HMM), see, for example, De Jong & Tickle (2006); Fung *et al.* (2017) for a discussion in a mortality context and Cappé *et al.* (2006); Durbin & Koopman (2012) for comprehensive introductions to HMMs (also known as state-space models.)

For a survey of various extensions of the Lee–Carter model, see Booth & Tickle (2008); Haberman & Renshaw (2011); Carfora *et al.* (2017) and the references therein. Another line of work is the Gaussian Bayesian extension of the Lee–Carter model treated in Pedroza (2006), where models with random drifts are discussed.

However, a major criticism of the Lee–Carter model is that it treats mortality by a two-step method; first, standard point estimates of mortality rates are obtained; and second, given these rates, a (discrete time) stochastic process is fitted. Compared to the MLE used in this paper, the two-step estimation in the Lee–Carter model ignores the uncertainty and estimation error from the first step, thus making the estimators inefficient. In the Lee–Carter model, it is also not possible to explicitly distinguish between the finite sample noise of the mortality estimates (Poisson) and the noise from the underlying latent variables. Thus, there may be a misattribution of variance that will affect the accuracy of the forecasts. Also this critique is addressed in the present paper. Other modelling approaches that discuss this issue are given in Brouhns *et al.* (2002); Ekheden & Hössjer (2014, 2015).

The main contribution of this paper is that a probabilistic model of mortality is introduced and that the MLE of the parameters in the model are computed using particle filter techniques. The model is simple; the main assumption is that mortality can be modelled as the first event of a Poisson process with intensity equal for all individuals of the same age in the same year. The intensity is described by a latent factor model where the latent variables are modelled as a Gaussian process in discrete time. Moreover, it is shown that this model corresponds to a model on a Lexis diagram where only the population level number of deaths and the exposure to risk is observed. These data are generally available on country level.

The model can be thought of as an HMM with non-Gaussian observations and therefore particle filters, in particular the forward filtering backward smoothing algorithm, can be used for the calculation of the posterior distribution of the latent variables. Stochastic approximation EM (SAEM) is then used to find estimates of the unknown parameters in the model.

The remainder of the paper is organised as follows: In section 2, the individual-level mortality model is introduced, and it is shown how this relates to the population level model. In section 3, the estimation of the model is discussed in detail. This includes the dimension reduction using exponential family principal component analysis (EPCA), particle filtering and smoothing together with the use of SAEM for the likelihood maximisation. In section 4, it is argued that model performance should be assessed with respect to death counts or scalings thereof. Therefore, model validation criteria which are based on (proper) scoring rules are discussed. They are applicable to both training and validation data, hence allowing for model selection based on predictive performance. In particular, an R^2-like measure defined in terms of deviance is introduced. In section 5, different ways of forecasting, depending on the application, are discussed. For illustration purposes, a number of examples based on Swedish and US data are given in section 6. It is seen that the forecasting performance is satisfactory with respect to both in-sample (training) and out-of-sample (validation) data. These numerical examples illustrate the importance of separating between the finite sample variation and the latent mortality rate process variation – for Swedish data (with smaller population), it is clearly seen that the majority of the variation stems from population variation. It is also illustrated that the Lee–Carter model in this situation will make an erroneous attribution of variation to the latent mortality rate process, which confirms the criticism of two-step methods that does not properly capture finite population dynamics.

2. Probabilistic Mortality Model

The probabilistic model that is introduced in the present paper is based on the population dynamics as it is summarised in a Lexis diagram, see Figure 1. On the horizontal axis is calendar year and on the vertical axis is age. An individual's life is represented by a 45° straight line. Since most individuals are not born on January 1, the time spent in each square will differ from individual

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

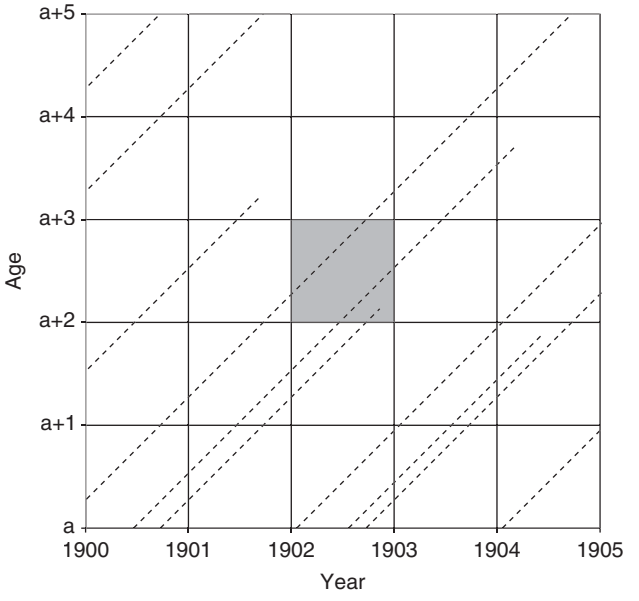

Figure 1. Example of a Lexis diagram.

to individual. As an example, consider the shaded area in Figure 1: The oldest individual, the one closest to the top-left corner of the shaded region is $a + 2$ years old at the beginning of calendar year 1902 and turn $a + 3$ during the year 1902. The second oldest individual will turn $a + 2$ during year 1902 and is alive at the end of the year 1902. The youngest individual will turn $a + 2$ during 1902 and will die before the end of 1902.

More specifically, with respect to mortality, the life of an individual, say i, can be characterised by the time of birth B_i and the time of death Q_i, where $0 \leq B_i \leq Q_i$, and time is measured in years.

Moreover, let $[\underline{t}, \overline{t}]$ be the time period when individuals are observed in the data set, and let n denote the total number of individuals that have been alive in $[\underline{t}, \overline{t}]$. That is, only individuals for whom $\left[\underline{t}, \overline{t}\right] \cap [B_i, Q_i] \neq \emptyset$ are considered. Further, the age of individual i at calendar time t is denoted by $A_i(t) := t - B_i, B_i \leq t \leq Q_i$.

The life history of individual i can be described by a counting process, $D_i(t) \in \{0, 1\}$, where 0 means that the individual is alive. The process can be defined in terms of a multiplicative intensity process (see Andersen *et al.* (1993); Aalen *et al.* (2008)): Let $m(a, t) \geq 0$, denote the hazard rate at age a and calendar time t. Using the introduced notation, the intensity process can be expressed as

$$\lambda_i(t) = m(A_i(t), t)\mathbb{1}(Q_i - B_i \geq A_i(t), B_i \leq t) = m(A_i(t), t)Y_i(t)$$

That is, $Y_i(t)$ is 1 if individual i is alive at t and is usually referred to as the "at-risk" indicator of individual i. That is, $Y_i(t) = 1 - D_i(t)$.

Hence, the total number of individuals who experience the event "death" up until t, $\underline{t} \leq t \leq \overline{t}$, denoted $D(t)$, is given by

$$D(t) = \sum_{i \in \mathcal{I}} D_i(t)$$

where $\mathcal{I}$ denotes the set of all individuals. This is a counting process with intensity process

$$\lambda(t) = \sum_{i \in \mathcal{I}} m(A_i(t), t)Y_i(t)$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Continuing, the main interest is to describe the process of deaths within yearly Lexis squares: The Lexis squares of interest are of the form

$$\mathcal{S}_{a,t} = [a, a+1) \times [t, t+1) \subset \mathbb{R}^2, \quad a, t \in \mathbb{N}, \quad \underline{t} \leq t \leq \overline{t}$$

Let $\mathcal{A}$ denote the set of relevant ages and let $\mathcal{T}$ denote the set of relevant calendar years. The collection of all relevant Lexis squares, $\bar{\mathcal{S}}$, is then given by

$$\bar{\mathcal{S}} := \{\mathcal{S}_{a,t} \mid (a,t) \in \mathcal{A} \times \mathcal{T}\}$$

A general lexis square, without specifying a and t, will be denoted as $\mathcal{S}$. Then, $D_i(t;\mathcal{S})$, which denotes the counting process which may register a single death for individual i in the Lexis square $\mathcal{S}$, has intensity process given by

$$\begin{aligned}\lambda_i(t;\mathcal{S}) &= m(A_i(t),t)Y_i(t)\mathbb{1}((A_i(t),t) \in \mathcal{S}) \\ &= m(A_i(t),t)Y_i\big(t;\mathcal{S}\big)\end{aligned}$$

where $Y_i(t;\mathcal{S}) := Y_i(t)\mathbb{1}\big((A_i(t),t) \in \mathcal{S}\big)$. The total number of observed death counts in $\mathcal{S}$ is given by the counting process

$$D(t;\mathcal{S}) = \sum_{i \in \mathcal{I}} D_i(t;\mathcal{S})$$

with intensity process

$$\lambda(t;\mathcal{S}) = \sum_{i \in \mathcal{I}} m(A_i(t),t)Y_i(t;\mathcal{S})$$

Moreover, due to that census data, typically, are only publicly available at integer ages and on yearly basis, the approach taken in the present paper is to model the hazard rates $m(a, t)$ as constants within yearly Lexis squares, that is, $m(a,t) = m_{\mathcal{S}}$ if $(a,t) \in \mathcal{S}$. That is, $D(t;\mathcal{S})$ has a multiplicative intensity process

$$\lambda(t;\mathcal{S}) = \sum_{i \in \mathcal{I}} m_{\mathcal{S}} Y_i(t;\mathcal{S})$$

Furthermore, by introducing $D_{\mathcal{S}}$, the stochastic number of deaths in $\mathcal{S}$, as

$$D_{\mathcal{S}} = \sum_{i \in \mathcal{I}} \mathbb{1}((Q_i - B_i, Q_i) \in \mathcal{S})$$

and the total amount of time that individuals have been alive in $\mathcal{S}$, the so-called "exposure to risk", $E_{\mathcal{S}}$, by

$$E_{\mathcal{S}} = \sum_{i \in \mathcal{I}} \int_{\underline{t}}^{\overline{t}} Y_i(t;\mathcal{S})\mathrm{d}t$$

it is possible to state the following lemma relating to observed data:

Lemma 2.1. *Assuming independence between individuals, the log-likelihood for the total population is,*

$$l(\mathcal{M}) = \sum_{\mathcal{S} \in \bar{\mathcal{S}}} (d_{\mathcal{S}} \log m_{\mathcal{S}} - e_{\mathcal{S}} m_{\mathcal{S}}) \tag{1}$$

where $\mathcal{M} = \{m_{\mathcal{S}} \mid \mathcal{S} \in \bar{\mathcal{S}}\}$ is the collection of unknown piecewise constant mortality rate parameters, $d_{\mathcal{S}}$ is the observed number of deaths and $e_{\mathcal{S}}$ is the observed exposure to risk in $\mathcal{S}$. This corresponds to the likelihood function of the Poisson distribution.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

For more on likelihood inference on Lexis diagrams, see Keiding (1991) and the references therein. Note that Lemma 2.1 and its derivation adjusts for partial information due to right censoring. The proof is given in Appendix C.

Now, consider the following probability model: For each $\mathcal{S}$, there is an independent Poisson process with constant intensity $m_{\mathcal{S}}$, running for time $e_{\mathcal{S}}$, during which $d_{\mathcal{S}}$ events are observed. The total log-likelihood of this model is equivalent to equation (1). Thus, by the likelihood principle, it is enough to consider this simpler model, where only the number of deaths and the exposure to risk in each Lexis square need to be observed, not the individual-level data. Also note that the model implied by (1) has no explicit dependence on the time of birth or death of specific individuals, since the exposure to risk summarises all this information. Thus, it is enough to have access to, for example, country-level mortality data. For later use, note that (1) gives the following maximum likelihood estimator (MLE) of $m_{\mathcal{S}}$

$$\widehat{m}_{\mathcal{S}} = \frac{d_{\mathcal{S}}}{e_{\mathcal{S}}} \tag{2}$$

In section 2.1, a state-space model is introduced, where $m_{\mathcal{S}}$ is treated as an unobservable, latent, stochastic process, $M_{\mathcal{S}}$, and the total number of deaths observed in $\mathcal{S}$ is Poisson distributed given $M_{\mathcal{S}} = m_{\mathcal{S}}$ and $e_{\mathcal{S}}$. A consequence of this modelling approach is that the latent $M_{\mathcal{S}}$ process is independent of population size. On the other hand, since $\widehat{m}_{\mathcal{S}}$ depends on the population size, these estimates will display more variation than that, typically, seen in the randomness of the latent $M_{\mathcal{S}}$ in itself. This effect is something that will be discussed further in section 6 where a numerical illustration is given.

2.1. Mortality model

In this section, the probabilistic model of mortality that will be used in our analysis will be defined. First, however, the notation will be introduced:

Number of age categories	k
Number of observation years	n
Number of factors	p
Number of deaths in $\mathcal{S}$	$D_{\mathcal{S}} \in \mathbb{N}_0$
Exposure to risk in $\mathcal{S}$	$e_{\mathcal{S}} \in [0, \infty)$
Death intensity in $\mathcal{S}$	$M_{\mathcal{S}} \in \mathbb{R}$
Factor loadings	$\Upsilon \in \mathbb{R}^{k \times p}$
Factor in year t	$X_t \in \mathbb{R}^p$
State transition matrix	$\Gamma \in \mathbb{R}^{p \times p}$
Transition covariance matrix	$\Sigma \in \mathbb{R}^{p \times p}$
Random mean in year t	$K_t \in \mathbb{R}^p$
Mean transition matrix	$\Gamma^K \in \mathbb{R}^{p \times p}$
Mean transition covariance matrix	$\Sigma^K \in \mathbb{R}^{p \times p}$
Mean level	$\mu \in \mathbb{R}^p$

All vectors are column vectors. Also, $D_t \in \mathbb{N}_0^k$ denotes the vector of number of deaths in year t, and similarly for e_t and m_t. The corresponding variables without subscripts are the matrices with the observation years in the columns and ages in the rows, for example, $D \in \mathbb{N}_0^{k \times n}$. The parameters Υ, Γ, Σ and μ will in general be unknown and are to be estimated. Also, Υ_a denotes the a:th row of Υ. The procedure for estimation is described in section 3.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

We define a Poisson model with linear Gaussian signal. For $t \in \{0, \ldots, n\}$,

$$\left.\begin{aligned} D_{\mathcal{S}} \mid M_{\mathcal{S}} &\sim \mathsf{Po}(e_{\mathcal{S}} M_{\mathcal{S}}) \\ M_{\mathcal{S}_{a,t}} &= \exp\{\Upsilon_a X_t\} \\ X_{t+1} &= \Gamma X_t + \mu + U_t,\ U_t \sim \mathsf{N}(0, \Sigma) \\ X_0 &\sim \mathsf{N}(\mu_0, \Sigma_0) \end{aligned}\right\} \tag{M1}$$

This is an HMM with non-Gaussian observations and linear Gaussian state equation. One can note that the dependence between ages is introduced by $M_{\mathcal{S}_{a,t}}$. That is, conditioned on $M_{\mathcal{S}_{a,t}}$, all ages are independent. Further, as argued in the previous section, under rather weak assumptions, a reasonable model for the number of deaths in a given year for a given age category is independent Poisson with intensity proportional to the exposure. The model specifies an exponential link function. It is certainly possible to choose a different link function and for the method described below, the exponential link function is not crucial. However, since this link function has been widely used in mortality studies going back to Lee & Carter (1992) (or even Gompertz (1825)) and also corresponds to the canonical link in terms of exponential families it is a natural choice. The matrix Υ contains the factor loadings associated with the time-varying factor scores X_t. This has two purposes: First, it seems intuitive that individuals of similar age at the same time should experience a similar mortality rate. Therefore, to model $M_{\mathcal{S}_{a,t}}$ independently for each a does not seem reasonable. Second, since we usually are concerned with a large number of ages (say about 100), it is impractical to estimate a mortality rate process for each age independently. Thus, Υ also provides a dimension reduction that simplifies the estimation problem and may be thought of as a non-parametric alternative to basis functions. The model for X_t is a linear Gaussian model, although non-linear models are certainly also possible to analyse, but they are not considered in this paper. However, even under the restriction of linear and Gaussian signals, (M1) should in many cases have enough flexibility so that it is possible to find a specific model that fits well with data.

For the purpose of the numerical illustration, a slight variation of Model (M1) will also be considered:

$$\left.\begin{aligned} D_{\mathcal{S}} \mid M_{\mathcal{S}} &\sim \mathsf{Po}(e_{\mathcal{S}} M_{\mathcal{S}}) \\ M_{\mathcal{S}_{a,t}} &= \exp\{\Upsilon_a X_t\} \\ X_{t+1} &= \Gamma X_t + K_t + \mu + U_t,\ U_t \sim \mathsf{N}(0, \Sigma) \\ K_{t+1} &= \Gamma^K K_t + V_t,\ V_t \sim \mathsf{N}(0, \Sigma^K) \\ X_0 &\sim \mathsf{N}(\mu_0, \Sigma_0) \\ K_0 &\sim \mathsf{N}(\mu_0^K, \Sigma_0^K) \end{aligned}\right\} \tag{M2}$$

Clearly (M1) is a special case of (M2). The difference is that in (M2) the drift is a stochastic process, while in (M1) it is a fixed parameter. Empirical evidence for including a random drift, in the context of mortality forecasts, is discussed in Pedroza (2006).

To make it easier to follow our numerical illustration, explicit formulas will, when necessary, be provided also for this model.

3. Model Fitting

The modelling approach taken in the present paper is based on a certain class of non-Gaussian HMMs, as described in section 2.1. In this section, the fitting of such models using maximum likelihood and particle filters is discussed. For easy comparison with the literature on HMMs, see, for example, Barber *et al.* (2011), we will adopt the standard notation $x_{0:n} = (x_0, \ldots, x_n)$.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

As defined in section 2.1, the following parameters are to be fitted: Υ, μ, Γ and Σ. Concerning, μ_0 and Σ_0, these will be set to deterministic values. For more on how this may be done, see the numerical illustration in section 6. Letting $\psi = (\mu, \Gamma, \Sigma)$, the *complete data likelihood* can be defined as

$$p_{\Upsilon,\psi}(x_{0:n}, d_{0:n}) = \nu(x_0) g_\Upsilon(d_0 \mid x_0) \prod_{t=1}^{n} f_\psi(x_t \mid x_{t-1}) g_\Upsilon(d_t \mid x_t) \tag{3}$$

where

$$\begin{cases} \nu(x_0) & = (2\pi)^{-\frac{p}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0)\right\}, \\ g_\Upsilon(d_t \mid x_t) & = \prod_{a=1}^{k} \exp\left\{-e_{a,t} \exp\left(\Upsilon_a x_t\right)\right\} \frac{\left(e_{a,t} \exp\left(\Upsilon_a x_t\right)\right)^{d_{a,t}}}{d_{a,t}!}, \\ f_\psi(x_t \mid x_{t-1}) & = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_t - \Gamma x_{t-1} - \mu)' \Sigma^{-1} (x_t - \Gamma x_{t-1} - \mu)\right\} \end{cases} \tag{4}$$

Note that $\nu(x_0)$ is the density of the starting point X_0, which we have assumed to be known. Since $x_{0:t}$ corresponds to unobservable, stochastic, state vectors, the likelihood function that we want to maximise is the one given by

$$p_{\Upsilon,\psi}(d_{0:n}) = \int p_{\Upsilon,\psi}(x_{0:n}, d_{0:n}) \mathrm{d}x_{0:n} \tag{5}$$

which in general is hard to evaluate. The present paper makes use of particle filter techniques and, in particular, the SAEM algorithm which is based on approximating (5) using simulation, see Cappé *et al.* (2006, Ch. 11.1.6). Apart from this, the SAEM procedure is closely related to the standard EM algorithm and will in this context correspond to iterating between sampling unknown states and updating of parameter estimates. A more detailed description of the particle filter techniques and sampling of unknown states is given in sections 3.2–3.4. Provided that the complete data likelihood will have low-dimensional sufficient statistics, the SAEM method can be described as a simple updating procedure in terms of these sufficient statistics. This is a nice feature of the method since it avoids the need to store all simulated trajectories. Therefore, before describing the SAEM technique in more detail, which is done in section 3.5, the properties of the complete data likelihood will be discussed.

First, one can note that the complete data likelihood from (3) may be written according to

$$p_{\Upsilon,\psi}(x_{0:n}, d_{0:n}) = g_\Upsilon(d_{0:n} \mid x_{0:n}) f_\psi(x_{0:n})$$

where

$$g_\Upsilon(d_{0:n} \mid x_{0:n}) := g_\Upsilon(d_0 \mid x_0) \prod_{t=1}^{n} g_\Upsilon(d_t \mid x_t) \tag{6}$$

$$f_\psi(x_{0:n}) := \nu(x_0) \prod_{t=1}^{n} f_\psi(x_t \mid x_{t-1}). \tag{7}$$

From the definition of $g_\Upsilon(d_{0:n} \mid x_{0:n})$, it is clear that $g_\Upsilon(d_{0:n} \mid x_{0:n})$ is a concave function in terms of Υ (see Lemma 3.2), but there is no low-dimensional statistic available for estimating Υ. Hence, the estimation of Υ is not suitable for inclusion in the SAEM algorithm. Note, however, that the role of Υ may be thought of as a non-parametric basis function used in order to introduce dependence across ages in X_t and to reduce the dimension of the problem. Thus, excluding the estimation of Υ from the SAEM algorithm and estimating Υ in isolation can be seen as conducting an EPCA. This is described in more detail in section 3.1. Consequently, the SAEM algorithm is used to estimate ψ by optimising $p_{\widehat{\Upsilon},\psi}(d_{0:n})$ via the corresponding complete data

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

likelihood $p_{\widehat{\Upsilon},\psi}(x_{0:n}, d_{0:n})$. Moreover, the Gaussian part of $p_{\widehat{\Upsilon},\psi}(x_{0:n}, d_{0:n})$, that is $f_\psi(x_{0:n})$, will produce estimators and low-dimensional statistics that can be written explicitly:

Lemma 3.1. *In both Models (M1) and (M2), the joint distribution of $x_{0:n}$ and $d_{0:n}$ defines a curved exponential family. The complete data maximum likelihood estimate, conditional on $x_{0:n}$ and $d_{0:n}$, can therefore be expressed in terms of low-dimensional sufficient statistics.*

The proof of Lemma 3.1 is given in Appendix C.2, where explicit formulas for the MLEs can be found. Recall from the beginning of section 3 that μ_0 and Σ_0 are being treated as constants, hence being outside of the estimation procedure. For more on how to assign values to μ_0 and Σ_0, see section 6.

3.1. Dimension reduction using EPCA – estimating Υ

As mentioned in sections 2.1 and 3, the matrix Υ may be thought of as a non-parametric choice of basis functions, that is, Υ is a matrix of factor loadings. The approach to estimate Υ in the present paper is closely connected to standard principal component analysis (PCA), but adapted to count data. The method that will be used consists of optimising the Poisson part of the complete data likelihood, that is, $g_\Upsilon(d_{0:n} \mid x_{0:n})$ from (6). Recall, from section 2.1, that both Υ and $x_{0:n}$ are unobservable quantities. One way of handling this is to estimate Υ and $x_{0:n}$ jointly, given $d_{0:n}$, by maximising $g_\Upsilon(d_{0:n} \mid x_{0:n})$. This is what is referred to as EPCA. There are many variations of EPCA, see Lu *et al.* (2016) for a review of some. Here the method introduced in Collins *et al.* (2001) is used (there called generalised PCA). One can, however, note the close connection to the approach taken in Brouhns *et al.* (2002), where a similar procedure is suggested within a Poisson GLM. A completely different interpretation of the approach from Collins *et al.* (2001) is to view the problem in a Bayesian setting and treat $x_{0:n}$ as having an improper ("flat") prior, that is, constant density, which is independent of Υ:

$$L_{\Upsilon,x_{0:n}}(d_{0:n}) \propto g_\Upsilon(d_{0:n} \mid x_{0:n})$$

Regardless of the interpretation of the objective function $g_\Upsilon(d_{0:n} \mid x_{0:n})$ in the EPCA optimisation, it is possible to show the following:

Lemma 3.2. *The function $-\log g_\Upsilon(d_{0:n} \mid x_{0:n})$ is convex in Υ given $x_{0:n}$, and convex in $x_{0:n}$ given Υ, but not jointly (globally) convex in both Υ and $x_{0:n}$.*

The proof of Lemma 3.2 is given in Appendix C. To see the effect of Lemma 3.2, one can note that for example $g_{\Upsilon/c}(d_{0:n} \mid cx_{0:n}) = g_\Upsilon(d_{0:n} \mid x_{0:n})$ for all $c \in \mathbb{R}_+$. Note that the above "marginal" convexity property corresponds to so-called "bi-convexity", see Gorski *et al.* (2007, Definition 1.1, 1.2). Moreover, in Gorski *et al.* (2007), conditions are given for when minimisation of a bi-convex function using so-called *alternate convex search* methods, which is a special case of *cyclic coordinate* methods, will converge, see (Gorski *et al.* 2007, Theorem. 4.7, 4.9, Corollary 4.10). For more on cyclic coordinate methods and convergence, see Bazaraa *et al.* (2006, Ch. 8.5).

Regarding the practical implementation, denote the fitted values by $\widehat{\Upsilon}$ and $\widehat{x}_{0:t}$ and also note the non-uniqueness of these. We therefore suggest to do as follows: Let $\widehat{\Sigma}$ denote the empirical $p \times p$-dimensional covariance matrix of $\widehat{x}_{0:t}$, and make a Cholesky factorisation of $\widehat{\Sigma}$ expressed in terms of $\widehat{\sigma}$, that is, $\widehat{\Sigma} = \widehat{\sigma}\widehat{\sigma}'$. Then, set

$$\widehat{\widehat{\Upsilon}} := \widehat{\Upsilon}\widehat{\sigma}$$

and

$$\widehat{\widehat{x}}_{0:t} := \widehat{\sigma}^{-1}\widehat{x}_{0:t}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

that is,

$$\widehat{\Upsilon}\widehat{x}_{0:t} = \widehat{\widehat{\Upsilon}}\widehat{\widehat{x}}_{0:t}$$

but $\widehat{\widehat{x}}_{0:t}$ will now have an empirical covariance which is a $p \times p$-dimensional identity matrix. Note that this procedure is a slight violation of the original EPCA optimisation: The suggested scaling does not affect the value of the loss function used in the optimisation, but it will affect its derivative. As a compromise, we propose to refit $\widehat{\Upsilon}$, given $\widehat{\widehat{x}}_{0:t}$. The advantage of this procedure is that it is natural to use the identity matrix as starting guess for Σ in the numerical optimisation that follow below.

Before ending this section, note that as opposed to classical PCA, the EPCA components are not orthogonal. Therefore, when increasing the number of components all the components may change.

3.2. Particle filtering

Having estimated Υ, as explained in the previous section, it remains to estimate ψ. Since the likelihood $p_\psi(d_{0:n})$ is not directly computable, approximations are needed. In the present paper, this will be done using simulation techniques, in particular, using particle filtering and smoothing. For more detailed accounts of these methods, see, for example, the book by Cappé *et al.* (2006) or the survey by Kantas *et al.* (2015).

In this section, it is assumed that all parameters are known, so that the task is to, for $0 \leq t \leq n$, find the *filtering distribution*

$$p(x_{0:t} \mid d_{0:t})$$

and, in the next section, the *smoothing distribution*

$$p(x_{0:t} \mid d_{0:n})$$

Note here that the difference between the filtering and smoothing distribution is up to what time it is conditioned, that is, the smoothing distribution assumes full knowledge of all available data – including future observations, seen from time t. The filtering distribution is the distribution of $x_{0:t}$ conditioned on the observations. On the other hand, the smoothing distribution conditions on all observations up to present time.

To start off, the filtering recursion equation can be written as,

$$p(x_{0:t} \mid d_{0:t}) = p(x_{0:t-1} \mid d_{0:t-1}) \frac{g(d_t \mid x_t) f(x_t \mid x_{t-1})}{p(d_t \mid d_{0:t-1})} \propto p(x_{0:t-1} \mid d_{0:t-1}) g(d_t \mid x_t) f(x_t \mid x_{t-1})$$

Further, assume that an approximation of $p(x_{0:t-1} \mid d_{0:t-1})$ of the form

$$\hat{p}(x_{0:t-1} \mid d_{0:t-1}) = \sum_{i=1}^{r} w_{t-1}^i \delta_{X_{0:t-1}^i}(x_{0:t-1}) \tag{8}$$

where w_{t-1}^i are the weights and where δ is the Kronecker-delta function, is available. Moreover, let $q(x_t \mid d_t, x_{t-1})$ denote an importance density function from which it is possible to draw samples from. It then follows that

$$\hat{p}(x_{0:t} \mid d_{0:t}) \propto \frac{g(d_t \mid x_t) f(x_t \mid x_{t-1})}{q(x_t \mid d_t, x_{t-1})} q(x_t \mid d_t, x_{t-1}) \hat{p}(x_{0:t-1} \mid d_{0:t-1})$$

As the above approximate recursion is iterated, the weights in (8) will be multiplied. Therefore, the variance of the method will increase rapidly with t. A partial remedy for this is to include an additional resampling step. That is, by introducing $\overline{X}_{0:t-1}$, denoting the random sample drawn from $\hat{p}(x_{0:t-1} \mid d_{0:t-1})$, the following approximation is obtained

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Algorithm 1. SISR

- At time $t = 0$, for all $i \in \{1, \ldots, r\}$:
  1. Sample: $X_0^i \sim q(\cdot \mid d_0)$.
  2. Compute: $w_0^i = \frac{g(d_0|X_0^i)v(X_0^i)}{q(X_0^i|d_0)}$.
  3. Resample: $\overline{X}_0^i \sim \sum_{i=1}^r w_0^i \delta_{X_0^i}(\cdot)$.
- At time $t \geq 1$, for all $i \in \{1, \ldots, r\}$:
  1. Sample: $X_t^i \sim q(\cdot \mid d_t, \bar{X}_{t-1}^i)$.
  2. Append: $X_{0:t}^i = (\bar{X}_{0:t-1}^i, X_t^i)$.
  3. Compute: $w_t^i = \frac{g(d_t|X_t^i)f(X_t^i|\bar{X}_{t-1}^i)}{q(X_t^i|d_t, \bar{X}_{t-1}^i)}$.
  4. Resample: $\overline{X}_{0:t}^i \sim \sum_{i=1}^r w_t^i \delta_{X_{0:t}^i}(\cdot)$.
- $\overline{X}_{0:t}^i$ is an approximate sample from $p(x_{0:t} \mid d_{0:t})$.

$$\hat{p}(x_{0:t-1} \mid d_{0:t-1}) = \sum_{i=1}^{r} \delta_{\overline{X}_{0:t-1}^i}(x_{0:t-1})$$

The recursion outlined above corresponds to the so-called sequential importance sampling resampling (SISR) algorithm, which is summarised in Algorithm 1. For more details concerning the derivation of this algorithm, see Cappé *et al.* (2006, Ch. 9.6) and Kantas *et al.* (2015).

Note that as a by-product of using the SISR algorithm, it follows that the likelihood may be estimated according to

$$\hat{p}(d_{0:n}) = \prod_{t=0}^{n} \frac{1}{r} \sum_{i=1}^{r} w_t^i \tag{9}$$

see Kantas *et al.* (2015).

3.3. Particle smoothing

In section 3.2, the SISR algorithm for obtaining the filtering distribution was described. Here one can recall that this algorithm was derived from Bayes' rule as a recursion going forward in time. Likewise, one could just as well consider similar recursive relationships based on that the time is *reversed*. This is what will be used in order to obtain the smoothing distribution,

$$p(x_{0:t} \mid d_{0:n})$$

First note that an application of Bayes' rule yields the following relation:

$$\begin{aligned} p(x_{0:n} \mid d_{0:n}) &= p(x_n \mid d_{0:n}) p(x_{0:n-1} \mid d_{0:n}, x_n) = p(x_n \mid d_{0:n}) p(x_{0:n-1} \mid d_{0:n-1}, x_n) \\ &= p(x_n \mid d_{0:n}) p(x_{n-1} \mid d_{0:n-1}, x_n) p(x_{0:n-2} \mid d_{0:n-2}, x_{n-1}) \\ &= p(x_n \mid d_{0:n}) \prod_{k=0}^{n-1} p(x_k \mid d_{0:k}, x_{k+1}) \end{aligned}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Algorithm 2. FFBS

- For $t = n$:
  1. Sample: $\widetilde{X}_n \sim \sum_{i=1}^{r} w_n^i \delta_{\widetilde{X}_n^i}(\cdot)$.
- For all $t = n-1, n-2, \ldots, 1$:
  1. Compute: $w_{t|t+1}^i \propto w_t^i f(\widetilde{X}_{t+1} \mid \bar{X}_t^i)$.
  2. Sample: $\widetilde{X}_t \sim \sum_{i=1}^{r} w_{t|t+1}^i \delta_{\bar{X}_t^i}(\cdot)$.
- $\widetilde{X}_{0:t}$ is an approximate sample from $p(x_{1:t} \mid d_{1:n})$.

where

$$p(x_k \mid d_{0:k}, x_{k+1}) = \frac{f(x_{k+1} \mid x_k) p(x_k \mid d_{0:k})}{p(x_{k+1} \mid d_{0:k})} \propto f(x_{k+1} \mid x_k) p(x_k \mid d_{0:k})$$

Recall from section 3.2 that Algorithm 1 produces an approximation of $p(x_k \mid d_{0:k})$. Thus, a combination of these observations suggests Algorithm 2 for sampling from the approximate smoothing distribution, which is the forward filtering backward sampling (FFBS) algorithm from Godsill *et al.* (2004).

3.4. Choosing the importance distribution

Recall that the particle filter algorithms, Algorithms 1 and 2, assume that there is an importance distribution $q(x_t \mid d_t, x_{t-1})$ from which it is possible to draw random samples. How to choose such a distribution is what will be discussed next. In order for Algorithms 1 and 2 to have small variances, the importance distribution should be chosen to be a close approximation of $g(d_t \mid x_t) f(x_t \mid x_{t-1})$. One way of doing this is as follows: Recall that as a by-product of the EPCA estimation of $\widehat{\Upsilon}$, an estimated state vector $\widehat{x}_{0:n}$ is produced. Given the estimated state vector, one can make a second-order Taylor expansion of $\log g(d_t \mid x_t)$ in x_t around $\widehat{x}_t$. For model (M1), this approach results in the following approximation

$$\log g(d_t \mid x_t) \approx \log g(d_t \mid \hat{x}_t) + \frac{1}{2}(x_t - \widehat{x}_t)' H_t (x_t - \widehat{x}_t) \propto \frac{1}{2}(x_t - \widehat{x}_t)' H_t (x_t - \widehat{x}_t) \tag{10}$$

where "$\propto$" corresponds to removing normalisation constants not depending on x_t, and where $H_t := H_t(d_t, \widehat{x}_t, \widehat{\Upsilon})$ denotes the Hessian of $\log g(d_t \mid \cdot)$, evaluated at $\widehat{x}_t$, which typically is obtained as a by-product from the EPCA optimisation. The first-order term is 0 since $\widehat{x}_t$ is obtained as the optimal value of the EPCA algorithm. Further, note that from section 3.1, it follows that $-H_t$ is positive semi-definite. Thus, (10) is the un-normalised log density, with x_t as argument, of a multivariate Gaussian distribution with mean $\widehat{x}_t$ and covariance $(-H_t)^{-1}$. Finally, by combining the above, the approximation of $\log\left(g(d_t \mid x_t) f(x_t \mid x_{t-1})\right)$ becomes

$$\begin{aligned}
&\log\left(g(d_t \mid x_t) f(x_t \mid x_{t-1})\right) \\
&\quad \approx \log g(d_t \mid \widehat{x}_t) - \frac{1}{2}(x_t - \widehat{x}_t)'(-H_t)(x_t - \widehat{x}_t) - \frac{1}{2}(x_t - \Gamma x_{t-1} - \mu)' \Sigma^{-1} (x_t - \Gamma x_{t-1} - \mu) \\
&\quad \propto -\frac{1}{2}\left(x_t - \left(-H_t \widehat{x}_t + \Sigma^{-1}(\Gamma x_{t-1} + \mu)\right)\right)' \left(-H_t + \Sigma^{-1}\right) \left(x_t - \left(-H_t \widehat{x}_t + \Sigma^{-1}(\Gamma x_{t-1} + \mu)\right)\right) \\
&\quad \propto \log\left(q(x_t \mid d_t, x_{t-1})\right)
\end{aligned}$$

where $q(x_t \mid d_t, x_{t-1})$ is the density of a multivariate Gaussian distribution with mean $-H_t \widehat{x}_t + \Sigma^{-1}(\Gamma x_{t-1} + \mu)$ and covariance $(-H_t + \Sigma^{-1})^{-1}$.

Analogously, for Model (M2), the approximation of $\log\left(g(d_t \mid x_t) f(x_t, k_t \mid x_{t-1}, k_{t-1})\right)$ instead becomes

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

$$
\begin{aligned}
&\log\left(g(d_t \mid x_t) f(x_t, k_t \mid x_{t-1}, k_{t-1})\right) \\
&\quad \approx \log g(d_t \mid \widehat{x}_t) - \frac{1}{2}(x_t - \widehat{x}_t)'(-H_t)(x_t - \widehat{x}_t) \\
&\qquad - \frac{1}{2}(x_t - \Gamma x_{t-1} - k_{t-1} - \mu)'\Sigma^{-1}(x_t - \Gamma x_{t-1} - k_{t-1} - \mu) \\
&\qquad - \frac{1}{2}(k_t - \Gamma^K k_{t-1})'(\Sigma^K)^{-1}(k_t - \Gamma^K k_{t-1}) \\
&\quad \propto -\frac{1}{2}\left(\begin{pmatrix} x_t \\ k_t \end{pmatrix} - \begin{pmatrix} \nu_t \\ \nu_t^K \end{pmatrix}\right)' \begin{pmatrix} -H_t + \Sigma^{-1} & 0 \\ 0 & (\Sigma^K)^{-1} \end{pmatrix} \left(\begin{pmatrix} x_t \\ k_t \end{pmatrix} - \begin{pmatrix} \nu_t \\ \nu_t^K \end{pmatrix}\right) \\
&\quad \propto \log\left(q(x_t, k_t \mid d_t, x_{t-1}, k_{t-1})\right)
\end{aligned}
$$

where $q(x_t, k_t \mid d_t, x_{t-1}, k_{t-1})$ is the density of a multivariate Gaussian distribution with mean $\widetilde{\nu}_t = (\nu_t, \nu_t^K)'$ and covariance $\widetilde{\Sigma}$ given by

$$
\widetilde{\nu}_t = \begin{pmatrix} -H\widehat{x}_t + \Sigma^{-1}\left(\Gamma x_{t-1} + k_{t-1} + \mu\right) \\ \Gamma^K k_{t-1} \end{pmatrix}
$$

$$
\widetilde{\Sigma} = \begin{pmatrix} -H_t + \Sigma^{-1} & 0 \\ 0 & (\Sigma^K)^{-1} \end{pmatrix}^{-1}
$$

In practice, to ensure a finite variance, a t-distribution with 3 degrees of freedom with location vector $\widetilde{\nu}_t$ and shape matrix $\widetilde{\Sigma}$ is used instead.

3.5. Parameter estimation

As described in the beginning of section 3, given that it is possible to obtain approximate samples from the smoothing distribution $p(x_{0:t} \mid d_{0:n})$, the suggested approach to fit the parameter vector ψ is to use the stochastic approximation expectation maximisation (SAEM) algorithm. The description of this method outlined below is primarily based on Cappé *et al.* (2006), where further references can be found.

To start off, recall the EM-algorithm: At step l:

1. E-step: $Q(\psi^{(l)}, \psi) = \int \log p_\psi(x_{0:n}, d_{0:n}) p_{\psi^{(l)}}(x_{0:n} \mid d_{0:n}) \mathrm{d}x_{0:n}$.
2. M-step: $\psi^{(l+1)} = \operatorname{argmax}_\psi Q(\psi^{(l)}, \psi)$.

Under certain conditions, the sequence $\psi^{(l)}$ is guaranteed to converge to the maximum likelihood estimate of ψ, see Cappé *et al.* (2006, Ch. 10.5). Here $p_\psi(x_{0:n}, d_{0:n})$ is given by (3). Therefore, disregarding terms not depending on ψ,

$$
Q(\psi^{(l)}, \psi) = \int \sum_{t=0}^{n} \log f_\psi(x_t \mid x_{t-1}) p_{\psi^{(l)}}(x_{0:n} \mid d_{0:n}) \mathrm{d}x_{0:n}
$$

Recall that the models in the current paper have multivariate Gaussian densities $f(x_t \mid x_{t-1})$, belonging to the exponential family, which makes it possible to write the M-step explicitly using the ML estimators from Lemma 3.1. That is, the M-step can be written in terms of a low-dimensional sufficient statistic $S(x_{0:n})$.

To obtain a better estimate $\widehat{Q}(\psi^{(l)}, \psi)$, one could draw a large number of replicates of $X_{0:n}$. This is often referred to as the Monte Carlo EM algorithm.

An alternative is to combine a stochastic approximation algorithm with the EM algorithm, which is known as the SAEM algorithm. In practice, this leads to an algorithm where the sufficient statistic is updated in each step by taking a weighted average of the current value and the sufficient statistic obtained by sampling from the smoothing distribution given the current estimates. The SAEM algorithm is described in Algorithm 3.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Algorithm 3. SAEM

- Initialise $\psi = \psi^{(0)}$, and $\widehat{S}^0 = 0$. Do for $l = 1, 2, \ldots, L$:
  1. Sample: $X_{0:n}^{l,i} \sim p_{\psi^{(l-1)}}(\cdot \mid d_{0:n})$, $i = 1, 2, \ldots, m$.
  2. Compute: $\widehat{S}^l = \widehat{S}^{l-1} + c_l\big[\frac{1}{r}\sum_{i=1}^{r} S(X_{0:n}^{l,i}) - \widehat{S}^{l-1}\big]$.
  3. New estimate: Using $\widehat{S}^l$, calculate ψ^l according to Lemma 3.1.
- $\psi^{(L)}$ approximates the MLE of ψ.

Here $c_l \geq 0$, $\sum_l c_l = \infty$ and $\sum_l c_l^2 < \infty$. Under certain assumptions, $\psi^{(l)}$ is guaranteed to almost surely converge to a stationary point of the log likelihood, as $l \to \infty$, see Cappé *et al.* (2006, Ch. 11.1.6) for details.

For easy reference, we summarise the proposed model fitting strategy in section A.

4. Model Validation

Recall from Lemma 2.1 that one may think of the data as coming from an experiment where a Poisson process is observed for a fixed time, the exposure to risk, during which d deaths occur. This is to be used when fitting the model and may also be used when evaluating the model. Therefore, when validating our model, the exposure to risk will be thought of as a fixed quantity, corresponding to a sample size. Then the observed d will be compared to the predictive distribution of the number of deaths in a Lexis square, denoted by P. In this setting, it is natural to consider splitting the data into a training and validation part, where parameters and state vectors are estimated based on the training data, and the model validation is based on the out-of-sample performance in the validation part of the data. That is, the out-of-sample performance is evaluated, given that the exposure to risk is assumed to be known.

On the other hand, when forecasting future values, the exposure to risk is yet to be observed, the situation is different and this problem is discussed in section 5.

In the present paper, the predictive distribution will be evaluated using (proper) scoring rules, see Gneiting & Raftery (2007); Czado *et al.* (2009). Loosely speaking, a scoring rule is a function that assigns a numerical value to the quality of a candidate predictive distribution, P, with respect to observed data, d. A scoring rule is said to be "proper" if there exists a unique optimal value, and it is "strictly proper" if the optimal value is attained for a unique P. It can be noted that many of the classical loss functions used for model evaluation are scoring rules. One such which will be used in the present paper is the absolute error

$$\mathrm{AE}(P, d) := |d - \mu|$$

where μ is a point prediction based on P. Observe that this is a proper, but not strictly proper, scoring rule since any predictive distribution with the same point forecast, for example, median, will give the same absolute error. Note that AE is a *negatively oriented* scoring rule, that is, the aim is to *minimise* AE. A more informative measure, which is a proper, negatively oriented, scoring rule, is the interval score (IS) defined according to

$$\mathrm{IS}_\gamma(P, d) := (u - l) + \frac{2}{\gamma}(l - d)1_{\{d<l\}} + \frac{2}{\gamma}(d - u)1_{\{d>u\}}$$

where l and u denote lower and upper $100(1-\gamma)\%$ percentiles, respectively, of the distribution P, see Gneiting & Raftery (2007). The IS measure is a generalisation of the standard probability coverage measure. In practice, the average of these losses will be analysed, corresponding to the "Mean AE" (MAE) and "Mean IS" (MIS), and it is, hence, clear that reducing MAE and MIS

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

still corresponds to improving model performance compared with observed data. Moreover, both measures may be used for model selection purposes based on both in-sample (training) and out-of-sample (validation) performance.

Furthermore, recall that the parameters in model (M1) (and model (M2)) are estimated by maximising the log likelihood, which is equivalent to maximising the logarithmic score (see Gneiting & Raftery 2007; Czado *et al.* 2009)

$$\text{logs}(P, d) := \log P(d)$$

where $P(d)$ is the probability mass of the observation d in a predictive distribution, which is as proper scoring rule. That is, maximising the likelihood is equivalent to maximising

$$l(d_{0:n}) = \sum_{\mathcal{S}} \text{logs}(P_{\mathcal{S}}, d_{\mathcal{S}})$$

which is a proper scoring rule. Here one can note that compared with (M)AE and (M)IS, where the optimal value is 0, the logarithmic score only tells us that a higher value is better. Still, scaling and translation of a proper scoring rule using constants is still a proper rule. One choice of scaling and translation of the logarithmic score suggested in Cameron & Windmeijer (1996) which produces an R-squared like measure is the following:

$$\text{R}^2_{\text{Dev}} := \frac{\sum_{\mathcal{S}} \text{logs}(P_{\mathcal{S}}, d_{\mathcal{S}}) - \sum_{\mathcal{S}} \text{logs}(\bar{P}_{\mathcal{S}}, d_{\mathcal{S}})}{\sum_{\mathcal{S}} \text{logs}(\widehat{P}_{\mathcal{S}}, d_{\mathcal{S}}) - \sum_{\mathcal{S}} \text{logs}(\bar{P}_{\mathcal{S}}, d_{\mathcal{S}})} \tag{11}$$

where "Dev" refers to that both the numerator and the denominator are deviance residuals. Further, $\bar{P}$ denotes the likelihood in a model with only a constant intercept. In the present case, this will be taken as the model with one constant death rate per age. The likelihood $\widehat{P}$ is the saturated model, that is, where the number of parameters are equal to the number of observations. The sums are taken over all the observed lexis squares. Note that it is clear that $\text{R}^2_{\text{Dev}} \leq 1$, but unlike a standard R^2 it is not certain that $\text{R}^2_{\text{Dev}} \geq 0$, since this will depend on whether $\bar{P}$ is a sub-model of P or not.

Moreover, R^2_{Dev} is possible to calculate both on the training and the validation part of the data, by calculating the model likelihood, $P_{\mathcal{S}}$, according to equation (9). From the calculation of $P_{\mathcal{S}}$ using (9) it follows that this can only be done easily for the *total* likelihood. That is, it is not possible to calculate logs or R^2_{Dev} for a particular age or calendar year – something which often is of interest when assessing predictive model performance. In order to, at least partly, overcome this shortcoming, the following scoring rule is suggested

$$\text{logs}^*(P, d) := \mathsf{E}_{X|D}[\log P(X, d)]$$

which is equivalent to the "E"-step of the EM algorithm, described in section 3.5. The core of the EM algorithm is that by improving $\text{logs}^*(P, d)$ it follows that logs is improved as well, see Dempster *et al.* (1977) or (Cappé *et al.* 2006, Ch. 10.1.2). Moreover, logs* is easy to calculate for a single age or calendar year, since it only amounts to drawing approximate samples from $p(x_{0:t} \mid d_{0:n})$, where $t \leq n'$, with n' being the last observed year in the *validation data and n being the last observed year in the* training data, that is, $n \leq n'$. Here it is important to note that the influence of $d_{0:n}$ on x_t, $n < t \leq n'$ is via the evolution of $x_{n'+1:t}$ based on the state vector $x_{0:n'}$ – an evolution entirely governed by the dynamics of the latent Gaussian X_t process. That is, given the observed exposure to risk, the Poisson variation is not time dependent. Furthermore, by using logs* it is natural to introduce

$$\text{R}^{2,*}_{\text{Dev}} := \frac{\sum_{\mathcal{S}} \text{logs}^*(P_{\mathcal{S}}, d_{\mathcal{S}}) - \sum_{\mathcal{S}} \text{logs}(\bar{P}_{\mathcal{S}}, d_{\mathcal{S}})}{\sum_{\mathcal{S}} \text{logs}(\widehat{P}_{\mathcal{S}}, d_{\mathcal{S}}) - \sum_{\mathcal{S}} \text{logs}(\bar{P}_{\mathcal{S}}, d_{\mathcal{S}})} \tag{12}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

which follows by noting that $\text{logs}^*(\cdot) = \text{logs}(\cdot)$ unless P is used, and again note that $\text{R}^{2,*}_{\text{Dev}} \leq 1$, but $\text{R}^{2,*}_{\text{Dev}}$ may be smaller than 0, due to the same reasons as for R^2_{Dev}.

Another method of validating the model is use a lookback window on which the estimates are based, comparing forecasts with observations, moving the window forward and repeating. The methods discussed in this section could then be used, or as in Dowd *et al.* (2010) by verifying that the residuals are consistent with the model assumptions. However, since the current model fitting approach is computationally expensive this latter approach is not practically feasible. Still, in section 6, apart from using the above measures, both in-sample and out-sample performances are illustrated graphically with respect to the main quantities of interest.

5. Forecasting

The main goal of the present paper is to forecast mortality. In this section, a number of complications related to this are discussed. The specifics of the forecast depend on what is assumed to be known and what one wants to forecast. For example:

1. The perhaps most basic quantity of interest when forecasting, regardless of the size of the population, is the mean or (distribution) of $M_{\mathcal{S}_{a,t}}$.
2. When making forecasts for sub-populations, for example, individuals in an insurance portfolio, the actual number of deaths is of interest and not only the mortality rate. In this situation, the randomness from the Poisson process should be taken into account. Here, typically, individual-level information is available.
3. When forecasting mortality in larger populations, the actual number of deaths may also be of interest, for example, when making country-level demographic forecasts. In this situation, however, information on individual level may not be available or may be impractical to incorporate.

Having fitted the model and obtained the filtering distribution of the state variables at present time, forecasting the state variables is simple. One may either use Monte Carlo simulation to iterate the recursion equation for the state variables, starting at the particle approximation of the present time filtering distribution, to obtain approximations of the distribution at a future time. Or, since the state variables are Gaussian, conditioned on each particle, the forecast will also be Gaussian for each particle, with mean and covariance recursively calculable. In this way, it is possible to obtain the predictive distribution of future $M_{\mathcal{S}_{a,t}}$, which covers Case 1 above.

Considering Case 2, assume that a forecast of $M_{\mathcal{S}_{a,t}} = m_{\mathcal{S}_{a,t}}$ has been produced, and the corresponding forecasted death count will be conditioned on this value. Further, recall that the current modelling approach is motivated by the structure of a Lexis diagram. This means that a single individual of age $a \geq 1$ at year t, can experience one of the following three events during year t:

(a) With probability $p^{\text{a}}_{a,t}$, die while being of age a.

(b) With probability $p^{\text{b}}_{a,t}$, live until becoming of age $a+1$, but die before the end of year t.

(c) With probability $p^{\text{c}}_{a,t}$, live throughout the entire year t.

Concerning the age $a = 0$, it is clear that $p^{\text{a}}_{0,t} = 1 - p^{\text{c}}_{0,t}$. Further, note that an individual i, born at calendar time b_i, which is a years old at the start of year t, was born in calendar year $y_i = \lfloor b_i \rfloor = t - (a+1)$. Thus, the time point during year y_i at which individual i was born is given by $u_i = b_i - y_i = b_i - t + (a+1) \in [0, 1]$. That is, in order to specify $p^{\text{a}}_{a,t}$, $p^{\text{b}}_{a,t}$ and $p^{\text{c}}_{a,t}$ explicitly, it suffices to know $a \in \mathbb{N}$, $t \in \mathbb{N}$, $u \in [0, 1]$, and $m_{\mathcal{S}_{a,t}}$:

Lemma 5.1. *The probabilities for a single individual born at time $b = t - (a+1) + u$, calculated under the assumptions underlying Lemma 2.1, conditional on $m_{\mathcal{S}_{a,t}}$ and u, are given by*

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

$$\begin{cases} p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, u) = 1 - e^{-um_{\mathcal{S}_{a,t}}} \\ p^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}, u) = e^{-um_{\mathcal{S}_{a,t}}}(1 - e^{-(1-u)m_{\mathcal{S}_{a+1,t}}}) \\ p^{\mathrm{c}}_{a,t}(m_{\mathcal{S}}, u) = e^{-um_{\mathcal{S}_{a,t}}} e^{-(1-u)m_{\mathcal{S}_{a+1,t}}} \end{cases}$$

for $a \geq 1$. *For* $a = 0$ *it holds that*

$$p^{\mathrm{a}}_{0,t}(m_{\mathcal{S}}, u) = 1 - e^{-um_{\mathcal{S}_{0,t}}}$$

and $p^{\mathrm{c}}_{0,t}(m_{\mathcal{S}}, u) = 1 - p^{\mathrm{a}}_{0,t}(m_{\mathcal{S}}, u)$.

The proof is a simple application of the probabilities used in the derivation of Lemma 2.1.

Therefore, given Lemma 5.1, it follows that an individual which is a years old at the start of calendar year t that experiences event (a) will contribute to the death count of a-year-olds. But if the same individual instead experiences event (b), will contribute to the death count of $a + 1$-year-olds. Thus, if $D_{a,t}$ denotes the total number of deaths in age group a during year t it follows that

$$D_{a,t} \mid m_{\mathcal{S}} \sim \sum_{i=1}^{n_{a,t}} \mathrm{Be}(p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, u_i)) + \sum_{i=1}^{n_{a-1,t}} \mathrm{Be}(p^{\mathrm{b}}_{a-1,t}(m_{\mathcal{S}}, u_i)) \tag{13}$$

where $\mathrm{Be}(p)$ denotes independent Bernoulli distributed random variables with probability of success p and $n_{a,t}$ is the number of a-year-old individuals alive at January 1st of year t. Note that this forecast is only applicable for one year ahead forecasts. After that, the number of individuals alive becomes random. But it is straightforward to implement multi-year forecasts either by doing bookkeeping of which individual is alive after each forecasted year, or by forecasting each individual's path in the Lexis diagram separately.

In Case 3, we do not assume complete information on each individual, and we are also only interested in the aggregate number of deaths each year. However, one needs to make assumptions on the distribution of time of birth of the individuals. A simplification commonly used in this situation is to assume that all individuals are born midyear, that is, $u_i \equiv 0.5$ for all individuals i. Another possible simplification is to assume that individuals are born uniformly during each year, see Wilmoth *et al.* (2017, Sect. 2). These assumptions can of course be questioned, but are in many situations satisfactory approximations. By assuming that the stochastic birth time during a year, U, is uniform, that is, $U \sim U(0, 1)$, it follows that $p^{\mathrm{a}}_{a,t}$ and $p^{\mathrm{b}}_{a,t}$ from Lemma 5.1 simplifies to

$$\begin{cases} \widetilde{p}^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}) = \mathsf{E}[p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, U) \mid m_{\mathcal{S}}] = 1 - \frac{1}{m_{\mathcal{S}_{a,t}}}\left(1 - e^{-m_{\mathcal{S}_{a,t}}}\right) \\ \widetilde{p}^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}) = \mathsf{E}[p^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}, U) \mid m_{\mathcal{S}}] = \frac{1}{m_{\mathcal{S}_{a,t}}}\left(1 - e^{-m_{\mathcal{S}_{a,t}}}\right) - \frac{1}{m_{\mathcal{S}_{a,t}} - m_{\mathcal{S}_{a+1,t}}}\left(e^{-m_{\mathcal{S}_{a+1,t}}} - e^{-m_{\mathcal{S}_{a,t}}}\right) \end{cases} \tag{14}$$

where the expectation is taken over U, which is assumed to be independent of $m_{\mathcal{S}}$

Note that for $m_{\mathcal{S}} \ll 1$ it follows that

$$\widetilde{p}^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}) \approx p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, 1/2) \approx \frac{1}{2} m_{\mathcal{S}_{a,t}}$$
$$\widetilde{p}^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}) \approx p^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}, 1/2) \approx \frac{1}{2} m_{\mathcal{S}_{a+1,t}}$$

by using a Taylor expansion. Both approximations are therefore approximately equal.

Further, another observation is that, by plugging in the expressions for $\widetilde{p}^{\mathrm{a}}_{a,t}$ and $\widetilde{p}^{\mathrm{b}}_{a,t}$ into relation (14), it follows that

$$D_{a,t} \mid m_{\mathcal{S}} \sim \mathrm{Bin}(n_{a,t}, \widetilde{p}^{\mathrm{a}}_{a,t}(m_{\mathcal{S}})) + \mathrm{Bin}(n_{a-1,t}, \widetilde{p}^{\mathrm{b}}_{a-1,t}(m_{\mathcal{S}})) \tag{15}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

For each simulated trajectory of $M_{\mathcal{S}}$ values, it is possible to forecast death counts, given the number of individuals alive. Note that the structure of (15) only relies on that all birth times are i.i.d., but not necessarily uniformly distributed.

We end this section by commenting on how to simulate in order to gain information on exposure to risk or when one wants to use analytically intractable assumptions on birth times. In these situations, one may use the following simulation procedure to simulate (a), (b) and (c):

(0) If the individual birth time of individual i is unknown, initialise individual i by drawing a random birth time B_i from the distribution of birth times. The distribution of birth times could be estimated from data or assumed to be, for example, uniform over the year.

(1) Draw a $T_{a,t} \sim \text{Exp}(m_{\mathcal{S}_{a,t}})$-distributed random variable.
   (a) If $T_{a,t} \leq B_i$, individual i died being of age a and contributed with $T_{a,t}$ to the exposure to risk $E_{a,t}$.
   (b) If $T_{a,t} > B_i$, individual i has survived age a during year t and contributes with B_i to the exposure to risk $E_{a,t}$.

(2) Draw a $T_{a+1,t} \sim \text{Exp}(m_{\mathcal{S}_{a+1,t}})$-distributed random variable.
   (b) If $T_{a+1,t} \leq 1 - B_i$, individual i died being of age $a+1$ and contributed with $T_{a+1,t}$ to the exposure to risk $E_{a+1,t}$.
   (c) If $T_{a+1,t} > 1 - B_i$, individual i has survived age $a+1$ during year t and contributes with $1 - B_i$ to the exposure to risk $E_{a+1,t}$.

6. Numerical Illustration

The purpose of this section is to illustrate how the models and methods introduced in the present paper can be applied. We will focus our attention on model (M2), which explicitly allows for a random drift term, a situation discussed in Pedroza (2006). The model is calibrated to Swedish and US mortality data, collected from the Human Mortality Database (HMD), see Human Mortality Database (2018). The reason for focusing on Sweden is due to its relatively small population size, which ought to make observations noisier and, hence, parameter estimation and prediction harder. The opposite argument apply to the US. Moreover, Swedish data are available from 1751, although with partly questionable quality until the end of the 1800s, whereas US data only are easily available from about 1935. In order to be able to use as much data as possible for out-of-sample evaluation, the initial focus for Swedish data will be to use the time period 1930–1960 for estimation and 1961–2017 for validation. For US data, we have avoided the second world war and primarily use data from the years 1950 to 1980 for estimation and use the period 1981–2017 for validation.

The parameter estimation is done as described in the summarised algorithm of section A, using the following configuration:

1. Run Algorithm 4, with $m = 50$ particles $L = 50$ iterations with $c_i := 1, i = 1, \ldots, 50$. As starting values, we set Σ to be the identity matrix, Σ_0 as a diagonal matrix with the value 100 along the diagonal. All other matrices are set as the identity matrix, and all mean vector are set to 0.
2. Use the estimated parameters from Step 1 as starting values for a second run using an m of 350 – 500 particles for $L = 100$ iterations, using $c_i := i^{-0.6}, i = 1, \ldots, 100$, where 350 particles were used for 1–3 EPCA components and 500 particles for four and five EPCA components.

The idea with using Step 1 is to hopefully avoid getting stuck close to possibly poor starting values.

Concerning the data to be used, there are known differences between female and male mortality, and our initial focus will be on Swedish males.

The convergence of $\psi^{(i)}$ in the SAEM algorithm, Algorithm 3, is illustrated in Figure 2 for the situation with three EPCA components fitted on Swedish male data from 1930 to 1960. From

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

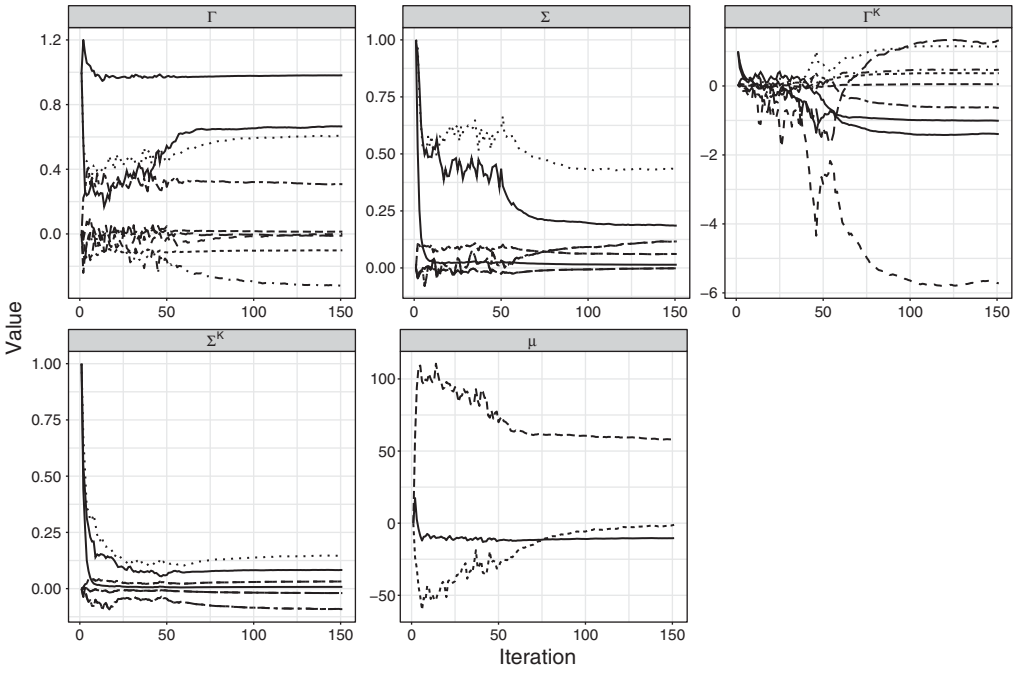

Figure 2. Trace of the estimate of ψ^l from the SAEM algorithm, Algorithm 3, fitted to Swedish male data from 1930 to 1960. Each line corresponds to a component of the respective vector/matrix of parameters.

Figure 2, there are no obvious signs of poor convergence of the SAEM algorithm, and we continue the analysis using the estimated ψ.

In Figure 3(a)–(c), the first three EPCA components are shown for Swedish males when the model has been fitted on data from 1930 to 1960. As seen from the figures, the behaviour of the EPCA components becomes increasingly irregular where the first component is the smoothest. This observed increase in irregularity is the reason for using more particles in the analyses of models containing more EPCA components.

In Figure 3(d)–(f), the in-sample validation results are displayed. The main conclusion here is that the model does perform better in-sample with increasing number of EPCA components. Also, the main improvement is seen when going from one to two components, while two to five components give similar results. In particular, in Figure 3(e) and (f), it is seen that both MAE and MIS favour increasing the number of EPCA components to be used for all ages. Note that the increase in MAE and MIS with increasing age is not surprising, since the mortality rate increases with age. This suggests that one instead could consider analysing MAE and MIS scaled with for example the expected or observed number of deaths. When turning to $R^{2,*}_{\text{Dev}}$ from (12), again, the measure is improved when increasing the number of EPCA components, see Figure 3(d). It is, however, clear that according to $R^{2,*}_{\text{Dev}}$, the performance is poorer for ages in the interval 45–65. This indicates that the performance of model (M2) does not outperform the constant mean mortality model, $\bar{P}$ from (12), for these ages during the years 1930–1960. On the other hand, recall that $R^{2,*}_{\text{Dev}}$ is an approximation of R^{2}_{Dev} from (11), introduced in order to be able to assess model performance within for example specific ages, whereas R^{2}_{Dev} only is possible to calculate over all ages and time periods in total. In Table 1, R^{2}_{Dev} is calculated for Swedish male data from 1930 to 1960, where it is seen that the in-sample performance is very good seen as a whole, and increases when increasing the number of EPCA components being used, as expected.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Table 1. Calculated values of R^2_{Dev} for model (M2) fitted to Swedish male data for the years 1930–1960

No. EPCA	1	2	3	4	5
R^2_{Dev}	0.9817	0.9953	0.9960	0.9964	0.9965

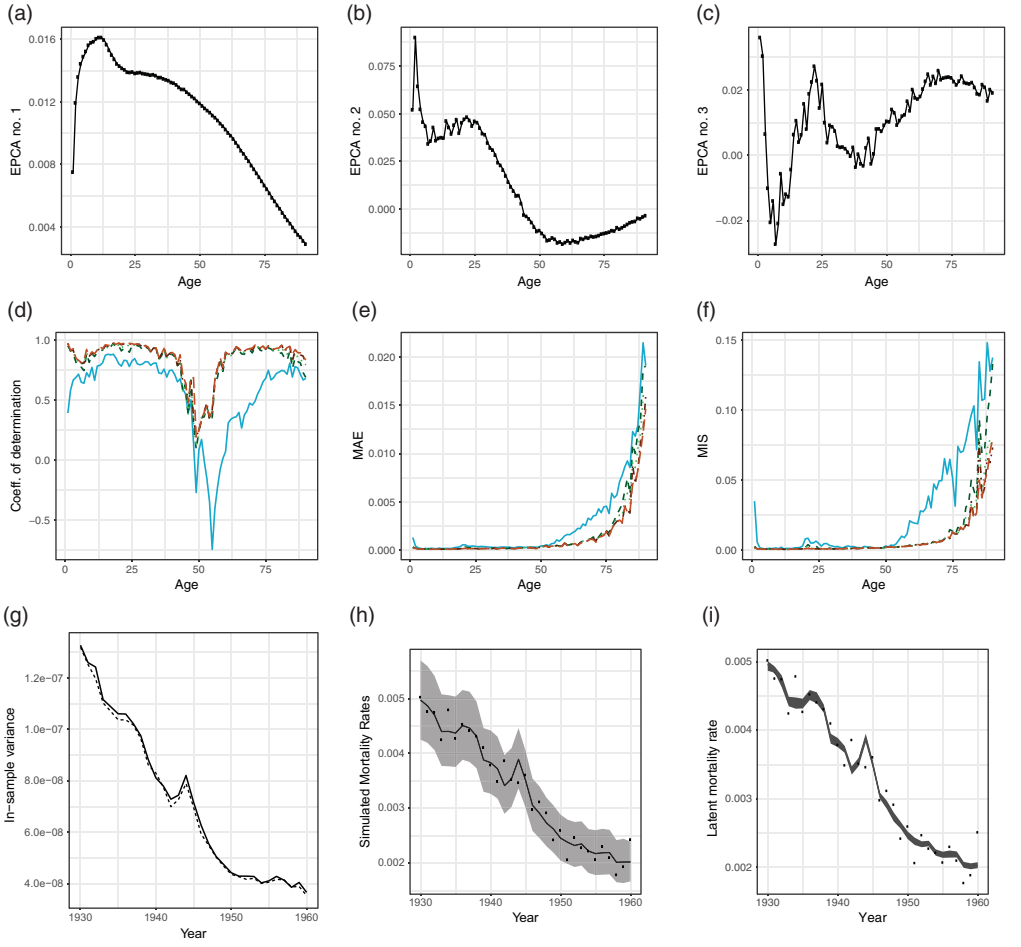

Figure 3. Fig. 3(a)–(c): First three EPCA components, Swedish males, 1930–1960. In Fig. 3(d)–(f), the number of EPCA components, 1–5, are indicated by lines that are solid/brown, short dashed/dark green, dotted/light green, dash-dotted/dark brown and long dashed/light brown, respectively. Fig. 3(d)–(f) shows, from left to right, $R^{2,*}_{Dev}$, MAE and MIS, calculated in-sample for the period 1930–1960 for Swedish males. Fig. 3(g): In-sample variance produced by the model for simulated mortality rates, Swedish males, age 40, three EPCA components; total variance (solid line), population variance (dashed line). Fig. 3(h): 95% yearly confidence levels for the simulated mortality rates M^* for Swedish males using three EPCA components (grey area), median (solid line), observed mortality rates, $\widehat{m}$, (circles). Fig. 3(i): Same as in Fig. 3(h), but for the simulated latent M-process.

The measures MAE, MIS and $R^{2,*}_{Dev}$ are all calculated using the model (M2), which is a model for death counts, whilst many practitioners are more used to considering models for mortality rates. Thus, from now on our focus will be on the latent M-process and the simulated mortality rate process M^* obtained according to

$$M^*_{\mathcal{S}} = \frac{D^*_{\mathcal{S}}}{e_{\mathcal{S}}} \tag{16}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

where $D^*_{\mathcal{S}}$ corresponds to the simulated death counts using $\widehat{m}_{\mathcal{S}}$ from (2). It is possible to compare $M^*_{\mathcal{S}}$ from (16) with the observed crude estimates $\widehat{m}_{\mathcal{S}}$ from (2), that is,

$$\widehat{m}_{\mathcal{S}} = \frac{d_{\mathcal{S}}}{e_{\mathcal{S}}}$$

and in particular, it is possible to decompose the observed variation into a population ("Poisson") variation and variation stemming from the latent M-process ("signal"):

$$\mathrm{Var}(M^*_{\mathcal{S}}) = \underbrace{\mathsf{E}[\mathrm{Var}(M^*_{\mathcal{S}} \mid M_{\mathcal{S}})]}_{=\text{"Poisson"}} + \underbrace{\mathrm{Var}(\mathsf{E}[M^*_{\mathcal{S}} \mid M_{\mathcal{S}}])}_{=\text{"signal"}}$$

In Figure 3(g), this is illustrated for Swedish males of age 40, from which it is seen that essentially all variation seen in $M^*_{\mathcal{S}}$ stems from the population part of the underlying process. Another way of illustrating this is given in Figure 3(h) and (i) where the $\widehat{m}_{\mathcal{S}}$ is compared with $M^*_{\mathcal{S}}$ and $M_{\mathcal{S}}$, respectively. From these figures, it is evident that the variation in the M-process does not capture the variation seen in the $\widehat{m}$:s – which is reasonable since the M-process is independent of population size. Moreover, this suggests that Lee–Carter type models which do not explicitly take the "Poisson" variation into consideration can lead to a misspecified latent mortality rate process. This is discussed further in Appendix B where numerical examples are given based on both the Lee–Carter model from Lee & Carter (1992) and the log-bilinear model from Brouhns *et al.* (2002) that have been implemented using publicly available R packages. The example from Appendix B clearly illustrates that using model (M2) together with the suggested estimation techniques may lead to a substantial reduction in the variation of the latent mortality rate process. This is particularly important when it comes to forecasting, since the predicted Gaussian variation will grow as a function of the number of time steps being forecasted. Therefore, a misattribution of variance will affect the forecast even if the purpose is to forecast death counts. Figure 4(a)–(c) shows the out-of-sample performance in the period 1961–2016. The main conclusion here is that MAE, MIS and $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ favour fewer EPCA components, compared to in-sample, where two or three EPCA components seem to be the best compromise for all ages. One can, however, note that $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ indicates a very poor out-of-sample performance for ages 40–80. For easier comparison with the corresponding in-sample performance, see Figure 4(d), where $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ is plotted for model (M2) using three EPCA components. Upon closer inspection, the lack of performance is due to a drastic decline in mortality occurring around the year 1980 in the mentioned age span, see Figure 4(g) for the model performance of 80-year-old Swedish males when using three EPCA components. Thus, in light of Figure 4(g), the poor model performance is to be expected. Further, Figure 4(h) shows the model performance when the model with three EPCA components has been fitted to data from 1930 to 1990, hence including the discussed mortality decline for ages 40–80. Even if the out-of-sample performance still is poor, one can note that the model behaves as expected: The first ten years of the sharp decline in the mortality rates for ages 40 to 80 is now included in the data being used for fitting, and the model reacts to these values as if they are part of a temporary observed anomaly, since the predictions strive to return to an evolution similar to the historical trend. Moreover, by inspecting $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ in Figure 4(d), it is also seen that by including parts of the mortality decline in the data used for fitting, the overall in-sample performance is improved, but at a cost of poorer predictive performance for a wider span of ages. In Figure 4(i), the model with three EPCA components has been fitted to the period 1970 to 2000, and it is now clear that the model has been able to adapt to the change in the observed mortality patterns. Still, the in-sample performance is somewhat poorer in general, but descent as a whole, see Figure 4(d). Concerning the out-of-sample $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ from Figure 4(d), the behaviour is highly erratic, but here one shall keep in mind that each $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ value is only calculated as an average of 16 years. In Figure 5(a)–(c), the simulated total, in-sample and out-of-sample trajectories for Swedish males of age 10, 40 and 80 are shown, using three EPCA components, fitted to the years 1970 to 2000, and it is seen that the

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

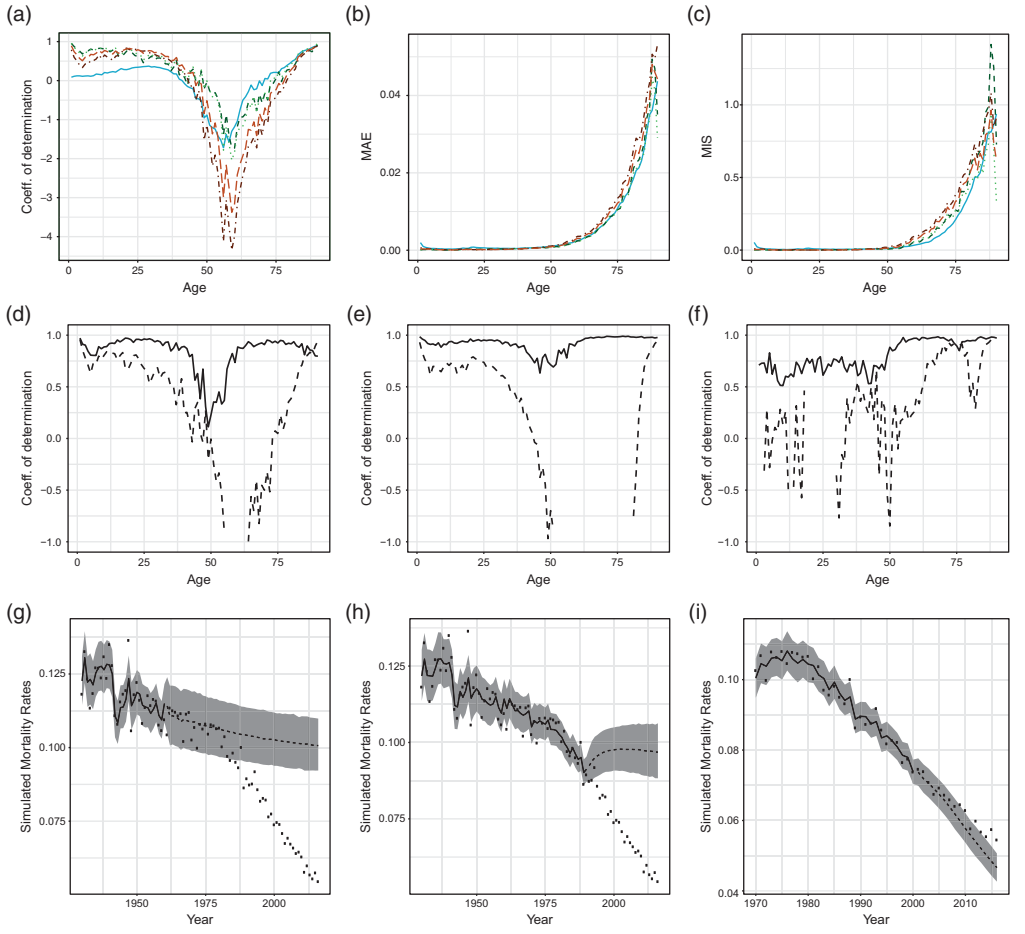

Figure 4. Fig. 4(a)–(c) shows the out-of-sample analog of Fig. 3(d)–(f), where all parameters have been estimated based on Swedish male data from 1930 to 1960, and the predictions are made for the period 1961 to 2016. Fig. 4(d)–(f): $R^{2,*}_{Dev}$ where solid lines correspond to in-sample performance and dashed lines correspond to out-of-sample performance when using three EPCA components – models fitted using data from 1930 to 1960, 1930 to 1990 and 1970 to 2000, respectively. Fig. 4(g)–(i): 95% yearly confidence/prediction intervals (grey area) for simulated mortality rates M^*, Swedish males, age 80, three EPCA components, median (solid/dashed line), observed mortality rates, $\widehat{m}$, (circles) – models fitted using data from 1930 to 1960, 1930 to 1990 and 1970 to 2000, respectively.

overall performance is satisfactory. One can also note that R^2_{Dev} from (11) increases, compared to using data from 1930 to 1960, when fitting the models to data from 1930 to 1990 and 1970 to 2000 – attaining the highest value for the latter time period.

Continuing in Figure 5(d)–(f), the model performance for Swedish females of ages 10, 40 and 80 is illustrated for model (M2) with two EPCA components fitted on data from 1950 to 1980, and Figure 5(g)–(i) shows the same situation for US females when using model (M2) with three EPCA components. From the figures for the Swedish females, it is seen that there is a similar decline in mortality for age 40, but less pronounced than the one seen in Figure 4(g) for 80-year-old Swedish male. Also note that no sudden drop in mortality is seen for 80-year-old Swedish female. Concerning the US females the mortality pattern is more irregular, and there are signs of a change in trend around 1990 where the mortality seems to increase, which is something not captured by the model. Moreover, when inspecting 80-year-old US female, the in-sample variation seems to be too small.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

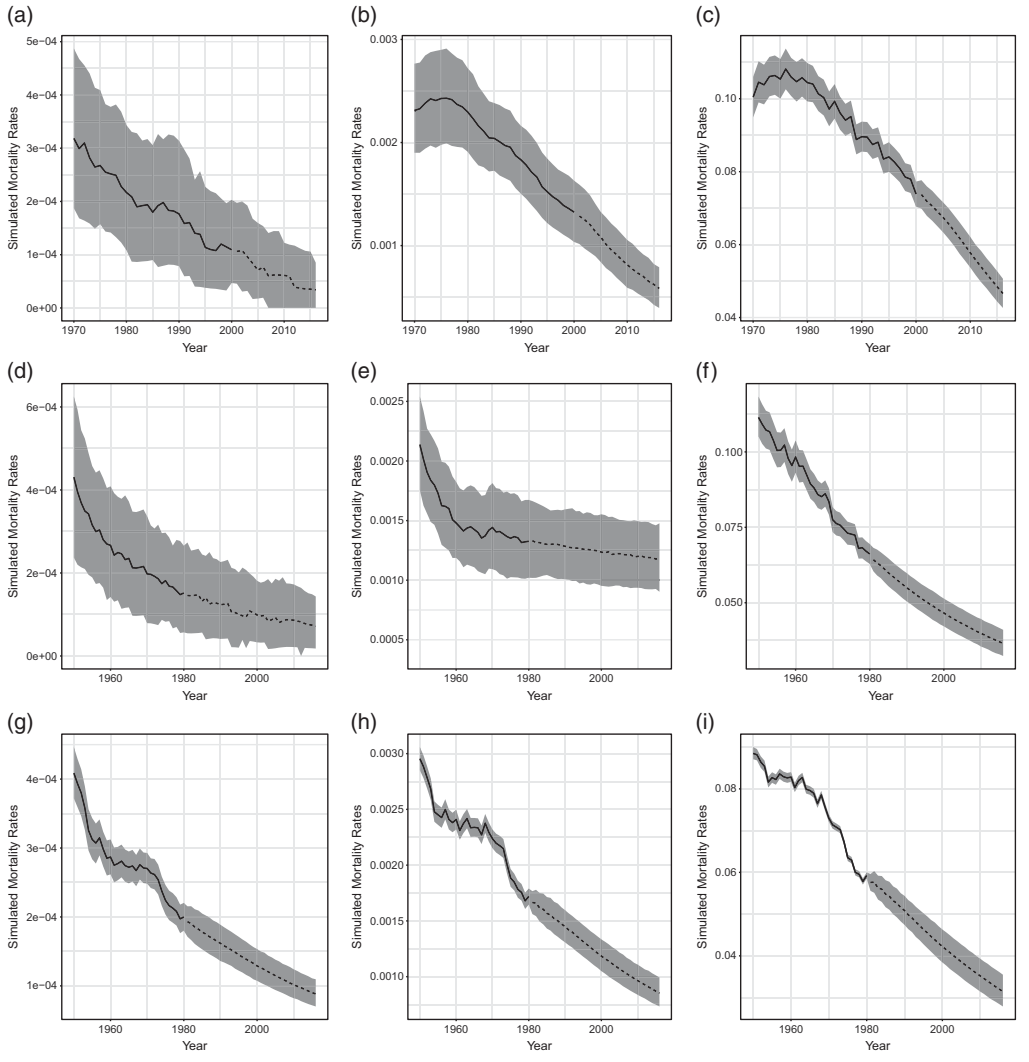

Figure 5. In all figures, 95% confidence/prediction intervals for the simulated centralised mortality rates (grey area), median (solid/dashed line) and observed centralised mortality rates (circles). In all figures, from left to right, age 10, age 40 and age 80, respectively. Fig. 5(a)–(c): Swedish males, three EPCA components, model fitted using data from 1970 to 2000. Fig. 5(d)–(f): Swedish females, five EPCA components, model fitted using data from 1950 to 1980. Fig. 5(g)–(i): US females, three EPCA components, model fitted using data from 1950 to 1980.

To summarise the above numerical illustration, it is seen the importance of using predictive measures for model selection, as well as the importance of assessing model performance based on death counts or scalings thereof (i.e. $\widehat{m}$ versus M^*). We have also described in detail how model (M2) may be used in practice and shown that the model is able to capture most of the relevant dynamics observed in the analysed historical data. Moreover, we have also seen that the model behaves as expected when data used for fitting contains drastic changes in mortality trends. Another important observation is that the analyses imply that by not explicitly (and correctly) accounting for the Poisson part of the variation, the variation attributed to the latent mortality rate process may become substantially misspecified. This may, hence, be a problem for Lee–Carter-type models.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

7. Conclusions

In the present paper, it has been argued for using a Poisson state-space model for mortality forecasting. The Poisson part of the model arises naturally from the mortality dynamics of a continuous time Lexis diagram. The unobservable state process corresponding to a mortality rate process is modelled as a multivariate Gaussian process inspired by the Lee–Carter model and its extensions, see Booth & Tickle (2008) and Haberman & Renshaw (2011). The suggested model class provides models for death counts, as opposed to, for example, Lee–Carter like models, which are models for mortality rates. Furthermore, most Lee–Carter like models are fitted in a two-step procedure, where first raw mortality estimates are obtained according to, for example (2), and then, in a second step, a stochastic process is fitted to the raw mortality rates. By using the suggested Poisson state-space models, estimation may be done coherently in a single step using particle filter techniques and the SAEM algorithm. Moreover, since all model parameters are estimated using maximum likelihood, it is argued that it is natural to use versions of logarithmic scores for model performance assessment. In particular, an R^2-like measure is introduced, which is closely connected to the "E"-step in the SAEM algorithm. This measure is possible to calculate both in-sample and out-of-sample for specific ages and time periods and is a proper scoring rule.

A large number of numerical illustration is also provided, where the necessary steps to fit the model and make forecasts have been discussed. In this numerical illustration, it was also shown that by using the Poisson state-space model for death counts it is possible to decompose the observed variability in terms of "population" (or Poisson) variation and "signal" (or mortality rate) variation. For the Swedish data, it is clear that the Poisson part of the variation is dominating in-sample. Further, the numerical examples illustrate that not explicitly accounting for these separate sources of variation, as in in the case of the Lee–Carter model, may lead to a misspecified latent mortality rate process, which will affect the forecasting ability of the model negatively.

References

Aalen, O., Borgan, O. & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag New York, Inc.

Barber, D., Cemgil, A.T. & Chiappa, S. (2011). *Bayesian Time Series Models [Elektronisk resurs]*. Cambridge University Press, Cambridge.

Bazaraa, M.S., Sherali, H.D. & Shetty, C.M. (2006). *Nonlinear Programming: Theory and Algorithms, Third Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey.

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

Boyd, S., Boyd, S.P. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Brouhns, N., Denuit, M. & Vermunt, J.K. (2002). A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**(3), 373–393.

Cameron, A.C. & Windmeijer, F.A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, **14**(2), 209–220.

Cappé, O., Moulines, E. & Rydén, T. (2006). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, New York.

Carfora, M. F., Cutillo, L. & Orlando, A. (2017). A quantitative comparison of stochastic mortality models on italian population data. *Computational Statistics & Data Analysis*, **112**, 198–214.

Collins, M., Dasgupta, S. & Schapire, R.E. (2001). A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems* (pp. 617–624).

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, **65**(4), 1254–1261.

De Jong, P. & Tickle, L. (2006). Extending Lee–Carter mortality forecasting. *Mathematical Population Studies*, **13**(1), 1–18.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dowd, K., Cairns, A.J., Blake, D., Coughlan, G.D., Epstein, D. & Khalaf-Allah, M. (2010). Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics* **47**(3), 255–265.

Durbin, J. & Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. vol. **38**. Oxford University Press.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Ekheden, E. & Hössjer, O. (2014). Analysis of the stochasticity of mortality using variance decomposition. In D. Silvestrov & A. Martin-Löf (Eds.), *Modern Problems in Insurance Mathematics* (pp. 199–222). EAA Series. Springer, Cham.

Ekheden, E. & Hössjer, O. (2015). Multivariate time series modeling, estimation and prediction of mortalities. *Insurance: Mathematics and Economics*, **65**, 156–171.

Fung, M.C., Peters, G.W. & 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.

Gneiting, T. & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

Godsill, S.J., Doucet, A. & West, M. (2004). Monte carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, **99**(465), 156–168.

Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. *Philosophical Transactions of the Royal Society of London*, **115**, 513–583.

Gorski, J., Pfeuffer, F. & Klamroth, K. (2007). Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, **66**(3), 373–407.

Haberman, S. & Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, **48**(1), 35–55.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Human Mortality Database. (2018). Available online at the address http://www.mortality.org or http://www.humanmortality.de [accessed 5-Dec-2018].

Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J., Chopin, N., et al. (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* **30**(3), 328–351.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **154**(3), 371–396.

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

Lu, M., Huang, J.Z. & Qian, X. (2016). Sparse exponential family principal component analysis. *Pattern Recognition*, **60**, 681–691.

Pedroza, C. (2006). A Bayesian forecasting model: predicting us male mortality. *Biostatistics*, **7**(4), 530–550.

Pitacco, E. (2018). Heterogeneity in mortality: a survey with an actuarial focus. ARC Centre of Excellence in Population Ageing Research Working Paper 2018/7.

Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V. & Vachon, P. (2017). Methods protocol for the human mortality database, November 27, 2017 (version 6). University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock Available online at the address http://www.mortality.org or http://www.humanmortality.de [accessed 5-Dec-2018].

Appendix A. Summary of Model Fitting

Algorithm A1. Summary

- Use EPCA to reduce the dimension. This gives the estimate $\hat{\Upsilon}$.
- Initialize $\psi = \psi^{(0)}$, and $\widehat{S}^0 = 0$. Do for $l = 1, 2, \ldots, L$:
  1. Use Algorithm 1 with $q(x_t \mid d_t, x_{t-1})$ being the t-distribution with location and shape as in section 3.4. This gives samples $\overline{X}^i_{0:t}$ from the filtering distribution.
  2. Use $\overline{X}^i_{0:t}$ as inputs to Algorithm 2. This gives samples $\widetilde{X}_{0:t}$ from the smoothing distribution.
  3. Calculate the sufficient statistics from $\widetilde{X}_{0:t}$ and update $\hat{S}$ and ψ as in Algorithm 3.
- ψ^L approximates the MLE of ψ.

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

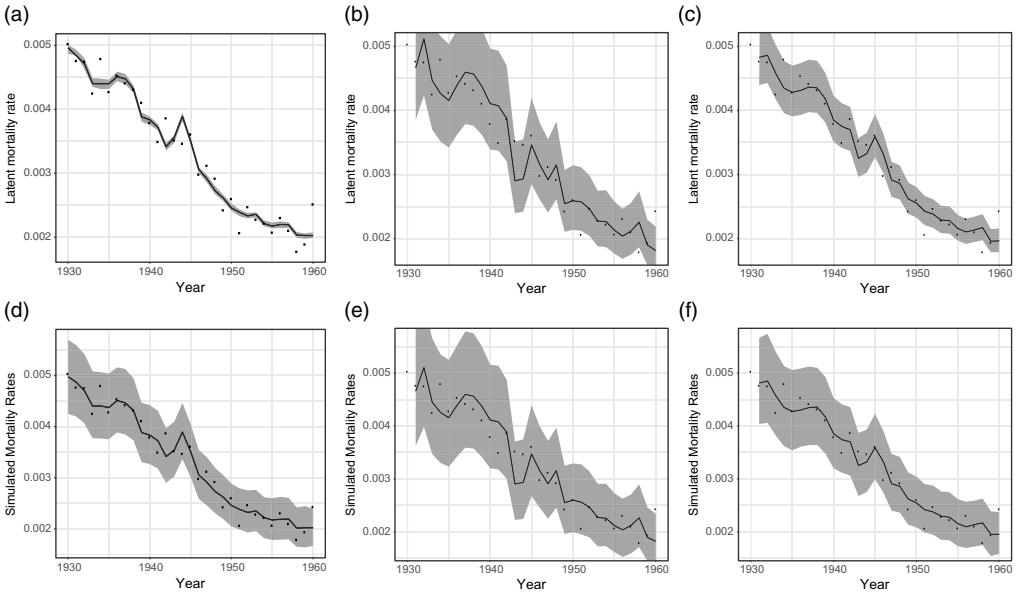

Figure B1. *Columns*: From left to right, model (M2) with three EPCA components, the Lee–Carter model, and the Poisson log bilinear model from Brouhns *et al.* (2002). *Rows*: First row shows the latent mortality process, and the second row shows the simulated mortality rates according to (16) when assuming that death counts are (conditionally) Poisson distributed. All figures show Swedish male mortality for the age 40, and all models have been fitted based on the ages 0–90 and the time period 1930–1960. The area in grey corresponds to 95% confidence regions, solid lines correspond to the median and the actual observations are indicated by circles.

Appendix B. Attribution of Mortality Rate Variation – A Numerical Illustration

In this section two examples of other mortality rate models will be considered and their ability to capture the variation seen in actual mortality rates given by the point estimates from (2) will be illustrated. For the first model, the classical Lee–Carter model from Lee & Carter (1992), the analysis is based on the `R` package `Demography`, using the `lca` model class. The second model is the Poisson log-bilinear model from Brouhns *et al.* (2002), and the analysis is based on the `R` package `StMoMo` using the `lc` model class. Note that it is clear that these models will have different numbers of parameters, and, hence, different ability to capture non-linear behavior. In order to account for this, all models have been estimated using a period and age where all models fit data well. The chosen data is Swedish male data, ages 0-90, during the years 1930–1960. The example will be restricted to in-sample performance, due to that we are here primarily interested in comparing the attribution of variance between the different models and not the actual predictive performance. In order to obtain in-sample variability for the Lee–Carter model first the lca model class was fitted to data, and then, in a second step, the `Arima` package was used. An example of the in-sample performance is given in Figure B1 for the age 40. Perhaps not surprisingly, by comparing the latent mortality rate process from the Lee–Carter model, see Figure B1(b), which has no explicit Poisson part, with the latent mortality rate process from model (M2), see Figure B1(a), the Lee–Carter models shows a clear tendency to attribute too much variation to the latent mortality rate process. Note that the Lee–Carter model needs one observation for initiation of the random walk process with drift governing the calendar time dynamics of the model. This explains why the first observation in Figure A1(b) lacks a confidence interval. Turning to the Poisson log-bilinear model from Brouhns *et al.* (2002), this model has an explicit Poisson part, and may be seen as

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

an approximation of model (M2). From Figure A1(f) it is clear that this model attributes much less variation to the latent mortality rate process than the Lee–Carter model, but still considerably more than seen in the latent Gaussian process from model (M2). Again, note that the first observation lacks a confidence interval due to the same reason as for the Lee–Carter model.

Furthermore, an illustration of simulated mortality rates following (16) are given in Figure B1(d)–(f). Again, it is seen that the standard Lee–Carter model's performance is the worst, but the Poisson log-bilinear model is now closer to model (M2). Still, even when it comes to forecasts of death counts, the Gaussian part of the model will after only a few predicted future time points dominate the variation. Consequently, when making predictions on moderate to long time horizons it is particularly important to make as correct attribution of variation as possible.

Appendix C. Proofs and Mathematical Motivations

C.1. Proof of Lemma 2.1

Following standard theory of counting processes with multiplicative intensity processes, see Andersen *et al.* (1993); Aalen *et al.* (2008): Let $\widetilde{t}_i$ denote the last time point when individual i was observed to be alive, $b_i \leq \widetilde{t}_i \leq q_i$, let $\delta_i = 1$ if individual i died at $\widetilde{t}_i$ and 0 otherwise. If we further assume that all individuals are independent, we get that the log-likelihood function is given by

$$\begin{aligned} l(m) &\propto \sum_{i=1}^{n} \delta_i \log m(a_i(\widetilde{t}_i), \widetilde{t}_i) - \sum_{i=1}^{n} \int_{\underline{t}}^{\widetilde{t}_i} \lambda_i(t) \mathrm{d}t \\ &= \sum_{i=1}^{n} \delta_i \log m(\widetilde{t}_i - b_i, \widetilde{t}_i) - \sum_{i=1}^{n} \int_{\underline{t}}^{\widetilde{t}_i} m(t - b_i, t) Y_i(t) \mathrm{d}t \end{aligned}$$

Thus, if we consider the situation with constant hazard rates on yearly Lexis squares, that is, $\mathcal{M} = \left\{ m_{\mathcal{S}} \mid \mathcal{S} \in \bar{\mathcal{S}} \right\}$, it follows that

$$\begin{aligned} l(\mathcal{M}) &\propto \sum_{i=1}^{n} \sum_{\mathcal{S} \in \bar{\mathcal{S}}} \delta_i 1_{\{(a_i, \widetilde{t}_i) \in \mathcal{S}\}} \log m_{\mathcal{S}} - \sum_{i=1}^{n} \sum_{\mathcal{S} \in \bar{\mathcal{S}}} m_{\mathcal{S}} \int_{\underline{t}}^{\widetilde{t}_i} Y_i(t; \mathcal{S}) dt \\ &= \sum_{\mathcal{S} \in \bar{\mathcal{S}}} (d_{\mathcal{S}} \log m_{\mathcal{S}} - e_{\mathcal{S}} m_{\mathcal{S}}) \end{aligned}$$

which is exactly the result from Lemma 2.1.

C.2. Proof of Lemma 3.1

These are standard results for VAR processes, see Hamilton (1994), which are included for the sake of completeness. This section is split into two parts, one for model (M1) and one for model (M2).

Here we use that for matrices X and Y of suitable dimensions,

$$\begin{aligned} \frac{\partial}{\partial X} \log |\det X| &= (X')^{-1} \\ \frac{\partial}{\partial X} \operatorname{tr}(XY) &= Y' \end{aligned}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

Model (M1)

By combining (7) and (4),

$$f_\psi(x_{0:n}) = \nu(x_0)\prod_{t=1}^{n} f_\psi(x_t \mid x_{t-1})$$
$$\propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{1}{2}\sum_{t=1}^{n}(x_t - \Gamma x_{t-1} - \mu)'\Sigma^{-1}(x_t - \Gamma x_{t-1} - \mu)\right\}$$

That it is an exponential family is clear since it is multivariate normal and that it is curved follows by finding the natural parameters and sufficient statistics. Since the quadratic form in the exponential function is a scalar and

$$\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA) = \mathrm{tr}\,(A'B) = \mathrm{vec}\,(A)\cdot\mathrm{vec}\,(B)$$

it follows that

$$-\frac{1}{2}\sum_{t=1}^{n}(x_t - \Gamma x_{t-1} - \mu)'\Sigma^{-1}(x_t - \Gamma x_{t-1} - \mu)$$
$$= \mathrm{vec}\begin{pmatrix} \Sigma^{-1} \\ -2\Sigma^{-1}\Gamma \\ \Gamma'\Sigma^{-1}\Gamma \\ -2\mu'\Sigma^{-1} \\ \mu'\Sigma^{-1}\Gamma \end{pmatrix} \cdot \mathrm{vec}\begin{pmatrix} \sum_{t=1}^{n} x_t x_t' \\ \sum_{t=1}^{n} x_{t-1}x_t' \\ \sum_{t=1}^{n} x_{t-1}x_{t-1}' \\ \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_{t-1} \end{pmatrix} + n\mu'\Sigma^{-1}\mu$$

We see that the dimension of the natural parameter is larger than that of ψ, and so the exponential family is curved. Let us denote the sufficient statistic as

$$S(x_{0:n}) = \begin{pmatrix} S_1(x_{1:n}) \\ S_2(x_{0:n}) \\ S_3(x_{0:n-1}) \\ S_4(x_{1:n}) \\ S_5(x_{0:n-1}) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} x_t x_t' \\ \sum_{t=1}^{n} x_{t-1}x_t' \\ \sum_{t=1}^{n} x_{t-1}x_{t-1}' \\ \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_{t-1} \end{pmatrix}$$

Towards finding the ML estimators, define

$$\widehat{\varepsilon}_t := x_t - \begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix}\begin{bmatrix} 1 \\ x_{t-1} \end{bmatrix}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

The log-likelihood as a function of Γ and μ is, ignoring constants, given by

$$-2l(\Gamma,\mu)=\sum_{t=1}^{n}\left(x_t-\begin{bmatrix}\mu & \Gamma\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)'\Sigma^{-1}\left(x_t-\begin{bmatrix}\mu & \Gamma\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)=\operatorname{tr}\left(\Sigma^{-1}\sum_{t=1}^{n}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'\right)$$
$$+\operatorname{tr}\left(\Sigma^{-1}\sum_{t=1}^{n}\left(\begin{bmatrix}\widehat{\mu}-\mu & \widehat{\Gamma}-\Gamma\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\left(\begin{bmatrix}\widehat{\mu}-\mu & \widehat{\Gamma}-\Gamma\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)'\right)$$
$$+2\operatorname{tr}\left(\Sigma^{-1}\sum_{t=1}^{n}\widehat{\varepsilon}_t\begin{bmatrix}1 & x_{t-1}'\end{bmatrix}\begin{bmatrix}\widehat{\mu}'-\mu'\\ \widehat{\Gamma}'-\Gamma'\end{bmatrix}\right)$$

If $\hat{\Gamma}$ is such that the third term above is 0, it is clear that $\Gamma=\hat{\Gamma}$ is a minimum. Therefore, the condition for a minimum is that

$$\sum_{t=1}^{n}\widehat{\varepsilon}_t\begin{bmatrix}1 & x_{t-1}'\end{bmatrix}=\sum_{t=1}^{n}\left(x_t-\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\begin{bmatrix}1 & x_{t-1}'\end{bmatrix}$$
$$=\begin{bmatrix}S_4 & S_2'\end{bmatrix}-\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}n & S_5'\\ S_5 & S_3\end{bmatrix}=0$$

with solution

$$\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}=\begin{bmatrix}S_4 & S_2'\end{bmatrix}\begin{bmatrix}n & S_5'\\ S_5 & S_3\end{bmatrix}^{-1}$$

Consequently, the log-likelihood as a function of Σ^{-1}, evaluated at $\widehat{\Gamma}$ and $\widehat{\mu}$, is given by

$$-2l(\Sigma^{-1};\widehat{\Gamma},\hat{\mu})=-n\log|\,\Sigma^{-1}\,|+\operatorname{tr}\Sigma^{-1}\sum_{t=1}^{n}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'$$

which yields

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma^{-1}}\left[-2l(\Sigma^{-1};\widehat{\Gamma},\widehat{\mu})\right]=-n\Sigma+\sum_{t=1}^{n}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'$$

resulting in the following MLE

$$\widehat{\Sigma}=\frac{1}{n}\sum_{t=1}^{n}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'=\frac{1}{n}\sum_{t=1}^{n}\left(x_t-\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\left(x_t-\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)'$$
$$=\frac{1}{n}\left(S_1-\begin{bmatrix}S_4 & S_2'\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix}-\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}S_4'\\ S_2\end{bmatrix}+\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}n & S_5'\\ S_5 & S_3\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix}\right)$$

Model (M2)

By using analogous arguments as those used for model (M1) the MLE of Γ^K is given by

$$\widehat{\Gamma}^K=S_2'(k_{0:n})S_3^{-1}(k_{0:n-1})$$

and of Σ^K by,

$$\widehat{\Sigma}^K=\frac{1}{n}\left(S_1(k_{1:n})+\widehat{\Gamma}S_3(k_{0:n-1})\widehat{\Gamma}'-S_2'(k_{0:n})\widehat{\Gamma}'-\widehat{\Gamma}S_2(k_{0:n})\right)$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

For Γ_X and μ, the condition for the MLE is that

$$\begin{aligned}\sum_{t=1}^{n}\left(x_t - k_{t-1} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}_X\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\begin{bmatrix}1 & x'_{t-1}\end{bmatrix}\\ = \left(\begin{bmatrix}S_4(x_{1:n}) & S'_2(x_{0:n-1})\end{bmatrix} - \begin{bmatrix}S_5(k_{0:n-1}) & \sum_{t=1}^{n} k_{t-1}x'_{t-1}\end{bmatrix}\right)\\ - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}_X\end{bmatrix}\begin{bmatrix}n & S'_5(x_{0:n-1})\\ S_5(x_{0:n-1}) & S_3(x_{0:n-1})\end{bmatrix} = 0\end{aligned}$$

with solution

$$\begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}_X\end{bmatrix} = \left(\begin{bmatrix}S_4(x_{1:n}) & S'_2(x_{0:n-1})\end{bmatrix} - \begin{bmatrix}S_5(k_{0:n-1}) & \sum_{t=1}^{n} k_{t-1}x'_{t-1}\end{bmatrix}\right)\begin{bmatrix}n & S_5(x_{0:n-1})'\\ S_5(x_{0:n-1}) & S_3(x_{0:n-1})\end{bmatrix}^{-1}$$

The MLE of Σ is then

$$\begin{aligned}\widehat{\Sigma} &= \frac{1}{n}\sum_{t=1}^{n}\left(x_t - k_{t-1} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\left(x_t - k_{t-1} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)'\\ &= \frac{1}{n}\Bigg(S_1(x_{1:n}) + S_3(k_{0:n-1}) + \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}n & S'_5(x_{0:n-1})\\ S_5(x_{0:n-1}) & S_3(x_{0:n-1})\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix}\\ &\quad - \sum_{t=1}^{n} x_t k'_{t-1} - \sum_{t=1}^{n} k_{t-1}x'_t - \begin{bmatrix}S_4(x_{1:n}) & S'_2(x_{0:n})\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}S'_4(x_{1:n})\\ S'_2(x_{0:n})\end{bmatrix}\\ &\quad + \begin{bmatrix}S_5(k_{1:n}) & \sum_{t=1}^{n} k_{t-1}x'_{t-1}\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix} + \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}S'_5(k_{1:n})\\ \sum_{t=1}^{n} x'_{t-1}k_{t-1}\end{bmatrix}\Bigg)\end{aligned}$$

C.3 Proof of Lemma 3.2

We will now show that $-\log g_\Upsilon(d_{0:n} \mid x_{0:n})$ is bi-convex in Υ and $x_{0:n}$ in the sense of (Gorski *et al.* 2007, Def. 1.2), but not jointly convex in $(\Upsilon, x_{0:n})$.

First, note that $\Upsilon \in \mathbb{R}^{m\times p}$ and $x_t \in \mathbb{R}^p$, $t = 0, \ldots, n$, are both elements of convex sets. Further, note that

$$-\log g_\Upsilon(d_{0:n} \mid x_{0:n}) = -\sum_{t=0}^{n} \log g_\Upsilon(d_t \mid x_t)$$

can be decomposed into a sum of terms of the form

$$h(z; k, d) = ke^z - dz$$

where $k > 0$ and $d > 0$ are constants, that is

$$-\log g_\Upsilon(d_t \mid x_t) = \sum_{i=1}^{k} h((\Upsilon x_t)_i; (e_t)_i, (d_t)_i)$$

By straightforward differentiation, it is clear that $h(z; k, d)$ is convex in z, but not monotone. Moreover, let $\underline{x}_t := (x_t, \ldots, x_t) \in \mathbb{R}^{m\times p}$ and let 1_i denote the $mp \times mp$ matrix whose off-diagonal elements are 0 with a diagonal consisting of zeros and ones defined so that the following relation holds

$$\operatorname{vec}(\Upsilon)'1_i \operatorname{vec}(\underline{x}_t) = (\Upsilon x_t)_i \in \mathbb{R}$$

https://doi.org/10.1017/S1748499520000275 Published online by Cambridge University Press

By using this representation it follows that

$$\begin{aligned}\operatorname{vec}(\Upsilon)' 1_i \operatorname{vec}(\underline{x}_t) &=: A_{\Upsilon,i} \operatorname{vec}(\underline{x}_t) \\ &=: A_{x_t,i} \operatorname{vec}(\Upsilon)\end{aligned}$$

and, in particular,

$$\begin{aligned}-\log g_\Upsilon(d_t \mid x_t) &= \sum_{i=1}^{k} h((\Upsilon x_t)_i; (e_t)_i, (d_t)_i) \\ &= \sum_{i=1}^{k} h(A_{\Upsilon,i} \operatorname{vec}(\underline{x}_t); (e_t)_i, (d_t)_i) \\ &= \sum_{i=1}^{k} h(A_{x_t,i} \operatorname{vec}(\Upsilon); (e_t)_i, (d_t)_i)\end{aligned}$$

which corresponds to compositions of affine mappings of a convex function. This shows that $-\log g_\Upsilon(d_t \mid x_t)$ is convex in Υ given x_t, as well as, convex in x_t given Υ, see Boyd (2004), Ch. 3.2. The argument can be repeated to show that $-\log g_\Upsilon(d_{0:n} \mid x_{0:n})$ is bi-convex with respect to Υ and $x_{0:n}$.

The following counter example shows that $-\log g_\Upsilon(d_{0:n} \mid x_{0:n})$ is not jointly convex:

$$h\big((pu_1 + (1-p)u_2)(pv_1 + (1-p)v_2); k, d\big) > ph\big(u_1 v_1; k, d\big) + (1-p)h\big(u_2 v_2; k, d\big)$$

when $k = d = 1$, $p = 0.8$ and $(u_1, v_1) = (-1.5, 1)$, $(u_2, v_2) = (-0.5, 1.5)$.

Consequently, $-\log g_\Upsilon(d_{0:n} \mid x_{0:n})$ is bi-convex in Υ and $x_{0:n}$ separately, but not jointly convex in both Υ and $x_{0:n}$.

Cite this article: Andersson P and Lindholm M (2021). Mortality forecasting using a Lexis-based state-space model, *Annals of Actuarial Science*, **15**, 519–548. https://doi.org/10.1017/S1748499520000275