Mathematical Statistics
Stockholm University

Mortality forecasting using a Lexis based state space model

Patrik Andersson, Mathias Lindholm

Research Report 2019:11

ISSN 1650-0377

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2019:11**,
http://www.math.su.se

Mortality forecasting using a Lexis based state space model

Patrik Andersson*, Mathias Lindholm†

July 2019

Abstract

A model of mortality is introduced which is based on a data generating process defined in terms of the continuous-time dynamics of a Lexis diagram. Through counting process arguments, the likelihood of death count data sampled at yearly intervals is shown to be equivalent to that of a certain Poisson likelihood. Therefore a hidden Markov model with Poisson distributed observations and a Gaussian state process is introduced. More specifically, the latent mortality rate process is driven by a low-dimensional autoregressive process, where the dimension reduction in a Poisson setting is based on so-called generalized principal component analysis (GPCA). The full likelihood of the Poisson state space model is not analytically tractable, but it is possible to derive explicit sufficient statistics when conditioning on the state of the latent mortality rate process. This makes it possible to estimate the model parameters using the stochastic approximation expectation-maximization (SAEM) algorithm, where sampling is made using particle filter techniques. This circumvents the two-step estimation procedure used for e.g. the Lee-Carter model.

Further, the constructive nature of the introduced model makes it easy to decompose the observed variation in terms of population ("Poisson") variation and variation due to the latent mortality rate process. Since all model parameters are estimated using maximum likelihood theory we argue that it is natural to assess the model performance using logarithmic scores. In particular, we introduce a proper scoring rule based on a transformation of a certain logarithmic score which is closely connected to the maximization step in the SAEM-algorithm. This scoring rule may be seen as a coefficient of determination like measure which can be used for assessing specific age and calendar year model performance, both in-sample and out-of-sample.

The versatility of the model is illustrated on Swedish and US data, where both in-sample and out-of-sample forecast performance is analyzed. We illustrate the convergence of the numerical routines being used and discuss initiation procedures. Further, the numerical illustrations indicate that by not explicitly taking the population part of the variation into account may lead to that too much variation is attributed to the mortality rates, consequently being a potential problem for Lee-Carter type models.

Keywords: Non-linear non-Gaussian state-space-models; Generalized Principal Component Analysis; log-concave likelihood; 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 e.g. pensions, taxation and public health. A closely related area

*Department of Statistics, Uppsala University. Email: patrik.andersson@statistics.uu.se
†Department of Mathematics, Stockholm University. Email: lindholm@math.su.se

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 w.r.t. time of birth).

One of the earliest contributions to the area of mortality forecasting is the so-called "Gompertz law of mortality" (Gompertz, 1825). For more on other mortality laws, see e.g. the survey Pitacco (2018) and the references therein. A more recent important contribution to the area is the Lee-Carter model (Lee and 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, in this way 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. Thus, an alternative, and very natural, interpretation of the Lee-Carter model is as a Gaussian hidden Markov model (HMM), see e.g. De Jong and Tickle (2006); Fung et al. (2017) for a discussion in a mortality context and e.g. Cappé et al. (2006); Durbin and 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 e.g. Booth and Tickle (2008); Haberman and Renshaw (2011) 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. One drawback with the Lee-Carter model is that it models the mortality rates directly, that is, firstly standard point estimates of mortality rates are obtained and secondly, given these rates, a (discrete time) stochastic process is fitted. However, in practice what is observed are death counts, and the mortality rates corresponds to unknown functions/processes to be estimated. Moreover, by instead using death count dynamics as a starting point, an additional source of randomness is introduced; randomness caused by the population being finite. This additional source of randomness allows for decomposing observed variation into population randomness and estimation randomness. Examples of more constructive modeling approaches which explicitly includes population variation are given in e.g. Brouhns et al. (2002); Ekheden and Hössjer (2014, 2015).

In the present paper a constructive HMM approach is taken. A Lexis diagram, which summarizes age and calendar time dynamics in continuous time per individual, will serve as the starting point. This allows us to use standard survival analysis techniques for right censored counting processes in order to arrive at a Poisson likelihood. Thus, by making use of the likelihood principle, it follows that a likelihood equivalent modelling approach is to model the mortality dynamics as a Poisson process. Moreover, it follows that this Poisson process is expressed in terms of exposure-to-risk, i.e. the total time individuals of a certain age has been alive (under risk) during a specific calendar time period, see Wilmoth et al. (2017, Sec. 2.2). This allows continuous time information to be summarized on e.g. a yearly basis. This is described in Section 2. Based on the likelihood equivalence with a Poisson process and inspired by the Lee-Carter model, we suggest that the mortality rates appearing in the Poisson process themselves are described by a multivariate Gaussian process. Hence, the model introduced in the present paper corresponds to a HMM with Poisson distributed observations, see Section 2.1. In Section 3 basic properties of the likelihood of the Poisson state space model is discussed, as well as how the model can be fitted. Moreover, in order to reduce the dimension of the multivariate Gaussian process governing the evolution of the mortality rates an approach known as generalized principal component analysis (GPCA) is suggested. This is described in Section 3.1, and is based on Collins et al. (2001).

Furthermore, the introduced Poisson state space model will have a likelihood function which is ana-

lytically intractable. Still, the model is based on a combination of Gaussian and Poisson dynamics, which makes it possible to obtain closed form low-dimensional *conditional* sufficient statistics. One efficient method for fitting models with these properties is to use the stochastic approximation EM (SAEM) technique, see Section 3.5. The "stochastic approximation" part of the SAEM algorithm is based on using particle filter techniques, and in particular forward filtering backward smoothing (FFBS), see e.g. Cappé et al. (2006); Kantas et al. (2015). This is described in Section 3.

Continuing, it is possible to obtain estimates of mortality rates without introducing an HMM, using e.g. central mortality rates which are defined as certain scalings of observed death counts. Still, by using an HMM it is possible to attribute the variation seen in these estimates to variation stemming from the underlying mortality rate process and to variation stemming from the population ("Poisson") variation. Due to this, we argue that model performance should be assessed w.r.t. death counts or scalings thereof. In Section 4 we discuss model validation criteria which are based on (proper) scoring rules that are applicable to both training and validation data, hence allowing for model selection based on predictive performance. In particular we use introduce an R^2-like measure defined in terms of deviance. Moreover, in Section 5 we discuss forecasting.

The paper concludes with a numerical illustration, where the methods from the present paper are illustrated based on Swedish and US data from Human Mortality Database (2018) and where predictions are made for 20 - 50 years, which corresponds to time intervals of interest for e.g. life insurers, designing of pension systems, and demographic applications. Based on the numerical illustration it is seen that the model forecasting performance is satisfactory w.r.t. both in-sample (training) data and out-of-sample (validation). This is illustrated by varying the length of the time period used for fitting. The analyses also illustrate the importance of separating between variation from the mortality rates and Poisson variation from the population – for Swedish data it is clearly seen that the majority of the variation stems from population variation. An implication of this is the potentially misleading results when using Lee-Carter type models in this type of situation.

2 Probabilistic mortality model

The probabilistic model which will be introduced in the present paper is based on the population dynamics as it is summarized 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 degree straight line. Since most individuals are not born on January 1, the time spent in each square will differ from individual to individual. As an example, consider the shaded area in Figure 1: The oldest individual, the one closest to the northwest corner of the Lexis diagram, 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 characterized 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 which $[\underline{t}, \overline{t}] \cap [B_i, Q_i] \neq \emptyset$ 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 corresponds to that the individual is alive. The process can be defined in terms of a multiplicative

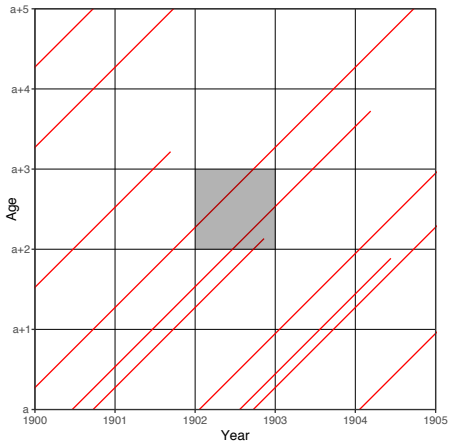

Figure 1: Example of a Lexis-diagram.

intensity process (see e.g. 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

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

i.e. $Y_i(t)$ 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 that experience the event "death" up until t, $\underline{t} \leq t \leq \bar{t}$, denoted $D(t)$, is given by

$$D(t) = \sum_{i=1}^{n} D_i(t),$$

which is a counting process with intensity process

$$\lambda(t) = \sum_{i=1}^{n} m(A_i(t),t)Y_i(t).$$

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

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

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

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

and the total number of observed death counts in $\mathcal{S}$ is given by the counting process

$$D(t;\mathcal{S}) = \sum_{i=1}^{n} D_i(t;\mathcal{S}),$$

with intensity process

$$\lambda(t;\mathcal{S}) = \sum_{i=1}^{n} m(A_i(t),t)Y_i(t;\mathcal{S}).$$

Moreover, due to that census data, typically, is 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, i.e. $m(a,t) = m_{\mathcal{S}}$ if $(a,t) \in \mathcal{S}$. That is, $D(t;\mathcal{S})$ has a multiplicative intensity process

$$\lambda(t;\mathcal{S}) = \sum_{i=1}^{n} m_S Y_i(t;\mathcal{S}).$$

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

$$D_{\mathcal{S}} = \sum_{i=1}^{n} \mathbb{1}\left((Q_i - B_i, Q_i) \in \mathcal{S}\right),$$

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

$$E_{\mathcal{S}} = \sum_{i=1}^{n} \int_{\underline{t}}^{\bar{t}} Y_i(t;\mathcal{S})\,\mathrm{d}t,$$

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 $\mathcal{M} = \left\{m_{\mathcal{S}} \mid \mathcal{S} \in \bar{\mathcal{S}}\right\}$ *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* $\mathcal{S}$*.*

For more on likelihood inference on Lexis diagrams, see e.g. 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 A.

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 needs 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 summarizes 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 ML estimator 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 which will be discussed further in Section 6 where a numerical illustration is given.

2.1 Mortality model

In this section we will define the probabilistic model of mortality that will be used in our analysis and describe the method of estimation and forecasting. We however begin by introducing the notation:

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, e.g. $D \in \mathbb{N}_0^{k \times n}$. The parameters Υ, Γ, Σ and μ will in general be unknown and are to be estimated. The procedure for estimation is described in Section 3.

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

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

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. That is, conditioned on M, 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 and 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: Firstly, 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. Secondly, 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 analyze, 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:

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

which explicitly allows for a random drift, something discussed in e.g. 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, we will adopt the standard notation $x_{0:n} = (x_0, \ldots, x_n)$.

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 large 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

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

where

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

Note that $v(x_0)$ is the density of the starting point X_0, which we have assumed to be known. In practice this is set to a Gaussian density with large variance. Moreover, since $x_{0:t}$ corresponds to unobservable, stochastic, state vectors, the likelihood function that we want to maximize 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 stochastic approximation EM algorithm (SAEM) which is based on approximating (5) using simulation, see e.g. (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 Section 3.2 – 3.4. Provided that the complete data likelihood will produce 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

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

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

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 a generalized principal component analysis (GPCA). This is described in more detail in Section 3.1. Consequently, the SAEM-algorithm is used to estimate ψ by optimizing $p_{\widehat{\Upsilon},\psi}(d_{0:n})$ via the corresponding complete data 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 Model* (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 A.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 Generalized principal component analysis – estimating Υ

As mentioned in Section 2.1 and Section 3, the matrix Υ may be thought of as a non-parametric choice of basis functions, i.e. 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 which will be used consists of optimizing the Poisson part of the complete data likelihood, i.e. $g_\Upsilon(d_{0:n} \mid x_{0:n})$ from (6). Recall, from Section 2.1, that both Υ and $x_{0:n}$ are unobservable quantities. One way of tackling this is to estimate Υ and $x_{0:n}$ jointly, given $d_{0:n}$, by maximizing $g_\Upsilon(d_{0:n} \mid x_{0:n})$. This is what is referred to as generalized principal component analysis (GPCA), and was introduced in Collins et al. (2001). 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, i.e. 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 GPCA-optimization, 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 A. Too see the effect of Lemma 3.2, one can note that e.g. $g_{\Upsilon/c}(d_{0:n} \mid cx_{0:n}) = g_\Upsilon(d_{0:n} \mid x_{0:n})$ for all $c \in \mathbb{R}_+$. Note that the above "marginal" convexity property corresponds to so-called "bi-convexity", see e.g. Gorski et al. (2007, Def. 1.1,1.2). Moreover, in Gorski et al. (2007) conditions are given for when minimization of a bi-convex function using so-called *alternate convex search* (ACS) methods, which is a special case of *cyclic coordinate* (CCM) methods, will converge, see (Gorski et al., 2007, Thm. 4.7, 4.9, Cor. 4.10). For more on cyclic coordinate methods and convergence, see e.g. Bazaraa et al. (2013, Ch. 8.5).

Before ending this section, note that as opposed to classical PCA, the GPCA 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 e.g. 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*

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

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 distribution 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

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

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

Algorithm 1 SISR

- At time $t = 0$, for all $i \in \{1, \ldots, r\}$:
  1. Sample: $X_0^i \sim q_m(\cdot \mid d_0)$.
  2. Compute: $w_0^i = \frac{g(d_0 \mid X_0^i)\nu(X_0^i)}{q_m(X_0^i \mid d_0)}$.
  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: $X_t^i \sim q_m(\cdot \mid d_t, \bar{X}_t^i)$.
  2. Append: $X_{0:t}^i = (\bar{X}_{0:t-1}^i, X_t^i)$.
  3. Compute: $w_t^i = \frac{g(d_t \mid X_t^i) f(X_t^i \mid \bar{X}_{t-1}^i)}{q(X_t^i \mid d_t, \bar{X}_{t-1}^i)}$.
  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 $p(x_{0:t} \mid d_{0:t})$.

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 e.g. 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.

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

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

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

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).

Algorithm 2 FFBS

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

3.4 Choosing the importance distribution

Recall that the particle filter algorithms, Algorithm 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 Algorithm 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 byproduct of the GPCA 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 $\widehat{x}_{0:n}$. For model (M1) this approach results in

the following approximation

$$\begin{aligned}\log g(d_t \mid x_t) &\approx \log g(d_t \mid \hat{x}_t) + (x_t - \widehat{x}_t)' D \log g(d_t \mid \widehat{x}_t) + \frac{1}{2}(x_t - \widehat{x}_t)' D^2 \log g(d_t \mid \widehat{x}_t)(x_t - \widehat{x}_t) \\ &= \log g(d_t \mid \widehat{x}_t) + \frac{1}{2}(x_t - \widehat{x}_t)' H_t (x_t - \widehat{x}_t) \\ &\propto \frac{1}{2}(x_t - \widehat{x}_t)' H_t (x_t - \widehat{x}_t), \end{aligned} \tag{10}$$

where "$\propto$" corresponds to removing normalization constants not depending on x_t, and where $H_t := H_t(d_t, \widehat{x}_{0:n}, \widehat{\Upsilon})$ denotes the Hessian, which typically is obtained as a byproduct from the GPCA optimization. Further, note that from Section 3.1 it follows that $-H_t$ is positive semi-definite. Thus, (10) is the un-normalized log-density, with x_t as argument, of a multivariate Gaussian distribution with mean $\widehat{x}_{0:n}$ and covariance $(-H_t)^{-1}$. Finally, by combining the above, the approximation of $\log(g(d_t \mid x_t) f(x_t \mid x_{t-1}))$ becomes

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

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

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

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

$$\begin{aligned}\widetilde{\nu}_t &= \begin{pmatrix} -H \widehat{x}_t + \Sigma^{-1}\left(\Gamma x_{t-1} + k_{t-1} + \mu\right) \\ \Gamma^K k_{t-1} \end{pmatrix}, \\ \widetilde{\Sigma} &= \begin{pmatrix} -H_t + \Sigma^{-1} & 0 \\ 0 & (\Sigma^K)^{-1} \end{pmatrix}^{-1}.\end{aligned}$$

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 maximization (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 ψ_l is guaranteed to converge to the maximum likelihood estimate of ψ, see e.g. (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,

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

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.

Algorithm 3 SAEM

- Initialize $\psi = \psi_0$, and $\widehat{S}^0 = 0$. Do for $l = 1, 2, \ldots$:
  1. Sample: $X_{0:n}^{l,i} \sim p_{\psi_{l-1}}(\cdot \mid d_{0:n})$, $i = 1, 2, \ldots, m$.
  2. Compute: $\widehat{S}^l = \widehat{S}^{l-1} + c_l \left[\frac{1}{r} \sum_{i=1}^{r} S(X_{0:n}^{l,i}) - \widehat{S}^{l-1} \right]$.
  3. New estimate: Using $\widehat{S}^l$, calculate ψ^l according to Lemma 3.1.
- ψ^l approximates the MLE of ψ.

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

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 e, the exposure-to-risk, during which d deaths occur. This is 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 evaluation 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 e.g. Gneiting and Raftery (2007); Czado et al. (2009). That is, loosely speaking, a scoring rule is a function which assigns a numerical value to the quality of a candidate predictive distribution, P, w.r.t. 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, e.g. median, will give the same absolute error. Note that AE is a *negatively oriented* scoring rule, i.e. the aim is to *minimize* 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 e.g. Gneiting and Raftery (2007). The IS measure is a generalization of the standard probability coverage measure. In practice the average of these losses will be analyzed, corresponding to the "Mean AE" (MAE) and "Mean IS" (MIS), and it is, hence, clear that reducing MAE and MIS 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 maximizing the log-likelihood, which is equivalent to maximizing the logarithmic score (see e.g. Gneiting and Raftery (2007); Czado et al. (2009))

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

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

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

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 and Windmeijer (1996) which produces an R-squared like measure is the following:

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

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, i.e. 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 $\mathrm{R}^2_{\mathrm{Dev}} \leq 1$, but unlike a standard R^2 it is not certain that $\mathrm{R}^2_{\mathrm{Dev}} \geq 0$, since this will depend on whether $\bar{P}$ is a sub-model of P or not.

Moreover, $\mathrm{R}^2_{\mathrm{Dev}}$ 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 $\mathrm{R}^2_{\mathrm{Dev}}$ for e.g. 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

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

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 $\mathrm{logs}^*(P, d)$ it follows that logs is improved as well, see e.g. Dempster et al. (1977) or (Cappé et al., 2006, Ch. 10.1.2). Moreover, logs^* is easy to calculate for e.g. 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, i.e. $n \leq n'$. 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. Furthermore, by using logs^* it is natural to introduce

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

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

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 depends on what is assumed to be known and what one wants to forecast. E.g.

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 subpopulations, e.g. 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, e.g. 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 $p^{\mathrm{a}}_{a,t}$, die while being of age a.

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

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

Concerning the age $a = 0$, it is clear that $p^{\mathrm{a}}_{0,t} = 1 - p^{\mathrm{c}}_{0,t}$. 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 $p^{\mathrm{a}}_{a,t}, p^{\mathrm{b}}_{a,t}$ and $p^{\mathrm{c}}_{a,t}$ 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*

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

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

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

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

$$D_{a,t} \mid m_{\mathcal{S}} \sim \sum_{i=1}^{n_{a,t}} \mathrm{Be}(p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, u_i)) + \sum_{i=1}^{n_{a-1,t}} \mathrm{Be}(p^{\mathrm{b}}_{a-1,t}(m_{\mathcal{S}}, u_i)), \tag{11}$$

where $n_{a,t}$ 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 mid year, i.e. $u_i \equiv 0.5$ for all individuals i. Another possible simplification is to assume that individuals are born uniformly during each year, see e.g. Wilmoth et al. (2017, Sec. 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, i.e. $U \sim U(0,1)$, it follows that $p^{\mathrm{a}}_{a,t}$ and $p^{\mathrm{b}}_{a,t}$ from Lemma 5.1 simplifies to

$$\begin{cases} \widetilde{p}^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}) & = \mathsf{E}[p^{\mathrm{a}}_{a,t}(m_{\mathcal{S}}, U) \mid m_{\mathcal{S}}] = 1 - \frac{1}{m_{\mathcal{S}_{a,t}}} \left(1 - e^{-m_{\mathcal{S}_{a,t}}}\right), \\ \widetilde{p}^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}) & = \mathsf{E}[p^{\mathrm{b}}_{a,t}(m_{\mathcal{S}}, U) \mid m_{\mathcal{S}}] = \frac{1}{m_{\mathcal{S}_{a,t}}} \left(1 - e^{-m_{\mathcal{S}_{a,t}}}\right) - \frac{1}{m_{\mathcal{S}_{a,t}} - m_{\mathcal{S}_{a+1,t}}} \left(e^{-m_{\mathcal{S}_{a+1,t}}} - e^{-m_{\mathcal{S}_{a,t}}}\right). \end{cases} \tag{12}$$

Note that for $m_t \ll 1$ it follows that

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

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 (12), it follows that

$$D_{a,t} \mid m_{\mathcal{S}} \sim \mathrm{Bin}(n_{a,t}, \widetilde{p}^{\mathrm{a}}_{a,t}(m_{\mathcal{S}})) + \mathrm{Bin}(n_{a-1,t}, \widetilde{p}^{\mathrm{b}}_{a-1,t}(m_{\mathcal{S}})). \tag{13}$$

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 (13) 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, initialize individual i by drawing a random birth-time B_i from a suitable distribution.

(1) Draw a $T_{a,t} \sim \mathrm{Exp}(m_{\mathcal{S}_{a,t}})$-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 aim with the current section is to illustrate how the proposed class of state space models defined by (M1) performs when calibrated to Swedish and US mortality data. All data have been collected from the Human Mortality Database (HMD), see Human Mortality Database (2018). The purpose of this section is *not* necessarily to find the best model for the analyzed data, but merely to illustrate how the models and methods introduced in the present paper applies. Still, we will focus our attention on model (M2), which explicitly allows for a random drift term, a situation treated in e.g. Pedroza (2006).

Concerning parameter estimation, the particle filter techniques described in Section 3.2 - 3.4, and the SAEM-algorithm, Section 3.5, have been used with the following configuration:

1. Run the FFBS-algorithm, see Algorithm 2, with 50 particles and use the SAEM-algorithm, see Algorithm 3, with 50 iterations with $c_i := 1, i = 1, \ldots, 50$.

2. Use the estimated parameters from Step 1 as starting values for a second run of the FFBS-algorithm using 350 – 500 particles for 100 iterations of the SAEM-algorithm, using $c_i := i^{-0.6}, i = 1, \ldots, 100$, where we used 350 particles for 1-3 GPCA components and 500 particles for 4 and 5 GPCA components.

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

Regarding the choice of starting values, recall the non-uniqueness of the estimates of Υ and the corresponding state vector $\widehat{x}_{0:t}$ from the GPCA, see the comments following Lemma 3.2. This suggests 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 factorization of $\widehat{\Sigma}$ expressed in terms of $\widehat{\sigma}$, i.e. $\widehat{\Sigma} = \widehat{\sigma}\widehat{\sigma}'$. Then, set

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

and

$$\widehat{\widehat{x}}_{0:t} := (\widehat{\sigma}')^{-1}\widehat{x}_{0:t},$$

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 GPCA-optimization: The suggested scaling

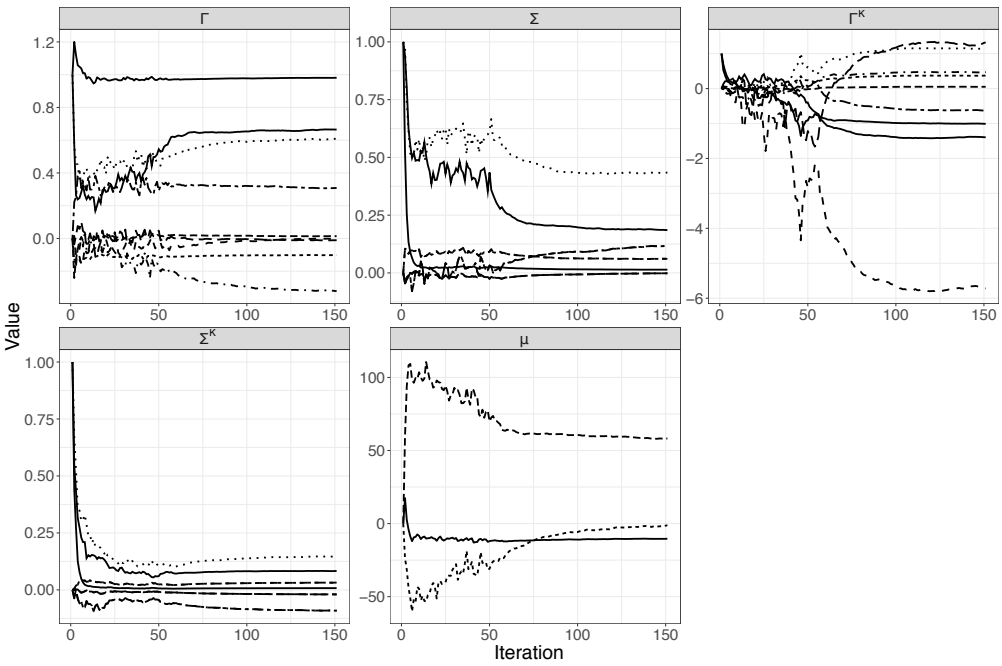

Figure 2: Estimates of ψ^l from the SAEM-algorithm, Algorithm 3, fitted to Swedish male data from 1930 to 1960.

does not affect the value of the loss function used in the optimization, but it will affect its derivative. As a compromise, we propose to re-fit $\widehat{\Upsilon}$, given $\widehat{\widehat{x}}_{0:t}$. By using this re-parametrization all covariance matrices are assumed to be diagonal: The covariance matrix Σ is assumed to be an identity matrix and Σ_0 a diagonal matrix with the value 100 along the diagonal. The initiation of Σ_0 is chosen in order to reduce the influence of a possibly bad starting value, and to avoid the optimization getting stuck near to such a value. Concerning the other starting values, all matrices are set to identity matrices and all mean vectors are set to 0.

Concerning the data to be used, there are known differences between female and male mortality, and our initial focus will be on Swedish males. In Figure 3(a)–3(c) the first three GPCA components are shown for Swedish males when the model has been fitted on data from 1930-1960. As seen from the figures, the behavior of the GPCA-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 GPCA-components. The convergence of ψ^i in the SAEM-algorithm, Algorithm 3, is illustrated in Figure 2 for the situation with three GPCA-components fitted on Swedish male data from 1930 to 1960. From 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(e) and 3(f) it is seen that both MAE and MIS favours increasing the number of GPCA-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 e.g. the expected or observed number of deaths. When turning to $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ from (4), again, the measure is improved when increasing the number of GPCA-components, see Figure 3(d). It is, however, clear that according to $\mathrm{R}^{2,*}_{\mathrm{Dev}}$, 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 (4), for these ages during the years 1930-1960. On the other hand, recall that $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ is an approximation of $\mathrm{R}^{2}_{\mathrm{Dev}}$ from (4),

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

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

introduced in order to be able to assess model performance within e.g. 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 1930-1960, where it is seen that the in-sample performance is very good seen as a whole, and increases when increasing the number of GPCA-components being used, as expected.

The measures MAE, MIS, and $R^{2,*}_{\text{Dev}}$ 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

$$M^*_{\mathcal{S}} = \frac{D^*_{\mathcal{S}}}{e_{\mathcal{S}}}. \tag{14}$$

It is possible to compare M^* from (14) 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"):

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

In Figure 3(g) this is illustrated for Swedish males of age 40, from which it is seen that essentially all variation seen in $M^*_{a,t}$ stems from the population part of the underlying process. Another way of illustrating this is given in Figure 3(h) and 3(i) where the $\widehat{m}_{a,t}$ is compared with $M^*_{a,t}$ and $M_{a,t}$, 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 will overestimate the mortality rate risk in these situations.

Regarding the out-of-sample performance in the period 1961-2016, Figure 4(a) - 4(c) show, as expected, that the MAE, MIS and $R^{2,*}_{\text{Dev}}$ favors fewer GPCA components, where 2 or 3 GPCA components seem to be the best compromise for all ages. One can, however, note that $R^{2,*}_{\text{Dev}}$ 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 $R^{2,*}_{\text{Dev}}$ is plotted for model (M2) using three GPCA-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 GPCA-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 GPCA-components has been fitted to data from 1930-1990, 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 40-80 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 $R^{2,*}_{Dev}$ 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 GPCA-components has been fitted to the period 1970-2000, 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 $R^{2,*}_{Dev}$ from Figure 4(d), the behaviour is highly erratic, but here one shall keep in mind that each $R^{2,*}_{Dev}$ value is only calculated as an average of 16 years. In Figure 5(a) - 5(c) the simulated total, in-sample and out-of-sample, trajectories for Swedish males of age 10, 40, and 80 are shown, using three GPCA-components, fitted to the years 1970-2000, and it is seen that the overall performance is satisfactory. One can also note that R^{2}_{Dev} from (4) increases, compared to using data from 1930-1960, when fitting the models to data from 1930-1990 and 1970-2000 – attaining the highest value for the latter time period.

Continuing, in Figure 5(d) - 5(f) the model performance for Swedish females of ages 10, 40, and 80, is illustrated for model (M2) with two GPCA components fitted on data from 1950-1980, and Figure 5(g) - 5(i) shows the same situation for US females when using model (M2) with three GPCA-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 Swedish male 80 year olds. Also note that no sudden drop in mortality is seen for Swedish female 80 year olds. 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 US female 80 year olds the in-sample variation seems to be too small.

To summarize the above numerical illustration, we have 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 accounting for the Poisson part of the variation too much variation may be attributed to the latent mortality rate process. This may, hence, be a problem for Lee-Carter type models.

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 e.g. Booth and Tickle (2008); Haberman and Renshaw (2011). The suggested model class provides models for death counts, as opposed to e.g. 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 e.g. (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 stochastic approximation EM-algorithm. Moreover, since all model parameters are estimated using maximum likelihood theory 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 sets it is clear that the Poisson part of the variation is dominating. This suggests that Lee-Carter type models will overestimate the mortality rate risk in these situations.

References

Aalen, O., Borgan, O., and 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., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons.

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

Boyd, S. (2004). *Convex Optimization*. Cambridge University Press.

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

Cameron, A. C. and 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., and Rydén, T. (2006). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer New York.

Collins, M., Dasgupta, S., and Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624.

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

De Jong, P. and Tickle, L. (2006). Extending lee–carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.

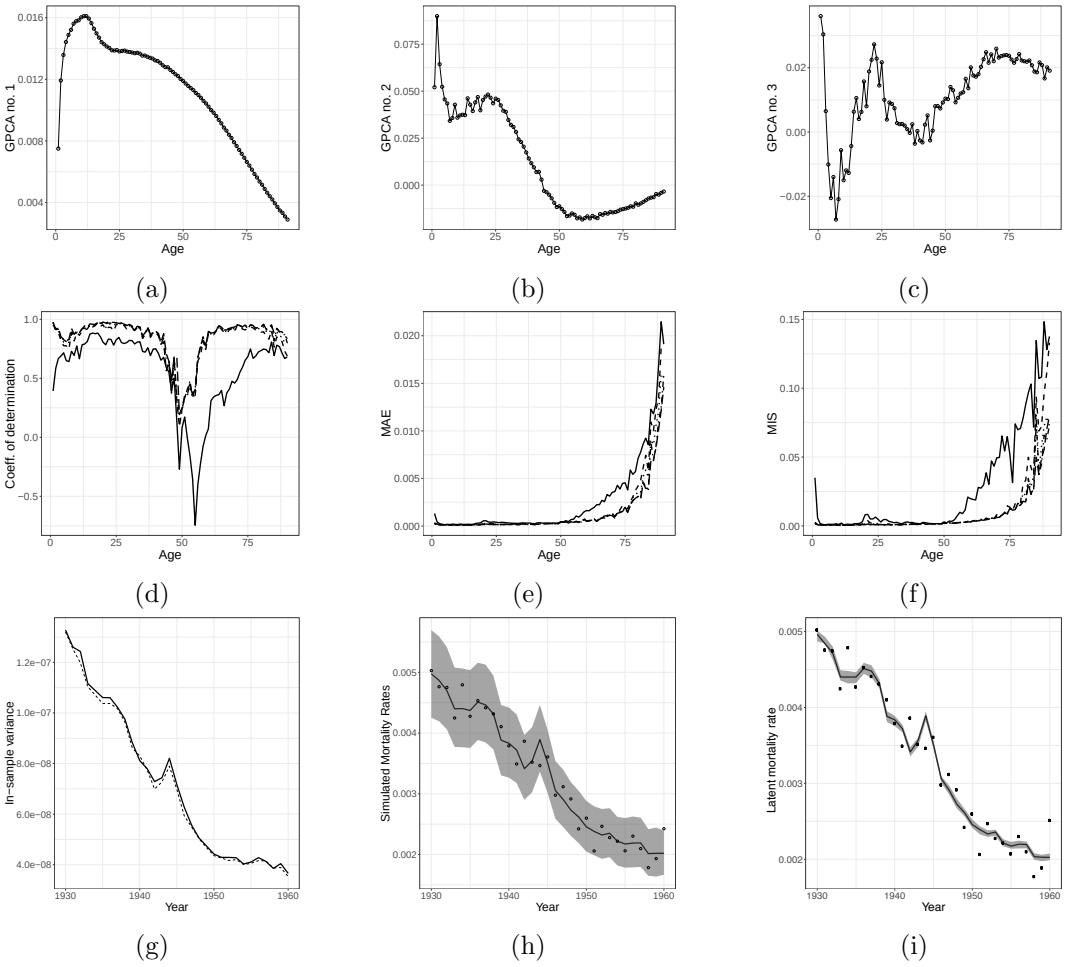

Figure 3: Fig. 3(a)–3(c): First three GPCA components, Swedish males, 1930-1960. In Fig. 3(d) – 3(f), the number of GPCA components, 1 – 5, are indicated by lines that are solid, short dashed, dotted, dash-dotted, and long dashed, respectively. Fig. 3(d) – 3(f) shows, from left to right, $\mathrm{R}^{2,*}_{\mathrm{Dev}}$, 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 GPCA 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 GPCA 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.

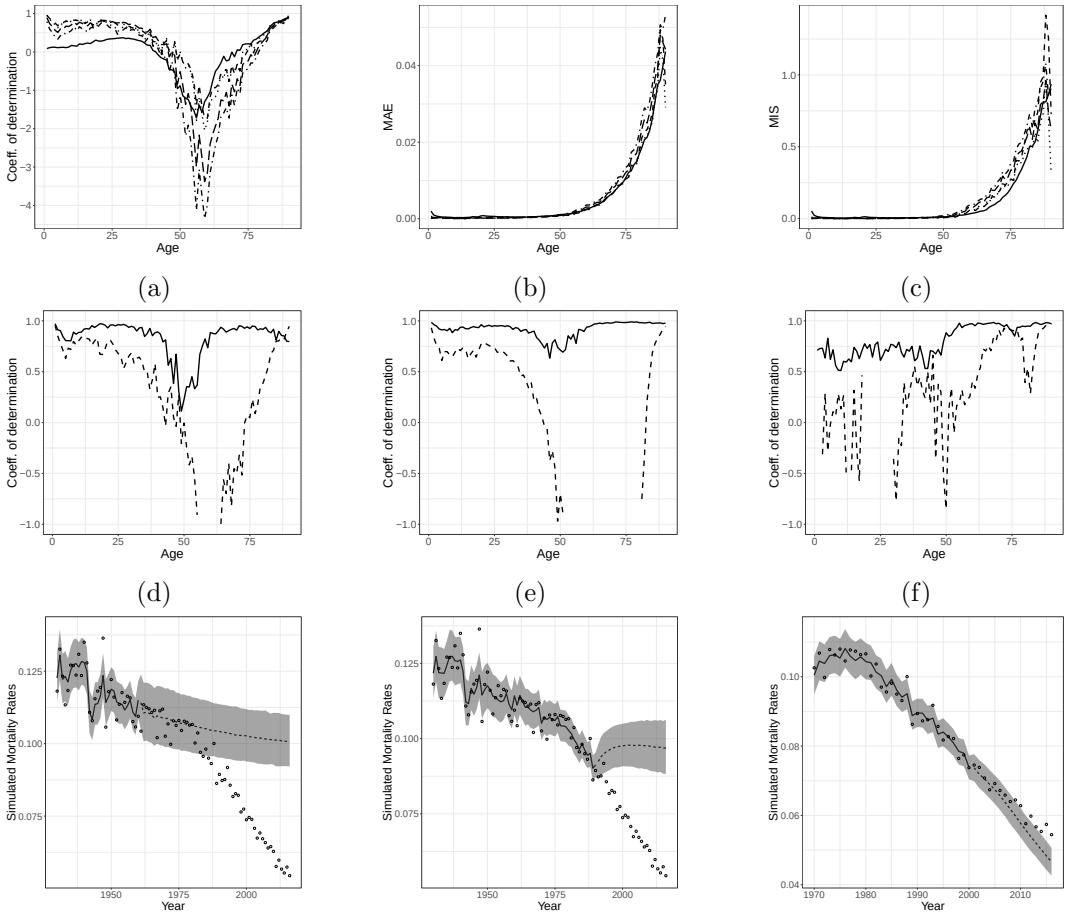

Figure 4: Fig. 4(a) – 4(c) shows the out-of-sample analog of Fig. 3(d) – 3(f), where all parameters have been estimated based on Swedish male data from 1930-1960, and the predictions are made for the period 1961-2016. Fig. 4(d) – Fig. 4(f): $\mathrm{R}^{2,*}_{\mathrm{Dev}}$ where solid lines corresponds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

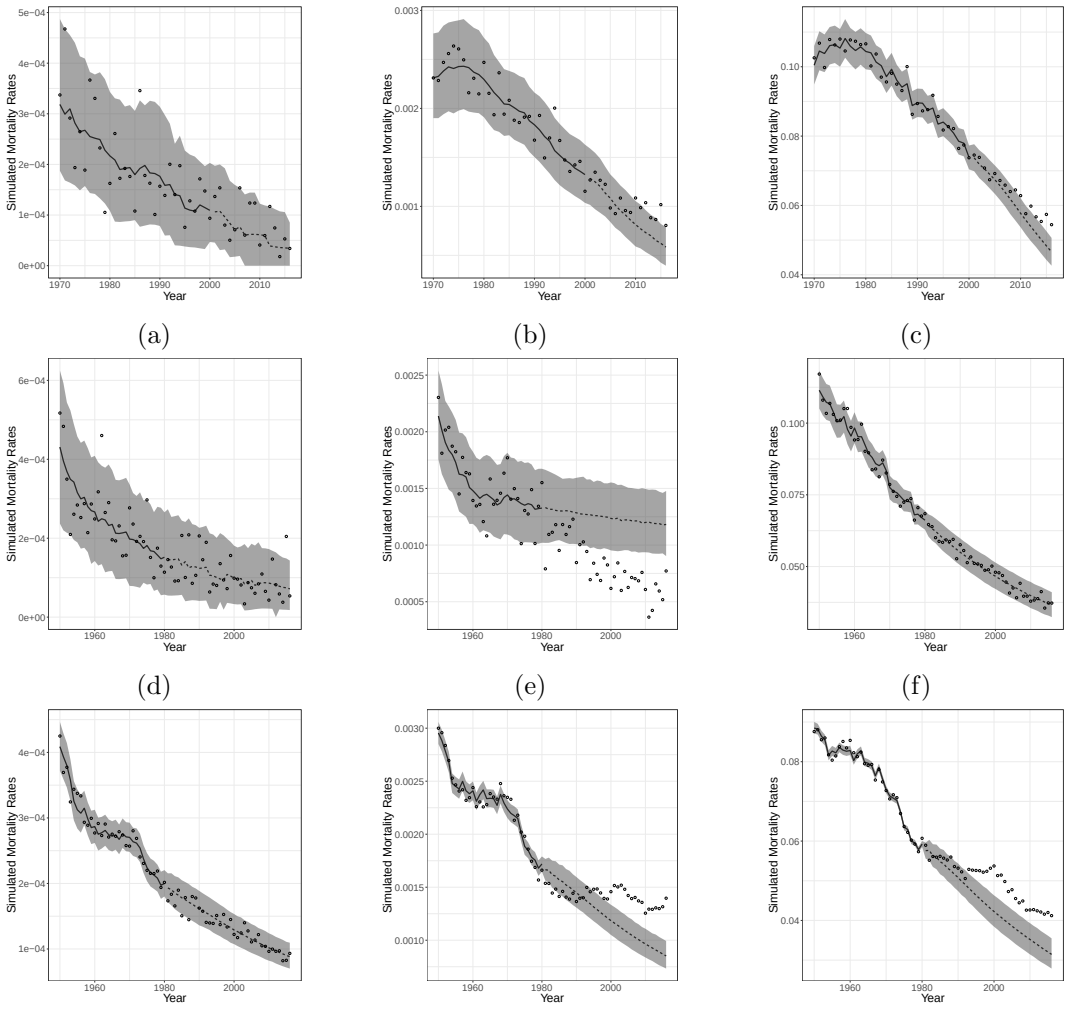

Figure 5: In all figures, 95% confidence/prediction intervals for the simulated centralized mortality rates (dashed lines), median (solid line), observed centralized mortality rates (circles). In all figures, from left to right, age 10, age 40, and age 80, respectively. Fig. 5(a)–5(c): Swedish males,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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Number 38. Oxford University Press.

Ekheden, E. and Hössjer, O. (2014). Analysis of the stochasticity of mortality using variance decomposition. In *Modern Problems in Insurance Mathematics*, pages 199–222. Springer.

Ekheden, E. and 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., and Shevchenko, P. V. (2017). A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. *Annals of Actuarial Science*, 11(2):343–389.

Gneiting, T. and 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., and West, M. (2004). Monte carlo smoothing for nonlinear time series. *Journal of the american statistical association*, 99(465):156–168.

Gompertz, B. (1825). 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. *Philosophical transactions of the Royal Society of London*, pages 513–583.

Gorski, J., Pfeuffer, F., and 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. and 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 Univ. Press, Princeton, N.J.

Human Mortality Database (2018). Available at `http://www.mortality.org` or `http://www.humanmortality.de` (downloaded on December 5, 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)*, pages 371–412.

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

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., and 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 at `http://www.mortality.org` or `http://www.humanmortality.de` (downloaded on December 5, 2018).

A Proofs and mathematical motivations

A.1 Proof of Lemma 2.1

Following standard theory of counting processes with multiplicative intensity processes, see e.g. 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

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

Thus, if we consider the situation with constant hazard rates on yearly Lexis-squares, i.e. $\mathcal{M} = \left\{ m_{\mathcal{S}} \mid \mathcal{S} \in \bar{\mathcal{S}} \right\}$, 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.

A.2 Proof of Lemma 3.1

These are standard results for VAR processes, see e.g. 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,

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

Model (M1)

By combining (7) and (4),

$$\begin{aligned} f_\psi(x_{0:n}) &= v(x_0)\prod_{t=1}^{n} f_\psi(x_t \mid x_{t-1}) \\ &\propto |\Sigma|^{-\frac{n}{2}} \exp\Big\{ -\tfrac{1}{2}\sum_{t=1}^{n} (x_t - \Gamma x_{t-1} - \mu)' \Sigma^{-1} (x_t - \Gamma x_{t-1} - \mu) \Big\}. \end{aligned}$$

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

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

it follows that

$$\begin{aligned} &-\tfrac{1}{2}\sum_{t=1}^{n} (x_t - \Gamma x_{t-1} - \mu)' \Sigma^{-1} (x_t - \Gamma x_{t-1} - \mu) \\ &= \operatorname{vec}\begin{pmatrix} \Sigma^{-1} \\ -2\Sigma^{-1}\Gamma \\ \Gamma'\Sigma^{-1}\Gamma \\ -2\mu'\Sigma^{-1} \\ \mu'\Sigma^{-1}\Gamma \end{pmatrix} \cdot \operatorname{vec}\begin{pmatrix} \sum_{t=1}^n x_t x_t' \\ \sum_{t=1}^n x_{t-1} x_t' \\ \sum_{t=1}^n x_{t-1} x_{t-1}' \\ \sum_{t=1}^n x_t \\ \sum_{t=1}^n x_{t-1} \end{pmatrix} + n\mu'\Sigma^{-1}\mu. \end{aligned}$$

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

$$S(x_{0:n}) = \begin{pmatrix} S_1(x_{1:n}) \\ S_2(x_{0:n}) \\ S_3(x_{0:n-1}) \\ S_4(x_{1:n}) \\ S_5(x_{0:n-1}) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^n x_t x_t' \\ \sum_{t=1}^n x_{t-1} x_t' \\ \sum_{t=1}^n x_{t-1} x_{t-1}' \\ \sum_{t=1}^n x_t \\ \sum_{t=1}^n x_{t-1} \end{pmatrix}. \tag{15}$$

Towards finding the ML estimators, define

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

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

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

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

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

with solution

$$\begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix} = \begin{bmatrix} S_4 & S_2' \end{bmatrix} \begin{bmatrix} n & S_5' \\ S_5 & S_3 \end{bmatrix}^{-1}. \tag{16}$$

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

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

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

$$\begin{aligned}\widehat{\Sigma} &= \frac{1}{n} \sum_{t=1}^{n} \widehat{\varepsilon}_t \widehat{\varepsilon}_t' = \frac{1}{n} \sum_{t=1}^{n} \left(x_t - \begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix} \begin{bmatrix} 1 \\ x_{t-1} \end{bmatrix} \right) \left(x_t - \begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix} \begin{bmatrix} 1 \\ x_{t-1} \end{bmatrix} \right)' \\ &= \frac{1}{n} \left(S_1 - \begin{bmatrix} S_4 & S_2' \end{bmatrix} \begin{bmatrix} \widehat{\mu}' \\ \widehat{\Gamma}' \end{bmatrix} - \begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix} \begin{bmatrix} S_4' \\ S_2 \end{bmatrix} + \begin{bmatrix} \widehat{\mu} & \widehat{\Gamma} \end{bmatrix} \begin{bmatrix} n & S_5' \\ S_5 & S_3 \end{bmatrix} \begin{bmatrix} \widehat{\mu}' \\ \widehat{\Gamma}' \end{bmatrix} \right). \end{aligned} \tag{17}$$

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,

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

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

$$\begin{bmatrix} \widehat{\mu} & \widehat{\Gamma}_X \end{bmatrix} = \left(\begin{bmatrix} S_4(x_{1:n}) & S_2'(x_{0:n-1}) \end{bmatrix} - \begin{bmatrix} S_5(k_{0:n-1}) & \sum_{t=1}^{n} k_{t-1} x_{t-1}' \end{bmatrix} \right) \begin{bmatrix} n & S_5(x_{0:n-1})' \\ S_5(x_{0:n-1}) & S_3(x_{0:n-1}) \end{bmatrix}^{-1} \tag{18}$$

The MLE of Σ is then

$$\begin{aligned}
\widehat{\Sigma} &= \frac{1}{n}\sum_{t=1}^{n}\left(x_t - k_{t-1} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)\left(x_t - k_{t-1} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}1\\ x_{t-1}\end{bmatrix}\right)' \\
&= \frac{1}{n}\Bigg(S_1(x_{1:n}) + S_3(k_{0:n-1}) + \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}n & S_5'(x_{0:n-1})\\ S_5(x_{0:n-1}) & S_3(x_{0:n-1})\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix} && (19)\\
&\quad - \sum_{t=1}^{n} x_t k_{t-1}' - \sum_{t=1}^{n} k_{t-1}x_t' - \begin{bmatrix}S_4(x_{1:n}) & S_2'(x_{0:n})\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix} - \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}S_4'(x_{1:n})\\ S_2(x_{0:n})\end{bmatrix} && (20)\\
&\quad + \begin{bmatrix}S_5(k_{1:n}) & \sum_{t=1}^{n} k_{t-1}x_{t-1}'\end{bmatrix}\begin{bmatrix}\widehat{\mu}'\\ \widehat{\Gamma}'\end{bmatrix} + \begin{bmatrix}\widehat{\mu} & \widehat{\Gamma}\end{bmatrix}\begin{bmatrix}S_5'(k_{1:n})\\ \sum_{t=1}^{n} x_{t-1}k_{t-1}'\end{bmatrix}\Bigg). && (21)
\end{aligned}$$

A.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

$$h(z; k, d) = k\mathrm{e}^z - dz,$$

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

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

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 e.g. (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 w.r.t. 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 $x_{0:n}$.