

Mortality Forecasting using Variational Inference

Patrik Andersson* Mathias Lindholm†

Abstract

This paper considers the problem of forecasting mortality rates. A large number of models have already been proposed for this task, but they generally have the disadvantage of either estimating the model in a two-step process, possibly losing efficiency, or relying on methods that are cumbersome for the practitioner to use. We instead propose using variational inference and the probabilistic programming library Pyro for estimating the model. This allows for flexibility in modelling assumptions while still being able to estimate the full model in one step.

The models are fitted on Swedish mortality data and we find that the in-sample fit is good and that the forecasting performance is better than other popular models.

Code is available at <https://github.com/LPAndersson/VImortality>.

Keywords: Non-linear state-space-models; Mortality forecasting; Hidden Markov model, Variational inference

1 Introduction

Attempts to forecast mortality go back at least as far as Gompertz (1825). A more recent example is the Lee-Carter model (Lee and Carter, 1992) and its extensions, see Booth and Tickle (2008); Haberman and Renshaw (2011); Carfora et al. (2017) for a survey. Applications of mortality forecasting can be found in for example demographic predictions and in the insurance industry.

The Lee-Carter model is a log-linear multivariate Gaussian model of mortality rates. A major criticism of Lee-Carter-type models is that the model training is done as a two-step process. In the first step, point estimates of the mortality rates are obtained, for example as the maximum likelihood estimate of a Poisson distribution, and in the second step, a latent process is fitted to these estimates. This method has the advantage that it is simple and fast to implement, but it is inefficient when compared to the simultaneous estimation of all unknown parameters. Also, it is not possible to distinguish between the finite population noise of the mortality estimates and the noise from the latent process. Both of these issues can potentially affect the quality of the forecasts.

Simultaneous estimation of parameters has been considered in Andersson and Lindholm (2021) where particle filtering methods are used to estimate a state-space model

*Uppsala University (patrik.andersson@statistics.uu.se)

†Stockholm University (lindholm@math.su.se)

with Poisson distributed observations, similar to Brouhns et al. (2002). However, this method has its drawbacks. It could be considered cumbersome for practitioners as it requires custom implementation and tuning, and since the particle filter methods are computationally expensive, the number of parameters must not be too large. The complexity of the modelling is reduced when changing the observational model from a Poisson distribution to a Gaussian, see e.g. De Jong and Tickle (2006) for a state-space model treatment of a Gaussian Lee-Carter model.

Recently it has been suggested to use models from deep learning, sometimes called deep factor models, to forecast high-dimensional multivariate time series. Some examples of this can be found in Nguyen and Quanz (2021); Wang et al. (2019); Salinas et al. (2020); Rangapuram et al. (2018). The applications presented in those articles differ from mortality forecasting in the scale of the problem. In mortality forecasting, the dimension of the time series is about 100 (the lifetime in years of a human) and the number of observations of the time series is also about 100 (although some countries do have reliable data for longer than that). As a consequence, to avoid overfitting, we need to consider simpler models. This includes simpler functions for mapping latent variables to the observed time series, linear Gaussian models instead of RNNs for propagating the latent variables forward in time and fewer latent factors.

Compared to previous mortality forecasting models, the novelty of this paper is therefore to use black-box variational inference (Ranganath et al., 2014) to solve the inference problem. This means that after specifying how to sample from the model and the approximate posterior, the inference is done automatically without any model-specific customisation. The family of models that can be handled is also expanded. For example, one can consider other families of functions for mapping the latent process to mortality rates. Also, in this case, all the parameters can be estimated simultaneously. This latter point is problematic when using particle filter techniques, where it is necessary to estimate the linear mapping from the latent process to the mortality rates first, before continuing with the estimation of the other parameters.

Other approaches that use machine learning techniques for forecasting mortality rates can be found in e.g. Richman and Wuthrich (2019); Richman and Wüthrich (2021); Perla et al. (2021) that consider various types of Gaussian recurrent neural network structures, Nigri et al. (2019); Marino and Levantesi (2020); Lindholm and Palmborg (2022) that consider univariate LSTM neural network, both with and without a Poisson population assumption, and Deprez et al. (2017) that consider tree-based techniques.

The model is implemented using the probabilistic programming language Pyro (Bingham et al., 2018) and the code is available at <https://github.com/LPAndersson/VImortality>.

The rest of the paper is organised as follows: In Section 2 we describe the probabilistic model that will be used for forecasting. In Section 3 we give a brief introduction to variational inference and in Section 4 we describe how to forecast the mortality once the model has been trained and how we validate the forecast. In Section 5 we demonstrate our method on an example and compare to other models. Section 6 concludes the paper.

2 Model

In this section, we describe the probabilistic model that defines the mortality dynamics. Uppercase letters will denote random variables and lowercase letters the corresponding observed value. Greek letters will denote unknown parameters that are to be estimated.

Mortality data can be aggregated in different forms and clearly the choice of model will have to be adjusted accordingly. For example, the data could contain the population size of each age group at the beginning of the year and the number of deaths during the year. In this case, a binomial model seems natural. We however consider data on the yearly number of deaths and the exposure to risk in each age group. The exposure to risk in this setting is the total time that the individuals in the population were a certain age in a certain year. The number of deaths in age $a \in \{0, 1, \dots, \bar{a}\}$, year $t \in \{0, 1, \dots, \bar{t}\}$ is denoted by $D_{a,t}$ and the exposure by $E_{a,t}$.

Our model is a state-space model that can be written as:

$$D_{a,t} \mid X_t, E_{a,t} \sim \text{Poisson} \left(E_{a,t} \exp \left(f_a^\psi(X_t) \right) \right), \quad (1)$$

$$X_{i,t} = X_{i,t-1} + K_{i,t-1} + U_{i,t}, \quad U_{i,t} \text{ iid } \mathbf{N}(0, \sigma_{X,i}^2), \quad (2)$$

$$K_{i,t} = \mu_i + \varphi_i(K_{i,t-1} - \mu_i) + V_{i,t}, \quad V_{i,t} \text{ iid } \mathbf{N}(0, \sigma_{K,i}^2). \quad (3)$$

Here $i = 1, 2, \dots, d$, where d is the dimension of the latent variables. We also require $0 \leq \varphi_i \leq 1$. The function f_a^ψ is the a :th component of $f^\psi : \mathbb{R}^d \rightarrow \mathbb{R}^{\bar{a}+1}$. In our examples in Section 5, f^ψ will be given by either an affine transformation or a sum of radial basis functions. That is either, $f^\psi(x) = Ax + b$, where A and b are trainable or the a th component of f^ψ is

$$f_a^\psi(x) = x^T \sum_{i=1}^p w_i e^{-\tau^2 \left(\frac{a - \mu_i}{\bar{a} + 1} \right)^2} + b_a.$$

We will fix τ^2 and therefore $\{w_i, \mu_i\}_{i=1}^p$ and $\{b_a\}_{a=0}^{\bar{a}}$ are the trainable parameters. Compared to the more general affine transformation, radial basis functions have the advantage of inducing a certain smoothness of f_a as a function of a , encoding a prior that similar ages should have similar mortality.

We remark here that the exact specification of the above model is not critical for the continuation. For example, the exponential link function in the Poisson distribution could be changed to some other positive differentiable function without complication. We are assuming that the components of the latent process are independent, instead, we let any dependence be captured by f . However, this latent process could be replaced with some other Markov process.

3 Variational inference

Here we explain the main ideas of variational inference in a general setting. At the end of the section, we connect this to our specific model. For more on variational inference in general we refer to Ranganath et al. (2014) and for the application to state space models, see Archer et al. (2015).

We are observing y , whose distribution depends on a latent variable x and an unknown parameter ψ . This is modelled by the joint distribution

$$p_\psi(y, x) = p_\psi(y | x)p_\psi(x).$$

The likelihood,

$$L(\psi) = p_\psi(y) = \int p_\psi(y, x)dx,$$

is in general not tractable and therefore approximations are needed in order to be able to estimate ψ . Consider a parametrised distribution, the approximate posterior, $q_\theta(x)$. Then observe that, due to Jensen's inequality, the log-likelihood is

$$\begin{aligned} l(\psi) &:= \log L(\psi) = \log \int \frac{p_\psi(y, x)}{q_\theta(x)} q_\theta(x) dx \\ &\geq \int (\log p_\psi(y, x) - \log q_\theta(x)) q_\theta(x) dx =: \mathcal{L}(\psi, \theta). \end{aligned}$$

The right-hand side is known as the evidence lower bound (ELBO). The idea of variational inference is to instead of maximising the log-likelihood, maximise the ELBO. Towards this, we calculate the gradients

$$\begin{aligned} \partial_\psi \mathcal{L}(\psi, \theta) &= \int \partial_\psi \log p_\psi(y, x) q_\theta(x) dx, \\ \partial_\theta \mathcal{L}(\psi, \theta) &= \int (\log p_\psi(y, x) - \log q_\theta(x)) \partial_\theta \log q_\theta(x) q_\theta(x) dx. \end{aligned}$$

We can then proceed to obtain unbiased estimates of the gradients by sampling from q_θ and maximise \mathcal{L} using stochastic optimisation algorithms. Once converged, $q_\theta(x)$ can be used as an approximation of the posterior distribution of the latent variables $p_\psi(x | y)$.

Further, to obtain faster convergence, various variance reduction techniques are often used. Here we only mention the so-called reparametrisation trick. Suppose that we can find functions x_θ such that

$$\int f(x) q_\theta(x) dx = \int f(x_\theta(z)) q(z) dz, \tag{4}$$

which makes the sampling distribution independent of θ . In particular, the gradient satisfies

$$\partial_\theta \int f(x) q_\theta(x) dx = \partial_\theta \int f(x_\theta(z)) q(z) dz = \int \partial_\theta f(x_\theta(z)) q(z) dz,$$

which usually improves the sampling variance, compared to differentiating the density directly. An important example of a distribution that allows for reparametrisation according to (4) is the Gaussian, since if $Z \sim \mathcal{N}(0, 1)$ then $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

In the numerical illustrations in Section 5, the approximate posterior is modelled as a Gaussian distribution with an autoregressive covariance. That is, the distribution of the process is given by

$$\begin{aligned}\tilde{X}_{i,t} &= \tilde{\mu}_{i,t}^X + \alpha_{i,t}\tilde{X}_{i,t-1} + \tilde{e}_{i,t}^X, & \tilde{e}_{i,t}^X &\text{iid } \mathbf{N}(0, \tilde{\sigma}_{X,i}^2), \\ \tilde{K}_{i,t} &= \tilde{\mu}_{i,t}^K + \beta_{i,t}\tilde{K}_{i,t-1} + \rho_{i,t}\tilde{X}_{i,t-1} + \tilde{e}_{i,t}^K, & \tilde{e}_{i,t}^K &\text{iid } \mathbf{N}(0, \tilde{\sigma}_{K,i}^2).\end{aligned}$$

4 Forecasting and validation

By maximizing the ELBO we have obtained estimates of φ and θ and the joint distribution of $(\tilde{X}_{i,\bar{t}}, \tilde{K}_{i,\bar{t}})$. This allows us to proceed with forecasting the mortality, as will be discussed in this section.

Since both the approximate posterior and the latent process is Gaussian, the forecasting distribution of the latent process is also Gaussian. That is, for $t > \bar{t}$,

$$\begin{pmatrix} \hat{X}_{i,t} \\ \hat{K}_{i,t} \end{pmatrix} \sim \mathbf{N} \left(\begin{pmatrix} \hat{\mu}_{i,t}^X \\ \hat{\mu}_{i,t}^K \end{pmatrix}, \begin{pmatrix} \hat{\sigma}_{X,t}^2 & \hat{\rho}_{i,t}\hat{\sigma}_{X,t}\hat{\sigma}_{K,t} \\ \hat{\rho}_{i,t}\hat{\sigma}_{X,t}\hat{\sigma}_{K,t} & \hat{\sigma}_{K,t}^2 \end{pmatrix} \right),$$

where the parameters can be calculated iteratively from (2) and (3), by using the initial value

$$\begin{pmatrix} \hat{X}_{i,\bar{t}} \\ \hat{K}_{i,\bar{t}} \end{pmatrix} := \begin{pmatrix} \tilde{X}_{i,\bar{t}} \\ \tilde{K}_{i,\bar{t}} \end{pmatrix},$$

and the forecast of mortality rates is given by $\exp(f^{\hat{\psi}}(\hat{X}_t))$.

If one wants to forecast the actual number of deaths, a forecast of the number of living at the beginning of the year is also needed, together with some assumption on the distribution of when in the year people are born. For a longer discussion on how this can be done, we refer to Andersson and Lindholm (2021).

The forecast is validated by calculating the logarithmic score of the forecast on the validation data set. The logarithmic score is the logarithm of the predicted density evaluated at the observed value, see for example Gneiting and Raftery (2007). That is, if $\mathcal{P}_{a,t}$ is the forecasted distribution of death counts from (1) and $d_{a,t}$ is the observation, the log-score is

$$\text{logs}(\mathcal{P}_t, d_t) := \sum_a \log \mathcal{P}_{a,t}(d_{a,t}).$$

We will evaluate the models using a rolling window of training and evaluation data and calculate the average score over these windows, which we call $\overline{\text{logs}(\mathcal{P}, d)}$. An alternative that we also calculate is

$$R^2 = \frac{\overline{\text{logs}(\mathcal{P}, d)} - \overline{\text{logs}(\overline{\mathcal{P}}, d)}}{\overline{\text{logs}(\hat{\mathcal{P}}, d)} - \overline{\text{logs}(\overline{\mathcal{P}}, d)}}.$$

Here, $\overline{\mathcal{P}}$ denotes the model with only a constant intercept and $\hat{\mathcal{P}}$ the saturated model, i.e. the model that forecasts with mean equal to the observation.

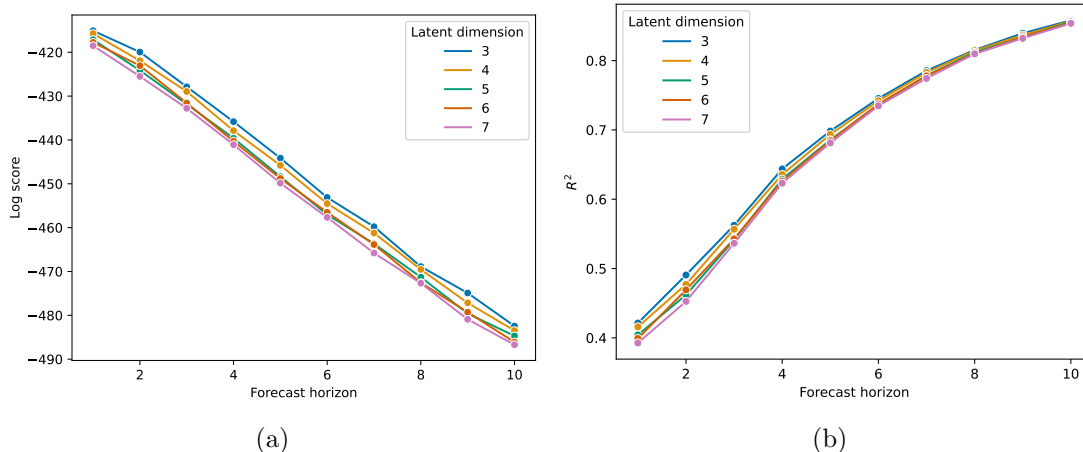


Figure 1: Affine model: Evaluation of the model when fitted on Swedish male mortality data with the first year from 1931 to 1952. Each model is fitted using 60 years of data and evaluated on the following 10 years. The figures show that 3 latent dimensions give the best out-of-sample performance.

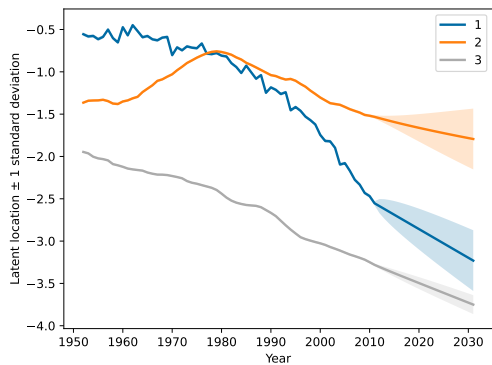
5 Results

In this section, we illustrate the performance of the models by fitting it to a dataset on the mortality of Swedish males. The dataset is collected from Human Mortality Database (2022). We evaluate the models using a rolling window; using 60 years to fit the model and 10 years to evaluate the forecast against the actual outcome. The training and evaluation windows are then rolled forward one year and the process repeats.

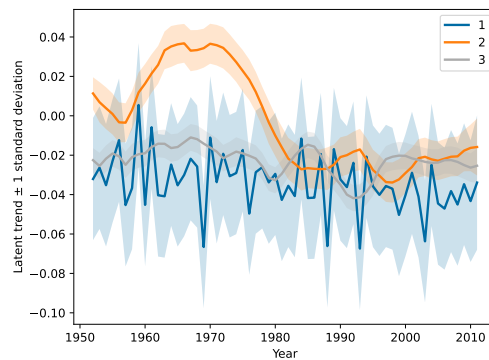
The model is as in (1) - (3) where f is either affine or a sum of radial basis functions. We begin by selecting hyperparameters for each model, e.g. dimension of the affine transformation or the number of radial bases. We then illustrate the fitted model and the out-of-sample forecasting performance. Finally, we compare our model with the Lee-Carter model and the model by Plat (2009).

5.1 Affine

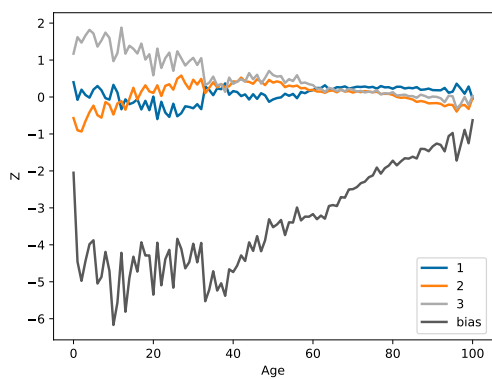
For model selection, we compare the out-of-sample log-score for varying dimensions of the latent process. In Figure 1 we see that 3 latent dimensions overall performs best. We have performed experiments also for dimensions 1 and 2, but the performance was considerably worse, and they are therefore excluded from the figure. Figure 2 illustrates the fitted model. In particular, we note in Figure 2d that the in-sample fit of the mortality rates are quite good. Figure 3 shows the smoothed and forecasted mortality rates for the ages 20, 40, 60 and 80. The shaded regions represent ± 1 standard deviation and the grey dots are the observed mortality rates. That is, we should expect that around 7 out of 10 observations are within the shaded region. The pictures seem to confirm this.



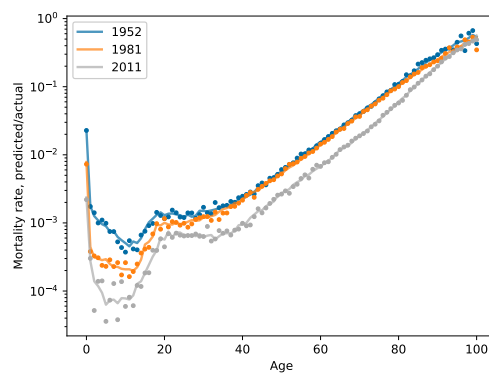
(a) Level of latent process in-sample and its forecast. Shaded region is ± 1 standard deviation. From 1952 to 2011 shows smoothed values, from 2012 shows forecast.



(b) Trend of latent process. The shaded region is ± 1 standard deviation.



(c) Factor loadings for the model with 3 latent dimensions



(d) In-sample fit of model mortality rates. Dots are observed rates and solid lines are the fitted model. This model is fitted on data from 1952 to 2011.

Figure 2: Affine model: Illustration of the model fitted to Swedish data from 1952 to 2011 with three latent dimensions.

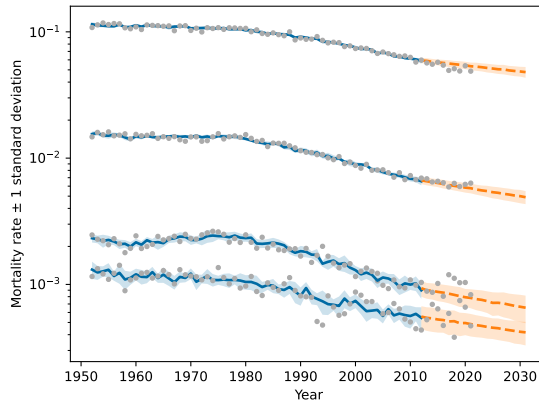


Figure 3: Affine model: Mortality forecasts for the age groups 20, 40, 60 and 80 (from the bottom up). The solid line is the smoothed mortality rate, the dashed line is the mean forecasted mortality rate and the shaded region is ± 1 standard deviation of the forecasted realised mortality rate. Dots indicate observed mortality rates.

5.2 Radial basis functions

This section follows the same pattern as the previous one. We choose hyperparameters for the radial basis functions and illustrate the fitted model. In all models we choose $\tau = 10$. Although this is a parameter that could be trained, our experiments show that it is difficult to train well. Our choice of τ corresponds to a typical width of the radial basis of about 14 years, which seems reasonable.

In Figure 4 we see that four latent dimensions and 15 radial basis functions give the best out-of-sample performance. In Figure 5 the model fit is illustrated. In particular, we note in figures 5c and 5d that the radial basis functions do indeed make both the factor loadings and the fitted mortality curves more smooth, compared to the affine model. In Figure 6 we see that the forecasted mortality rates are quite similar to the affine model.

5.3 Comparison

In this section we compare our two models to two other commonly used mortality models, the Lee-Carter model with Poisson distribution from Brouhns et al. (2002) and the model from Plat (2009). Both are fitted using the StMoMo package in R (Villegas et al., 2018). They both model mortality as

$$D_{a,t} \sim \text{Poisson}(E_{a,t} \exp(\eta_{a,t})),$$

where in the Lee-Carter model,

$$\eta_{a,t} = \alpha_a + \beta_a \kappa_t,$$

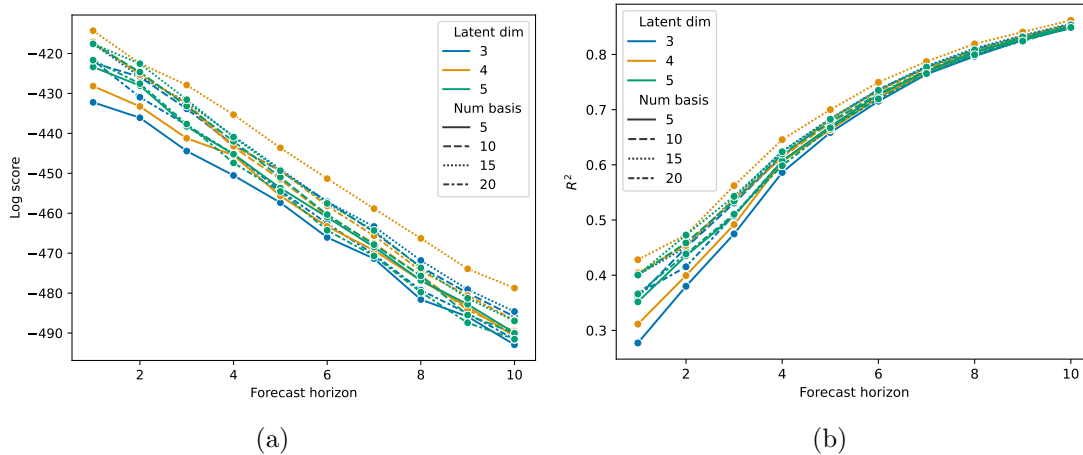


Figure 4: Radial basis model: Evaluation of the model when fitted on Swedish male mortality data with the first year from 1931 to 1952. Each model is fitted using 60 years of data and evaluated on the following 10 years. The figures show that 4 latent dimensions and 15 radial basis functions give the best out-of-sample performance.

where α_a and β_a are factor loadings and κ_t is the dynamic factor, modelled as a random walk with drift. In the Plat model,

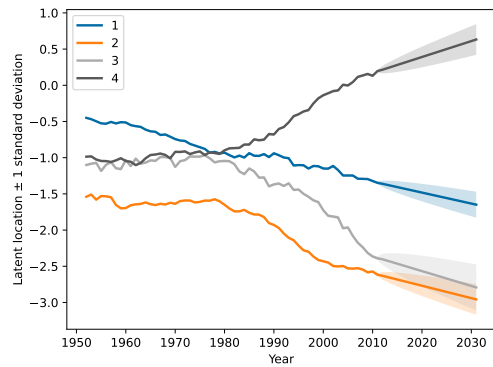
$$\eta_{a,t} = \alpha_a + \kappa_t^{(1)} + (\bar{a} - a)\kappa_t^{(2)} + (\bar{a} - a)^+\kappa_t^{(3)} + \gamma_{t-a}.$$

Here the $\kappa_t^{(\cdot)}$ s follow a multivariate random walk with drift and γ_{t-a} is ARIMA(2,0,0) with intercept. The unknowns are estimated using maximum likelihood, and then the dynamic factors are modelled and forecasted.

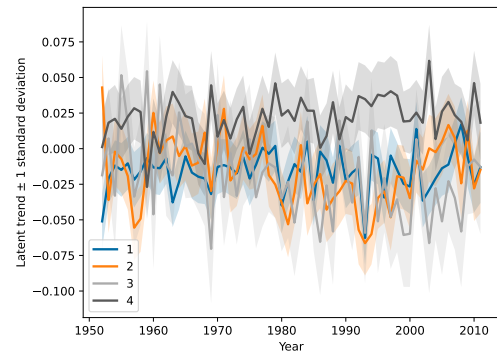
The model performance is summarized in Table 1 and Figure 7 where the log-score and R^2 are shown for each model. We see that our two models perform almost identically and improve substantially on the two compared models.

Model	Log-score	R^2
Affine	-448.2	0.686
Radial basis	-447.3	0.687
Lee-Carter	-584.6	0.263
Plat	-477.6	0.610

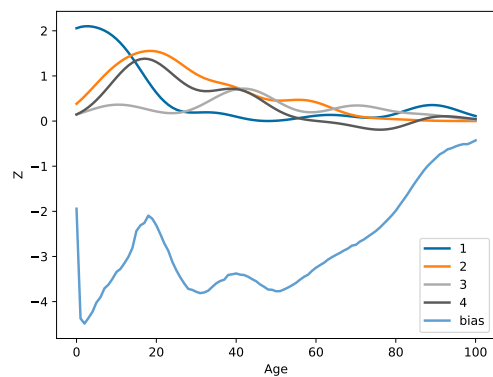
Table 1: Out-of-sample model evaluation metrics for Sweden. Numbers are averaged over all forecast horizons and rolling windows.



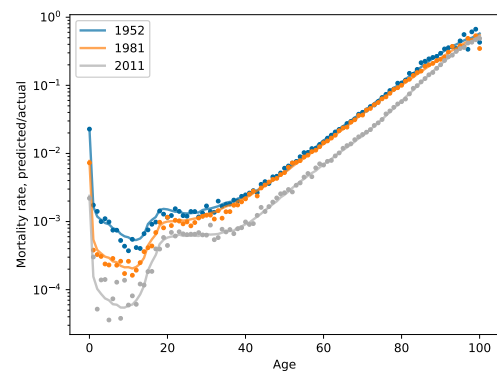
(a) Level of latent process in-sample and its forecast. Shaded region is ± 1 standard deviation. From 1952 to 2011 shows smoothed values, from 2012 shows forecast.



(b) Trend of latent process. The shaded region is ± 1 standard deviation.



(c) Factor loadings for the model with 4 latent dimensions



(d) In-sample fit of model mortality rates. Dots are observed rates and solid lines are the fitted model. This model is fitted on data from 1952 to 2011.

Figure 5: Radial basis model: Illustration of the model fitted to Swedish data from 1952 to 2011 with 4 latent dimensions.

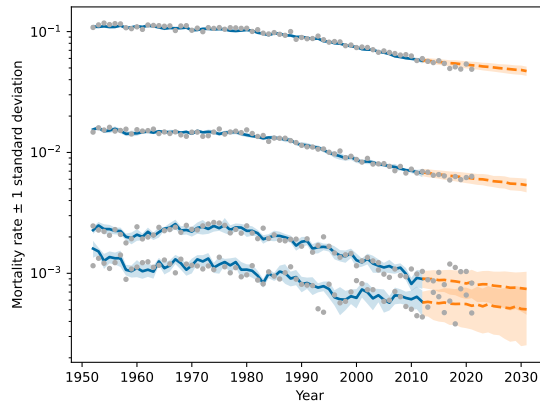


Figure 6: Radial basis model: Mortality forecasts for the age groups 20, 40, 60 and 80 (from the bottom up). The solid line is the smoothed mortality rate, the dashed line is the mean forecasted mortality rate and the shaded region is ± 1 standard deviation of the forecasted realised mortality rate. Dots indicate observed mortality rates.

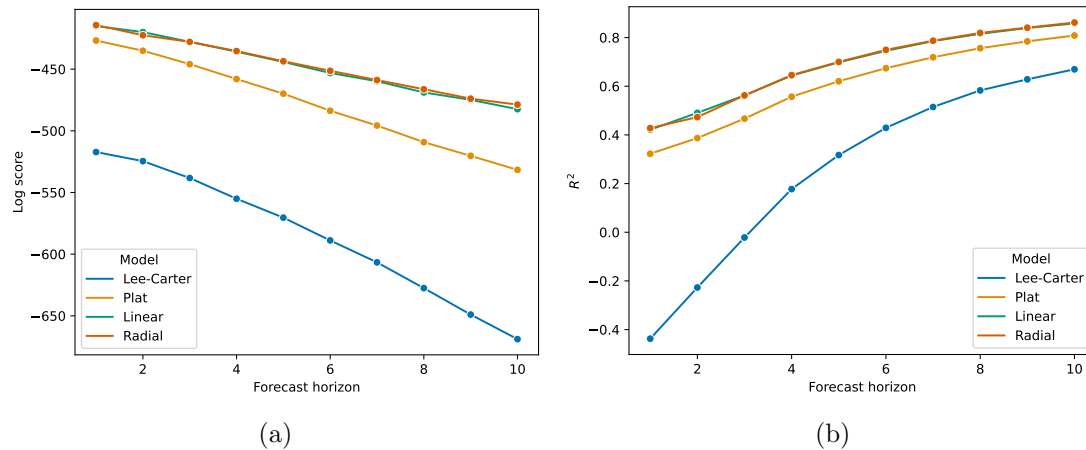


Figure 7: Comparison of our model and the Lee-Carter and Plat models. Each model is fitted using 60 years of data and evaluated on the following 10 years. The figures show that both the affine and radial basis models perform well.

6 Conclusions

In this paper, we have considered a state-space model for mortality forecasting and we have shown how it is possible to fit such a model using variational inference. Using variational inference it is possible to not only use a Poisson likelihood for the observed number of deaths but also to estimate the complete model in one step. The model is also flexible in that, for example, we can consider different functions for projecting the latent variables to the mortality curve. We considered both affine functions and radial basis functions, but the practitioner has the freedom to choose other classes of functions without complication. Another advantage is that the model is implemented in Pyro, so that very little custom code is needed. Finally, we show that our model and inference method outperform other popular methods.

Data availability statement

The data used in this paper can be downloaded free of charge from <https://www.mortality.org>.

References

- Andersson, P. and Lindholm, M. (2021). Mortality forecasting using a lexis-based state-space model. *Annals of actuarial science*, 15(3):519–548.
- Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. (2015). Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*.
- Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*.
- Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of actuarial science*, 3(1-2):3–43.
- 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.
- Carfora, M. F., Cutillo, L., and Orlando, A. (2017). A quantitative comparison of stochastic mortality models on italian population data. *Computational Statistics & Data Analysis*, 112:198–214.
- De Jong, P. and Tickle, L. (2006). Extending lee–carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.
- Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7:337–352.

- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- 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.
- Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.
- Human Mortality Database (2022). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at <http://www.mortality.org> (downloaded on January 22, 2022).
- Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.
- Lindholm, M. and Palmborg, L. (2022). Efficient use of data for lstm mortality forecasting. *European Actuarial Journal*, pages 1–30.
- Marino, M. and Levantesi, S. (2020). Measuring longevity risk through a neural network lee-carter model. *Available at SSRN 3599821*.
- Nguyen, N. and Quanz, B. (2021). Temporal latent auto-encoder: A method for probabilistic multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9117–9125.
- Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated lee-carter model. *Risks*, 7(1):33.
- Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7):572–598.
- Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.
- Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*.
- Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. (2018). Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31:7785–7794.
- Richman, R. and Wuthrich, M. V. (2019). Lee and carter go machine learning: recurrent neural networks. *Available at SSRN 3441030*.

- Richman, R. and Wüthrich, M. V. (2021). A neural network extension of the lee–carter model to multiple populations. *Annals of Actuarial Science*, 15(2):346–366.
- Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.
- Villegas, A. M., Kaishev, V. K., and Millossovich, P. (2018). StMoMo: An R package for stochastic mortality modeling. *Journal of Statistical Software*, 84(3):1–38.
- Wang, Y., Smola, A., Maddix, D., Gasthaus, J., Foster, D., and Januschowski, T. (2019). Deep factors for forecasting. In *International Conference on Machine Learning*, pages 6607–6617. PMLR.