Bayesian Mortality Forecasting with Overdispersion

Jackie S. T. Wong [a,*], Jonathan J. Forster [a], Peter W. F. Smith [b]

[a] Mathematical Sciences , University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[b] Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Abstract

The ability to produce accurate mortality forecasts, accompanied by a set of representative uncertainty bands, is crucial in the planning of public retirement funds and various life-related businesses. In this paper, we focus on one of the drawbacks of the Poisson Lee-Carter model that imposes mean-variance equality, restricting the mortality variations across individuals. Specifically, we present two models to potentially account for overdispersion. They are fitted within a Bayesian paradigm for coherency. Markov Chain Monte Carlo (MCMC) methods are implemented to carry out parameter estimation. Several comparisons are made with the Bayesian Poisson Lee-Carter model to highlight the importance of accounting for overdispersion. We demonstrate that the methodology we developed prevents over-fitting and yields better calibrated prediction intervals for the purpose of mortality projections. Bridge sampling is used to approximate the marginal likelihood of each candidate model to perform model determination.

Keywords: Mortality forecast; Overdispersion; Bayesian methods; MCMC; Bridge sampling.

[*]Corresponding Author.

E-mail addresses: jstw1g09@soton.ac.uk (J.S.T. Wong), J.J.Forster@soton.ac.uk (J.J. Forster), P.W.Smith@soton.ac.uk (P.W.F. Smith)

1 Introduction

Mortality forecasting is becoming an increasingly important issue especially recently in a wide variety of areas: funding of public retirement systems, planning of social security, medical health care systems, and actuarial applications (pricing and reserving of annuity portfolios). It is well established that mortality has been improving over the years. This poses an immediate threat to the government and various institutions because calculation of the expected present values of numerous life-related products using life annuities functions relies on an accurate projection of the mortality rates (longevity risk). Hence, development of appropriate models to model and forecast mortality is crucial to avoid adverse costs.

Stochastic models have gained a lot of popularity in mortality projection due to their ability to produce probabilistic intervals that encapsulate uncertainties associated with the forecasts, thereby facilitating informed decision making within an acceptable risk margin. The first stochastic model was pioneered by Lee and Carter (henceforth LC) in 1992, and has since then becomes the focus of most of the subsequent research in this regard. This model has gained worldwide acceptance too and is often applied in the context of stochastic mortality forecasting (Tuljapurkar et al. (2000)). For instance, it is used by the US census Bureau as a benchmark in their population forecasts. Lee and Miller (2001) demonstrated that the LC based forecasts led to a systematic underestimation of future life expectancies in the United States (see Girosi and King (2008) for more criticisms). Various modifications of the LC approach began to emerge thereafter. Brouhns et al. (2002) proposed a Poisson-equivalent version of the LC model by introducing Poisson random variation onto the number of deaths rather than an additive error term on the logarithm of mortality rates. Cairns et al. (2005) developed the CBD mortality model, which is a simple two-factor model that imposes a log-linear relationship between the death probabilities (in their definition) and age-time covariates. They demonstrated that the CBD model fits UK mortality data for ages above 60 and years 1961-2002 substantively well. For a comprehensive review of the recent development of mortality forecasting, readers are referred to Booth and Tickle (2008).

In this paper, we focus on one of the drawbacks of the Poisson LC model in Brouhns et al. (2002) that the mean and variance are restricted to be the same. This problem has been considered by several papers, which mainly recommend using mixed Poisson models. Renshaw and Haberman (2005) introduced a single dispersion parameter into the quasi-Poisson likelihood to increase the flexibility of their model specification, but made no attempt to assess the impact of this parameter on the prediction intervals. Their approach also suffers from the issue that the relationship between the expectation, variance and probability function of death data under the model are internally inconsistent (see Li et al. (2009)). Delwarde et al. (2007) then proposed a direct extension of the Poisson LC model to form the negative binomial LC model (again, they did not consider the construction of prediction intervals). In addition, Li et al. (2009) attempted to account for mortality variations by introducing an age-specific latent variable that accounts for heterogeneity of individuals, which upon marginalisation, leads to the negative binomial LC model as well. They also extended the parametric bootstrap approach in Brouhns et al. (2002) for the generation of prediction intervals. All these approaches considered model fitting within the classical framework. Our aim is to modify their methodology by fitting the mixed Poisson models within a Bayesian paradigm, on top of developing a new mixed Poisson model to account for overdispersion. The advantages of adopting Bayesian modelling will be discussed in details in Section 4. Bayesian mortality forecasting has generated some literature in its own right. For instance, Girosi and King (2008) introduced Bayesian modelling of mortality data in the presence of some exogenous covariates. On the other hand, Czado et al. (2005) fitted the Poisson LC model within the Bayesian framework.

We begin this paper by briefly reviewing the Poisson LC model in Section 2. The existence of overdispersion in the England and Wales female mortality data is also illustrated through a heat map. In Section 3, two extensions of the Poisson Lee-Carter model to account for overdispersion are presented. Section 4 discusses a coherent modelling approach by implementing the Bayesian methodology. The prior distributions of each of the unknown parameters are then provided. In Section 5, approaches to computation are given. In particular, we describe the Markov Chain Monte Carlo (MCMC) algorithm for posterior sample generation by deriving the conditional posterior distributions. Additionally, bridge sampling, a rather efficient and accurate approximation method for computing marginal likelihoods is considered. Some numerical results, including the fitted/projected parameters and Bayesian model determination, are presented in Section 7.

2 The Poisson LC (PLC) Model

Let D_{xt} denotes the number of deaths of age group x in year t, where $x = x_1, x_2, \ldots, x_A$ and $t = t_1, t_2, \ldots, t_T$ represent a set of A different age groups and T years respectively. Also let e_{xt} and μ_{xt} be the corresponding central exposed to risk and central mortality rate of age group x in year t.

Then, as proposed by Brouhns et al. (2002), the PLC model is given by

$$D_{xt} \sim \text{Poisson}(e_{xt}\mu_{xt}) \quad \text{with} \quad \log \mu_{xt} = \alpha_x + \beta_x \kappa_t. \tag{1}$$

For model identifiability, the constraints

$$\sum_x \beta_x = 1 \quad \text{and} \quad \sum_t \kappa_t = 0$$

are adopted as the model parameters are invariant to the following transformations:

$$\begin{aligned} \beta_x &\mapsto \frac{\beta_x}{b} \\ \kappa_t &\mapsto b(\kappa_t - k) \\ \alpha_x &\mapsto \alpha_x + k\beta_x, \end{aligned}$$

for any $b \in \mathbb{R}\backslash\{0\}$ and $k \in \mathbb{R}$. After imposing the constraints, the parameters can be interpreted as followed:

α_x : is the average of the logarithm of mortality rates over time (i.e. $\alpha_x = \frac{\sum_t \log \mu_{xt}}{T}$).

β_x : is the age-specific pattern of mortality improvement, measuring the sensitivity of each age's mortality to overall changes in mortality on the log scale.

κ_t : captures the overall time trend of mortality change (after being appropriately modulated by β_x).

In order to fit this model, weighted least squares (with D_{xt} as the weights) or Newton's iterative updating scheme can be used to obtain the maximum likelihood estimates $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$ (see Renshaw and Haberman (2005) for details). The ordinary generalized regression method does not work here due to the bilinear terms in Equation 1. One can however, fit this model within the generalized linear model (GLM) framework by iteratively conditioning on one of beta or kappa (so the parameters are now log-linear with respect to μ_{xt}) and estimating the remaining parameters until convergence. Note that there is no need to perform second stage estimation of κ_t to match the fitted deaths with observed deaths as in the original LC approach because Poisson variations automatically adjust for these discrepancies by modelling D_{xt} directly instead of μ_{xt}.

The key advantage of LC based models is that age and time components are partitioned such that the age components remain constant through time, while the time component intrinsically forms the stochastic part of the model to be projected forward in time. Hence, in terms of projection, the time parameter, κ_t, is simply modelled and projected using any autoregressive integrated moving average (ARIMA) time series model (e.g. random walk with drift).

2.1 Data

The data chosen for illustrative purposes are the female death data and the corresponding exposures of England and Wales, extracted from the Human Mortality Database (HMD (2000)). They are classified by single year of age from 0 to 99, and years ranging from 1961 to 2002. Hence, here we have $\{x_1, \ldots, x_A\} = \{0, \ldots, 99\}$ and $\{t_1, \ldots, t_T\} = \{1961, \ldots, 2002\}$ with $A = 100$ and $T = 42$. We intentionally held back the data for years $2003 - 2013$ as the validation set, see Section 7.

2.2 Overdispersion

The PLC model induces mean-variance equality ($\mathbb{E}[D_{xt}] = \mathrm{Var}[D_{xt}] = e_{xt}\mu_{xt}$), which implies a rigid model structure with strong assumption of homogeneity within each age-period cell. In other words, individuals born in the same year (same x at any given time) have the exact same mortality experience. This is an undesirable mortality assumption in reality since it is well established that other factors such as smoking prevalence, income, ethnicity, genetic backgrouds etc. have significant impacts on mortality (see Brown (2003)), thereby causing extra mortality variations across the individuals, a phenomenon known as overdispersion.

To further illustrate this point, we monitor the square of Pearson residuals under the PLC model given as:

$$r_{xt}^2 = \left. \frac{(d_{xt} - \mathbb{E}[D_{xt}])^2}{\mathrm{Var}[D_{xt}]} \right|_{\mu_{xt}=\hat{\mu}_{xt}} = \frac{(d_{xt} - e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t))^2}{e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t)}, \tag{2}$$

where d_{xt} is the observed number of deaths at age x in year t, and $\hat{\alpha}_x$, $\hat{\beta}_x$, $\hat{\kappa}_t$ are the maximum likelihood estimates of the model parameters. To visualize overdispersion in our mortality data, a colour-coded heat map of r_{xt}^2 can be constructed, as depicted in Figure 1.

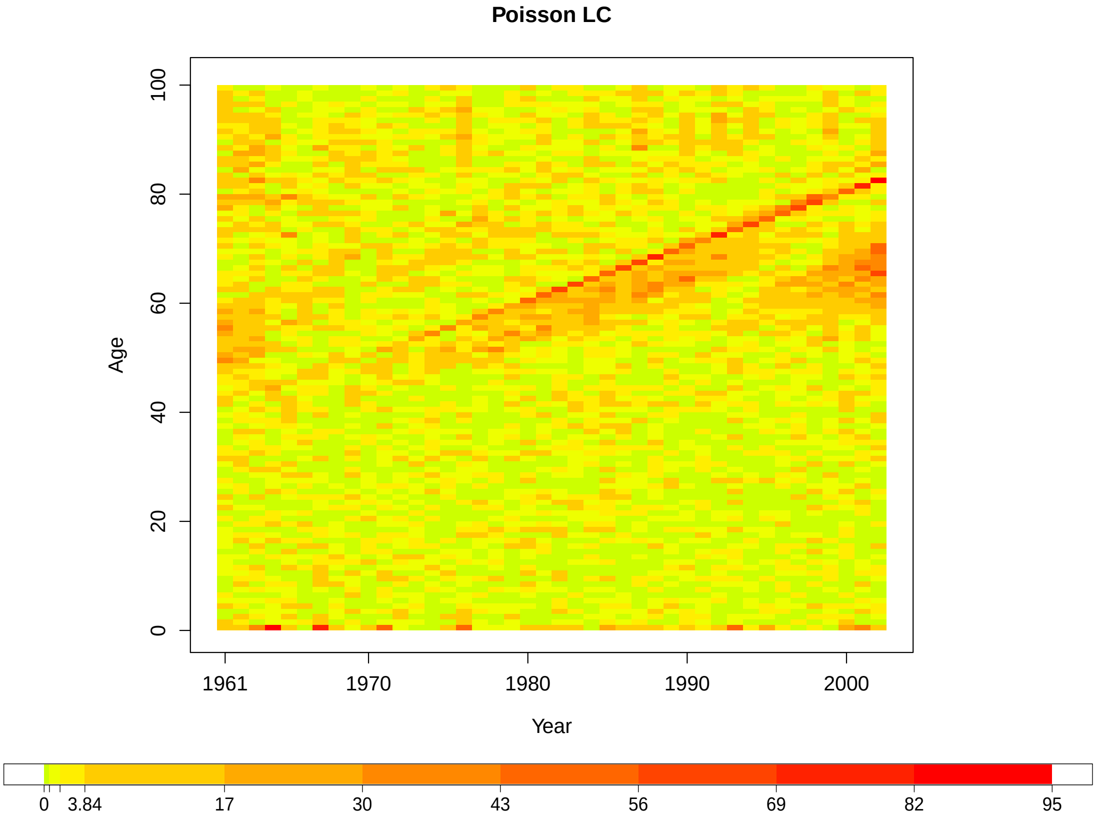


Figure 1: Heat map of r^2_{xt} for the PLC model, accompanied by the corresponding colour code. Green/yellow coloured rectangular cells indicate areas with good fit; while orange/red coloured cells indicate areas with significantly poor fit.

Under the null hypothesis that the PLC model is the true underlying model (and some mild conditions), each r^2_{xt} is approximately chi-squared distributed with degrees of freedom one (χ^2_1). Ideally, we expect only around 5% of the rectangular cells ($AT \times 0.05 = 210$) to have poor fit (defined as $r^2_{xt} > 3.84$). However, it is evident from Figure 1 that the heat map is scattered with more than the expected amount of orange/red cells (about 25%), and is especially obvious for age 0 and ages above 40, suggesting model inadequancy in accounting for extra variations in the data. Moreover, the model deviance computed as the sum of r^2_{xt},

$$r^2 = \sum_{x,t} r^2_{xt} = \sum_{x,t} \frac{(d_{xt} - e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t))^2}{e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t)},$$

has a value of 15378.73, which is substantially larger than the critical value of the conventional chi-squared statistics (i.e. the 95^{th} percentile of $\chi^2_{(A-1)(T-2)}$ is 4107.51). Note that the obvious red/orange diagonal lines displayed in Figure 1 corresponds to possible cohort effects which we do not attempt to address in this paper, but will do so in our future work.

Failure to account for overdispersion typically leads to under-smoothing and over-optimistic forecast because the extra source of uncertainty due to heterogeneity is effectively neglected. Appropriately accounting for overdispersion, on the other hand, provides a better calibration of the unexplained variations. This prevents over-fitting, thereby producing a much more representative prediction interval for the associated mortality forecast.

3 Overdispersion Models

In this section, we propose two models to account for overdispersion by extending the PLC model in a rather straightforward manner. Both these models introduce a general dispersion parameter to relax the assumption of a Poisson distribution.

3.1 Poisson Log-Normal Lee-Carter (PLNLC) Model

The first model we propose, is essentially a direct combination of the original LC model with its Poisson based equivalent, which we refer to as the Poisson Log-Normal LC model. In particular, a normal perturbation term is added onto $\log \mu_{xt}$ for an extra layer of variability in the model:

$$\begin{aligned} D_{xt}|\mu_{xt} &\overset{\text{ind}}{\sim} \text{Poisson}(e_{xt}\mu_{xt}) \\ \log \mu_{xt} &= \alpha_x + \beta_x \kappa_t + \nu_{xt} \\ \nu_{xt}|\sigma_\mu^2 &\overset{\text{ind}}{\sim} N(0, \sigma_\mu^2). \end{aligned} \tag{3}$$

Here, σ_μ^2 is regarded as the general dispersion parameter, whose role is to capture the global level of extra variability in the data. The likelihood function now consists of two parts:

i.

$$f(\boldsymbol{d}|\log \boldsymbol{\mu}) = \prod_{x,t} \left[\frac{\exp(-e_{xt}\mu_{xt})(e_{xt}\mu_{xt})^{d_{xt}}}{d_{xt}!} \right] \propto \exp\left(-\sum_{x,t} e_{xt}\mu_{xt} \right) \prod_{x,t} \mu_{xt}^{d_{xt}}.$$

ii.

$$\begin{aligned} f(\log \boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \sigma_\mu^2) &= \prod_{x,t} \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{1}{2\sigma_\mu^2} (\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2 \right] \\ &\propto (\sigma_\mu^2)^{-\frac{AT}{2}} \exp\left[-\frac{1}{2\sigma_\mu^2} \sum_{x,t} (\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2 \right], \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A)^\top$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_A)^\top$ and $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_T)^\top$ are vectors of parameters, while $\boldsymbol{\mu}$ and $\boldsymbol{d}$ are matrices of the latent variables, μ_{xt} and the observed death data, d_{xt} respectively. Under this model,

$$\mathbb{E}[D_{xt}] = \mathbb{E}_{\mu_{xt}}(\mathbb{E}_{D_{xt}}[D_{xt}|\mu_{xt}]) = e_{xt} \exp\left(\alpha_x + \beta_x \kappa_t + \frac{1}{2}\sigma_\mu^2 \right)$$

and

$$\begin{aligned} \text{Var}[D_{xt}] &= \mathbb{E}_{\mu_{xt}}(\text{Var}_{D_{xt}}[D_{xt}|\mu_{xt}]) + \text{Var}_{\mu_{xt}}(\mathbb{E}_{D_{xt}}[D_{xt}|\mu_{xt}]) \\ &= \mathbb{E}[D_{xt}] \times \left\{ 1 + \mathbb{E}[D_{xt}](\exp(\sigma_\mu^2) - 1) \right\} > \mathbb{E}[D_{xt}]. \end{aligned}$$

Hence, this model possesses a larger variance than its mean in general, with σ_μ^2 governing the relative excess spread, providing more flexibility in our model specification.

3.2 Negative Binomial Lee-Carter (NBLC) Model

The second model is a classic extension of the Poisson distribution to incorporate overdispersion. Specifically, it is a gamma mixture of Poisson as followed:

$$\begin{array}{rcl} D_{xt}|\mu_{xt} & \overset{\text{ind}}{\sim} & \text{Poisson}(e_{xt}\mu_{xt}) \\ \log\mu_{xt} & = & \alpha_x+\beta_x\kappa_t+\log\nu_{xt} \\ \nu_{xt}|\phi & \overset{\text{ind}}{\sim} & \text{Gamma}(\phi,\phi), \end{array} \qquad (4)$$

where ϕ is regarded as the general dispersion parameter in this case. Similarly, the expectation and variance of this model are given by

$$\mathbb{E}[D_{xt}] = e_{xt}\exp(\alpha_x+\beta_x\kappa_t)$$

and

$$\text{Var}[D_{xt}] \;=\; \mathbb{E}[D_{xt}]\times\left[1+\frac{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)}{\phi}\right] > \mathbb{E}[D_{xt}].$$

Therefore, this model possesses the same mean as the PLC model (as opposed to the PLNLC model), while at the same time has a larger variance depending on the value of ϕ. In particular, the smaller the value of ϕ, the larger the variance, and hence the stronger the evidence of overdispersion; while the larger the ϕ, the more this model approaches the PLC model, with exact resemblance when $\phi\to\infty$. In other words, $1/\phi$ represents the overall magnitude of overdispersion in the data.

One attractive feature about this model is that the latent variables, μ_{xt}, can be conveniently integrated out, producing its equivalent version, which we call the Negative Binomial LC model. That is,

$$D_{xt}|\alpha_x,\beta_x,\kappa_t,\phi \sim \text{ Neg-Bin }\left(\phi,\frac{\phi}{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi}\right). \qquad (5)$$

The likelihood function now consists of only 1 part:

$$\begin{array}{rcl} f(\boldsymbol{d}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa},\phi) & = & \displaystyle\prod_{x,t}\left\{\frac{\Gamma(d_{xt}+\phi)}{\Gamma(\phi)\Gamma(d_{xt}+1)}\left[\frac{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)}{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi}\right]^{d_{xt}}\left[\frac{\phi}{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi}\right]^{\phi}\right\} \\ & \propto & \displaystyle\frac{\phi^{AT\phi}}{[\Gamma(\phi)]^{AT}}\prod_{x,t}\frac{\Gamma(d_{xt}+\phi)\exp[d_{xt}(\alpha_x+\beta_x\kappa_t)]}{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi]^{d_{xt}+\phi}}. \end{array}$$

The prominent advantage of the marginalisation is that we avoid the need to simulate the high-dimensional μ_{xt} (dimension=AT=4200 in our case), at the expense of having a slightly more complicated likelihood function. In particular, we found in our preliminary study that the computational gain from marginalising μ_{xt} substantially outweighs the burden of dealing with the more complicated negative binomial likelihood (by comparing the effective number of samples generated per unit time). Note that this model has already been considered by Delwarde et al. (2007), but within the classical framework.

4 Bayesian Modelling

We perform the model fitting within a Bayesian paradigm. The rationale for considering Bayesian methodology is it provides a natural framework in which prior knowledge can be incorporated and various sources of uncertainty (due to inherent random variation, parameter estimation, projection and model misspecification) can be coherently included to provide a more representative prediction interval. Classical LC approach often ignores uncertainty due to parameter estimation. Although it has been shown in Lee and Carter (1992) that the forecast uncertainty will dominate over parameter uncertainty in long term projection, the same is not true for short to moderate term projection. In Bayesian framework, parameter uncertainty is incorporated in the form of probability distributions through prior specification for each of the unknown parameters. In addition, we also acknowledge the presence of model uncertainty by performing Bayesian model determination using posterior model probabilities, instead of assuming in advance, a single underlying model.

Moreover, a major criticism on the traditional LC approach is the potential inconsistencies that may arise due to its two-stage model fitting procedures: the parameters are first estimated using maximum likelihood approach, they are then separately fitted using the ARIMA time series model solely for the purpose of projection. Technically, the ARIMA model, being part of the model specification, should have contributed directly in the parameter estimation stage. Bayesian modelling solves this issue by directly specifying an ARIMA prior on κ_t, forming a single framework of a hierarchical model. Parameter estimation then proceeds simultaneously through the computation of joint posterior distribution. Additionally, this allows for the possibility of performing smoothing over time (as mentioned in Czado et al. (2005)), depending on the ARIMA model fitted. Projection of mortality then follows naturally within the Bayesian framework based on the ARIMA model chosen (see Section 5.5).

Furthermore, carefully calibrated percentiles of the posterior predictive distribution carry valuable information necessary to characterize the uncertainties we encounter during forecasting. In practice, any percentile can be used as a point estimate other than the posterior mean or median in context of probabilistic forecasts. This provides more flexibility to the users in their decision making.

4.1 Prior Distributions

In this section, we provide the prior distributions used for each parameters. Ideally, the prior distributions chosen should reflect our uncertainty/prior knowledge about mortality (e.g. smoothness of mortality rates across age). However, we do not pursue this matter here. Rather, we specify some commonly used priors rendered sufficiently diffuse for data-dominated inference. It is also worth mentioning that we incorporated the identifiability constraints directly through the prior distributions. This is in contrast to Czado et al. (2005), who modified their MCMC algorithm with some deterministic adjustments to account for the constraints, but did not present any theoretical justification that the constructed chain converges to the correct target distribution. In addition, we also attempt to be indifferent in terms of prior specification under both overdispersion models to facilitate model comparison later on.

4.1.1 Prior Distribution for α_x, β_x, σ^2_β, σ^2_μ, and ϕ

From here on, we denote $\mathbf{1}_n$ as a length$-n$ vector of ones, while $\boldsymbol{J}_n$ and $\boldsymbol{I}_n$ as a matrix of ones and the identity matrix respectively of dimension $n \times n$. For simplicity, we assign an independent normal prior on α_x, i.e.

$$\boldsymbol{\alpha} \sim N(\alpha_0 \mathbf{1}_A, \sigma^2_\alpha \boldsymbol{I}_A).$$

Here, we set $\alpha_0 = 0$, while σ^2_α is chosen to be relatively large, say $\sigma^2_\alpha = 100$ for a vague prior. Similarly, we impose, a priori

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_\beta \boldsymbol{I}_A),$$

subject to the constraint $\sum_x \beta_x = 1$. Applying the constraint on the marginal prior of β_x, and using the conditional property of a normal distribution, we obtain the following prior for $\boldsymbol{\beta}_{-1} = (\beta_2, \beta_3, \ldots, \beta_A)^\top$,

$$\boldsymbol{\beta}_{-1} \sim N\left(\frac{1}{A}\mathbf{1}_{A-1}, \sigma_\beta^2\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)\right).$$

That way, the constraint is automatically accounted for by the above prior with β_1 deterministically computed from $\beta_1 = 1 - \beta_2 - \ldots - \beta_A$. Moreover, the hierarchical variance, σ_β^2 is now treated as a hyperparameter with the conventional prior

$$\sigma_\beta^{-2} \sim \text{Gamma}(a_\beta, b_\beta),$$

where $a_\beta = b_\beta = 0.001$. The result of this is a heavier-tailed Student's t-distribution on β_x a priori, characterizing our larger uncertainty in β_x due to its more erratic behaviour as compared to α_x empirically.

As pointed out in Section 3.1 and 3.2, σ_μ^2 and ϕ serve as the dispersion parameter in each model. Since we have no knowledge on the appropriate extent of overdispersion in our data, we assign the conditional conjugate (see Gelman (2006)) prior

$$\sigma_\mu^{-2} \sim \text{Gamma}(a_\mu, b_\mu),$$

with $a_\mu = b_\mu = 0.0001$ for computational purposes under the PLNLC model. In order to specify a prior with similar amount of information embedded within the distribution for ϕ, we need to establish a relationship between the two dispersion parameters. By Using a Taylor Series approximation to $\log \mu_{xt}$ under the NBLC model, and ignoring the variabilities due to α_x, β_x, and κ_t, we have

$$\text{Var}(\log \mu_{xt}) = \text{Var}(\log \nu_{xt}) \approx \left(\frac{d \log z}{dz}\right)^2\Bigg|_{z=\mathbb{E}(\nu_{xt})} \times \text{Var}(\nu_{xt}) = \frac{1}{\phi}.$$

Knowing that $\text{Var}(\log \mu_{xt}) = \sigma_\mu^2$ (conditional upon α_x, β_x and κ_t) under the PLNLC model, this implies that a sensible prior for ϕ could be

$$\phi \sim \text{Gamma}(a_\phi, b_\phi),$$

where $a_\phi = b_\phi = 0.0001$.

4.1.2 Prior Distribution for κ_t

κ_t represents the overall mortality level at time t, which forms the crucial element for stochastic forecasts. For reason mentioned in Section 4, an ARIMA time series model is imposed on κ_t, which can then be straightforwardly extrapolated forward in time for mortality projection. In various occasions, a random walk with drift was empirically found to provide an adequate fit for κ_t (see Tuljapurkar et al. (2000)). Following Czado et al. (2005) though, we fit a first order autoregressive (AR(1)) model with linear drift. Specifically,

$$\left\{\begin{array}{ll} \kappa_t - \eta_t = \rho(\kappa_{t-1} - \eta_{t-1}) + \epsilon_t, & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \eta_1 + \epsilon_1 & \end{array}\right., \tag{6}$$

where $\eta_t = \psi_1 + \psi_2 t$ denotes the linear drift and $\epsilon_t \overset{\text{ind}}{\sim} N(0, \sigma_\kappa^2)$ are random errors. Note that Equation (6) includes random walk with drift as a special case when $\rho = 1$, provided we do not use a truncated prior on ρ. In other words, we allow the data to choose either an AR(1) or random walk with drift instead of specifying beforehand the appropriate model since it is entirely possible that random walk with drift fits our data poorly. We also adopt a different constraint for κ_t, $\kappa_1 = 0$ as compared to the conventional $\sum_t \kappa_t = 0$. This changes the interpretation of α_x slightly, where α_x now represents the log mortality rates in the base year. Elsewhere, the impact of this is purely computational, the fitted values of $\log \mu_{xt}$ will not be affected.

This model can be equivalently expressed in its multivariate form (with the constraint) as

$$\begin{cases} \boldsymbol{\kappa}_{-1} \sim N(\boldsymbol{Y}_{-1}\boldsymbol{\psi} - \rho \boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi}, \sigma_\kappa^2 \boldsymbol{Q}^{-1}) \\ \kappa_1 = 0 \end{cases}, \tag{7}$$

where

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \rho & 0 & & & \vdots \\ 0 & \rho & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho & 0 \end{pmatrix}_{(T-1)\times(T-1)}, \boldsymbol{Y}_{-1} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & T \end{pmatrix}_{(T-1)\times 2}, \boldsymbol{Y}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}_{(T-1)\times 2},$$

$\boldsymbol{R} = \boldsymbol{I}_{T-1} - \boldsymbol{P}$, $\boldsymbol{Q} = \boldsymbol{R}^\top \boldsymbol{R}$, $\boldsymbol{\psi} = (\psi_1, \psi_2)^\top$, and $\boldsymbol{\kappa}_{-1} = (\kappa_2, \kappa_3, \ldots, \kappa_T)^\top$. For complete specification of the model on κ_t, the unknown parameters ρ, σ_κ^2 and $\boldsymbol{\psi}$ are treated as hyperparameters with the following standard vague priors:

$$\begin{aligned} \rho &\sim N(0, \sigma_\rho^2), \\ \sigma_\kappa^{-2} &\sim \text{Gamma}(a_\kappa, b_\kappa), \\ \boldsymbol{\psi} &\sim N(\boldsymbol{\psi}_0, \boldsymbol{\Sigma}_\psi), \end{aligned}$$

where $\sigma_\rho^2 = 100$, $a_\kappa = b_\kappa = 0.001$, $\boldsymbol{\psi}_0 = (0,0)^\top$, and $\boldsymbol{\Sigma}_\psi = \begin{pmatrix} 1000 & 0 \\ 0 & 10 \end{pmatrix}$. These priors are chosen to be conditionally conjugate with respect to the AR(1) model, which ease the subsequent computation of the conditional posterior distributions as we shall see later in Section 5.

5 Computation

5.1 MCMC Method

The MCMC method we propose is the variable-at-a-time Metropolis-Hastings (MH) algorithm as described in O'Hagan and Forster (2004), where each component of the parameters are updated sequentially through MH algorithm in each iteration, conditional on the rest of the parameters. In the case where the conditional posterior distributions are tractable, typically where conditional conjugate priors are used, the Gibbs algorithm is undertaken (MH algorithm with acceptance probability equals to 1).

In addition, we will be adopting the idea of blocking of parameters wherever possible within our MCMC updating scheme. The motivation of considering blocking is the fact that it enables the MCMC algorithm to acknowledge the correlation structure of the parameters in order to make informed movements/jumps across the parameter spaces, facilitating the exploration of posterior distributions. For instance, Roberts and Sahu (1997) suggest that blocking, if done efficiently, is capable of improving the convergence rate of the resulting MCMC sampler substantially. However, the efficacy of performing blocking is clearly dictated by the dimensions of parameters involved and the resulting complexity of the conditional posterior distributions of the respective blocks. Therefore, our general strategy of blocking is to allocate highly-correlated parameters in a single block such that the correlations between blocks are reduced (rather than allocating all in one block).

5.2 MCMC Scheme for the PLNLC Model

Suppose we allocate the α_x, β_x and κ_t each in one separate block, and the rest of the parameters updated univariately. Due to the model structure, the conditional posterior distributions of all of the parameters can be conveniently recognized as standard distributions (Appendix A), except for the $\log \mu_{xt}$. Hence, the MCMC updating scheme can be easily implemented by iterating through a series of Gibbs steps. We describe in details the MH step for the remaining $\log \mu_{xt}$ in the next subsection.

5.2.1 MH Step for $\log \mu_{xt}$

We forfeited the concept of blocking here due to the immense dimensionality involved. Instead, each $\log \mu_{xt}$ is updated univariately using random walk MH algorithm (see for example O'Hagan and Forster (2004)). In particular, using the assumption that $\boldsymbol{D}$ are mutually independent given $\log \boldsymbol{\mu}$, and $\log \boldsymbol{\mu}$ are independent elementwise given $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2)$, the conditional posterior density of $\log \mu_{xt}$ can be expressed as

$$f(\log \mu_{xt}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}_{-xt}, \sigma_\kappa^2, \sigma_\beta^2, \rho, \boldsymbol{\psi}, \sigma_\mu^2) \propto \mu_{xt}^{d_{xt}} \exp\left[-e_{xt}\mu_{xt} - \frac{1}{2\sigma_\mu^2}(\log \mu_{xt} - \alpha_x - \beta_x\kappa_t)^2\right],$$

where $\boldsymbol{\mu}_{-xt} = (\mu_{11}, \mu_{21}, \ldots, \mu_{x-1\,t}, \mu_{x+1\,t}, \ldots, \mu_{AT})^\top$ is a vector of all the mortality rates excluding the xt^{th} component. Next, we propose a value at the i^{th} iteration,

$$\log \mu_{xt}^* \sim N(\log \mu_{xt}^{i-1}, \sigma_{\mu_{xt}}),$$

where $\log \mu_{xt}^{i-1}$ is the current value of $\log \mu_{xt}$, and $\sigma_{\mu_{xt}}^2$ are the proposal variances to be specified deterministically. The proposal is then accepted according to the following probability,

$$\begin{aligned} a(\log \mu_{xt}^* | \log \mu_{xt}^{i-1}) &= \min\left\{1, \left(\frac{\mu_{xt}^*}{\mu_{xt}^{i-1}}\right)^{d_{xt}} \exp\left[-e_{xt}(\mu_{xt}^* - \mu_{xt}^{i-1})\right.\right. \\ &\qquad \left.\left. -\frac{1}{2\sigma_\mu^2}((\log \mu_{xt}^* - \alpha_x - \beta_x\kappa_t)^2 - (\log \mu_{xt}^{i-1} - \alpha_x - \beta_x\kappa_t)^2)\right]\right\}. \end{aligned}$$

The choice of $\sigma_{\mu_{xt}}^2$ is arbitrary, but has a direct impact on the speed of convergence of the constructed chain. In practice, $\sigma_{\mu_{xt}}^2$ are carefully chosen such that the acceptance rates of $\log \mu_{xt}$ are within the recommended range 0.15-0.45 (Roberts and Rosenthal (2001)). Following Czado et al. (2005), we develop a simple automatic trial and error search algorithm for tuning $\sigma_{\mu_{xt}}^2$, which starts off with a crude search:

i. Set initial values of $\sigma_{\mu_{xt}}^2 = 0.01$ for all x and t.

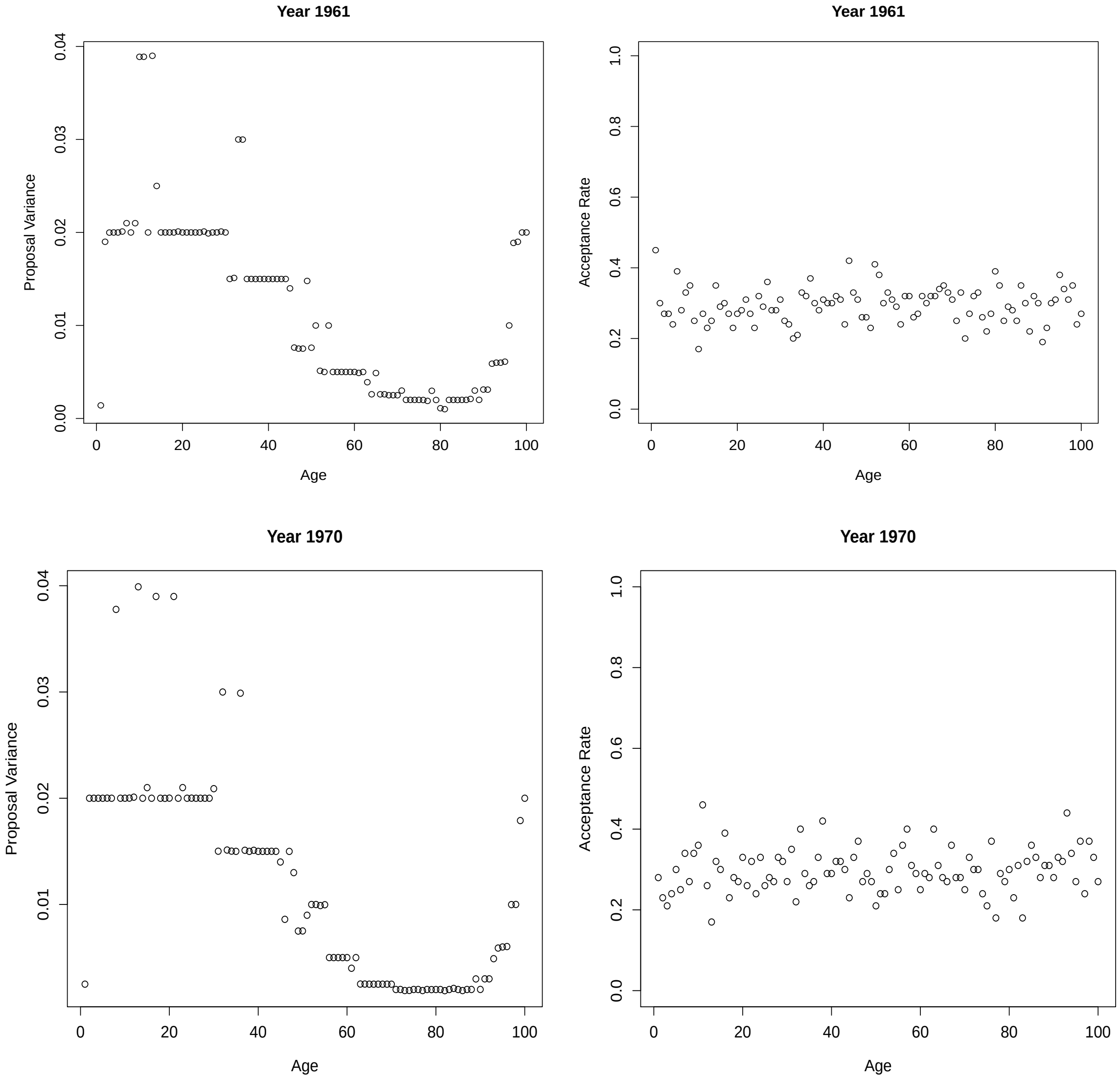
ii. A pilot run of 100 iterations is executed.

iii. Proposal variances that correspond to acceptance rates smaller than 0.15 are halved.

iv. Proposal variances that correspond to acceptance rates exceeding 0.45 are doubled.

v. Repeat steps ii-iv until a predefined threshold is achieved (e.g. when 4000 of the acceptance rates are within 0.15-0.45).

The search can then be further refined by shrinking the increments (or decrements) of the adjustments within the above algorithm, so instead of a multiplicative factor of two, we can add (or subtract) a small amount, say 0.001 during the tuning of the proposal variances. As a result, the $\sigma^2_{\mu_{xt}}$ can be numerically determined and are depicted in Figure 2.



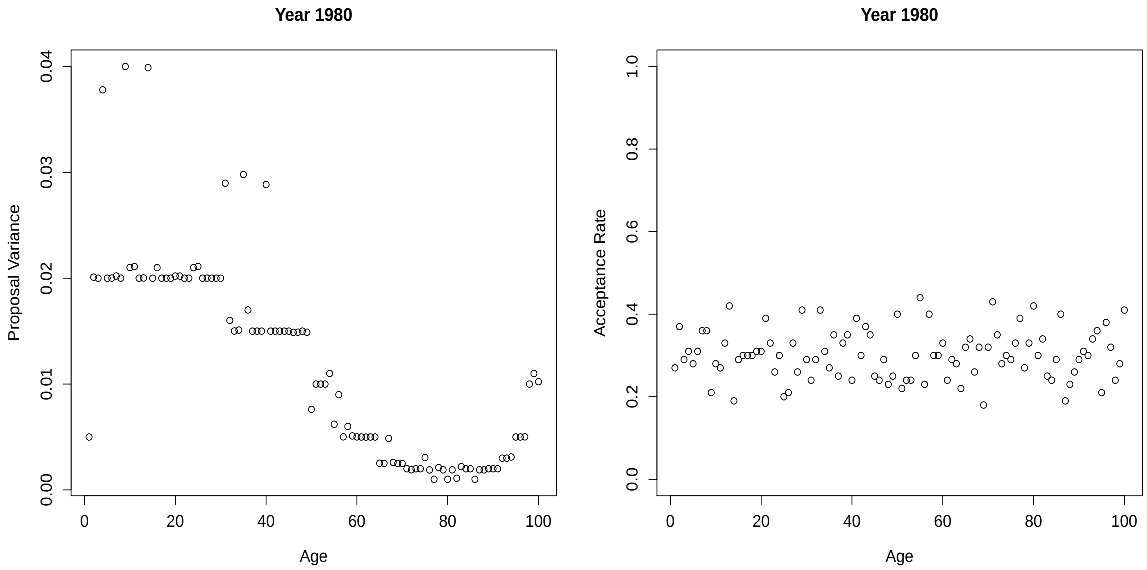


Figure 2: Plots of proposal variances, $\sigma^2_{\mu_{xt}}$ (left panels) and the corresponding acceptance rates of μ_{xt} (right panels) for years 1961, 1970 and 1980.

Interestingly, $\sigma^2_{\mu_{xt}}$ exhibit a consistent age pattern across the years. It turns out that the rough pattern of posterior variances of $\log \mu_{xt}$ in a given year can potentially be deduced from this set of approximate optimal proposal variances, which we shall verify later. This can be attributed to the finding in Roberts and Rosenthal (2001) that the optimal proposal variance for a MH algorithm with a univariate normal distribution as its target is proportional to the posterior variance (with 2.38^2 as the proportionality constant).

5.3 MCMC Scheme for the NBLC Model

Here, we apply the random walk MH algorithm on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{-1}$ and $\boldsymbol{\kappa}_{-1}$ instead because the normal priors are no longer conditionally conjugate. Nevertheless, the Gibbs steps for $\rho, \sigma^2_\kappa, \sigma^2_\beta, \boldsymbol{\psi}$ are unaffected (refer to Appendix A) because they belong to the lower part of the hierarchical model, hence their conditional posterior distributions remain the same conditional upon $\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}$, and $\boldsymbol{\kappa}_{-1}$. Note that our preliminary study also revealed that performing the sequential updating scheme univariately without blocking is more efficient here in terms of the effective number of posterior samples generated per unit time. This is because the negative inverse of Hessian matrix, which is only an approximation of the posterior variance, was used to derive the optimal proposal variance for the corresponding multivariate MH algorithm.

5.3.1 MH Steps for α_x, β_x, κ_t, and ϕ

The conditional posterior densities and expressions for the MH acceptance probabilities are displayed in Appendix B. Using obvious notation, a set of numerically determined proposal variances for the random walk MH algorithm (derived from similar search algorithm as above), $\sigma^2_{\alpha_x}$, $\sigma^2_{\beta_x}$, and $\sigma^2_{\kappa_t}$ are illustrated in Figure 3.

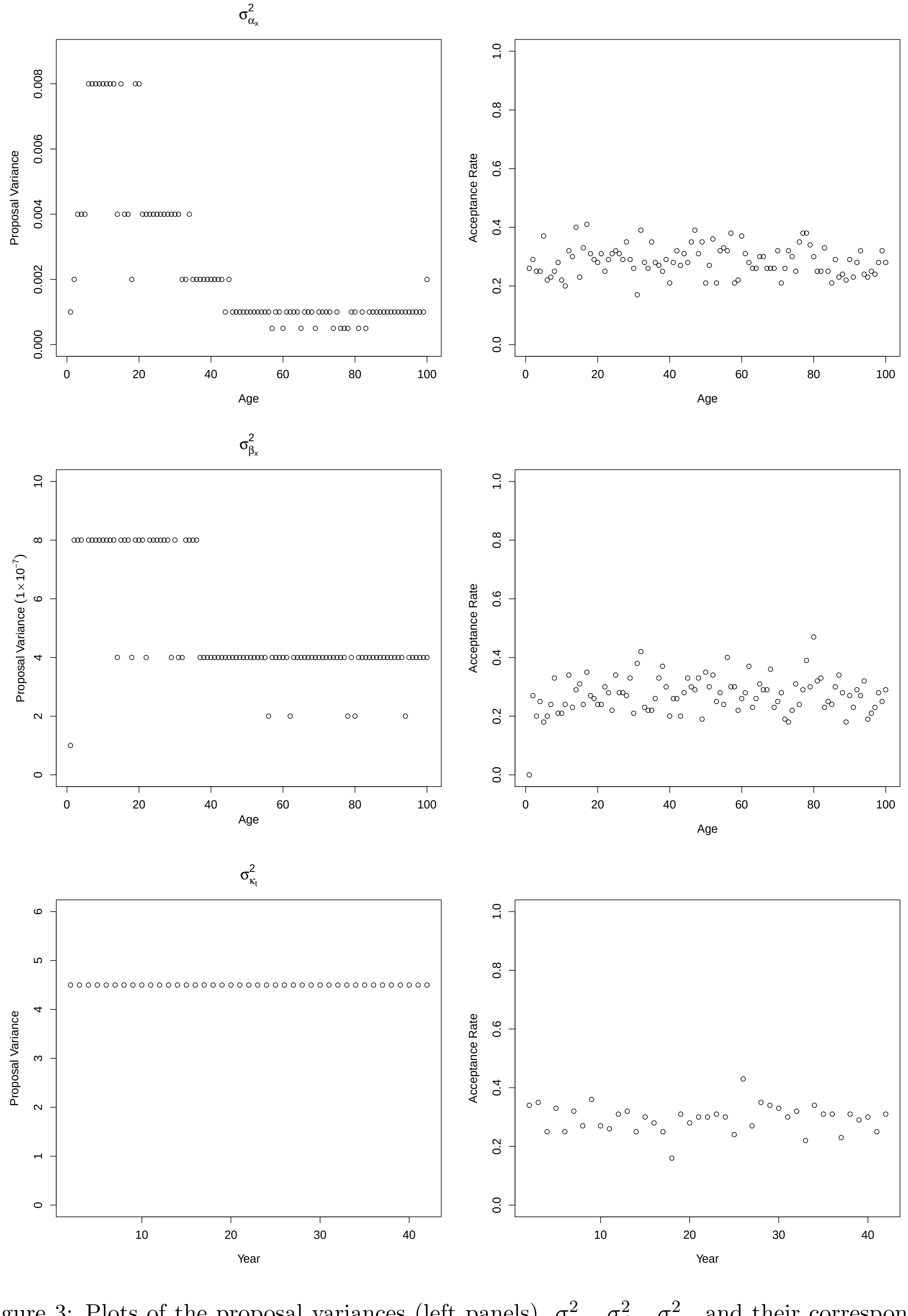


Figure 3: Plots of the proposal variances (left panels), $\sigma^2_{\alpha_x}$, $\sigma^2_{\beta_x}$, $\sigma^2_{\kappa_t}$, and their corresponding acceptance rates (right panels).

According to Figure 3, $\sigma^2_{\alpha_x}$ demonstrates a rather similar age pattern to $\sigma^2_{\mu_{xt}}$ at any given time as before. This is perhaps not so surprising since α_x represent the log mortality rates in the base year. However, the age pattern exhibited for $\sigma^2_{\beta_x}$ is less obvious than that of $\sigma^2_{\alpha_x}$, albeit still having a rather similar pattern. On the other hand, the $\sigma^2_{\kappa_t}$ derived from the search algorithm, are strikingly identical across the years. This signifies that the marginal posterior variances of κ_t are very similar, in contrast to α_x and β_x, where their proposal variances vary substantially across age.

A proposal variance of $\sigma^2_\phi = 0.08$ will return an acceptance rate of approximately 0.30 for ϕ.

5.4 Generating μ_{xt} under the NBLC Model

Although the mortality rates, μ_{xt}, have been integrated out for the NBLC model, it can still be useful to simulate them to potentially learn about their posterior distributions. The latent variables can be retrieved by noting that for any $x = 1, \ldots, A$ and $t = 1, \ldots, T$,

$$f(\mu_{xt}|\boldsymbol{d}) = \int f(\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d}) f(\alpha_x, \beta_x, \kappa_t, \phi|\boldsymbol{d}) \mathrm{d}\alpha_x \mathrm{d}\beta_x \mathrm{d}\kappa_t \mathrm{d}\phi,$$

where $f(\alpha_x, \beta_x, \kappa_t, \phi|\boldsymbol{d})$ is the joint posterior density of α_x, β_x, κ_t, and ϕ, while $f(\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d})$ can be derived as

$$f(\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d}) \propto \mu_{xt}^{(d_{xt}+\phi)-1} \exp\left[-\left(e_{xt} + \frac{\phi}{\exp(\alpha_x + \beta_x\kappa_t)}\right)\mu_{xt}\right],$$

implying that

$$\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d} \sim \text{Gamma}\left(d_{xt} + \phi, e_{xt} + \frac{\phi}{\exp(\alpha_x + \beta_x\kappa_t)}\right). \tag{8}$$

Therefore, the posterior samples of μ_{xt} can be generated by simulating from the expression in (8), where the joint posterior samples of α_x, β_x, κ_t, and ϕ (which are readily available from our MCMC outputs) are substituted wherever applicable.

5.5 Mortality Forecast

Projection within the Bayesian framework is particularly natural through the derivation of posterior predictive distribution. Specifically, the posterior predictive distribution of 1-year ahead log mortality rates for each age group (with the age parameters held fixed), under the PLNLC model for instance, can be written as

$$\begin{aligned} f(\log\mu_{xT+1}|\boldsymbol{d}) &= \int f(\log\mu_{xT+1}|\alpha_x, \beta_x, \kappa_{T+1}, \sigma^2_\mu) f(\alpha_x, \beta_x, \sigma^2_\mu|\boldsymbol{d}) f(\kappa_{T+1}|\kappa_T, \rho, \sigma^2_\kappa, \boldsymbol{\psi}) \\ &\quad \times f(\kappa_T, \rho, \sigma^2_\kappa, \boldsymbol{\psi}|\boldsymbol{d}) \mathrm{d}\alpha_x \mathrm{d}\beta_x \mathrm{d}\kappa_T \mathrm{d}\kappa_{T+1} \mathrm{d}\rho \mathrm{d}\sigma^2_\kappa \mathrm{d}\boldsymbol{\psi} \mathrm{d}\sigma^2_\mu, \end{aligned} \tag{9}$$

where $f(\alpha_x, \beta_x, \sigma^2_\mu|\boldsymbol{d})$ and $f(\kappa_T, \rho, \sigma^2_\kappa, \boldsymbol{\psi}|\boldsymbol{d})$ are the joint posterior distributions. Hence, posterior uncertainties, with respect to the model likelihood, prior distributions and projection, is fully integrated in the posterior predictive distribution. The density in (9) is analytically intractable, but can be empirically estimated using our MCMC samples as follows. Essentially, generation of the posterior samples of $\log\mu_{xT+1}$ proceeds in two steps:

1. Generate κ_{T+1} from the AR(1) model,

$$\kappa_{T+1} \sim N(\psi_1 + \psi_2(T+1) + \rho(\kappa_T - \psi_1 - \psi_2 T), \sigma^2_\kappa),$$

   where joint posterior samples of $(\kappa_T, \rho, \sigma^2_\kappa, \psi_1, \psi_2)$ from the MCMC output are substituted into the expression.

2. Generate $\log \mu_{x\,T+1}$ from

$$\log \mu_{x\,T+1} \sim N(\alpha_x + \beta_x \kappa_{T+1}, \sigma_\mu^2),$$

where κ_{T+1} is from step 1 and $\alpha_x, \beta_x, \sigma_\mu^2$ are joint posterior samples from the MCMC output.

By analogy, h-years ahead projections can be obtained by recursive implementation of the above generation procedures. Having generated a set of posterior predictive samples, a fanplot of carefully calibrated percentiles (see Abel (2015)) can then be constructed to better visualize the underlying uncertainty associated with our probabilistic forecast.

Once the future underlying mortality rates, for instance $\log \mu_{x\,T+h}$, have been simulated, we can generate the h-years ahead number of deaths simply through

$$D_{x\,T+h} \sim \text{Poisson}(e_{x\,T+h}\mu_{x\,T+h}),$$

where $e_{x\,T+h}$ is the future exposure at age x in year $T+h$ (which we assumed known). The future crude mortality rates can subsequently be obtained by

$$\hat{\mu}_{x\,T+h} = \frac{D_{x\,T+h}}{e_{x\,T+h}}.$$

The key difference between them is that the projected crude mortality rates include the Poisson variation in their prediction intervals, whereas the projected underlying mortality rates do not. The choice of which one to use depends on the users' preference, whether or not they prefer to base their policy making on the underlying rates, or the "observed" rates. Indeed, it should be noted that computation of the future crude death rates requires the availability of future exposures, which can be an unrealistic assumption at times.

6 Bayesian Model Comparison

Formal Bayesian model comparison proceeds through the computation of posterior model probabilities. For a set of models $M \in M^S$ under consideration, the posterior model probability of model M, $f(M|\boldsymbol{d})$ is given by

$$f(M|\boldsymbol{d}) = \frac{f(M)f_M(\boldsymbol{d})}{\sum_{j\in M^S} f(j)f_j(\boldsymbol{d})},$$

where $f_M(\boldsymbol{d})$ is the marginal likelihood of model M and $f(M)$ is the prior model probability of model M. Typically, we assume equal prior model probabilities so that models are compared directly using their marginal likelihoods.

6.1 Bridge Sampling

Bridge sampling is a sampling-based technique originally developed by Meng and Wong (1996) to estimate ratio of two normalising constants. It can be applied in the context of approximation of marginal likelihood $f_M(\boldsymbol{d})$ of model M, if we construct the algorithm such that the second normalising constant is known. In particular, bridge sampling estimator of marginal likelihood is given by

$$\hat{f}^\omega_M(\boldsymbol{d}) = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} f_M(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}^i_M)f_M(\tilde{\boldsymbol{\theta}}^i_M)\omega_M(\tilde{\boldsymbol{\theta}}^i_M)}{\frac{1}{N_1}\sum_{i=1}^{N_1} g_M(\boldsymbol{\theta}^i_M)\omega_M(\boldsymbol{\theta}^i_M)},$$

where $\{\boldsymbol{\theta}_M^i\}_{i=1}^{N_1}$ is a sample of size N_1 from the posterior distribution with density $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$, $\{\tilde{\boldsymbol{\theta}}_M^i\}_{i=1}^{N_2}$ is a sample of size N_2 from a normalised distribution with density $g_M()$, and $\omega_M()$ is the so called bridge function that satisfies $0 < |\int f_M(\boldsymbol{\theta}_M|\boldsymbol{d})g_M(\boldsymbol{\theta}_M)\omega_M(\boldsymbol{\theta}_M)\mathrm{d}\boldsymbol{\theta}_M| < \infty$.

Meng and Wong (1996) proposed that an optimal choice of $\omega_M()$, in the sense of minimizing the asymptotic relative mean square error, is given by

$$\omega_M^*(\boldsymbol{\theta}_M) \propto \left[N_1 \frac{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)f_M(\boldsymbol{\theta}_M)}{f_M(\boldsymbol{d})} + N_2 g_M(\boldsymbol{\theta}_M)\right]^{-1},$$

provided draws from both distributions are independent. In the case where dependent samples are available, as in our MCMC generated posterior, effective sample size should be used in place of N_1. Alternatively, j^{th} thinning (retaining samples every j iteration) can be applied to obtain a set of approximately independent MCMC samples, where j is chosen such that the sample autocorrelations are close to 0.

Indeed, $\omega_M^*()$ still depends on the unknown marginal likelihood, $f_M(\boldsymbol{d})$, so Meng and Wong (1996) suggest an iterative procedure for estimating $f_M(\boldsymbol{d})$:

$$\hat{f}_M^*(\boldsymbol{d})^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{\tilde{l}_i}{N_1\tilde{l}_i + N_2\hat{f}_M^*(\boldsymbol{d})^{(t)}}\right]}{\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\frac{1}{N_1 l_i + N_2\hat{f}_M^*(\boldsymbol{d})^{(t)}}\right]}, \tag{10}$$

where $l_i = \frac{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i)f_M(\boldsymbol{\theta}_M^i)}{g_M(\boldsymbol{\theta}_M^i)}$ and $\tilde{l}_i = \frac{f_M(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}_M^i)f_M(\tilde{\boldsymbol{\theta}}_M^i)}{g_M(\tilde{\boldsymbol{\theta}}_M^i)}$. Starting with an initial guess, $\hat{f}_M^*(\boldsymbol{d})^{(0)}$, the bridge sampling estimate, $\hat{f}_M^*(\boldsymbol{d})$, of the marginal likelihood can be obtained by iterating (10) until convergence.

The choice of the density $g_M()$ is entirely arbitrary. In practice though, bridge sampling tends to perform most efficiently when $g_M()$ resembles the posterior density, $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$. An obvious candidate would be a normal distribution with its first two moments chosen to match those from the posterior distribution. Posterior mode (as first moment) and negative inverse of Hessian matrix (as second moment) appear to be an option here. However, these quantities can be difficult to derive when the dimension involved is huge. They also appear to provide insufficient information for bridge sampling to work efficiently when the distribution is heavy-tailed. Therefore, a better alternative is to use sample mean and variance computed directly from our posterior sample for the moment-matching.

Unfortunately, the use of posterior sample statistics induces a correlation between the sample from $g_M()$, $\{\tilde{\boldsymbol{\theta}}_M^i\}$ and the posterior samples, $\{\boldsymbol{\theta}_M^i\}$ through the sample moments, which then manifests itself in the form of a systematic underestimation of the corresponding marginal likelihood (see Overstall and Forster (2010)). Hence, we adopt the idea of Overstall and Forster (2010) to use only half of the posterior sample to estimate the posterior moments, while the remainder of the posterior sample is used to evaluate the bridge sampler in (10).

Finally, the allocation of sample sizes, N_1 and N_2 is influential here due to their appearance in the optimal $\omega()$ as the mixture proportions of $f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)f_M(\boldsymbol{\theta}_M)$ and $g_M()$. However, Chen et al. (2000) (pg. 129) stated that the optimal choice of $\omega()$ itself is often more crucial than the optimal allocation of sample sizes. Thus, we propose to allocate $N_1 = N$ and $N_2 = 2N_1 = 2N$ because it is relatively easy to sample from the density $g_M()$ and the resulting bridge sampling estimate has a lower mean squared error than that using $N_2 = N_1 = N$.

The general algorithm for computing the marginal likelihood using bridge sampling is

1. Generate a sample, $\{\boldsymbol{\theta}_M^1, \ldots, \boldsymbol{\theta}_M^N, \boldsymbol{\theta}_M^{N+1}, \ldots, \boldsymbol{\theta}_M^{2N}\}$, of size $2N$ from the posterior distribution, $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$.

2. Compute the sample mean and covariance matrix of $\{\boldsymbol{\theta}_M^1, \ldots, \boldsymbol{\theta}_M^N\}$, denoted as $\boldsymbol{\mu}_M$ and $\boldsymbol{\Sigma}_M$ respectively. Let $g_M()$ be the density of a p_M-dimensional normal distribution, $N(\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$, where p_M is the number of unknown parameters under model M.

3. Generate a sample, $\{\tilde{\boldsymbol{\theta}}_M^1, \ldots, \tilde{\boldsymbol{\theta}}_M^{2N}\}$, of size $2N$ from the density $g_M()$.

4. Obtain the bridge sampling estimate of marginal likelihood, $\hat{f}_M^*(\boldsymbol{d})$, using (10), evaluated at the samples $\{\boldsymbol{\theta}_M^{N+1}, \ldots, \boldsymbol{\theta}_M^{2N}\}$ and $\{\tilde{\boldsymbol{\theta}}_M^1, \ldots, \tilde{\boldsymbol{\theta}}_M^{2N}\}$.

For our overdispersion models, the marginal likelihood of the PLNLC model, dropping subscript M wherever applicable, can be expressed as

$$\begin{aligned} f_{\text{PLNLC}}(\boldsymbol{d}) &= \int f(\boldsymbol{d}|\log\boldsymbol{\mu}) f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\sigma_\mu^2) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}_{-1}|\log\sigma_\beta^2) f(\boldsymbol{\kappa}_{-1}|\rho, \log\sigma_\kappa^2, \boldsymbol{\psi}) \\ &\quad \times f(\rho) f(\log\sigma_\kappa^2) f(\log\sigma_\beta^2) f(\boldsymbol{\psi}) f(\log\sigma_\mu^2) \mathrm{d}\boldsymbol{\theta}_{\text{PLNLC}}, \end{aligned} \tag{11}$$

where $\boldsymbol{\theta}_{\text{PLNLC}} = (\log\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \rho, \log\sigma_\kappa^2, \log\sigma_\beta^2, \boldsymbol{\psi}, \log\sigma_\mu^2)$ is the full set of parameters under this model, which is of dimension $p_{\text{PLNLC}} = 4446$. Note that the relevant components of the parameters are log-transformed for the normal approximation to work better. Similarly, the marginal likelihood of the NBLC model is given by

$$\begin{aligned} f_{\text{NBLC}}(\boldsymbol{d}) &= \int f(\boldsymbol{d}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\phi) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}_{-1}|\log\sigma_\beta^2) f(\boldsymbol{\kappa}_{-1}|\rho, \log\sigma_\kappa^2, \boldsymbol{\psi}) \\ &\quad \times f(\rho) f(\log\sigma_\kappa^2) f(\log\sigma_\beta^2) f(\boldsymbol{\psi}) f(\log\phi) \mathrm{d}\boldsymbol{\theta}_{\text{NBLC}} \end{aligned} \tag{12}$$

where $\boldsymbol{\theta}_{\text{NBLC}} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \rho, \log\sigma_\kappa^2, \log\sigma_\beta^2, \boldsymbol{\psi}, \log\phi)$ and is of dimension $p_{\text{NBLC}} = 246$.

7 Numerical Results

In this section, we compare our proposed models with the Bayesian PLC model (Czado et al. (2005)) to highlight the importance of accounting for overdispersion. We also provide a comparison of our proposed models with each other.

7.1 Estimated Parameters

The Bayesian PLC model (without overdispersion) is fitted using Czado's methodology, except we adopt the same prior specification as in Section 4.1 to facilitate model comparison later on.

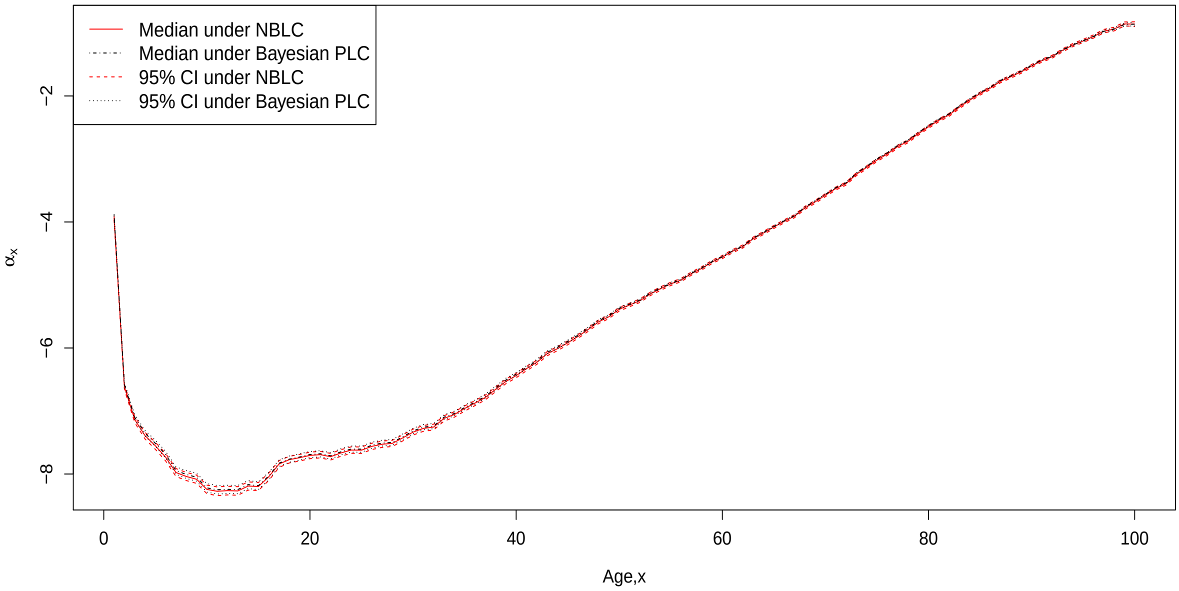


Figure 4: Plot of estimated α_x against age with their 95% credible intervals under the Bayesian Poisson LC model and the NBLC model.

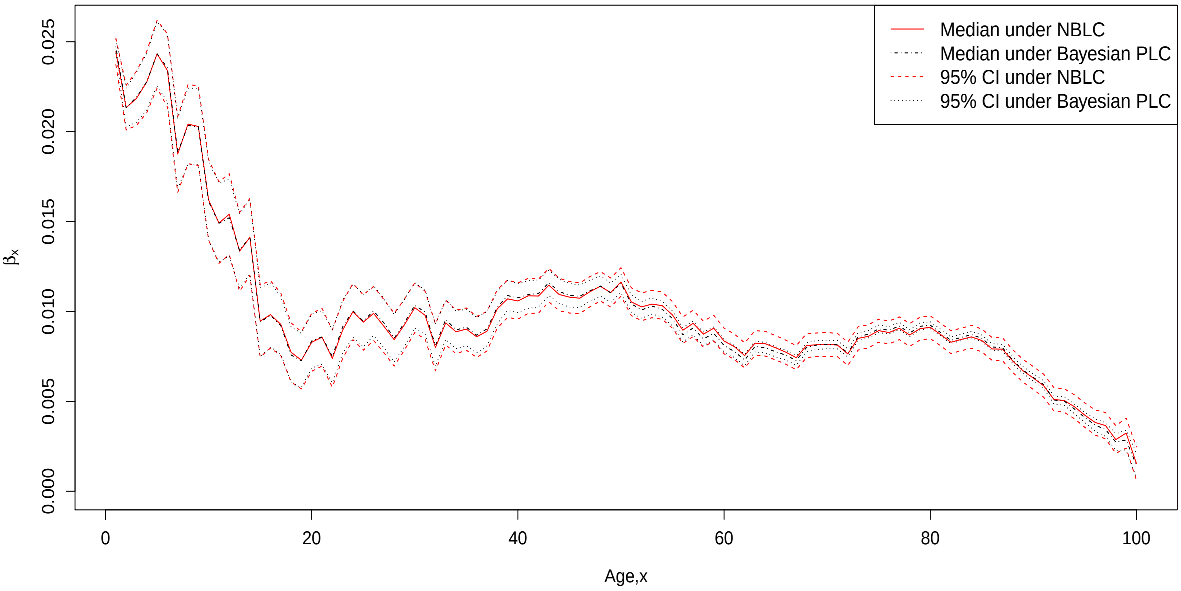


Figure 5: Plot of estimated β_x against age with their 95% credible intervals under the Bayesian PLC model and the NBLC model.

Figure 4 and 5 depict the fitted values (posterior medians) of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (accompanied by the associated 95% credible interval) under the Bayesian PLC and NBLC model. Note that the fitted values under the PLNLC model are not displayed for some of the plots here because they almost coincide with those of the NBLC model, and hence are excluded for a better visualization. According to Figure 4 and 5, the fitted values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under these models are rather similar, with the overdispersion models producing slightly wider credible intervals in general. Additionally, the width of the credible intervals also appear to be noticeably different as age increases.

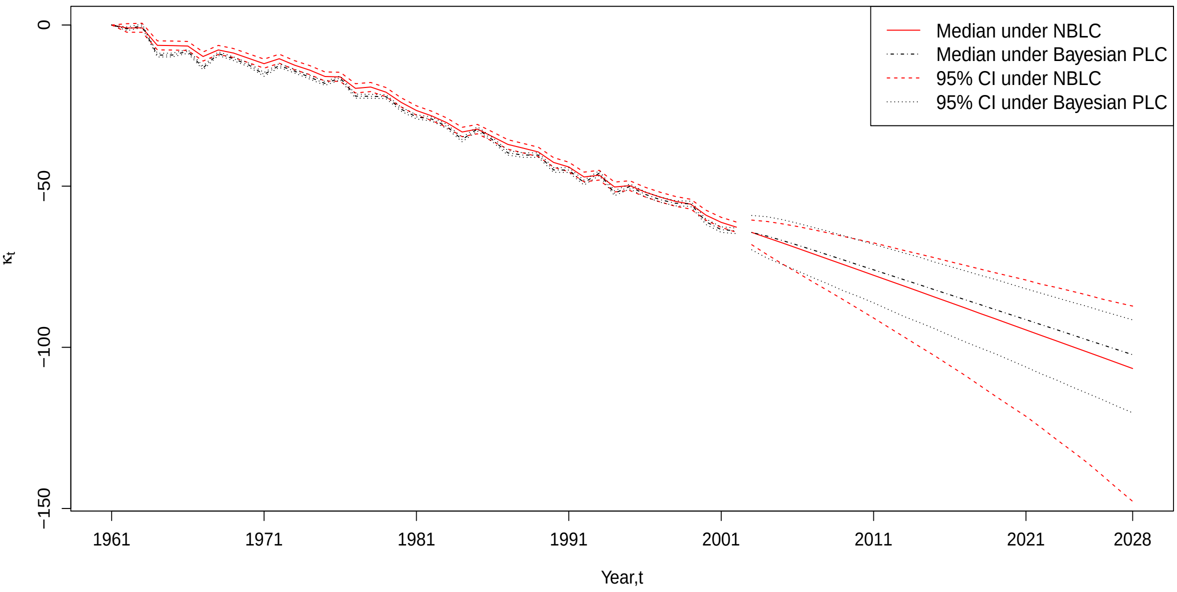


Figure 6: Plot of estimated κ_t and their 26−years ahead projection against years, accompanied by the corresponding 95% intervals under the Bayesian PLC model and the NBLC model.

The main difference arises from the parameter, $\boldsymbol{\kappa}$. As evident from Figure 6, the fitted values are larger and much smoother under the overdispersion models (with arguably wider credible intervals yet again). Furthermore, in terms of projection, not only do the overdispersion models forecast a better mortality improvement, the corresponding prediction intervals for the projected κ_t are also substantially wider. This is perhaps a little surprising considering that AR(1) prior is imposed on κ_t under all approaches. An intuitive explanation for this is that the overdispersion parameter provides more flexibilities for the model to "better" describe the data. In other words, models with overdispersion allow more priority to be put on fitting the AR(1) prior, hence the smoother fitted values. The exact reason behind this finding will be further explored when the marginal posterior distribution of ρ is examined in the next paragraph.

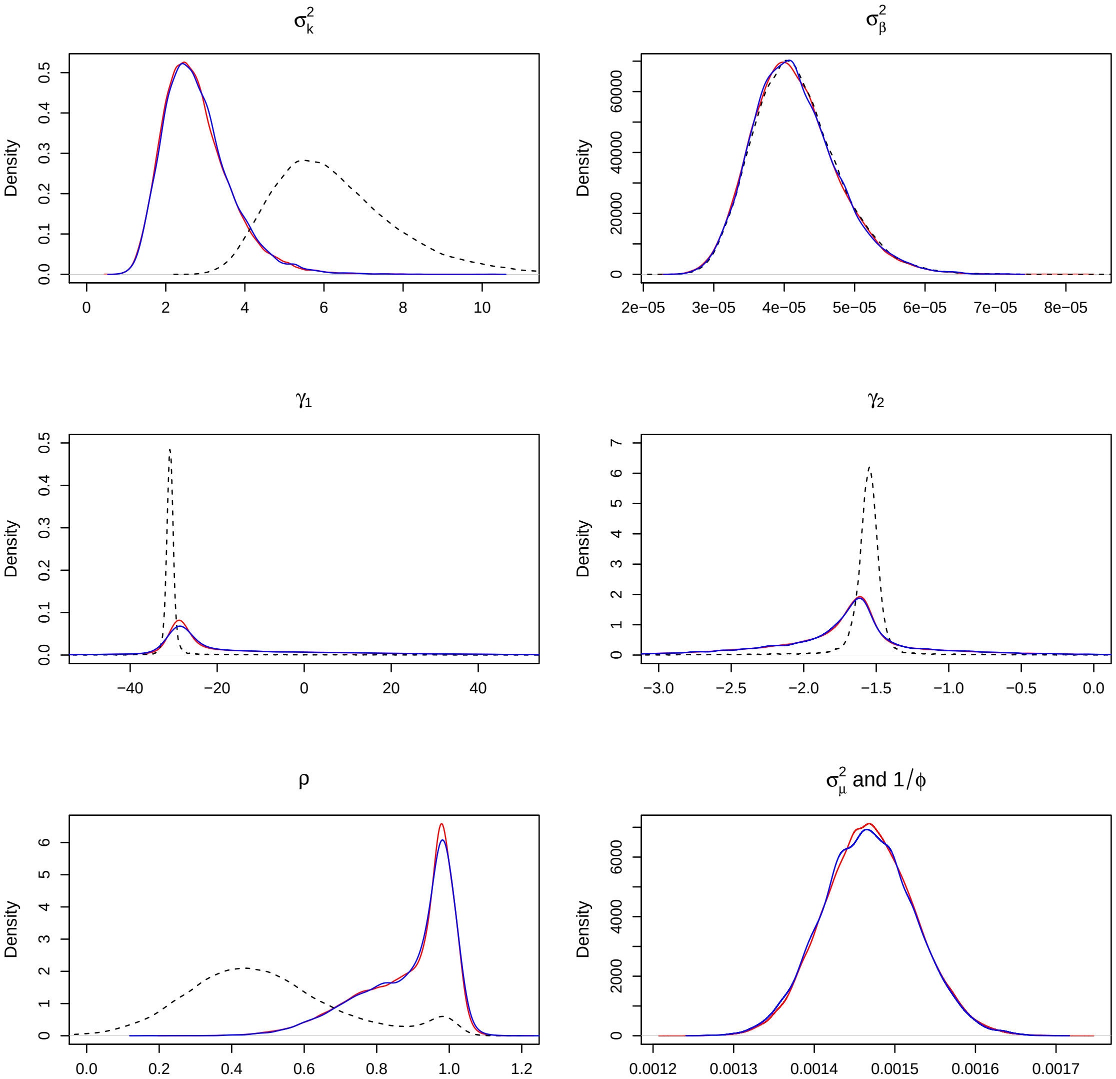


Figure 7: Kernel density plots of σ^2_κ, σ^2_β, ψ_1, ψ_2, ρ, and ϕ under the Bayesian PLC (black dotted), PLNLC (blue solid) and NBLC model (red solid).

Kernel estimates of the marginal posterior density of the rest of the parameters, derived from the posterior samples, are presented in Figure 7. The kernel densities of σ^2_β are almost identical. For ψ_1 and ψ_2, there are slight departures between the models, with the overdispersion models yielding heavier tails in both cases. The most apparent discrepancies occur at the marginal posterior of σ^2_κ and ρ. Specifically, the density of σ^2_κ for the Bayesian PLC model concentrates more at higher values, suggesting larger residuals for κ_t under this model. Interestingly, the marginal posterior of ρ is a mixture distribution under all models. In particular, the mixture distribution has 2 peaks, one at $\rho = 1$, and another one at somewhere between 0 and 1. This indicates that the projection model fitted on κ_t is a mixture of a stationary AR(1) and a random walk with drift model. Closer inspection shows that peaks of the marginal posterior of ρ occur at 0.42 and 1 for Bayesian PLC model; while for the overdispersion models, the peaks are at 0.80 and 1. In addition, the allocation of mixture proportion is also different, with the overdispersion models allocating a higher proportion for $\rho = 1$ (random walk) than their counterpart.

The marginal posterior of ρ enables us to justify our earlier findings on κ_t. Firstly, as the fitted ρ inclines towards larger values for the overdispersion models, the stationary AR(1) part of the model imposes a stronger smoothing on κ_t (as explained earlier). Hence, the smoother fitted κ_t for this model as observed. Secondly, the prediction intervals associated with the projection of κ_t are wider under these models because their projection model is largely dominated by the random walk with drift ($\rho = 1$), and a random walk model is known to produce relatively wider intervals than a stationary AR(1) model. Note also that this effect overshadows the fact that the residual variance, σ^2_κ, is larger for the Bayesian PLC model. Nevertheless, the projections of κ_t into the future under these models are expected to exhibit less explosive behaviour than what would be obtained if a pure random walk with drift was used.

Recall that the Poisson distribution is the limiting case of a negative binomial distribution as $\phi \to \infty$ (or $1/\phi \to 0$). Based on the MCMC sample generated, the posterior median of ϕ is approximately 681 ($1/\phi = 0.001468$), implying that the level of overdispersion is non-negligible. Moreover, for the PLNLC model, the Bayesian PLC model can be retrieved when $\sigma^2_\mu = 0$. Since the posterior median of σ^2_μ is around 0.001465, this confirms again the presence of overdispersion.

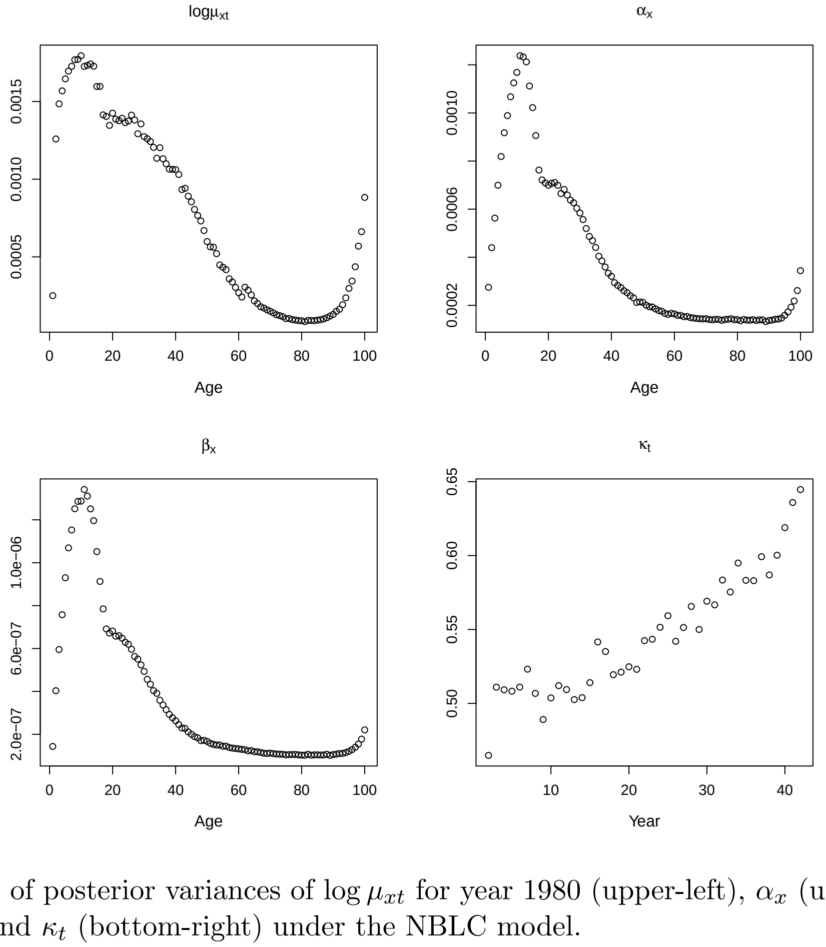


Figure 8: Plots of posterior variances of $\log \mu_{xt}$ for year 1980 (upper-left), α_x (upper-right), β_x (bottom-left) and κ_t (bottom-right) under the NBLC model.

The plots of posterior variances of $\log \mu_{x\,20}$, α_x, β_x, and κ_t under the NBLC model, computed from the MCMC samples are displayed in Figure 8. In particular, they demonstrate somewhat consistent patterns as their proposal variances, subject to small variations, (except maybe κ_t which shows a slight increasing trend) as described in Section 5, supporting our conjecture about the relationship between optimal proposal variances and posterior variances.

7.2 Mortality Forecast

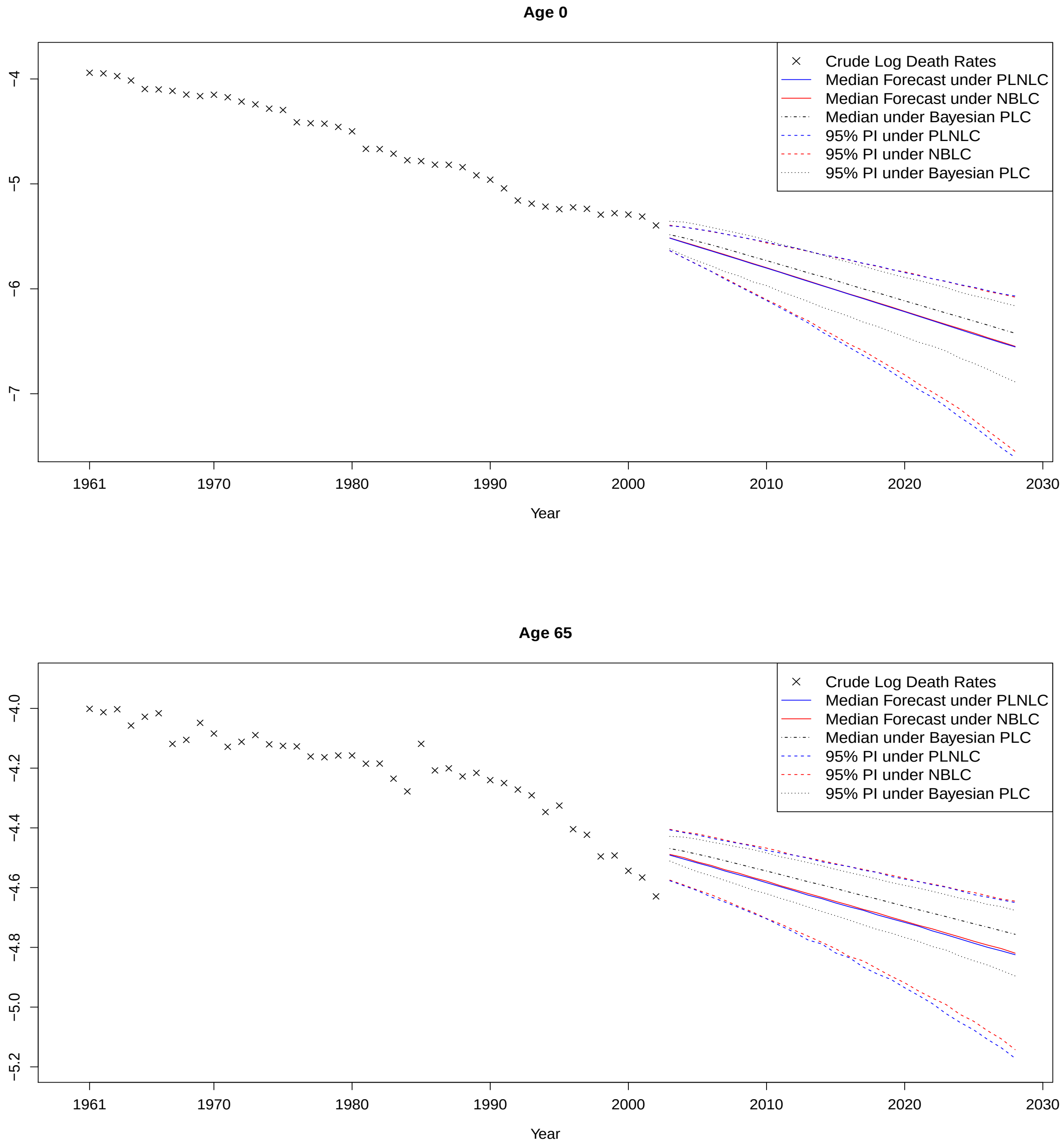


Figure 9: Plots of the logarithm of crude death rates, $\log(d_{xt}/e_{xt})$ and their associated 25-years ahead forecast for age 0 (upper panel) and 65 (lower panel) under the Bayesian PLC and the overdispersion models, accompanied by 95% prediction intervals.

Figure 9 shows the projected log mortality rates for newborns and age 65, 25 years into the future. In both cases (and for the rest of the ages), the overdispersion models clearly forecast a better improvement in the mortality rates, and also produces considerably wider prediction intervals. This is a sensible result as Lee and Miller (2001) illustrated that the original LC approach has a tendency to underestimate mortality improvement, which may well be inherited by the Bayesian PLC model. Moreover, the prediction intervals under the Bayesian PLC model also appear to be implausibly narrow, which is consistent with the findings by Alho (1992).

Hence, the inclusion of dispersion parameters provides a more sensible improvement in rates as well as better calibrated probabilistic intervals in terms of the projection.

7.3 Model Assessment

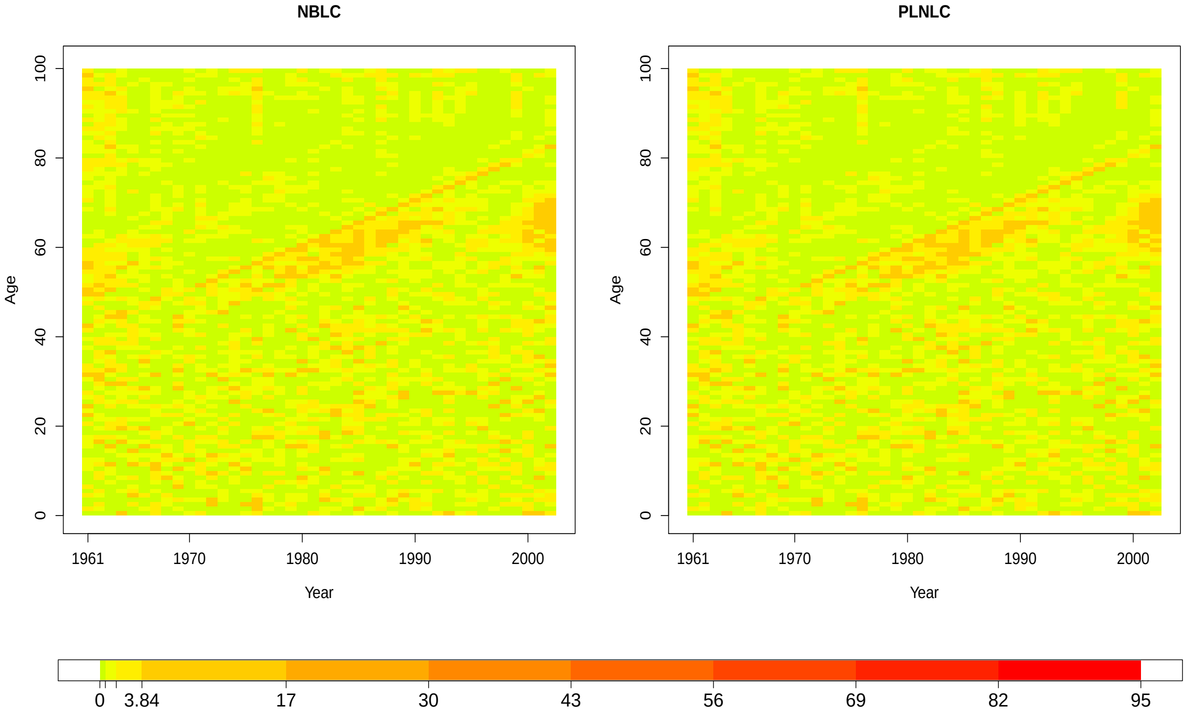


Figure 10: Heat map of squared Pearson residuals, r_{xt}^2, under the PLNLC (left panel) and the NBLC model (right panel).

We can similarly construct a heat map of the squared Pearson residuals, r_{xt}^2, where now the posterior mean of the parameters α_x, β_x, κ_t, and ϕ, are substituted into the expression in (2) for an estimate. As illustrated in Figure 10, the heat maps of the overdispersion models are much "greener" than before (Figure 1), indicating an overall improvement in goodness of fit. The sum of r_{xt}^2 ($\hat{r}$) for the PLNLC and the NBLC model are now 4235.24 and 4235.83 respectively, which are considerably smaller than 15378.73 of the original PLC model, and 15379.91 of the Bayesian PLC model. The improvement is substantial, but is still not ideal mostly because of the uncaptured cohort effects, emerged as yellow/orange diagonal lines in Figure 10. Nevertheless, it is rather obvious that the overdispersion models outperformed both the original PLC and Bayesian PLC model by a considerable margin.

Note that the distribution of the sum of squared Pearson residuals, r^2, is no longer Chi-squared, but can be properly calibrated against its empirical distribution to then carry out posterior predictive checking. Following Gelman et al. (1995), we first generate a set of replicated data, $\boldsymbol{d}^{\text{rep}}$, which has a density representation

$$f(\boldsymbol{d}^{\text{rep}}|M) = \int f(\boldsymbol{d}^{\text{rep}}|\boldsymbol{\theta}_M, M)f(\boldsymbol{\theta}_M|\boldsymbol{d}, M)d\boldsymbol{\theta},$$

from the posterior samples of $\boldsymbol{\theta}$ under each model. Next, we define our test quantity as

$$T(\boldsymbol{d}, \boldsymbol{\theta_M}) = \sum_{x,t} \frac{(d_{xt} - \mathbb{E}[D_{xt}|\boldsymbol{\theta}_M, M])^2}{\mathrm{Var}[D_{xt}|\boldsymbol{\theta}_M, M]},$$

which depends on both the data and parameters. An expression of $T(\boldsymbol{d}, \boldsymbol{\theta_M})$ for each of the models under consideration is presented in Appendix C. The test quantity is then evaluated at the replicated data to yield $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta_M})$, from which histograms can be constructed (Figure 11).

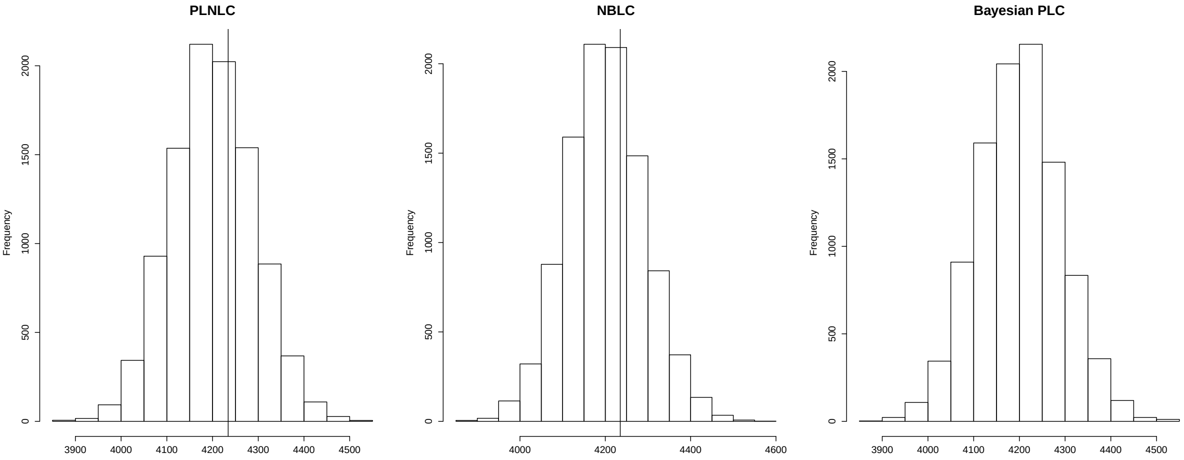


Figure 11: Histograms of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta_M})$ for the PLNLC, NBLC, and Bayesian PLC model, with their corresponding sum of squared Pearson residuals, r^2 included as the vertical solid lines.

The value of r^2 for each model is displayed in Figure 11 to highlight the magnitude of its discrepancy with the $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$. It can be seen that the r^2 for the overdispersion models lies somewhere in the middle of the histograms; while the r^2 for the Bayesian PLC model (15379.91) is completely off the charts. Moreover, the posterior predictive p-value, defined as

$$p_B = \Pr(T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M) \geq T(\boldsymbol{d}, \boldsymbol{\theta}_M)|\boldsymbol{d}),$$

can be used to assess statistical significance formally. In practice, it is easily computed as the proportion of the predictive test quantity, $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$, which equals or exceeds the realized test quantity, $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$. The posterior predictive p-values of the Bayesian PLC, PLNLC and NBLC model are 0.0161, 0.0156 and 0.00 respectively. Therefore, there is no evidence at 1% level that the overdispersion models are inadequate in this aspect of the data; while the extreme p-value of the Bayesian PLC model strongly indicate model inadequancy.

7.4 Out-out-Sample Validation

In this section, we validate the candidate models against the holdout data based on an aggregate mortality quantity, the life expectancy at birth, derived from the projected crude mortality rates (Section 5.5).

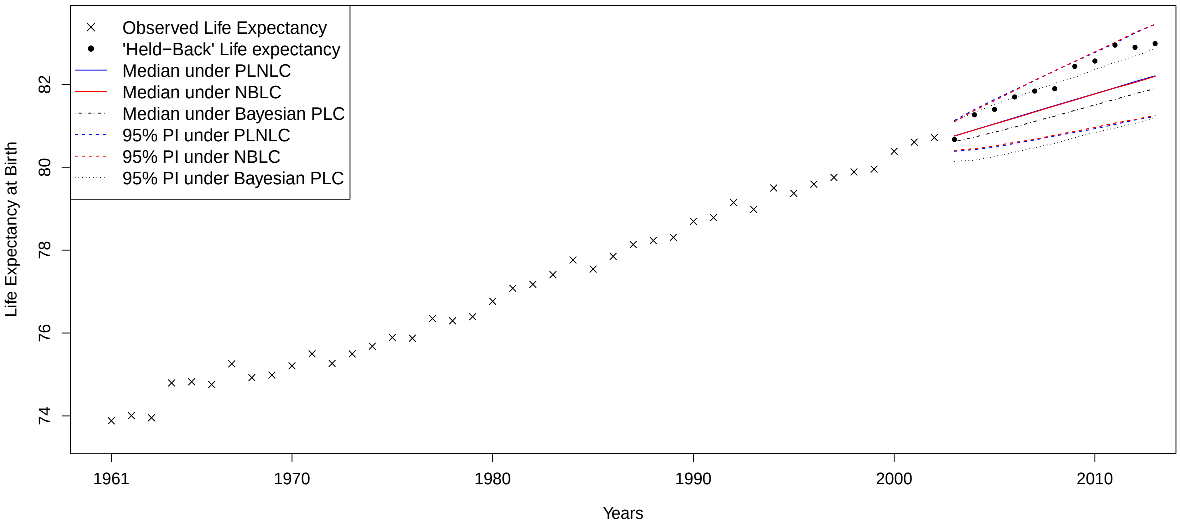


Figure 12: Plots of the observed life expectancy at birth and the associated 11-years ahead forecast under the Bayesian PLC and the overdispersion models, accompanied by the 95% prediction intervals.

As illustrated in Figure 12, the overdispersion models forecast larger life expectancies at birth consistently and produce wider prediction intervals than the Bayesian PLC model. Moreover, the holdout life expectancies at birth all lie well within the 95% prediction intervals of the overdispersion models; while the Bayesian PLC model clearly underestimates the gains in the future life expectancy at birth, as well as producing an overly narrow prediction interval. All in all, the overdispersion models offer a better predictive power than their counterpart. One concern is that the overdispersion models seemingly also yield a systematic underestimation of the life expectancy, even though their prediction intervals provide satisfactory coverages.

7.5 Investigating Model Similarity

Throughout the previous subsections, most of the results suggest that the two overdispersion models are very similar. This prompts the initiative to compare the fitted log mortality rates using sample quantiles-quantiles (QQ) plots.

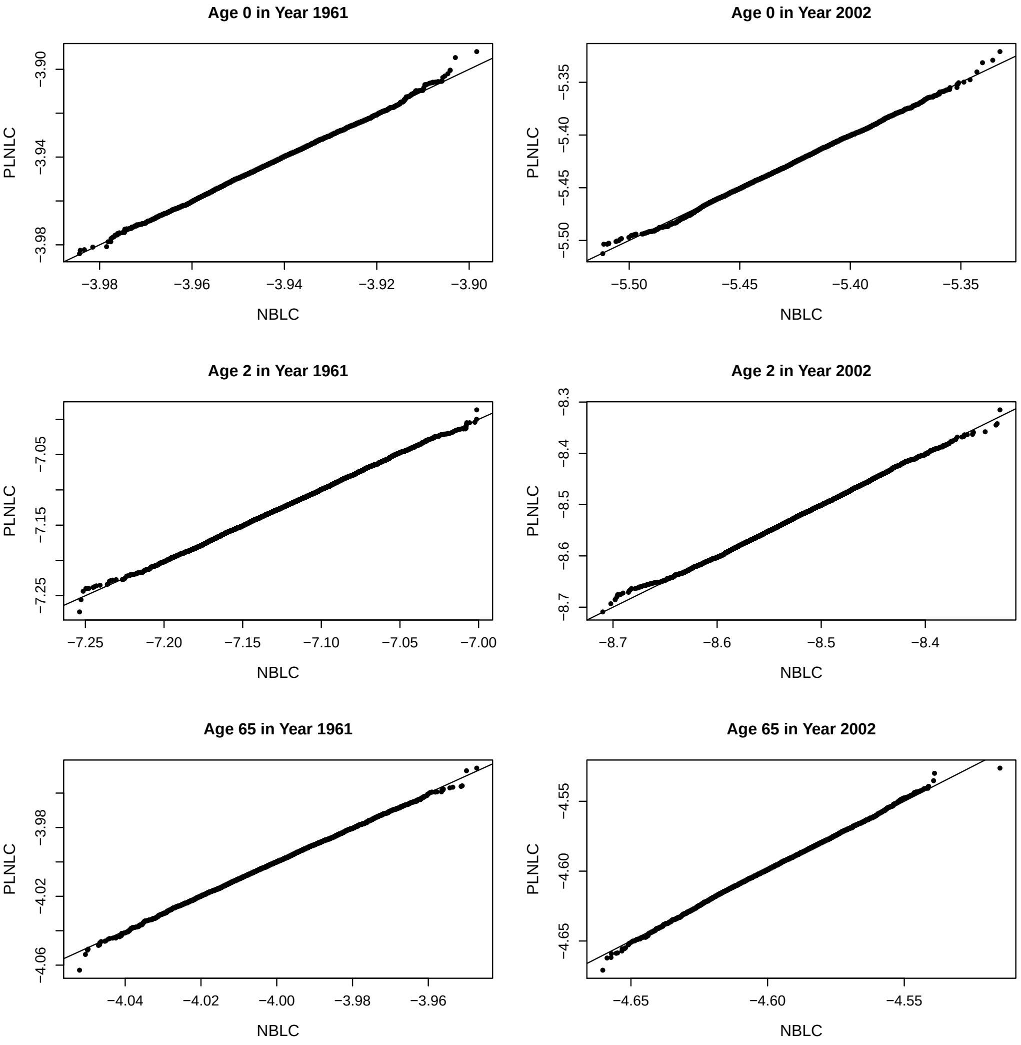


Figure 13: QQ-plots of the posterior sample of $\log \mu_{1\,1}$, $\log \mu_{1\,43}$, $\log \mu_{2\,1}$, $\log \mu_{2\,43}$, $\log \mu_{66\,1}$, $\log \mu_{66\,42}$ for the overdispersion models, with solid lines denoting equality.

From Figure 13, it is evident that all of the sample QQ-plots appear to lie reasonably close to the reference line, with no peculiar behaviour (no U or S-shape). This suggests that the posterior distributions of $\log \mu_{xt}$ have similar skewness and tail distributions under both overdispersion models. Arguably, the NBLC model produces slightly heavier tail for some of the mortality rates (for example, $\log \mu_{66\,42}$). This may be due to the feature of a gamma distribution, which generally possesses a heavier tail in comparison to a log-normal distribution.

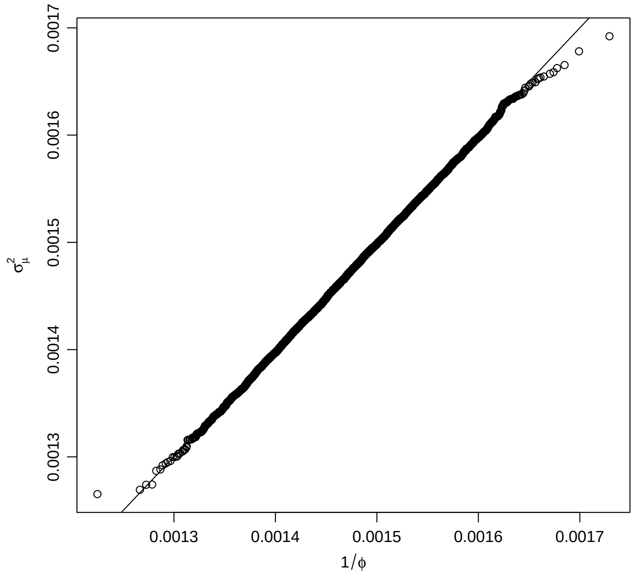


Figure 14: QQ-plot of posterior sample of σ_μ^2 against $1/\phi$, with black solid line denoting equality.

Furthermore, the QQ-plot of σ_μ^2 against $1/\phi$ is remarkably close to the reference line as depicted in Figure 14, suggesting that their posterior distributions are essentially the same. In other words, the overall level of overdispersion indicated under both models are virtually the same, supporting our conjecture derived from Taylor's approximation (Section 4.1). Again, this signifies model similarity.

7.6 Bayesian Model Determination

Here, we perform Bayesian model comparison using marginal likelihoods. The marginal likelihoods of each model, approximated using bridge sampling are presented in Table 1.

Table 1: The marginal likelihoods (in logarithmic scale) of each model approximated from bridge sampling.

Bayesian Poisson LC	Poisson Log-normal LC	Negative Binomial LC
−26684.10	−23723.65	−23727.48

As expected, the marginal likelihoods of both the overdispersion models are appreciably larger than the Bayesian PLC model. Recall also that the exploratory analyses in the previous subsection suggest that the PLNLC and the NBLC model are very similar. In particular, the marginal likelihoods of the overdispersion models are exceptionally close to each other, verifying again the similarity between the two proposed models. However, it should be pointed out that we experienced major difficulty during the computation of bridge sampling estimate of the marginal likelihood for the PLNLC model due to high dimensionality. Without marginalising the log mortality rates, $\log \mu_{xt}$, this model has a dimensionality of 4446, as compared to 246 of the NBLC model. With the MCMC algorithm only generating dependent posterior samples, it implies that a relatively large sample size is essential to learn about the posterior distributions under this model. Our hypothesis on the failure of bridge sampling in accurately estimating the marginal likelihood in this case is the lack of sufficiently long samples to obtain a good approximation of the posterior moments (especially the variance matrix). That being said, we were still able to get the bridge sampling estimate to attain convergence after devoting an immense computational effort.

Even though both the overdispersion models provide similar fit qualitatively, the NBLC model is to be recommended due to its computational advantage over its counterpart by having a lower dimension after integrating out the latent variables, μ_{xt}.

8 Conclusion

In this paper, we focused on the importance of accounting for overdispersion in modelling a mortality data. In particular, we presented two models, the PLNLC and the NBLC model, both of which extended the original PLC model by introducing a single dispersion parameter. These models were then fitted within the Bayesian framework for coherency. Vague priors were used for illustrative purposes, but elicitation of expert mortality knowledge can be carried out in practice wherever applicable. In general, we demonstrated that neglecting overdispersion not only leads to over-confident probabilistic intervals, but in our case also gives rise to overfitting, both of which are detrimental for the subsequent mortality projection. Specifically, our results showed that both the overdispersion models forecast a better mortality improvement in the future, as well as yielding much more representative prediction intervals than the Bayesian PLC model (as indicated by the out-of-sample validation). Moreover, various model assessment tools suggested that the overdispersion models provide significantly better fit than the Bayesian PLC model. Between the two proposed models, they provide rather similar qualitative fit, with the NBLC model producing slightly heavier-tailed posterior distributions. Formal Bayesian model comparison using posterior model probabilities also showed that they are very similar. However, we recommend the NBLC model over the PLNLC model mainly due to computational reasons. Finally, the overdispersion models provide pronounced improvement in fit, but can be further refined by including the cohort components. Until then, the dispersion parameters do not represent heterogeneity entirely in the sense that it is contaminated with the cohort effects.

Appendix A Conditional Posterior Distributions for the PLNLC Model

Denote $\boldsymbol{\mu}_x = (\mu_{x1}, \mu_{x2}, \ldots, \mu_{xT})^\top$ and $\boldsymbol{\mu}_t = (\mu_{1t}, \mu_{2t}, \ldots, \mu_{At})^\top$ as vectors of mortality rates corresponding to age group x and year t respectively. Also, denote $\boldsymbol{\mu}_{-x,t} = (\mu_{1t}, \ldots, \mu_{x-1\,t}, \mu_{x+1\,t}, \ldots, \mu_{At})^\top$ as a vector of mortality rates corresponding to year t, excluding the x^{th} component, and $\boldsymbol{\mu}_{x,-t} = (\mu_{x1}, \ldots, \mu_{x\,t-1}, \mu_{x\,t+1}, \ldots, \mu_{xT})^\top$ as a vector of mortality rates corresponding to age group x, excluding the t^{th} component.

i. $\boldsymbol{\alpha}|\boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma^2_\kappa, \sigma^2_\beta, \rho, \boldsymbol{\psi}, \sigma^2_\mu \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_\alpha &= \left(\frac{T}{\sigma^2_\mu} + \frac{1}{\sigma^2_\alpha}\right)^{-1} \boldsymbol{I}_A, \\
\boldsymbol{\mu}_\alpha &= \boldsymbol{\Sigma}_\alpha \times \left(\frac{\sum_t (\log\boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)}{\sigma^2_\mu} + \frac{\alpha_0 \mathbf{1}_A}{\sigma^2_\alpha}\right).
\end{aligned}$$

ii. $\boldsymbol{\beta}_{-1}|\boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma^2_\kappa, \sigma^2_\beta, \rho, \boldsymbol{\psi}, \sigma^2_\mu \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_\beta &= \left[\frac{\sum_t \kappa_t^2}{\sigma^2_\mu}(\boldsymbol{I}_{A-1} + \boldsymbol{J}_{A-1}) + \frac{1}{\sigma^2_\beta}\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\right]^{-1}, \\
\boldsymbol{\mu}_\beta &= \boldsymbol{\Sigma}_\beta \times \left[\frac{1}{\sigma^2_\mu}\sum_t \kappa_t(\log\boldsymbol{\mu}_{-1,t} - \boldsymbol{\alpha}_{-1}) + \frac{1}{\sigma^2_\mu}\sum_t \kappa_t(-\log\mu_{1t} + \alpha_1 + \kappa_t)\mathbf{1}_{A-1}\right. \\
&\quad \left. + \frac{1}{A\sigma^2_\beta}\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\mathbf{1}_{A-1}\right].
\end{aligned}$$

iii. $\boldsymbol{\kappa}_{-1}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma^2_\kappa, \sigma^2_\beta, \rho, \boldsymbol{\psi}, \sigma^2_\mu \sim N(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_\kappa &= \left[\frac{\sum_x \beta_x^2}{\sigma^2_\mu}\boldsymbol{I}_{T-1} + \frac{1}{\sigma^2_\kappa}\boldsymbol{Q}\right]^{-1}, \\
\boldsymbol{\mu}_\kappa &= \boldsymbol{\Sigma}_\kappa \times \left[\frac{1}{\sigma^2_\mu}\sum_x \beta_x(\log\boldsymbol{\mu}_{x,-1} - \alpha_x\mathbf{1}_{T-1}) + \frac{1}{\sigma^2_\kappa}\boldsymbol{Q}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi})\right].
\end{aligned}$$

iv. $\sigma^2_\kappa|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{\mathbf{-1}}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma^2_\beta, \rho, \psi, \sigma^2_\mu$
$\sim \text{Gamma}\left(a_\kappa + \frac{T-1}{2}, b_\kappa + \frac{1}{2}\sum_{t=2}^T[\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1})]^2\right)$.

v. $\sigma^2_\beta|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{\mathbf{-1}}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma^2_\kappa, \rho, \psi, \sigma^2_\mu$
$\sim \text{Gamma}\left(a_\beta + \frac{A-1}{2}, b_\beta + \frac{1}{2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right)$.

vi. Defining $a_\rho = \sum_{t=2}^T(\kappa_{t-1} - \eta_{t-1})^2$ and $b_\rho = \sum_{t=2}^T(\kappa_t - \eta_t)(\kappa_{t-1} - \eta_{t-1})$, then $\rho|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{\mathbf{-1}}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma^2_\kappa, \sigma^2_\beta, \psi, \sigma^2_\mu \sim N(\rho^*, \sigma^{*2}_\rho)$, where

$$\rho^* = \frac{b_\rho}{a_\rho + \frac{\sigma^2_\kappa}{\sigma^2_\rho}} \quad \text{and} \quad \sigma^{*2}_\rho = \frac{\sigma^2_\kappa}{a_\rho + \frac{\sigma^2_\kappa}{\sigma^2_\rho}}.$$

vii. $\boldsymbol{\psi}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{\mathbf{-1}}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma^2_\kappa, \sigma^2_\beta, \rho, \sigma^2_\mu \sim N(\boldsymbol{\psi}^*, \boldsymbol{\Sigma}^*_\psi)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}^*_\psi &= \left[\frac{1}{\sigma^2_\kappa}(\boldsymbol{Y}_{-1} - \rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)^\top\boldsymbol{Q}(\boldsymbol{Y}_{-1} - \rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1) + \boldsymbol{\Sigma}_0^{-1}\right]^{-1} \\
\psi^* &= \boldsymbol{\Sigma}^*_\psi \times \left[\frac{1}{\sigma^2_\kappa}(\boldsymbol{Y}_{-1} - \rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)^\top\boldsymbol{Q}\boldsymbol{\kappa}_{-1} + \boldsymbol{\Sigma}_0^{-1}\psi_0\right].
\end{aligned}$$

Appendix B Conditional Posterior Densities And MH Acceptance Probabilities under the NBLC model

B.1 For α_x

The conditional posterior density of α_x is given by

$$\begin{aligned}
& f(\alpha_x|\boldsymbol{\alpha}_{-x}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \rho, \sigma^2_\kappa, \sigma^2_\beta, \boldsymbol{\psi}, \phi) \\
\propto & \prod_t \left\{\frac{\exp(d_{xt}\alpha_x)}{[e_{xt}\exp(\alpha_x + \beta_x\kappa_t) + \phi]^{d_{xt}+\phi}}\right\} \times \exp\left[-\frac{(\alpha_x - \alpha_0)^2}{2\sigma^2_\alpha}\right].
\end{aligned}$$

Using similar notation as before, the acceptance probability of α_x for our MH algorithm is

$$\begin{aligned} a(\alpha_x^*|\alpha_x^{i-1}) &= \min\left\{1, \exp\left[(\alpha_x^* - \alpha_x^{i-1})\sum_t d_{xt} - \frac{(\alpha_x^* - \alpha_0)^2 - (\alpha_x^{i-1} - \alpha_0)^2}{2\sigma_\alpha^2}\right]\right. \\ &\qquad \left. \times \prod_t \left[\frac{e_{xt}\exp(\alpha_x^{i-1} + \beta_x\kappa_t) + \phi}{e_{xt}\exp(\alpha_x^* + \beta_x\kappa_t) + \phi}\right]^{d_{xt}+\phi}\right\}, \end{aligned}$$

where $\alpha_x^* \sim N(\alpha_x^{i-1}, \sigma_{\alpha_x}^2)$ is a random walk proposal centered around the current iteration and with the proposal variance, $\sigma_{\alpha_x}^2$.

B.2 For β_x

The conditional posterior density of β_x for $x = 2, \ldots, A$ is given by

$$\begin{aligned} & f(\beta_x|\beta_2, \ldots, \beta_{x-1}, \beta_{x+1}, \ldots, \beta_A, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \rho, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \phi) \\ \propto\ & \exp\left\{-\phi\sum_t \frac{\mu_{xt}}{\exp(\alpha_x + \beta_x\kappa_t)} - \phi\sum_t \frac{\mu_{1t}}{\exp\left[\alpha_1 + \kappa_t * (1 - \sum_{j\neq 1}\beta_j)\right]}\right. \\ & \left. -\frac{1}{2\sigma_\beta^2}\left(1 - \sum_{j\neq 1}\beta_j\right)^2 - \frac{1}{2\sigma_\beta^2}\beta_x^2\right\}. \end{aligned}$$

With a random walk proposal, $\beta_x^* \sim N(\beta_x^{i-1}, \sigma_{\beta_x}^2)$, the acceptance probability of β_x for our MH algorithm is

$$\begin{aligned} a(\beta_x^*|\beta_x^{i-1}) &= \min\left\{1, \frac{\exp\left[-\phi\sum_t \frac{\mu_{xt}}{\exp(\alpha_x+\beta_x^*\kappa_t)} - \phi\sum_t \frac{\mu_{1t}}{\exp[\alpha_1+\kappa_t(1-\beta_x^*-\sum_{j\neq 1,x}\beta_j)]}\right]}{\exp\left[-\phi\sum_t \frac{\mu_{xt}}{\exp(\alpha_x+\beta_x^{i-1}\kappa_t)} - \phi\sum_t \frac{\mu_{1t}}{\exp[\alpha_1+\kappa_t(1-\beta_x^{i-1}-\sum_{j\neq 1,x}\beta_j)]}\right]}\right. \\ &\qquad \left. \times \frac{\exp\left[-\frac{1}{2\sigma_\beta^2}\left(1 - \beta_x^* - \sum_{j\neq 1,x}\beta_j\right)^2 - \frac{1}{2\sigma_\beta^2}(\beta_x^*)^2\right]}{\exp\left[-\frac{1}{2\sigma_\beta^2}\left(1 - \beta_x^{i-1} - \sum_{j\neq 1,x}\beta_j\right)^2 - \frac{1}{2\sigma_\beta^2}(\beta_x^{i-1})^2\right]}\right\}. \end{aligned}$$

B.3 For κ_t

The conditional posterior distribution of κ_t is given by

$$\begin{aligned} & f(\kappa_t|\boldsymbol{\kappa}_{-t}\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \rho, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \phi) \\ \propto\ & \exp\left\{-\phi\sum_x\left[\frac{\mu_{xt}}{\exp(\alpha_x + \beta_x\kappa_t)}\right] - \phi\kappa_t - \frac{1}{2\sigma_\kappa^2}[\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1})]^2\right. \\ & \left. -\frac{1}{2\sigma_\kappa^2}[\kappa_{t+1} - \eta_{t+1} - \rho(\kappa_t - \eta_t)]^2\right\}, \end{aligned}$$

for $t = 2, \ldots, T-1$, and

$$\begin{aligned} & f(\kappa_T|\boldsymbol{\kappa}_{-T}\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \rho, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \phi) \\ \propto\ & \exp\left\{-\phi\sum_x\left[\frac{\mu_{xt}}{\exp(\alpha_x + \beta_x\kappa_t)}\right] - \phi\kappa_t - \frac{1}{2\sigma_\kappa^2}[\kappa_T - \eta_T - \rho(\kappa_{T-1} - \eta_{T-1})]^2\right\}. \end{aligned}$$

With a random walk proposal, $\kappa_t^* \sim N(\kappa_t^{i-1}, \sigma_{\kappa_t}^2)$, the acceptance probabilities of κ_t for our MH algorithm can be expressed as

$$\begin{aligned} a(\kappa_t^*|\kappa_t^{i-1}) &= \min\left\{1, \exp\left\{-\phi \sum_x \frac{\mu_{xt}}{e^{\alpha_x}}\left(\frac{1}{e^{\beta_x \kappa_t^*}} - \frac{1}{e^{\beta_x \kappa_t^{i-1}}}\right) - \phi(\kappa_t^* - \kappa_t^{i-1})\right.\right. \\ &\qquad -\frac{1}{2\sigma_\kappa^2}(\kappa_t^* - \kappa_t^{i-1})[\kappa_t^* + \kappa_t^{i-1} - 2\eta_t - 2\rho(\kappa_{t-1} - \eta_{t-1})] \\ &\qquad \left.\left. -\frac{1}{2\sigma_\kappa^2}\rho(\kappa_t^{i-1} - \kappa_t^*)[2\kappa_{t+1} - 2\eta_{t+1} - \rho(\kappa_t^* + \kappa_t^{i-1} - 2\eta_t)]\right\}\right\}, \end{aligned}$$

for $t = 2, \dots, T-1$, and

$$\begin{aligned} a(\kappa_T^*|\kappa_T^{i-1}) &= \min\left\{1, \exp\left\{-\phi \sum_x \frac{\mu_{xT}}{e^{\alpha_x}}\left(\frac{1}{e^{\beta_x \kappa_T^*}} - \frac{1}{e^{\beta_x \kappa_T^{i-1}}}\right) - \phi(\kappa_T^* - \kappa_T^{i-1})\right.\right. \\ &\qquad \left.\left. -\frac{1}{2\sigma_\kappa^2}(\kappa_T^* - \kappa_T^{i-1})[\kappa_T^* + \kappa_T^{i-1} - 2\eta_t - 2\rho(\kappa_{t-1} - \eta_{t-1})]\right\}\right\}. \end{aligned}$$

B.4 For ϕ

The conditional posterior density of ϕ is given by

$$\begin{aligned} & f(\phi|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho) \\ \propto\; & \frac{1}{\Gamma(\phi)^{AT}}\phi^{AT\phi + a_\phi - 1}\exp(-b_\phi \phi) \times \prod_{x,t}\left\{\frac{\Gamma(d_{xt} + \phi)}{[e_{xt}\exp(\alpha_x + \beta_x \kappa_t) + \phi]^{d_{xt} + \phi}}\right\}. \end{aligned}$$

Thus, we have

$$\begin{aligned} a(\phi^*|\phi^{i-1}) &= \min\left\{1, \frac{(\phi^*)^{\phi^* AT + a_\phi - 1}}{(\phi^{i-1})^{\phi^{i-1}AT + a_\phi - 1}}\left[\frac{\Gamma(\phi^{i-1})}{\Gamma(\phi^*)}\right]^{AT}\left(\prod_{x,t}\mu_{xt}\right)^{\phi^* - \phi^{i-1}}\right. \\ &\qquad \left. \times \exp\left[-(\phi^* - \phi^{i-1})\left(b_\phi + \sum_{x,t}\left(\alpha_x + \beta_x \kappa_t + \frac{\mu_{xt}}{\exp(\alpha_x + \beta_x \kappa_t)}\right)\right)\right]\right\}, \end{aligned}$$

as the acceptance probability for a proposal $\phi^* \sim N(\phi^{i-1}, \sigma_\phi^2)$.

Appendix C The Test Quantity

Expression of the test quantity under each model is given by

$$\begin{aligned} T(\boldsymbol{d}, \boldsymbol{\theta}_{\text{PLC}}) &= \sum_{x,t}\frac{(d_{xt} - e_{xt}\exp(\alpha_x + \beta_x \kappa_t))^2}{e_{xt}\exp(\alpha_x + \beta_x \kappa_t)}, \\ T(\boldsymbol{d}, \boldsymbol{\theta}_{\text{PLNLC}}) &= \sum_{x,t}\frac{(d_{xt} - e_{xt}\exp(\alpha_x + \beta_x \kappa_t + \frac{1}{2}\sigma_\mu^2))^2}{e_{xt}\exp(\alpha_x + \beta_x \kappa_t + \frac{1}{2}\sigma_\mu^2)[1 + (e^{\sigma_\mu^2} - 1)e_{xt}\exp(\alpha_x + \beta_x \kappa_t + \frac{1}{2}\sigma_\mu^2)]}, \\ T(\boldsymbol{d}, \boldsymbol{\theta}_{\text{NBLC}}) &= \sum_{x,t}\frac{(d_{xt} - e_{xt}\exp(\alpha_x + \beta_x \kappa_t))^2}{e_{xt}\exp(\alpha_x + \beta_x \kappa_t)\left[1 + \frac{e_{xt}\exp(\alpha_x + \beta_x \kappa_t)}{\phi}\right]}. \end{aligned}$$

References

Abel, G. J. (2015). fanplot: An R package for Visualizing Sequential Distributions. *R JOURNAL, 7*(1), 15–23.

Alho, J. M. (1992). Modelling and Forecasting the Time Series of U.S. mortality. *Journal of American Statistical Associantion, 87*, 673–674.

Booth, H. and L. Tickle (2008). Mortality Modelling and Forecasting: A review of methods. *The Australian Demographic & Social Research Institute Working Paper,* (3).

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

Brown, J. R. (2003). Redistribution and Insurance: Mandatory Annuitization with Mortality Heterogeneity. *The Journal of Risk and Insurance, 70*(1), 17–41.

Cairns, A., D. Blake, and K. Dowd (2005). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty. *Centre for Risk & Insurance Studies Discussion Paper Series,* **5**.

Chen, M. H., Q. M. Shao, and J. G. Ibrahim (2000). *Monte Carlo Methods in Bayesian Computation.* Springer-Verlag New York, Inc.

Czado, C., A. Delwarde, and M. Denuit (2005, January). Bayesian Poisson Log-Bilinear Mortality Projections. *Insurance: Mathematics and Economics,* **36**, 260–284.

Delwarde, A., M. Denuit, and C. Partrat (2007). Negative Binomial Version of the Lee-Carter Model for Mortality Forecasting. *Applied Stochastic Models in Business and Industry, 23*, 381–401.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis,* **1**(3), 515–533.

Gelman, A., D. B. Rubin, J. B. Carlin, and H. S. Stern (1995). *Bayesian Data Analysis* (1st ed.). Chapman and Hall Ltd.

Girosi, F. and G. King (2008). *Demographic Forecasting.* Princeton University Press.

HMD (2000). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org. Accessed: 2016-03-01.

Lee, R. D. and L. R. Carter (1992). Modelling and Forecasting u.s. Mortality. *Journal of the American Statistical Association,* **87**(419), 659–671.

Lee, R. D. and T. Miller (2001). Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality. *Demography,* **38**(4), 537–549.

Li, S. H., M. R. Hardy, and K. S. Tan (2009). Uncertainty in Mortality Forecasting: An extension to the classical lee-carter approach. *Astin Bulletin,* **39**(1), 137–164.

Meng, X. L. and W. H. Wong (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica,* **6**(4), 831–860.

O'Hagan, A. and J. Forster (2004). *Kendall's Advanced Theory of Statistics* (2nd ed.), Volume 2B. Kendall's Library of Statistics.

Overstall, A. M. and J. J. Forster (2010). Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis,* **54**(12), 3269–3288.

Renshaw, A. E. and H. Haberman (2005). A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance: Mathematics and Economics,* **38**(3), 556–570.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science,* **16**(4), 351–367.

Roberts, G. O. and S. K. Sahu (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistics Society, Series B (Methodological),* **59**(2), 291–317.

Tuljapurkar, S., N. Li, and C. Boe (2000). A Universal Pattern of Mortality Decline in The G7 Countries. *Letters to Nature,* **405**, 789–792.