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 (Brouhns et al., 2002) that imposes mean-variance equality, restricting mortality variations across individuals. Specifically, we present two models to potentially account for overdispersion. We propose to fit these models within the Bayesian framework for various advantages, but primarily 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 (Czado et al., 2005) 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 compare the models quantitatively.

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

*Corresponding Author.

E-mail addresses: jstw1r17@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 mortality model was pioneered by Lee and Carter (henceforth LC) in 1992, and has since then become 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 for the number of deaths rather than an additive error term for the logarithm of mortality rates. Cairns et al. (2006) 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 Poisson gamma/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, which suffers from issues of the inconsistent two-stage model fitting procedure (see Section 4 for more details) and the inability to account for multiple sources of uncertainties coherently. Czado et al. (2005) partially solved these issues by implementing a fully integrated Bayesian approach of fitting the Poisson LC model, but did not consider the presence of overdispersion. Therefore, our main aim is to combine their methodologies, that is to fit the mixed Poisson LC models within a Bayesian paradigm, which has the primary advantage of producing properly calibrated uncertainty bands that incorporate

various sources of uncertainties. More advantages of Bayesian mortality modelling/forecasting will be discussed in detail 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, Pedroza (2006) innovatively performed mortality forecasting using a Bayesian state-space model (treating ages as "space") using Kalman's filtering estimation procedure, with a built-in ability to handle missing data. Li (2014) applied Bayesian methods in their mortality projections for countries with limited data by appropriately modifying the original LC method. For more, see Antonio et al. (2015), Wiśniowski et al. (2015), Raftery and Chunn (2013) etc.

On top of fitting the Poisson gamma LC model (Delwarde et al., 2007) to deal with overdispersion, we also consider another mixed Poisson LC model, the Poisson log-normal LC model, as a possible alternative candidate model. This is because we would like to investigate which of these two distributions better describes the variability due to overdispersion, that clearly depends on the underlying shape of the tail distributions (unknown a priori). This is apparently a classic dilemma in the specification of error distributions within the generalised linear model framework. For instance, Firth (1988) investigated the efficiency of the modelling procedure under reciprocal misspecification of the multiplicative errors. Cox (1961) discussed the application of Neyman-Pearson maximum likelihood ratio test to compare between the two models (see also Cox, 1962 and Atkinson, 1970). For more on gamma versus log-normal errors, see Wiens (1999), Dick (2004), Alzaid and Sultan (2009), Cho et al. (2004) etc.

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. 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} denote 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 follows:

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 the mortality at each age to overall changes in the mortality on the log scale.

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

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 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 an appropriate autoregressive integrated moving average (ARIMA) 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)[1]. 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 ($\mathbb{E}[D_{xt}] = \text{Var}[D_{xt}] = e_{xt}\mu_{xt}$), 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) are assumed to 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:

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

[1]See http://www.mortality.org.

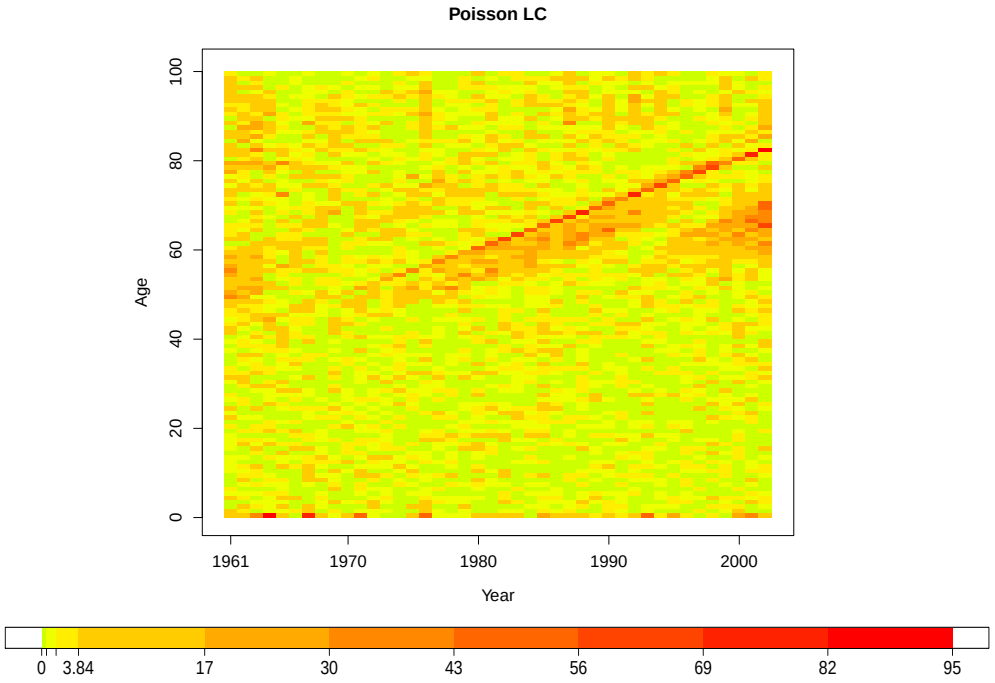

Figure 1: Heat map of r_{xt}^2 for the PLC model, accompanied by the corresponding colour code. Green/yellow rectangular cells indicate areas with good fit, while orage/red coloured cells indicate areas with significantly poor fit.

where d_{xt} is the observed number of deaths at age x in year t and $\hat{\mu}_{xt} = \exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t)$ is the maximum likelihood estimate (MLE) of the underlying mortality rate. A colour-coded heat map of r_{xt}^2 can then be constructed to visualise the lack of fit of the PLC model to our mortality data, as depicted in Figure 1.

Under the null hypothesis that the PLC model is the true underlying model (and some mild conditions), each r_{xt}^2 has an approximate chi-squared distribution with degrees of freedom one (χ_1^2) asymptotically. Ideally, we should expect only around 5% of the rectangular cells ($AT \times 0.05 = 210$) to have poor fit (defined as $r_{xt}^2 > 3.84$, where 3.84 is the 95^{th} precentile of χ_1^2). 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 the infants and ages above 40, suggesting model inadequancy in accounting for extra variations in the data. Additionally, we can also perform the Pearson's chi-squared overall goodness of fit test. In particular, the model deviance computed as the sum of r_{xt}^2,

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

has a value of 15378.73. Again, under the null hypothesis that this model is a good fit to the data, the r^2 should follow an approximate chi-squared distribution with degrees of freedom (df) given as $\text{df} = (A-1)(T-2) = 3960$ (see Renshaw and Haberman, 2005). Since the model deviance of 15378.73 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), this clearly suggests that the PLC model does not provide a satisfactory fit to the data. Note that the obvious red/orange diagonal lines displayed in Figure 1 correspond to possible cohort effects which we

do not attempt to address in this paper, but will do so in our future work. Setting aside the lack of fit of the PLC model as evidenced by the systematic pattern of orange/red cells in the heat map (mainly due to the uncaptured cohort effect), there is still a considerable amount of orange/red cells scattering around various regions in the heat map, particularly at older ages, indicating the presence of overdispersion.

In general, failure to account for overdispersion typically leads to under-smoothing, where the variance imposed by a model that ignores overdispersion forces the fitted values to adhere more closely to the data. Fundamentally, this is because the likelihood function of a model with smaller variance heavily penalizes fitted values that are distant from the observed values. This forces the fitted values to be undesirably close to the observed values which prevents an accurate description of the underlying process. Ignoring overdispersion also leads to over-optimistic forecast uncertainty because the extra source of uncertainty due to heterogeneity is effectively neglected. Appropriately accounting for overdispersion, on the other hand, provides a greater flexibility for the fitted values to adhere less to the observed data and allow for the possibility of greater smoothing, potentially resulting in an improved description of the underlying mortality trend. This prevents over-fitting and offers a better calibration of the unexplained variation, thereby producing a much more representative prediction interval for the associated mortality forecast (see Section 7.2 for more details).

3 Overdispersion Models

In this section, we present two models to account for overdispersion, both of which extend the PLC model in a rather straightforward manner. Both these models introduce a general dispersion parameter to relax the stringent assumption of a Poisson distribution.

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

The first model we introduce 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,

$$\mathbb{E}[D_{xt}] = \mathbb{E}_{\mu_{xt}}(\mathbb{E}_{D_{xt}}[D_{xt}|\mu_{xt}]) = e_{xt} \exp\left(\alpha_x + \beta_x \kappa_t + \frac{1}{2}\sigma_\mu^2\right), \tag{4}$$

and

$$\text{Var}[D_{xt}] \quad = \quad \mathbb{E}[D_{xt}] \times \left\{1 + \mathbb{E}[D_{xt}](\exp(\sigma_\mu^2) - 1)\right\} > \mathbb{E}[D_{xt}]. \tag{5}$$

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 the model specification. Note that equation (4) implies that the mean of D_{xt} under the PLNLC model is slightly different from the PLC model (due to the extra term $\sigma_\mu^2/2$). Some researchers (e.g. Dick, 2004) apply a correction by directly subtracting $\sigma_\mu^2/2$ from the rate model in equation (3). However, we chose to retain a similar model structure between the overdispersion models for easy interpretation and comparison, since the magnitude of the correction term is small relative to the overall magnitude of $\log \mu_{xt}$.

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 follows:

$$\begin{aligned} D_{xt}|\mu_{xt} &\overset{\text{ind}}{\sim} \text{Poisson}(e_{xt}\mu_{xt}) \\ \log \mu_{xt} &= \alpha_x + \beta_x \kappa_t + \log \nu_{xt} \\ \nu_{xt}|\phi &\overset{\text{ind}}{\sim} \text{Gamma}(\phi, \phi), \end{aligned} \tag{6}$$

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

$$\mathbb{E}[D_{xt}] = e_{xt} \exp(\alpha_x + \beta_x \kappa_t) \tag{7}$$

and

$$\text{Var}[D_{xt}] \quad = \quad \mathbb{E}[D_{xt}] \times \left[1 + \frac{\mathbb{E}[D_{xt}]}{\phi}\right] > \mathbb{E}[D_{xt}]. \tag{8}$$

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 NBLC model. That is,

$$D_{xt}|\alpha_x, \beta_x, \kappa_t, \phi \sim \text{ Neg-Bin } \left(\phi, \frac{\phi}{e_{xt} \exp(\alpha_x + \beta_x \kappa_t) + \phi}\right). \tag{9}$$

The likelihood function now consists of only 1 part:

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

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 a classical framework. Hence, one of our contributions in this paper is to fit this model within a Bayesian paradigm for an integrated modelling procedure.

4 Advantages of Bayesian Mortality Modelling/Forecasting

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. Computing parameter uncertainty within the frequentist framework typically necessitates bootstrapping (see for example Brouhns et al., 2005). 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 the context of probabilistic forecasts (see for example Berger, 2013). This provides more flexibility to users who are involved in risk-controlled 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. In addition, we also attempt to be indifferent in terms of prior specification under both overdispersion models to facilitate model comparison later on. Note that even though our prior specification differs considerably from that of Czado et al. (2005), this difference should not be consequential in terms of the parameter estimation, given the size of our mortality data.

4.1.1 Prior Distribution for α_x, β_x, σ_β^2, σ_μ^2, 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 independent normal priors on α_x, i.e.

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

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

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

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 $\beta_1 = 1 - \beta_2 - \ldots - \beta_A$. This corresponds to transforming the constraint into a proper point mass prior on the unidentified β parameters, which automatically yields proper posterior inference, as stated by Gelfand and Sahu (1999). They also note the issue of a slower rate of convegence of this constraint-handling approach (due to the correlations induced), which we propose to solve using the blocking strategy (see Section 5). 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

For reasons 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. On 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,

$$\begin{cases} \kappa_t - \eta_t = \rho(\kappa_{t-1} - \eta_{t-1}) + \epsilon_t, & \text{for } t = 2, 3, \ldots, T, \\ \kappa_1 = \eta_1 + \epsilon_1, \end{cases} \tag{10}$$

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 (10) includes random walk with drift as a special case when $\rho = 1$, provided that it is not ruled out a priori. 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 log mortality rates in the base year. Fixing $\kappa_1 = 0$ also has the effect of setting the first year as the baseline year, where values of κ_t for the remaining years are estimated relative to the value of κ_1. In other words, κ_t should be interpreted as the parameter that represents the overall mortality trend with respect to the baseline year. Elsewhere, the impact of this is purely computational, the posterior distribution of $\log \mu_{xt}$ will not be affected.

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

$$\begin{cases} \boldsymbol{\kappa}_{-1} \sim N(\boldsymbol{Y}_{-1}\boldsymbol{\psi} - \rho \boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi}, \sigma_\kappa^2 \boldsymbol{Q}^{-1}) \\ \kappa_1 = 0 \end{cases}, \tag{11}$$

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:

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

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 for the PLNLC model can be easily implemented by iterating through a series of Gibbs steps, together with some MH steps for the remaining $\log \mu_{xt}$. We describe in detail 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,

$$\log \mu_{xt}^* \sim N(\log \mu_{xt}^{i-1}, \sigma_{\mu_{xt}}^2),$$

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,

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

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 $\sigma_{\mu_{xt}}^2$ can be numerically determined and are depicted in Figure 2.

Interestingly, $\sigma_{\mu_{xt}}^2$ 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 can be verified by referring to Appendix C. 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 $\rho, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}$ 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.

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 in Section 5.2.1), $\sigma_{\alpha_x}^2$, $\sigma_{\beta_x}^2$, and $\sigma_{\kappa_t}^2$ are illustrated in Figure 3.

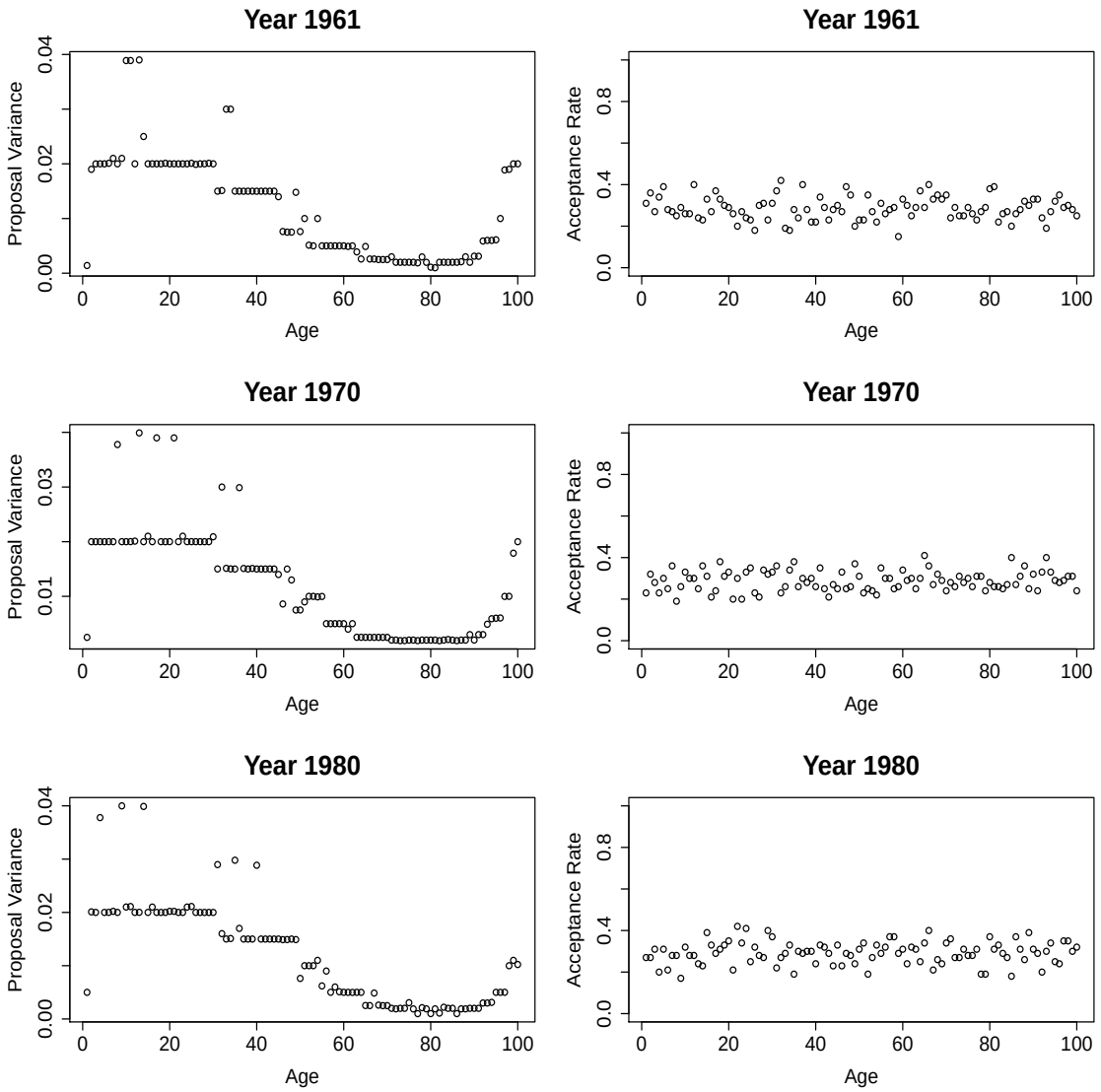

Figure 2: Plots of proposal variances, $\sigma^2_{\mu_{xt}}$ (left panels) and the corresponding acceptance rates of μ_{xt} (right panels) for years 1961, 1970 and 1980 under the PLNLC model.

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 by $\sigma^2_{\beta_x}$ is less sensitive to age than those of the $\sigma^2_{\mu_{xt}}$ as well as $\sigma^2_{\alpha_x}$, 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 ages. To verify that the marginal posterior variances of these parameters do exhibit similar shapes as the chosen proposal variances, please refer to Appendix C.

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

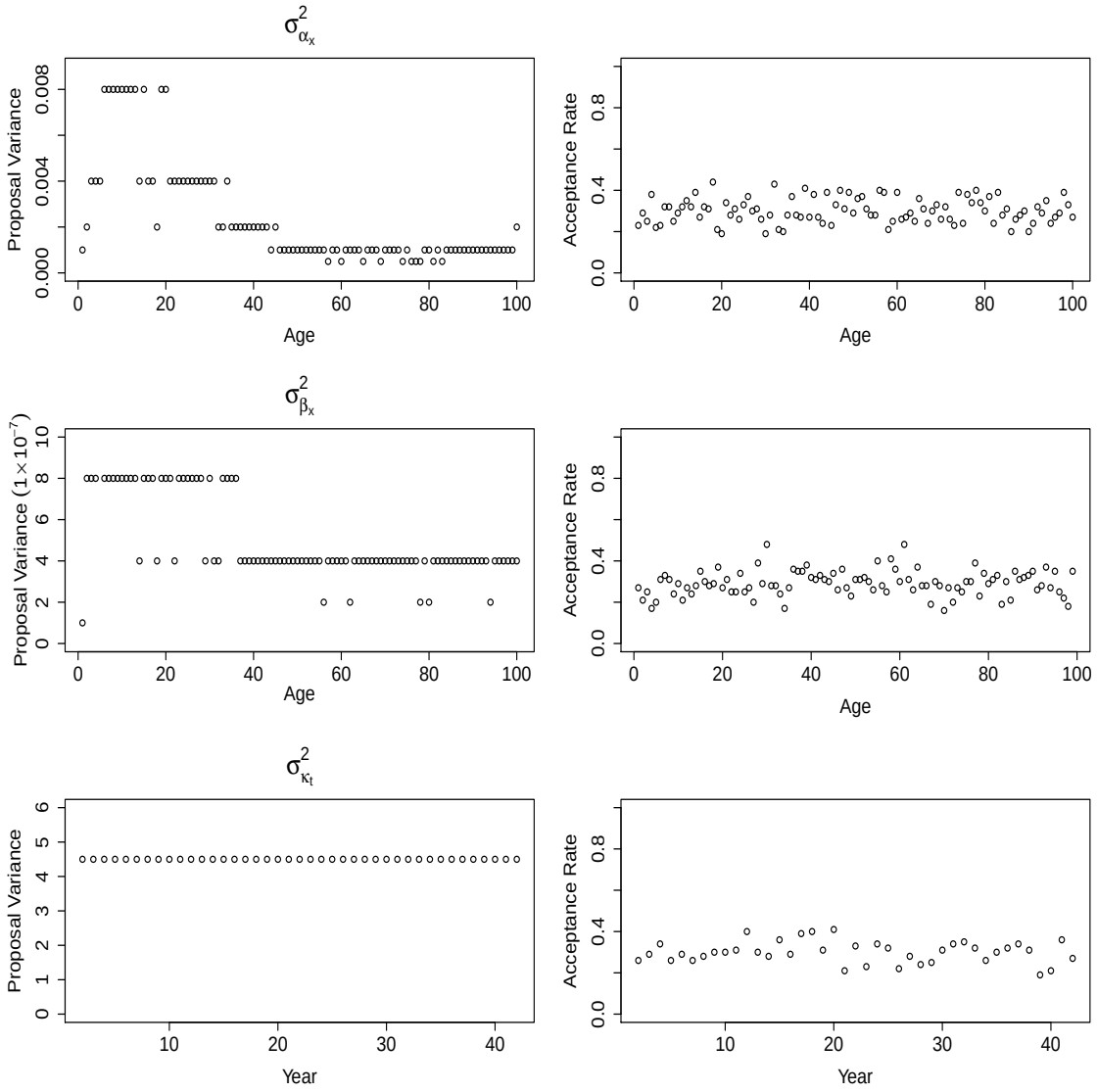

Figure 3: Plots of the proposal variances (top panels), $\sigma^2_{\alpha_x}$, $\sigma^2_{\beta_x}$, $\sigma^2_{\kappa_t}$, and their corresponding acceptance rates (bottom panels) for the NBLC model.

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

$$\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d} \sim \text{Gamma}\left(d_{xt} + \phi, e_{xt} + \frac{\phi}{\exp(\alpha_x + \beta_x \kappa_t)}\right). \tag{12}$$

Therefore, the posterior samples of μ_{xt} can be generated by simulating from the expression in (12), 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

$$\begin{aligned} f(\log \mu_{x\,T+1}|\boldsymbol{d}) \;=\; & \int f(\log \mu_{x\,T+1}|\alpha_x, \beta_x, \kappa_{T+1}, \sigma_\mu^2) f(\alpha_x, \beta_x, \sigma_\mu^2|\boldsymbol{d}) f(\kappa_{T+1}|\kappa_T, \rho, \sigma_\kappa^2, \boldsymbol{\psi}) \\ & \times f(\kappa_T, \rho, \sigma_\kappa^2, \boldsymbol{\psi}|\boldsymbol{d}) \mathrm{d}\alpha_x \mathrm{d}\beta_x \mathrm{d}\kappa_T \mathrm{d}\kappa_{T+1} \mathrm{d}\rho \mathrm{d}\sigma_\kappa^2 \mathrm{d}\boldsymbol{\psi} \mathrm{d}\sigma_\mu^2, \end{aligned} \tag{13}$$

where $f(\alpha_x, \beta_x, \sigma_\mu^2|\boldsymbol{d})$ and $f(\kappa_T, \rho, \sigma_\kappa^2, \boldsymbol{\psi}|\boldsymbol{d})$ are the joint posterior distributions. Hence, posterior uncertainties, with respect to the model likelihood, prior distributions and the projection model, are fully integrated in the posterior predictive distribution. The density in (13) is analytically intractable, but can be empirically estimated using our MCMC samples. Essentially, generation of the posterior samples of $\log \mu_{x\,T+1}$ proceeds in two steps:

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

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

   where joint posterior samples of $(\kappa_T, \rho, \sigma_\kappa^2, \psi_1, \psi_2)$ 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-year 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 visualise 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-year 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 (unobservable), or the crude rates (observable). We chose to present the projected crude mortality rates in the result section purely because plots of observable quantities provide a more sensible visualisation in terms of validating the models against the observed crude death rates (see Section 7.3). Indeed, it should also 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 Initialization and Convergence Diagnostics

For initialization of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{-1}$ and $\boldsymbol{\kappa}_{-1}$, we use the MLEs obtained using Goodman's method (see Renshaw and Haberman, 2005). On the other hand, the initial values of σ_κ^2 and ρ are obtained by fitting an AR(1) with linear drift model on $\boldsymbol{\kappa}$ (using the 'arima' function within the 'forecast' package in R), while σ_β^2 is initialised by the empirical variance of the MLEs of $\boldsymbol{\beta}$. Finally, $\boldsymbol{\psi}$ is initialised as $(0,0)^\top$, while the overdispersion parameters, σ_μ^2 and ϕ, are initialised by 0.01 and 100 respectively. Under the PLNLC model, the latent parameters, μ_{xt}, are initialized using the empirical death rates, d_{xt}/e_{xt}. Note that the initialization is proposed based on values close to the MLEs to possibly speed up convergence, but should not be impactful in terms of the parameter estimation. Ideally, multiple chains with different initializations should be run to ascertain the convergence of the chains. Specifically, Gelman and Rubin (1992) proposed the use of multiple sequences with starting values initialised from an overdispersed distribution, and developed a quantity as a function of within and across chains variance to assess convergence. Instead, we assume here that a burn-in phase of 10000 iterations is sufficiently long to mitigate the effect of initialization. In addition, we applied 100^{th} posterior sample thinning (collecting one realization every 100 iterations) for each of the parameters to reduce the autocorrelations of these series. After discarding the burn-in iterations and applying thinning, we obtain a sample of size 10000 for each of the paramaters under the NBLC model and a sample of size 100000 for those under the PLNLC model (a larger sample size is required to learn about the posterior under the PLNLC model due to high-dimensionality).

Before making any inferential comparisons, trace plots and auto-correlation plots (see for example Lunn et al., 2013) can be used as diagnostic tools for detecting anomalies in the MCMC generated posterior samples. By referring to Appendix D, the trace plots of some of the randomly selected parameters emerge as if convergence has been attained, with proper mixing and no apparent anomaly. The sample auto-correlations also appear to decay fairly quickly after applying thinning, except perhaps κ_t, which are relatively more correlated. In summary, the MCMC generated posterior samples seem to be well-behaved and, thus, are ready to be used to perform subsequent computations for accurate inferences to be drawn.

7 Numerical Results

In this section, we compare our proposed models with the Bayesian PLC model (i.e. the PLNLC model or NBLC model without the overdispersion component, ν_{xt}) by Czado et al. (2005) to highlight the importance of accounting for overdispersion. The data used for this purpose are as described in Section 2.1. The Bayesian PLC model is fitted using Czado's methodology, except we adopt the same prior specification as in Section 4.1 (for all the parameters and hyperparameters involved) to facilitate model comparison later on. We also provide a comparison of our

proposed models with each other.

7.1 Estimated Parameters

Figures 4 depicts the fitted values (posterior medians) of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$, accompanied by the associated 95% credible intervals (computed from the sample quantiles) under the Bayesian PLC and NBLC models. Also included is the projection of κ_t, 25 years into the future (until year 2027), for illustrative purposes. 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 visualisation. According to Figure 4, the fitted values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under these models are rather similar (because the same vague priors are specified across the models), with the overdispersion models producing slightly wider credible intervals in general. This is the general feature of a model which accounts for overdispersion, where the responses (D_{xt}) are allowed to have more variabilities due to the extra flexibility offered by the model likelihood, permitting the parameters to be more volatile, and hence, the wider credible intervals. Additionally, the width of the credible intervals also appears to be noticeably different as age increases.

The main difference arises from the parameter $\boldsymbol{\kappa}$, where 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 larger mortality improvement (indicated by more negative values of the projected κ_t), the corresponding prediction intervals for the projected κ_t are also substantially wider. This is perhaps a little interesting considering that the same AR(1) prior is imposed on κ_t under all approaches. An intuitive explanation for this is that the dispersion parameter provides more flexibility for the model to describe the variabilities present in the data (where the model likelihood penalizes less on fitted values that are distant from the observed data due to the larger variance postulated), thereby allowing more priority to be put on fitting the AR(1) prior during the Bayesian estimation, hence the smoother fitted values. On the other hand, with less variabilities imposed for D_{xt} under the Bayesian PLC model, their fitted values are restricted to stay close to the observed values, implying that less smoothing is applied. The exact reason behind this finding will be further explored when the marginal posterior distribution of ρ is examined in the next paragraph.

Kernel estimates of the marginal posterior density of the rest of the parameters, derived from the posterior samples, are presented in Figure 5. The kernel densities of σ_β^2 are almost identical. The most apparent discrepancies occur at the marginal posterior of σ_κ^2 and ρ. Specifically, the density of σ_κ^2 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 has the same characteristics as a two-component mixture distribution under all models, consisting of a stationary AR(1) component ($\rho < 1$) and a non-stationary component close to a random walk ($\rho = 1$). 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 0.85 and 1. This indicates that the projection model fitted on κ_t, in some sense, resembles a mixture of a stationary AR(1) model and a random walk with drift model. In addition, the allocation of proportion is also different, with the overdispersion models allocating a higher proportion for the peak at around $\rho = 1$ than the Bayesian PLC model.

The marginal posterior of ρ enables us to justify our earlier findings on κ_t. Firstly, as the fitted ρ increases towards larger values for the overdispersion models, the fitted time series model imposes a stronger smoothing on κ_t. 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

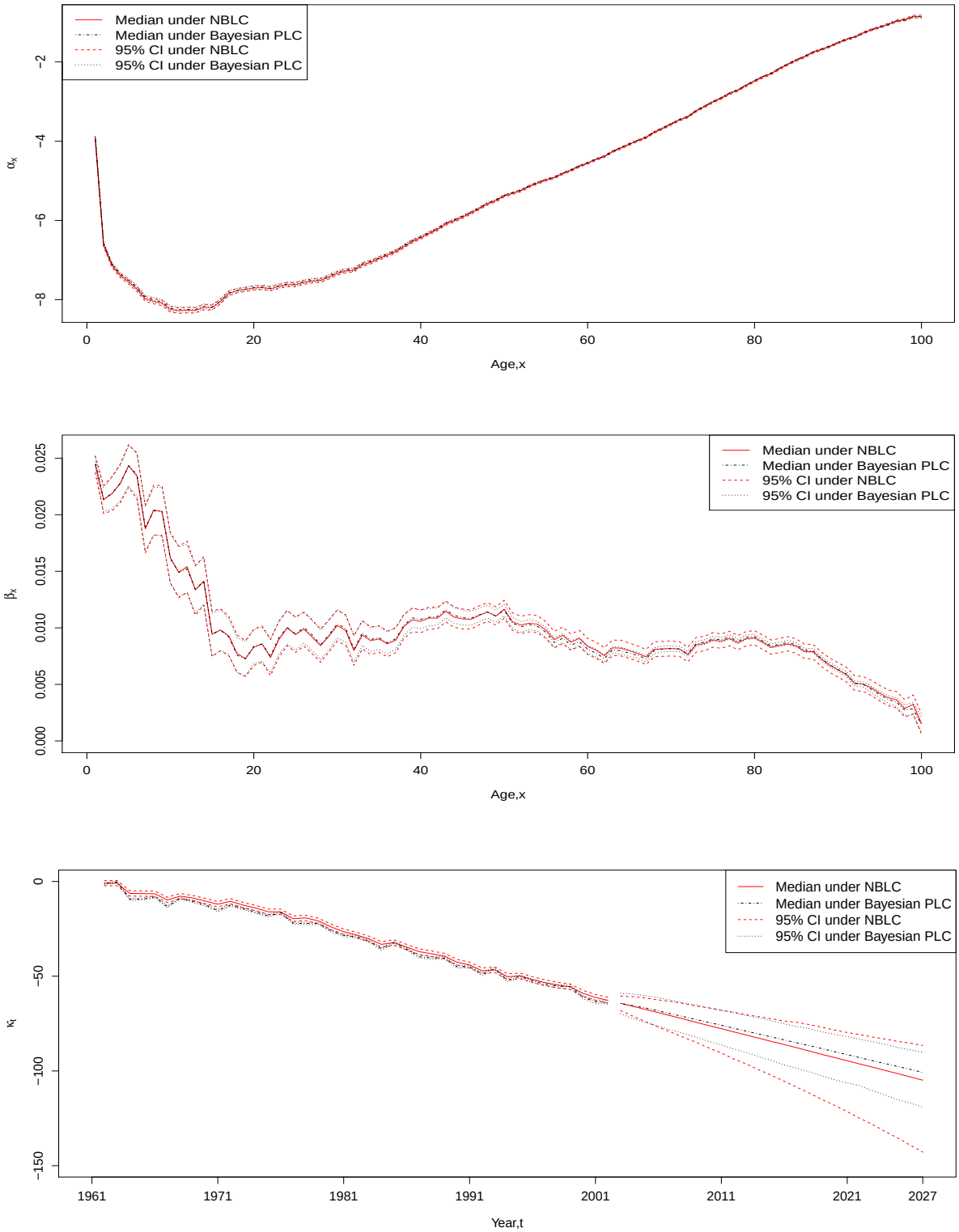

Figure 4: Plots of the estimated α_x, β_x and κ_t with their 95% credible intervals under the Bayesian PLC model and the NBLC model. The 25-years ahead projection of κ_t, accompanied by the corresponding 95% intervals, is also presented.

these models because their projection model is largely dominated by the values of $\rho \approx 1$ that almost correspond to a random walk model ($\rho = 1$), 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_κ^2, 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.

There are also slight differences for ψ_1 and ψ_2 between the models, with the overdispersion models yielding heavier tails in both cases. Fundamentally, this is directly related to the mixture posterior distribution of ρ, where when a random walk model ($\rho = 1$) is used, the model on κ_t reduces to

$$\kappa_t = \kappa_{t-1} + \psi_2 + \epsilon_t,$$

where ψ_2 is now the drift term, and ψ_1 becomes a redundant parameter that is non-identifiable under the model, hence, the large uncertainty. To be more specific, the conditional posterior distribution of ψ_1 reduces analytically to $N(0, 1000)$ when $\rho = 1$, which does not depend on the data and other parameters. This implies that its marginal posterior distribution is indeed $N(0, 1000)$, which is exactly the same as its prior distribution. This happens because ψ_1 and ψ_2 are assumed to be independent a priori, so nothing is learned about ψ_1 given that it is a non-identifiable parameter as far as the likelihood is concerned. In other words, the posterior distribution of ψ_1 also behaves like a mixture distribution, formed by mixing its prior distribution (which is relatively vague) and the posterior distribution when $\rho < 1$. On the other hand, all the uncertainties regarding the drift of κ_t are now absorbed by ψ_2 since it is the only remaining drift parameter when ρ is very close to 1, hence a heavier tail for the marginal posterior distribution of ψ_2. Therefore, with the overdispersion models highly favouring values of ρ that are close to 1 (corresponding to a random walk model), the much heavier-tailed posterior distributions for ψ_1 and ψ_2 are justified.

Regarding the overdispersion parameters, there is a substantial amount of Bayesian learning for both σ_μ^2 and $1/\phi$, as indicated by the obvious shifts of their posterior distributions (proper unimodal distributions with 95% quantiles of around $[0.00136, 0.00158]$) from the arbitrarily diffuse prior distributions (which have close to negligible densities for the region of values presented in Figure 5). Recall also 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 samples generated, the posterior median of ϕ is approximately 681 ($1/\phi = 0.001468$), implying that the level of overdispersion is non-negligible. To further strengthen this argument, we can assess the practical significance of the magnitude of this value of ϕ estimated using the expression for the variance of D_{xt} under the NBLC model, given in Equation (8). The term $\frac{\mathbb{E}[D_{xt}]}{\phi}$ can be interpreted as the relative increase in the variance of D_{xt} with respect to its mean, which measures the extent of overdispersion in the mortality data. For the purpose of a simple illustration of the level of overdispersion implied, a crude calculation can be carried out by replacing $\mathbb{E}[D_{xt}]$ with observed deaths. For example, using the mean observed number of deaths and the median of ϕ, we obtain a value of $2846.945/681 \approx 4$, implying that there is a roughly four times increase in the variance of D_{xt} (relative to the mean) on average under the NBLC model. More importantly, for the age and time with the largest observed number of deaths, the relative increase is $12399/681 \approx 18$, which is massive. Both these examples indicate that the extent of overdispersion implied by the value of ϕ fitted is rather substantial, and hence, should not be ignored. On the other hand, for the PLNLC model, the Bayesian PLC model can be retrieved when $\sigma_\mu^2 = 0$. Since the posterior median of σ_μ^2 is around 0.001465, this indicates again the presence of non-negligible overdispersion. Similar calculation as above can be undertaken for an interpretation of the magnitude of the σ_μ^2 estimated. According to Equation (5), $\mathbb{E}[D_{xt}](\exp(\sigma_\mu^2) - 1)$ represents the relative increase

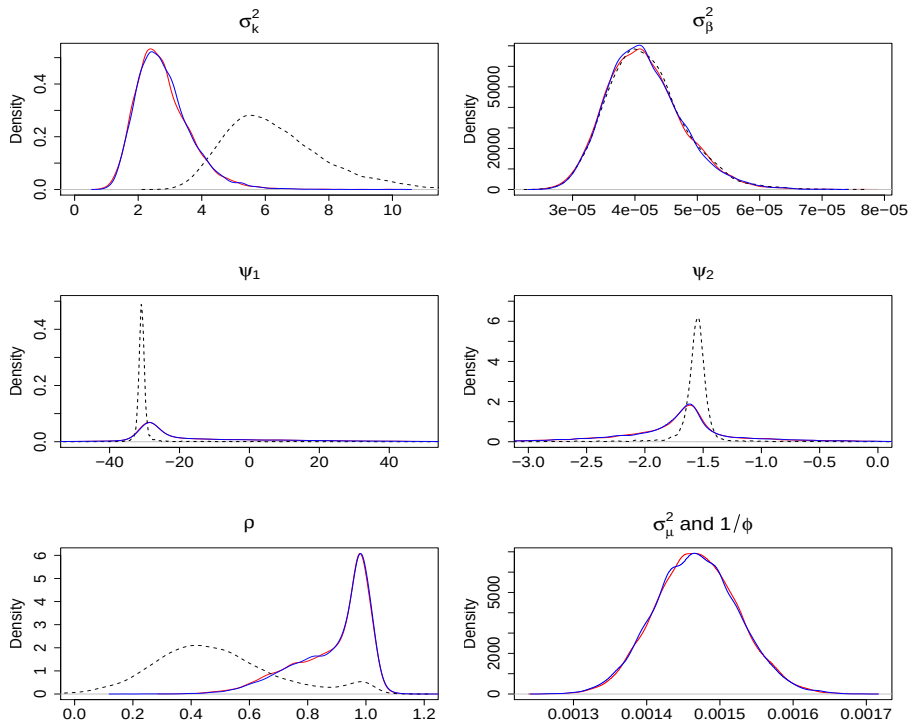

Figure 5: Kernel density plots of σ_κ^2, σ_β^2, ψ_1, ψ_2, ρ, and ϕ under the Bayesian PLC (black dotted), PLNLC (blue solid) and NBLC model (red solid).

in the variance of D_{xt} over its mean. It is straightforward to see that the variance of D_{xt} can easily increase by several folds for this value of σ_μ^2 ($= 0.001465$). For instance, replacing $\mathbb{E}[D_{xt}]$ with $\max[d_{xt}] = 12399$ yields a relative increase of $12399 \times (\exp(0.001465) - 1) \approx 18$, suggesting the practical significance of accounting for overdispersion.

7.2 Fitted Crude Mortality Rates

Figure 6 shows the fitted and projected log mortality rates for ages 30, 55 and 80 plotted against time, 11 years into the future. According to the figure, there are considerable differences between the Bayesian PLC model and the overdispersion models in terms of the fitted rates. Firstly, the median fitted rates for the overdispersion models are slightly smoother than the Bayesian PLC model across the ages. More crucially, the credible intervals of fitted rates for the overdispersion models are substantially wider than that of the Bayesian PLC model. These are consistent with our conjecture before on the failure to account for overdispersion, where the fitted values are generally under-smoothed due to the model's rigid structure as evidenced by the zig-zag patterns of the medians and are accompanied by over-optimistic credible intervals due to the lower variance by construction. In other words, we witness here that ignoring overdispersion has the tendency to force the fitted values to adhere more closely to the data due to the smaller variance imposed by the model (over-fitting), causing under-smoothing and narrower intervals. Both of these properties, when projected into the future, are detrimental to the resulting mortality forecasts due to the poor description of underlying trends and variabilities. On the contrary, the greater flexibility of the overdispersion models allow the fitted values to adhere less to the data (encouraging more smoothing), where the residuals due to the unexplained variations are then absorbed into the dispersion parameters, resulting in wider intervals in general. The trade-off

between adherence to the data and smoothness clearly favours the overdispersion models here, where their credible intervals provide reasonably good coverages of the observed rates across the ages, with most points lying within the intervals, while the credible intervals for the Bayesian PLC model appear to be overly narrow, with a large number of points still lying outside the intervals (particularly for age 55).

7.3 Projected Crude Mortality Rates and Out-of-Sample Validation

According to Figure 6, the overdispersion models clearly forecast a larger improvement in the mortality rates, and also produce considerably wider prediction intervals in all cases (and for the rest of the ages). 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). This can also be explained by the time series model fitted on κ_t, where the overdispersion models favour a random walk with drift model (which is known to produce wide prediction intervals). Hence, the inclusion of dispersion parameters provides a more sensible improvement in the rates as well as better calibrated probabilistic intervals in terms of the projection.

Then, we validate the candidate models against the holdout data to assess their predictive abilities. First, this is undertaken based on a disaggregate mortality quantity, the projected age-specific crude mortality rates as shown in Figure 6, derived using the projected underlying mortality rates and the holdout exposure data (see Section 5.5). The performances of the models in terms of their coverages vary across ages. In particular, the median projections of mortality improvement and the associated 95% prediction intervals for age 30 are rather similar across all three candidate models, with good predictive properties (appropriately projected past trend with good coverages for the prediction intervals) when assessed against the holdout samples. On the contrary, the projected mortality improvements for age 55 are somewhat pessimistic, especially for the Bayesian PLC model. In particular, the coverage of the 95% prediction intervals for the Bayesian PLC model is rather low due to the overly narrow intervals, while the overdispersion models yield prediction intervals that are wide enough to cover most of the holdout rates. For age 80 (where it is rich in death data), the coverages of all of the models are satisfactory, with the overdispersion models slightly outperforming the Bayesian PLC model by having smaller biases and better coverages. Overall, the validation process using the disaggregate mortality quantity indicates that the overdispersion models outperform the Bayesian PLC model in terms of predictive ability for this particular dataset.

It is perhaps more useful to perform the validation based on an aggregate mortality quantity (instead of focusing on a specific age), the life expectancy at birth, derived from the projected crude mortality rates (where the holdout central exposed to risks are used). As illustrated in Figure 7, 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 for this particular dataset. 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.

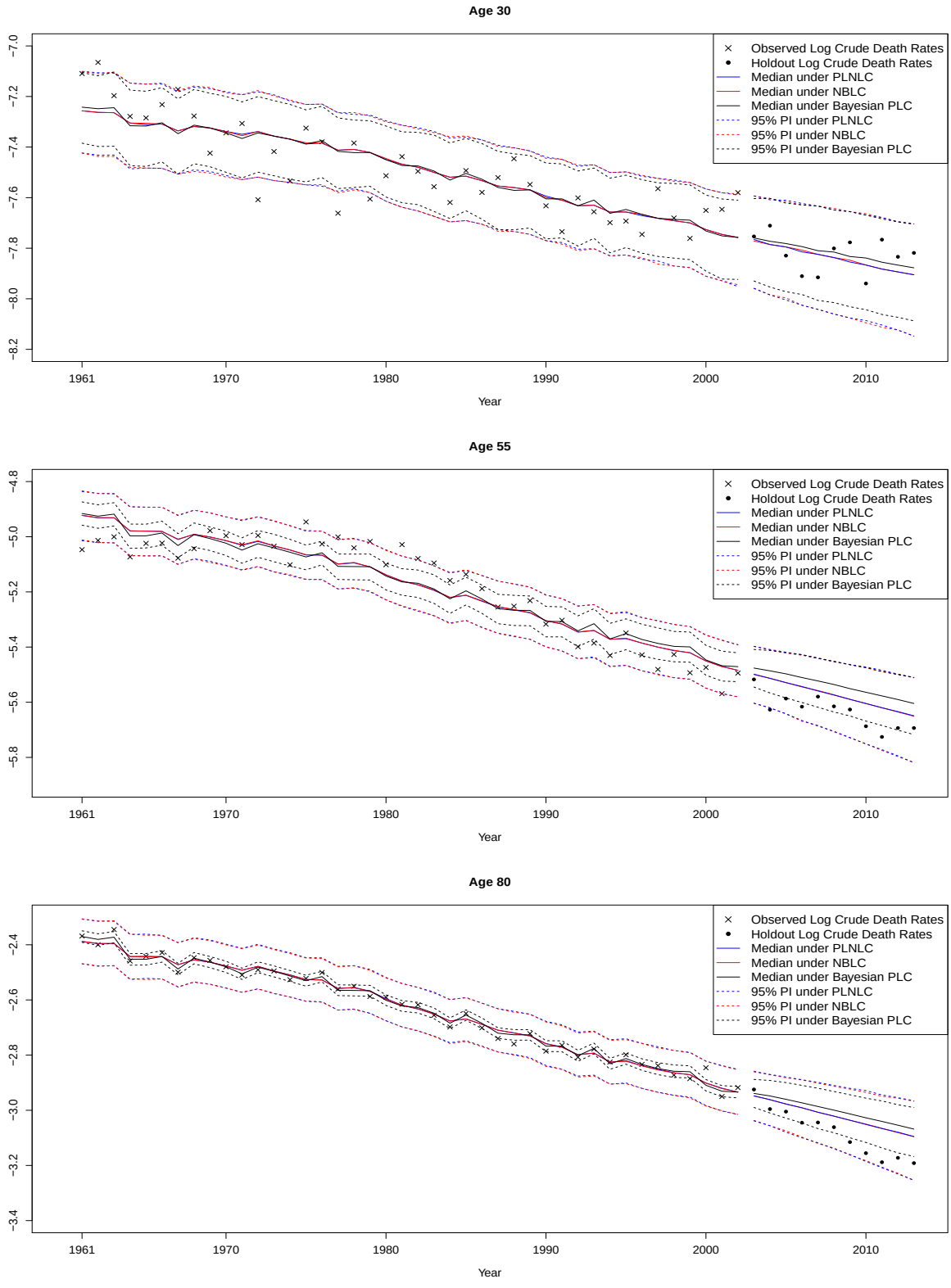

Figure 6: Plots of the observed log crude death rates, $\log(d_{xt}/e_{xt})$, fitted log crude death rates and the associated 11-years ahead projection of the crude log death rates for age 30 (upper panel), age 55 (middle panel) and 80 (lower panel) under the Bayesian PLC model and the overdispersion models, accompanied by 95% credible intervals.

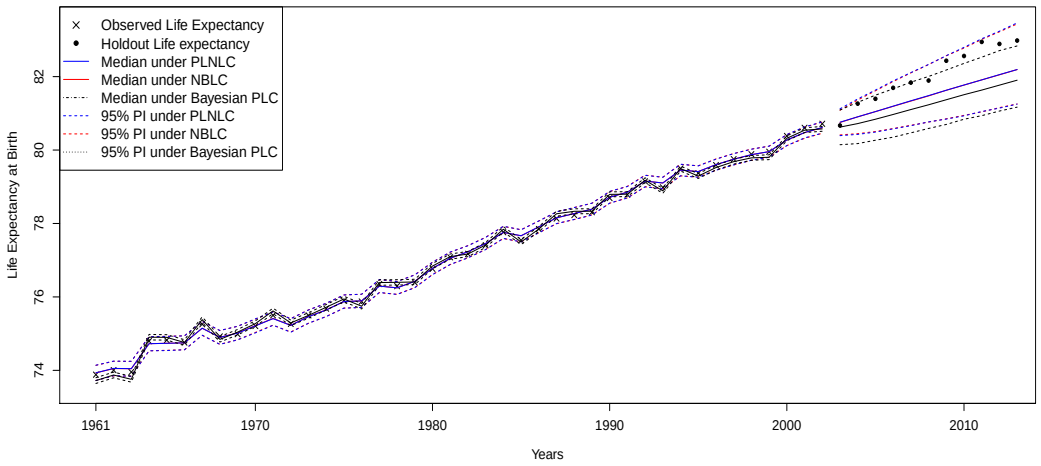

Figure 7: Plots of the observed, fitted life expectancy at birth and the associated 11-years ahead forecast under the Bayesian PLC and the overdispersion models, accompanied by the 95% prediction intervals.

7.4 Model Assessment

We can similarly construct a heat map of the squared Pearson residuals for the overdispersion models. Expressions of the squared Pearson residuals for the PLNLC and NBLC models are given respectively as

$$\frac{[d_{xt} - e_{xt}\exp(\alpha_x + \beta_x\kappa_t + \sigma_\mu^2/2)]^2}{e_{xt}\exp(\alpha_x + \beta_x\kappa_t + \sigma_\mu^2/2) + e_{xt}^2[\exp(\sigma_\mu^2) - 1]\exp(2(\alpha_x + \beta_x\kappa_t) + \sigma_\mu^2)},$$

and

$$\frac{[d_{xt} - e_{xt}\exp(\alpha_x + \beta_x\kappa_t)]^2}{e_{xt}\exp(\alpha_x + \beta_x\kappa_t)\left[1 + e_{xt}\frac{\exp(\alpha_x + \beta_x\kappa_t)}{\phi}\right]},$$

where now the posterior mean of the parameters α_x, β_x, κ_t, σ_μ^2 and ϕ are substituted into the expression for an estimate. As illustrated in Figure 8, 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 squared Pearson residuals 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 un-captured cohort effects, emerged as yellow/orange diagonal lines in Figure 8. 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, 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

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

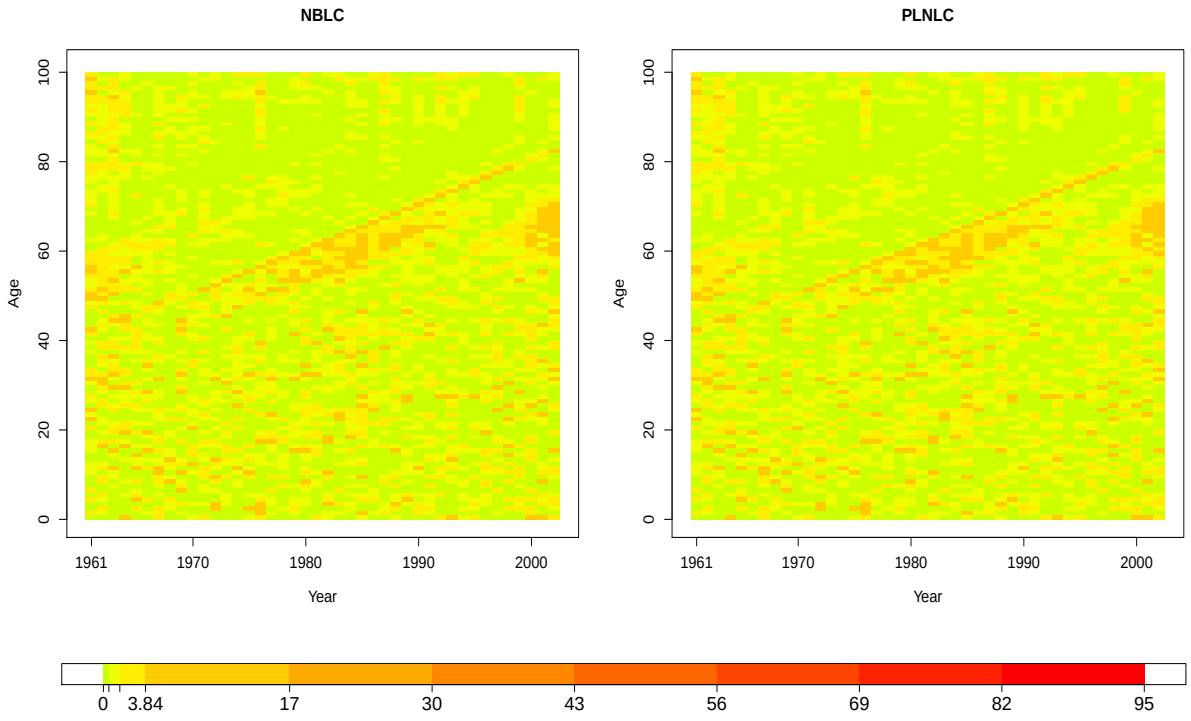

Figure 8: Heat map of squared Pearson residuals, r_{xt}^2, under the PLNLC model (left panel) and the NBLC model (right panel), accompanied by the corresponding colour code.

from the posterior samples of $\boldsymbol{\theta}_M$ under each model, where M is the model indicator, $f_M(\boldsymbol{d}^{\text{rep}}|\boldsymbol{\theta}_M)$ is the likelihood function, $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$ is the posterior of $\boldsymbol{\theta}_M$. Next, we define our test quantity as

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

which is the usual χ^2 discrepancy (that 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 E. 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 9). The sum of squared Pearson residuals (or equivalently $T(\boldsymbol{d}, \bar{\boldsymbol{\theta}}_M)$, where $\bar{\boldsymbol{\theta}}_M$ is the posterior mean under model M) for each model is displayed in Figure 9 to highlight the magnitude of its discrepancy with the $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$. It can be seen that the sum of squared Pearson residuals for the overdispersion models lies somewhere in the middle of the histograms; while that of 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 PLNLC, NBLC and Bayesian PLC models 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 inadequacy.

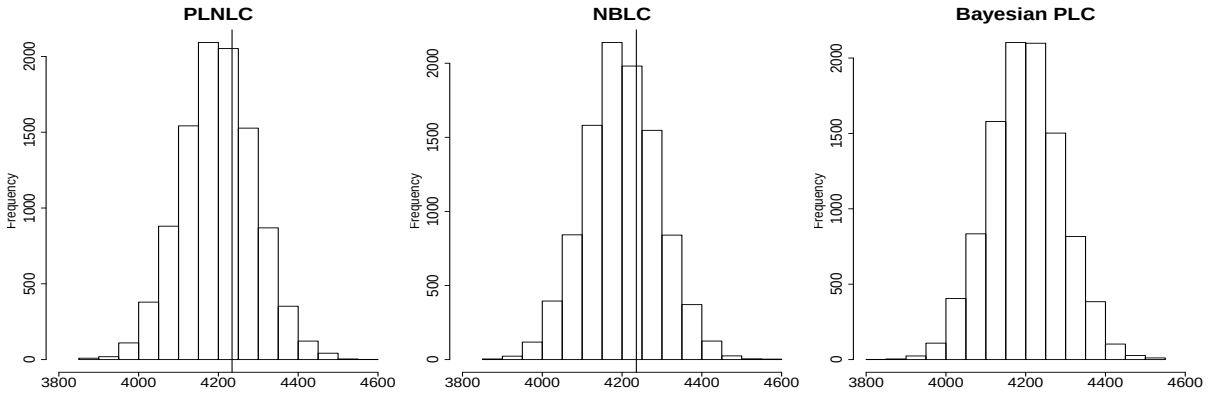

Figure 9: Histograms of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta_M})$ for the PLNLC, NBLC, and Bayesian PLC model, with their corresponding sum of squared Pearson residuals, r^2 included as the vertical solid lines.

7.5 Bayesian Model Determination

Most of the previous 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. It is evident from Appendix F (Figure F.1) 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. Furthermore, the QQ plot of σ_μ^2 against $1/\phi$ is remarkably close to the reference line as depicted in Appendix F (Figure F.2), 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. Therefore, Bayesian model comparison is carried out to ascertain this observation.

Formal Bayesian model comparison proceeds through the computation of posterior model probabilities (e.g. Kass and Raftery, 1995). 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 (ML) 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 MLs, expressed as

$$f_M(\boldsymbol{d}) \quad = \quad \int f_M(\boldsymbol{d}|\boldsymbol{\theta}_M) f_M(\boldsymbol{\theta}_M) \mathrm{d}\boldsymbol{\theta}_M, \tag{14}$$

which is effectively the normalising constant of the joint posterior distribution. For the computation of MLs, we use bridge sampling (Meng and Wong, 1996), which is an efficient method of approximating ratio of normalising constants. In the context of approximating ML, we construct the bridge sampling algorithm such that the second normalising constant is known. In particular, the asymptotically optimal iterative formula for bridge sampling suggested by Meng

and Wong (1996) is used,

$$\hat{f}_M^{(t+1)}(\boldsymbol{d}) = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{\tilde{l}_i}{N_1\tilde{l}_i+N_2\hat{f}_M^{(t)}(\boldsymbol{d})}\right]}{\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\frac{1}{N_1 l_i+N_2\hat{f}_M^{(t)}(\boldsymbol{d})}\right]}, \tag{15}$$

where $\hat{f}_M^{(t)}(\boldsymbol{d})$ is the t^{th} iteration of the estimator, $l_i = \frac{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i)f_M(\boldsymbol{\theta}_M^i)}{g_M(\boldsymbol{\theta}_M^i)}$, $\tilde{l}_i = \frac{f_M(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}_M^i)f_M(\tilde{\boldsymbol{\theta}}_M^i)}{g_M(\tilde{\boldsymbol{\theta}}_M^i)}$, $\{\boldsymbol{\theta}_M^i\}_{i=1}^{N_1}$ is a sample of size N_1 from the posterior distribution with density $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$, and $\{\tilde{\boldsymbol{\theta}}_M^i\}_{i=1}^{N_2}$ is a sample of size N_2 from an arbitrary distribution (normalised) with density $g_M()$. Starting with an initial guess $\hat{f}_M^{(0)}(\boldsymbol{d})$, the bridge sampling estimate, $\hat{f}_M(\boldsymbol{d})$, of the ML can be obtained by iterating (15) until convergence. The choice of the density $g_M()$ is entirely arbitrary, but we set it to be a normal distribution (of the same dimensionality) with its first two moments chosen to match those from the posterior distribution under each model M for higher efficiency. Also, we set $N_1 = N_2$ equal to the respective sample size of the posterior under each model as given in Section 6 for simplicity. Thus, the MLs of each model approximated using bridge sampling are presented in Table 1.

Table 1: The marginal likelihoods (on 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, indicating the superiority of the overdispersion models in terms of goodness of fit. Recall also that the exploratory analyses using QQ plots 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 them. 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 distribution of 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.

Our results here agree with Firth (1988) to some extent, where the misspecification due to gamma or log-normal error is non-impactful in terms of the goodness of fit, other criteria should be used to discriminate between the two. Hence, even though both the overdispersion models provide similar fit qualitatively for this dataset, 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 models, both of which extended the original PLC model by introducing a single dispersion parameter. Another main contributions of this paper is fitting these overdispersion models within a Bayesian paradigm, which offers a natural framework for integrating over various sources of uncertainty in a coherent manner. 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 produce smoother estimates of mortality rates and forecast larger mortality improvements 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. Choosing between the two overdispersion models is essentially the classic discrimination problem between the log-normal and gamma multiplicative error distributions. We illustrated that they provide rather similar qualitative fit. Formal Bayesian model comparison using posterior model probabilities also showed that they are very similar. Hence, the NBLC model is to be recommended 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. The inclusion of cohort effects further improves the calibration between data signals and errors, which together with the incorporation of overdispersion, is expected to yield more accurate mortality projections. Until then, the dispersion parameters do not represent heterogeneity entirely in the sense that it is contaminated with the cohort effect.

Acknowledgements

The authors greatly appreciate the comments from any anonymous referee involved, who helped improving the quality of the paper.

Appendix: Supplementary Material

Supplementary material related to this article can be found online at [to be included if accepted].

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 Association, 87*, 673–674.

Alzaid, A. and K. S. Sultan (2009). Discriminating between gamma and lognormal distributions with applications. *Journal of King Saud University - Science* **21**(2), 99–108.

Antonio, K., A. Bardoutsos, and W. Ouburg (2015). Bayesian Poisson log-bilinear models for mortality projections with multiple populations. *European Actuarial Journal,* **5**(2), 245–281.

Atkinson, A. C. (1970). A Method for Discriminating Between Models. *Journal of the Royal Statistical Society. Series B (Methodological),* **32**(3), 323–353.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis.* Springer Science & Business Media.

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

Brouhns, N., M. Denuit, and I. V. Keilegom (2005). Bootstrapping the poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal,* **2005**(3), 212–224.

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 (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk & Insurance,* **73**(4), 687–718.

Cho, H. K., K. P. Bowman, and G. R. North (2004). A Comparison of Gamma and Lognormal Distributions for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission. *Journal of Applied Meteorology,* **43**(11), 1586–1597.

Cox, D. R. (1961). Tests of Separate Families of Hypotheses. *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability,* Berkeley, University of California Press, 105–123.

Cox, D. R. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological),* **24**(2), 406–424.

Czado, C., A. Delwarde, and M. Denuit (2005). Bayesian Poisson Log-Bilinear Mortality Projections. *Insurance: Mathematics and Economics,* **36**(3), 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**(5), 381–401.

Dick, E. J. (2004). Beyond 'lognormal versus gamma': discrimination among error distributions for generalized linear models. *Fisheries Research,* **70**(2-3), 351–366.

Firth, D. (1988). Multiplicative Errors:Log-normal or Gamma? *Journal of the Royal Statistical Society. Series B (Methodological),* **50**(2), 266–268.

Gelfand, A. E. and S. K. Sahu (1999). Identifiability, Improper Priors and Gibbs Sampling for Generalized Linear Models. *Journal of the American Statistical Association,* **94**(445), 515–533.

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

Gelman, A. and D. B. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science,* **7**(4), 457–511.

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.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association,* **90**(430), 773–795.

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, J. (2014). An application of MCMC simulation in mortality projection for populations with limited data. *Demography,* **30**(1), 1–48.

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.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis.* Chapman and Hall/CRC.

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.

Pedroza, C. (2006). A Bayesian Forecasting Model: Predicting U.S. Male Mortality. *Biostatistics,* **7**(4), 530–550.

Raftery, A. E. and J. L. Chunn (2013). Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography,* **50**(3), 777–801.

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.

Wiens, B. L. (1999). When Log-Normal and Gamma Models Give Different Results: A Case Study. *The American Statistician,* **53**(2), 89–93.

Wiśniowski, A., P. W. F. Smith, J. Bijak, J. J. Forster, and J. Raymer (2015). Bayesian population forecasting: Extending the Lee-Carter Method. *Demography,* **52**(3), 1035–1059.