UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Bayesian Estimation and Model Comparison for Mortality Forecasting

by

Jackie Siaw Tze Wong

Thesis for the degree of Doctor of Philosophy

March 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

BAYESIAN ESTIMATION AND MODEL COMPARISON FOR MORTALITY FORECASTING

by Jackie Siaw Tze Wong

The ability to perform mortality forecasting accurately is of considerable interest for a wide variety of applications to avoid adverse costs. The recent decline in mortality poses a major challenge to various institutions in their attempts to forecast mortality within acceptable risk margins. The ultimate aim of our project is to develop a methodology to produce accurate mortality forecasts, with carefully calibrated probabilistic intervals to quantify the uncertainty encountered during the forecasts.

Bayesian methodology is mainly implemented throughout the thesis for various benefits, but primarily due to its ability to provide a coherent modelling framework. Our contributions in this thesis can be divided into several parts. Firstly, we focus on the Poisson log-bilinear model by Brouhns et al. (2002), which induces an undesirable property, the mean-variance equality. A Poisson log-normal and a Poisson gamma log-bilinear models, fitted using arbitrarily diffuse priors, are presented as possible solutions. We demonstrate that properly accounting for overdispersion prevents over-fitting and offers better calibrated prediction intervals for mortality forecasting. Secondly, we carry out Bayesian model determination procedures to compare the models, using marginal likelihoods computed by bridge sampling (Meng and Wong, 1996). To achieve our goal of approximating the marginal likelihoods accurately, a series of simulation studies is conducted to investigate the behaviour of the bridge sampling estimator.

Next, a structurally simpler model which postulates a log-linear relationship between the mortality rate and time is considered. To provide a fair comparison between this model and the log-bilinear model, we carry out rigorous investigations on the prior specifications to ensure consistency in terms of the prior information postulated for the models. We propose to use Laplace prior distributions on the corresponding parameters for the log-linear model. Finally, we demonstrate that the inclusion of cohort components is crucial to yield more accurate projections and to avoid unnecessarily wide prediction intervals by improving the calibration between data signals and errors.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Jackie Siaw Tze Wong , declare that the thesis entitled *Bayesian Estimation and Model Comparison for Mortality Forecasting* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:...

Date:...

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors, Professor Jonathan J. Forster and Professor Peter W.F. Smith, without whom this project would not have been successful. Their invaluable knowledge and continuous guidances are incredibly useful and are greatly appreciated. I also wish to thank the department of Mathematical Sciences of the University of Southampton for funding this project via Mayflower Scholarship, where I was given the opportunity to also gain more teaching experience.

Finally, I would like to thank my parents, Wong Heng Ming and Chew Leh Swang, and the rest of my family and friends for their constant encouragement.

Chapter 1

Introduction

Mortality forecasting refers to the act of determining the mortality rates, or the life expectancies of a certain population in the future. It has become 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). For example, the World Health Organization (WHO) has regularly forecast mortality and morbidity to provide useful statistics for the use of international institutions all over the world, thus facilitating efficient allocation of funds. On the other hand, planning of social security to fund a nation's retirement income system by government depends on accurate forecasts of its intergenerational trust fund (which depends on the forecast of mortality rates). Insurance companies also rely heavily on these quantities (future death rates/life expectancies) in the pricing and reserving of annuities as well as pension portfolios. Therefore, the ability to forecast mortality accurately is crucial to avoid adverse costs.

It is well-established (from past data) that mortality has been declining over the years for most of the countries (the world population is ageing). Actuaries often call this **longevity risk**, because it poses an immediate threat to the actuarial applications (and other institutions) as calculation of the expected present values of numerous life-related products using life annuities functions relies on an accurate projection of the mortality rates. Ultimately, models developed for forecasting have to be able to capture the relevant features in the reduction of mortality, as well as capable of producing appropriately calibrated uncertainty bands associated with the forecasts so that users are aware of the relevant risks involved in their decision making. Recently, researchers have also been moving towards developing probabilistic forecasts rather than point forecasts (that ignore uncertainty) because the importance of being able to quantify uncertainty in an automated and transparent manner is beginning to be recognised by practitioners. These motivated our research, which is to develop a methodology that produces accurate mortality forecasts, accompanied by probabilistic intervals that are representative of the underlying uncertainties encountered. First of all, we investigate the plausibility of extending the model proposed by Brouhns et al. (2002) to account for overdispersion in the UK mortality data. This has the potential to prevent over-fitting by providing more flexibility for the model to describe extra variabilities within the data. Next, model comparison is carried out to acknowledge the model uncertainty encountered (rather than assuming a priori a single underlying model). A structurally simpler model which postulates a log-linear relationship between the mortality rate and time is then introduced as an alternative candidate model for projecting mortality. Finally, the inclusion of cohort components is considered to further improve the calibration between data signals and errors. In order to achieve our main objective of developing a coherent modelling framework for mortality forecasting, we prioritize the implementation of Bayesian methodology throughout the thesis.

Throughout the entire thesis, we use the terms forecasting and projection interchangeably to describe the act of acquiring future quantities based on the information derived from past data (without making further assumptions about the future). In general, there are three broad ways in which mortality forecasting can be carried out:

- **Expectation**: Created from subjective opinions from experts, usually accompanied by alternative high and low scenarios (serving as a "prediction interval") that are constructed from informed assumption of certain future quantities, e.g. fertility rates, ultimate mortality reduction factors etc. This approach has the privilege of user acceptability compared to other methods due to the fact that most users are ill-equipped to interpret the output from more sophisticated methods, notably those involving time series modelling or those producing probabilistic prediction intervals. It also allows the incorporation of relevant knowledge from various disciplines, such as epidemiology, demography, medical health etc. Therefore, it was previously favoured by most official statistical agencies and actuaries, who are known to hold substantive demographic and epidemiological expertise as well as having the general public as the target audience, before switching into extrapolative methods more recently (Waldron, 2005). Note that this does not necessarily imply that the expectation forecasting approach exclusively rules out probabilistic projections, in fact, Lutz (1996) proposed a fully probabilistic method of incorporating expert opinions for population (including mortality) forecasting (see Section 1.3). The expectation approach possesses numerous disadvantages: justifiability, potential for bias, conservativeness, "assumption drag" and expert flocking (for a detailed explanation of these terms, see Booth and Tickle, 2008).

- **Explanation**: Based on structural or causal epidemiological models of certain causes of death involving disease processes and known risk factors (Booth and Tickle, 2008). This approach attempts to describe mortality by using several causes of death (e.g. lung cancer and cardiovascular diseases) and their relationships with risk factors (e.g. smoking prevalences, diets etc.). Many of the models used for explanatory forecasting belong to the Generalised Linear Models (GLM) framework, where the response variables are mortality quantities (mortality rates or life expectancies), while the explanatory variables are the risk factors. Its main advantage relies on the fact that feedback mechanisms and limiting factors can be taken into account. Therefore, this method is particularly useful to deduce the potential relationships between mortality rates and the risk factors as well as to explain the mechanisms underlying various causes of death, and hence, is favoured by the epidemiologists. However, this approach is currently underdeveloped and also subject to the problem of data availability.

- **Extrapolation**: This approach will be the focus of our research. It assumes future trends will be a continuation of past patterns, that the future mortality rates will evolve in a similar pattern as the past mortality rates. Basically, a statistical model is developed to fit the data, any subsequent inferences including forecasting can then be undertaken based on the selected model (more generally referred to as modelling). There are two general types of models, **parametric** and **non-parametric** models. Parametric models attempt to describe the general mortality age profile using several parameters. Each of the parameters is then made time-dependent and projected into the future. Some examples include the Gompertz model, Makeham's model, Perks model and Heligman-Pollard model (see Section 1.4). Non-parametric models on the other hand, do not specify a priori the structure of the model. Orbanz and Teh (2010) defined a non-parametric model as one that uses only a finite subset of the parameter dimensions to explain a finite sample of observations with the set of dimensions chosen on the basis of the sample, such that the effective complexity of the model adjusts according to the data. The main difference between a parametric model and a non-parametric model is that a parametric model has a predefined (and finite) number of parameters (given the model), while the latter has number of parameters that depends on the size of the data involved, making them more flexible (typically the size of a non-parametric model grows with the complexity of the data).

1.1 Data and Notations

In this thesis, we denote D_{xt} as 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 d_{xt}, e_{xt} and μ_{xt} be the corresponding observed number of deaths, central exposed to risk and central mortality rate of age group x in year t.

The data chosen for illustrative purposes are the female death data and the corresponding exposures of England and Wales (EW), 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 $\{x_1, x_2, \ldots, x_A\} = \{0, 1, \ldots, 99\}$ and $\{t_1, t_2, \ldots, t_T\} = \{1961, 1962, \ldots, 2002\}$ with $A = 100$ and $T = 42$. Note that for simplicity, we use $t = \{1, 2, \ldots, T = 43\}$ to represent the set of years $\{t_1 = 1961, t_2 = 1962, \ldots, t_T = 2003\}$ and $x = \{1, 2, \ldots, A = 100\}$ to represent the set of age groups $\{x_1 = 0, x_2 = 1, \ldots, x_A = 99\}$ henceforth to avoid dealing with double subscripts later. We intentionally held back the data for years $2003 - 2013$ as validation set; see Section 3.8.6. So the matrix of observed death data, $\boldsymbol{d}$ is of dimensionality 100×42:

$$\begin{array}{lc}
\text{Age Group/Year} & \begin{array}{ccccc} 1961 & 1962 & 1963 & \ldots & 2002 \end{array} \\
\begin{array}{l} 0 \\ 1 \\ 2 \\ \vdots \\ 99 \end{array} & \begin{pmatrix} d_{11} & d_{12} & d_{13} & \ldots & d_{1\,42} \\ d_{21} & d_{22} & d_{23} & \ldots & d_{2\,42} \\ d_{31} & d_{32} & d_{33} & \ldots & d_{3\,42} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{100\,1} & d_{100\,2} & d_{100\,3} & \ldots & d_{100\,42} \end{pmatrix}.
\end{array}$$

By analogy, the death rates matrix, $\boldsymbol{\mu}$ takes similar form but with death rates as entries. Vectors and matrices are written in bold form in general. A general notation is:

$\mathbf{1}_n$	–	A $n \times 1$ vector of ones,
$\boldsymbol{I}_n$	–	A $n \times n$ identity matrix ,
$\boldsymbol{J}_n$	–	A $n \times n$ matrix of ones,
$N(\mu, \sigma^2)$	–	A univariate normal distribution with mean μ and variance σ^2,
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	–	A p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$,
Poisson(μ)	–	A Poisson distribution with mean parameter μ,
Exp(λ)	–	An exponential distribution with rate parameter λ,
Gamma(a, b)	–	A gamma distribution with shape parameter a and rate parameter b,
Inverse Gamma(a, b)	–	A distribution with its reciprocal as Gamma(a, b),
Laplace(a, b)	–	A Laplace distribution with location parameter a and scale parameter b.
Neg-Bin (a, b)	–	A negative binomial distribution with parameters a and b.

Unless otherwise stated, log denotes the natural logarithm (a logarithm to the base exp(1)). Other new variables/parameters introduced later in this thesis will be defined accordingly.

1.2 The Mortality Rates

Figure 1.1 shows the observed age-specific log mortality rates from age 0 to age 99 of England and Wales females in year 1961.

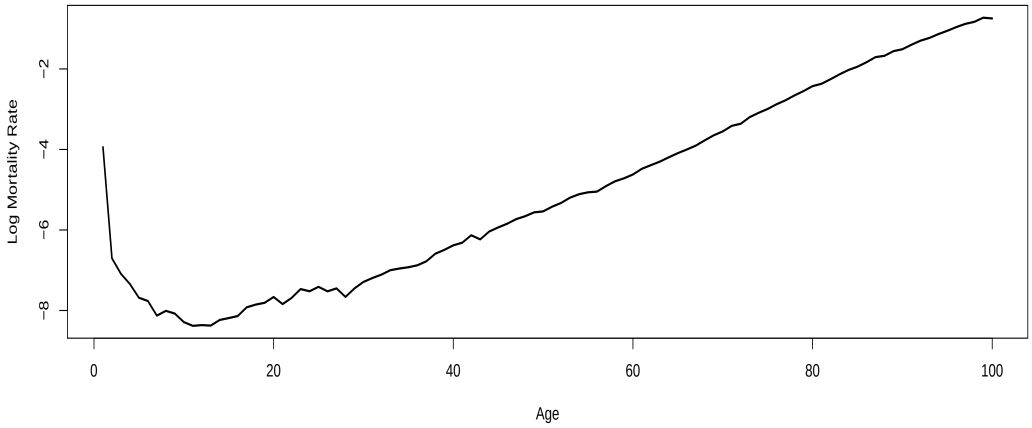


Figure 1.1: Observed female age-specific log mortality rates of EW in year 1961.

The general pattern for the mortality age profile can be described in the following way:

1. Initially, infant mortality rate is extremely high because babies are more vulnerable to diseases (e.g. measles and pneumonia), and also due to various other issues such as birth defects, complications during childbirth etc (see Hunt, 2001 for more details).
2. It then declines rapidly during childhood and early teenage phase.
3. Within the teenage age range, it rises again until a hump can be seen (often known as the "accident hump"). This is where teenagers (especially males) start being involved in hazardous activities such as reckless driving and alcohol consumption, resulting in high death rates. This effect is less apparent for females but is still somewhat discernible from the plot.
4. After that, it stabilises for a few years and then increases gradually (roughly exponentially).

To visualise the recent time trend of the mortality rates, we obtain plots of log mortality rates against age and time, as depicted in Figure 1.2 and 1.3 respectively.

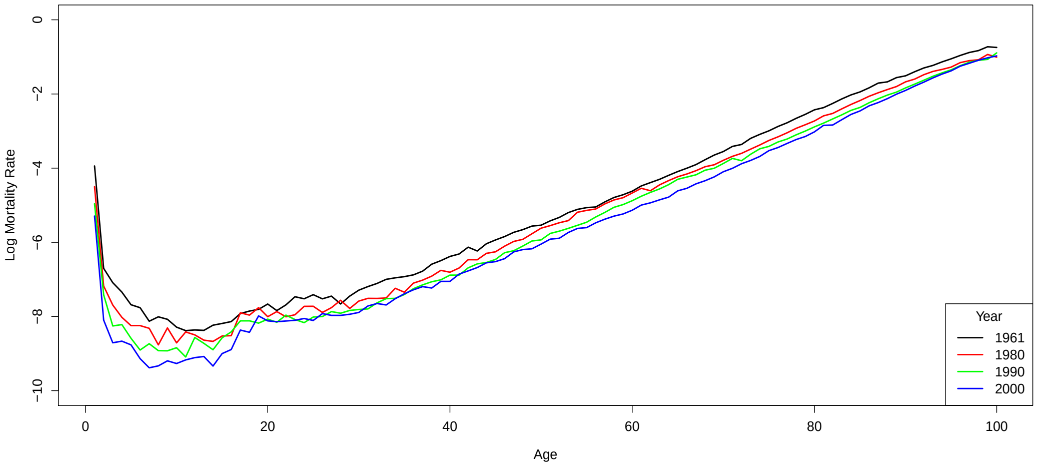


Figure 1.2: Plot of observed log mortality rates against age for several years.

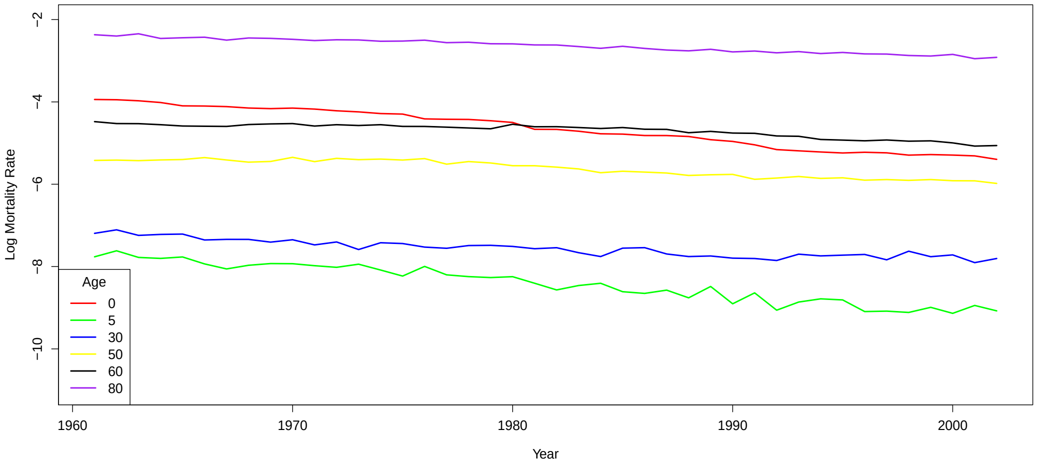


Figure 1.3: Plot of observed log mortality rates against time for various ages.

It is evident from Figure 1.2 that mortality has been improving over the years (shown by the downward shifting of the curves throughout the years). Also, notice that the general mortality age pattern persisted throughout the years. On the other hand, two things should be pointed out for Figure 1.3. Firstly, each of the lines is a decreasing function of time, portraying again the improvement in mortality. Secondly, each line has different rate of decrease. For instance, the decrease in infant mortality (red line) is more pronounced than the decrease in mortality for age 60 (black line), implying that the extent to which each age's mortality had improved differs.

1.3 A Review of the Expectation and Explanation Forecasting Approach

A brief overview of the expectation and explanation forecasting approaches is first provided in this section, followed by an extensive review of the extrapolative approach in Section 1.4. For a comprehensive review of the recent mortality forecasting methods, readers are recommended to refer to Booth and Tickle (2008) , Wong-Fupuy and Haberman (2004) or Tabeau et al. (2001).

The expectation approach to forecasting is traditionally undertaken by targeting a specific mortality quantity (e.g. life expectancy, mortality reduction factor) in the future. As described in Hollmann et al. (2000), the United States (US) Census Bureau projects the US mortality by assuming convergence of the base life tables to long term target life tables, using the projected life expectancies in Lee and Tuljapurkar (2001) as benchmarks and a set of relative speeds of mortality decline by age formulated based on expert opinion. Alders and de Beer (2005) employed a combination of a deterministic model and scenario forecasting, where the high and low variants are constructed by targeting the boundaries of forecast intervals (based on four main expert arguments) to accompany their forecast of mortality in Netherlands. Soneji and King (2012) presented the use of a combination of linear extrapolation and the targeting of ultimate reduction factors in the cause-specific mortality forecasting by the Social Security Administration (SSA). Previously, the Continuous Mortality Investigation Bureau (CMIB), a mortality research team of the United Kingdom (UK) Institute and Faculty of Actuaries, also implemented the targeting of ultimate reduction factors (formulated by the actuaries) before switching to stochastic mortality forecasting (see Continuous Mortality Investigation Bureau, 2004 and Continuous Mortality Investigation Bureau, 2007). Alternatively, a fully probabilistic approach of mortality forecasting in the presence of experts' opinion was described in Lutz (1996, p. 397-428), where subjective opinions/assumptions derived from experts are converted into probability distributions by assuming normality and linearity over time (see also Chapters 2 and 3 of Lutz et al., 2004). On the other hand, Heathcote and McDermid (1994) suggested the inclusion of extra judgemental factors from experts for forecasting the descriptive regression model that they developed for describing the relationship between mortality quantities and some polynomial functions of time and age (i.e. by appropriately changing values of regression coefficient of the fitted model for forecasting purposes and properly justifying the underlying reasons). Despite being widely applied for most official forecasts, the expectation forecasting approach was found by various researchers to be too conservative. Shaw (2007) discovered that the mortality assumptions used for the UK projections are overly pessimistic. For example, Willets (1999) warns against a potential underestimation of mortality improvement in the methodology developed by the CMIB. According to Oeppen and Vaupel (2002), experts' assertion that the life expectancy is reaching its maximum have repeatedly been proven wrong by historic increases in the life expectancies of various countries, which have not display signs of slowing down. Alho and Spencer (1990) also found that there is a consistent underestimation of mortality improvement by the Office of the Actuary and SSA in the US, on top of criticizing the unjustifiability of the high-low mortality variants in a probabilistic sense. Generally speaking, another difficulty of expert based forecasting approach is that it is challenging to conceptualise and then formally incorporating such subjective opinions in the forecasts. Alho (1992a) proposed to characterise the contribution of judgement by using the mixed estimation and forecasting procedure, where certain future values relating to mortality as specified by expert judgement are adjoined to historical observations, forming an augmented data set to be fitted and projected with the relative importance of judgemental factors automatically accounted for through a partitioning of the variance.

The literature available for the explanatory approach is rather limited primarily due to its demand for high-quality data and an imperfect understanding of the relationships between the risk factors and mortality in general. Typically, the main purpose of developing explanatory mortality models is to establish a relationship between several risk factors and mortality/morbidity risks, and hence, these models are extremely popular in epidemiological study. In epidemiological modelling, two particularly common models are used, namely the statistical regression models and dynamic multistate model. Statistical regression models serve to determine the association between mortality risks and explanatory variables. One of the most popular model in this context is the Cox proportional hazards model regression (Cox, 1972), given as

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i),$$

where $h_i(t)$ is the hazard for individual i, $h_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{x}_i$ is the vector of covariates for individual i. An important feature of this model is that the baseline hazard, $h_0(t)$, is assumed to only vary with time (irrespective of the individual under consideration), implying that the ratio of hazards for two individuals with different risk factors is constant over time. Alternatively, various GLM (McCullagh and Nelder, 1989) can also be applied. The Poisson regression model is typically used to model death counts, where the expected value depends on the covariates through a log-linear link function (see for example Frome, 1983). On the other hand, the logistic regression model can prove useful in modelling rates/probabilities of dying, where regression is carried out using a logistic function linking to the covariates (see Butler and Park, 1987; Altomare et al., 1990; Zhu and Li, 2013). The accelerated failure time model (Philip, 1999) postulates that the time until death can be expressed as the product of a baseline time variable and a time multiplicative factor, where the baseline time variable has some pre-specified parametric form (e.g. Weibull distributed) and the time multiplicative factor relates to the covariates through a log-linear relationship. The role of the time multiplicative factor is essentially to capture the extent to which the baseline time axis is stretched/shrunk correspondingly as the covariates vary. Regarding dynamic multistate models, they offer a flexible framework to account for multiple events (other than total mortality) to describe different aspects of morbidity and mortality simultaneously as functions of age and time. Due to the potential complexity of the multistate set up, microsimulation is a technique that is often employed here, where life histories are generated in the form of stochastic continuous processes for separate individuals using Monte Carlo techniques (see Gunning-Schepers, 1988; Van Oortmarssen et al., 1981; Wolfson, 1994). One major concern with the application of statistics in epidemiology is that statistical associations discovered may not necessarily be causal due to the complex interaction of various underlying risk factors. Therefore, conclusions drawn from these modelling approaches (and projections eventually) have to be interpreted with extreme caution, given the limitations involved.

The following are some other papers on explanatory mortality forecasting which might also be of interest. Murray and Lopez (1997b) developed explanatory models using several known risk factors which can be used for scenario forecasting subsequently (see also Murray and Lopez, 1997d; Murray and Lopez, 1997a). Manton et al. (1991) described a methodology to classify causes of death into exogenous (treatable) and endogenous (untreatable) causes to estimate the limit of future human life expectancy. In particular, they proposed to use two interrelated processes, consisting of a regression model (describing the dynamic of the time-varying covariates) and a combination of a Gompertz function of age and the above regression function, to model the dependency of mortality and several time-varying covariates. The limit of human life expectancy is then deduced by eliminating the exogenous causes of mortality (the cause-elimination method). Manton and Stallard (1992) developed a projection model based on a multivariate continuous state, stochastic process that allows multiple time-varying covariates to be used. The resulting model can be used to forecast changes in mortality at specific ages over time and, more importantly (in the epidemiological context), to understand the effects of specific medical interventions. The model considered in Girosi and King (2008) is simply a regression of mortality rates against time-varying covariates (e.g. Gross Domestic Product and smoking prevalence) using the GLM. Interestingly, the methodology described by Girosi and King (2008) is an example of an explanatory model, used in conjunction with the expectation approach as they extensively investigate the possibility of incorporating a valuable piece of information, smoothness of mortality rate across ages. Most of these methods involve the use of time-varying covariates/risk factors. They suffer the drawback of needing to project the covariates, which may well be no simpler to project than the mortality rates themselves. Therefore, these methods are usually limited to short term usage.

Decomposition of mortality by cause of death is an efficient tool for understanding the evolution of diseases and justifying the underlying contribution of each diseases in mortality improvement. It often utilises explanatory models in its analysis, although it exclusively does not belong to the explanatory forecasting approach. Ultimate reduction factors are then postulated based on expert opinion and historical data to produce a forecast (Wong-Fupuy and Haberman, 2004). Crimmins (1981) carried out a thorough investigation of the causes of death in the US population and proceeded to perform a mortality projection in year 2000. Despite being potent in explaining mortality, cause of death analysis is not particularly useful for forecasting because it is difficult to unify over the causes of death to produce an aggregated death rates due to the underlying correlations (Stoto and Durch, 1993). Only in cases where the relationship between the leading causes of death and the overall mortality are rather regular/linear or that no sharp changes in trends are expected in the future, then forecasting results from causes of death analysis would be similar to those produced from trend extrapolation of mortality (Murphy, 1990). It is also typically used for short term applications only (as pointed out in Tabeau et al., 2001) because the mortality improvement due to most of the diseases are seen to be governed by some sort of limiting factors, as demonstrated by Wilmoth (1995). To be specific, Murray and Lopez (1997c) discovered that there was an epidemiological shift of leading causes of death in most countries from previous experience, where occurences of some chronic diseases are now postponed to later stages of life (due to vaccination and other medical interventions) and also depend on different risk factors (see also Mackenbach, 1988). In addition, Alho (1991) stated that this method is very likely to fail if highly non-linear models are used or when modelling error is considered, where implausible figures may emerge for rapidly changing causes of death. Hence, most institutions and actuaries recommended against the use of this method. An exception is Soneji and King (2012), who specifically discussed the role of various causes of death in reducing mortality and presented the expected behaviour of each of them in the long run for the purpose of mortality forecasting (see also Caselli and Egidi, 1992; Lopez and Crujisen, 1991).

1.4 A Review of the Extrapolation Forecasting Approach

There is a wealth of literature on the extrapolative forecasting approach, and is blooming rapidly especially recently when various institutions are being brought to the attention of the potential negative impacts of longevity risk across the world. Specifically, stochastic models have gained a lot of popularity in mortality projection due to their abilities to produce probabilistic intervals that encapsulate uncertainties associated with the forecasts, thereby facilitating informed decision making within an acceptable risk margin. In what follows, we focus on reviewing stochastic models that are capable of yielding probabilistic intervals.

As stated before, extrapolative forecasting approach is broadly classified into parametric and non-parametric modelling. **Parametric** mortality models aim to represent mortality over the whole age range parsimoniously using mathematical curves with several parameters. Commonly used parametric models include the Gompertz's Law, Makeham's Law, Perks model and so forth, each with their own merits. In particular, the Gompertz's Law as proposed by Gompertz (1825) is

$$h(x) = AB^x,$$

where $h()$ is the hazard function; A and B are parameters to be projected forward. Makeham (1860) slightly modified the Gompertz model by introducing an age-independent parameter to account for the accident hump. The Gompertz model has been found to describe mortality for middle ages (30-90) reasonably well (see Spiegelman, 1968, p. 164 and Wetterstrand, 1978), but fails to represent mortality at very old ages. Specifically, Thatcher (1999) pointed out that Gompertz model tends to overestimate the old age mortality, who then proceeded to suggest the Perks model (Perks, 1932), which is a logistics model that assumes mortality eventually reaches a plateau at the oldest ages (in contrast to the Gompertz function, which increases indefinitely). Cairns et al. (2006a) demonstrated the use of the Perks model on the UK mortality data, the model of which is given by

$$1 - p(t+1, t, t+1, x) = \frac{\exp[A_1(t+1) + (x+t)A_2(t+1)]}{1 + \exp[A_1(t+1) + (x+t)A_2(t+1)]}, \tag{1.1}$$

where $p(t, T_0, T_1, x)$ represents the probability as measured at t that an individual aged x at time 0 and still alive at T_0 survives until time $T_1 > T_0$, $A_1(t)$ and $A_2(t)$ are the time-varying parameters to be projected using time series methods. Unfortunately, both the Gompertz and Perks models focus on a specific range of ages only, so Heligman and Pollard (1980) proposed

$$\frac{q_x}{1 - q_x} = \lambda_1^{(x+\lambda_2)^{\lambda_3}} + \lambda_4 \exp[-\lambda_5(\log(x) - \log(\lambda_6))^2] + \lambda_7 \lambda_8^x, \tag{1.2}$$

where q_x is the probability of dying within 1 year for an individual aged x, $\{\lambda_1, \dots, \lambda_8\}$ are eight parameters of the model. The Heligman-Pollard model has the potential to account for three stages of the life span: infancy, young adulthood (the accident hump) and senescence. In terms of mortality forecasting, Cramer and Wold (1935) fitted (using minimum chi-square) and projected (by extrapolating relevant parameters) Swedish mortality for years 1800-1930 using the Makeham curve McNown and Rogers (1989) demonstrated the use of the Heligman-Pollard model for projecting the US mortality data by modelling and extrapolating each of the eight parameters independently using univariate auto-regressive integrated moving average (ARIMA) models. Despite describing the Australian mortality fairly well, the Heligman-Pollard model was demonstrated by Carriere (1992) to be not universally applicable (it does not fit US female mortality for instance). Instead, Carriere (1992) developed a general law of mortality which is a mixture of Weibull, Inverse Weibull/Gompertz and Gompertz survival functions to represent childhood mortality, the accident hump and adult mortality respectively. The model postulated by Carriere (1992) was found to have clearer parameter interpretation than the Heligman-Pollard model and is easily generalised by including more survival functions.

Albeit the simplicity of implementation, it is challenging to develop parametric models that are able to describe the mortality age profile adequately without requiring high dimensional parameters. A high dimensional parametric model faces serious issues during parameter interpretation and projection. For example, the methodology proposed by McNown and Rogers (1989) described above (to model and project the parameters of the Heligman-Pollard model separately using univariate ARIMA models) ignores the interdependencies among the parameters (Hartmann, 1987), and hence, was found later by McNown and Rogers (1992) to be insufficient. Another shortcoming of this approach is that it is rather challenging to combine the prediction intervals of each of the parameters to form an overall probabilistic prediction interval for the resulting forecast. One possible solution is to use multivariate vector autoregressive (VAR) models to account for the potential correlations among the parameters; constructing the aggregated prediction intervals is still an issue though. Another more promising solution was demonstrated by Cairns et al. (2006a), where the Bayesian methodology was implemented to construct an aggregated prediction interval, incorporating parameter uncertainty (and other sources of uncertainty) through the computation of joint posterior distribution (using sampling based methods).

Non-parametric models (typically used for the rate models) are also used predominantly for mortality forecasting. The first stochastic model was pioneered by Lee and Carter (1992), and has since then becomes the focus of most of the subsequent research in this regard. The well-known Lee-Carter (henceforth LC) model is

$$\log \mu_{xt} = \alpha_x + \beta_x \kappa_t + \nu_{xt}, \tag{1.3}$$

where α_x and β_x are age-specific components, κ_t is a time-varying parameter, while ν_{xt} are residuals with mean 0 and variance σ_ν^2. Singular Value Decomposition (SVD) can then be used to provide a least squares solution, which is pointed out by Girosi and King (2008) to be consistent with the assumption that $\nu_{xt} \sim N(0, \sigma_\nu^2)$. Additionally, Lee and Carter (1992) also carried out a **second stage estimation** of κ_t to ensure that the parameters estimated do indeed lead to the observed number of deaths in the data by matching the fitted number of deaths with the observed total deaths in a given year, keeping the estimated α_x and β_x. Finally, the re-estimated κ_t, $\tilde{\kappa}_t$ is projected forward using a random walk with drift,

$$\tilde{\kappa}_t = \tilde{\kappa}_{t-1} + \text{drift} + \epsilon_t,$$

where ϵ_t has a zero-centered normal distribution. This model has been widely applied across the globe. For example, Andreozzi et al. (2008) and Haberman and Russolillo

(2005) applied the LC methodology to forecast Argentinian and Italian mortality respectively. It is also used as a benchmark by the US Bureau of census and recommended by the US Social Security Technical Advisory Panels. Note that the LC model as illustrated in (1.3) can be used to generate a one-parameter family of life tables by letting $\kappa = 0$ representing one (known) life table, and $\kappa = 1$ representing another life table. Varying the parameter κ between 0 and 1 geometrically interpolates between the two life tables, while varying κ outside the interval extrapolates from the life tables (Tabeau et al., 2001). Hence, this provides users the option to generate missing age profiles of mortality in countries with limited data, although this method is known to be less efficient in terms of the estimation when more than two life tables are available.

Despite being a popular approach, there has been a lot of criticisms against the LC approach. Various modifications of the LC approach began to emerge thereafter. In particular, Girosi and King (2008) stated that the LC model will eventually produce implausible forecasts (see Section 1.5) in the long run due to the time-invariant age pattern β_x. In response, Lee and Miller (2001) proposed to restrict the fitting period from 1950 to ensure consistency with the model assumption of the time-invariant β_x for the US mortality data. They also proposed to re-estimate κ_t by matching the observed life expectancy rather than the total deaths to avoid the need for death data. They then criticised the jump off discontinuity at the forecast origin, which has the consequence of causing a bias in the projected life expectancy. Even after the adjustments, it was still concluded that LC model based forecasts led to a systematic underestimation of future life expectancy in the US (although slightly better than the official forecasts). Booth et al. (2002) proposed the Booth-Maindonald-Smith (BMS) model, which explicitly treats the problem of selecting appropriate fitting period using formal statistical goodness of fit criteria to ensure that the model assumptions of time-invariant β_x and linearity of κ_t are met. They also point out that the use of only one principal component by the LC model is insufficient in describing the variations of mortality data (despite explaining up to 90% of the variation in the data, see Girosi and King, 2008), and proceeded to extend the model to allow for more than one principal component in their functional data analysis. Li and Chan (2005) slightly modified the original LC approach to forecast mortality rates of EW and Scandinavian data, with specific focus on outlier detection and adjustment to overcome the sensitivity of this model with respect to outliers, particularly those near the forecast origin.

Brouhns et al. (2002) criticised the normal assumption on the residuals, ν_{xt}, which they reasoned that the observed log mortality rates have more variabilities at younger ages, where there are smaller number of deaths. While Lee and Carter (1992) informally addressed this issue using the second stage estimation of κ_t, it is claimed that this method does not possess a properly defined minimization criterion which can lead to incoherency in subsequent inferences, as pointed out by Booth et al. (2002). Therefore, 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 log mortality rates. We refer to this model as the Poisson LC model, which is given by

$$D_{xt} \sim \text{Poisson}(e_{xt}\mu_{xt}) \text{ with } \log \mu_{xt} = \alpha_x + \beta_x \kappa_t. \tag{1.4}$$

The Poisson LC model was found to fit death data reasonably well (see Brilinger, 1986 and McDonald, 1996), and also renders the minimization criterion clear. This model is discussed in detail in Section 1.5 as the contributions of this thesis are mostly related to it.

One drawback of this particular model is 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 to relax the stringent model structure of a Poisson distribution. 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. These motivated our research which constitutes the first part of the thesis. 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. Bayesian mortality modelling/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 and also recommended methods of forecasting the covariates. Murphy and Wang (2001) demonstrated the use of Bayesian model averaging approach on Chinese infant's mortality data to improve the predictive performance of their logistic regression model based on several explanatory variables. However, they did not consider the projection of the model to predict future evolution of infant's mortality for that country, which can be rather troublesome (requires forecasts of each of the covariates selected). On the other hand, Czado et al. (2005) fitted the Poisson LC model for French male mortality within the Bayesian framework. 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. Wiśniowski et al. (2015) adopted the Bayesian version of the original LC method, but also extended the model to incorporate the cohort effect and to allow for overdispersion in their mortality projections. Additionally, they proposed to use a bivariate vector autoregressive process with drift for the projection of the time-variant parameters for both sexes to ensure long term coherency. Favouring to model life expectancy rather than death rates, Raftery and Chunn (2013) performed Bayesian probabilistic mortality projections on an aggregate level by pooling across several countries to borrow strength. The main advantage of Bayesian methods is it provides a coherent modelling framework for multiple sources of uncertainty to be integrated. The rest of the advantages of adopting Bayesian modelling will be discussed in detail in Section 3.2.

The next part of our contribution focuses on the log-bilinear structure of the Poisson LC rate model. Specifically, we consider the possibility of using a log-linear (in time) structure, which is a simpler model formulation to describe mortality. This is inspired by Wong-Fupuy and Haberman (2004), that “a log-linear relationship between mortality rates and time is present in most actuarial applications and in the Lee-Carter model”, quoted directly from them. The inclusion of time as an explanatory variable by Renshaw et al. (1996) in their GLM approach of modelling mortality exemplifies the log-linearity of the mortality rate in time. While it may not be obvious, the complementary use of the Poisson LC model with a random walk model on κ_t implicitly assumes a log-linear relationship between the mortality rates and time. Thus, we explicitly postulate a rate model with log-linearity in time, and compare it with the original Poisson LC model (see Chapter 5).

Though cohort effects are typically less apparent than the period effects, they are nevertheless clearly noticeable for some countries. To exemplify this fact, cohort effects have been identified in the US (McNown and Rogers, 1989), Japan (Willets, 2004), Norway (MacMinn, 2003), France (Wilmoth, 1990) and so on. The existence of cohort effects in the UK mortality data is well established and is anticipated to persist over the next few decades (see Government Actuary’s Department, 2001 and Government Actuary’s Department, 2006). The age-period-cohort (APC) model is among the most popular models used for cohort modelling (see for example Fu, 2008), and is given by

$$\log \mu_{xt} = \alpha_x + \kappa_t + \gamma_{t-x},$$

where α_x, κ_t and γ_{t-x} are model parameters. Currie (2012) applied this model for mortality forecasting (see also Cairns et al., 2010), and also examined some of the pitfalls of this model, particularly the identification problem. Another model which is closer to our application is proposed by Renshaw and Haberman (2005), with the rate model given by

$$\log \mu_{xt} = \alpha_x + \beta_x^1 \kappa_t + \beta_x^2 \gamma_{t-x} + \nu_{xt},$$

where α_x, β_x^1, β_x^2 are age-specific parameters, κ_t is a time varying parameter, γ_{t-x} are the cohort components, and ν_{xt} are residuals. Continuous Mortality Investigation Bureau (2006) demonstrated the use of P-spline regression to smooth and forecast UK age-cohort data for ages 20-100, which was later found to work well in recent years by Continuous Mortality Investigation Bureau (2009). It was also mentioned that a major issue with using the P-splines method for forecasting is the so called "edge" effect, where the projected trajectories are very sensitive to the particular order of penalty used. Currie et al. (2004) addressed this issue by examining the impact of different orders of penalty on the projections and then making a sensible choice according to subjective opinions. Again, all these cohort models were fitted using the frequentist approach. Therefore, we investigate the feasibility of incorporating the cohort effect within a Bayesian paradigm in the final part of this thesis.

Other than those described above, there are a lot of existing methodologies that are relevant. Reichmuth and Samad (2008) applied a spatio-temporal modelling concept, with a minor extension to account for the inclusion of covariates and the use of more than one principal component. Favouring to treat mortality projections in the context of time series modelling, Hagnell (1991) applied vector autoregressive and moving averages (VARMA) models to forecast mortality for Swedish data. On the other hand, Giacometti et al. (2012) proposed the use of autoregression(1)-autoregressive conditional heteroskedasticity(1), AR(1)-ARCH(1) model with Student's t innovations and illustrated the model on Italian mortality data. Their methodology also allows the possibility to project in the "direction" of age to counteract the issue of unreliable mortality data at older ages (Coale and Kisker, 1990). Claiming that quantitative comparisons had never been undertaken, Cairns et al. (2007) compared several stochastic mortality models quantitatively using data from England and Wales and the US, which they discovered that no single model prevails on the basis of all the criteria considered. Specifically, if the Bayesian Information Criterion (BIC) is used, then England & Wales data favours the model extended from Cairns et al. (2006b) while US data favours the model by Renshaw and Haberman (2005). On the other hand, if robustness of parameter estimates is considered, then the preferred model is a different extension of the Cairns et al. (2006b), which includes both a cohort effect and a period effect that is quadratic in age. Hyndman and Ullah (2007) advocated the use of functional data analysis, which is claimed to be the improved version of LC method (a functional analysis with one principal component) by allowing for the possibility of using multiple principal components to explain more variabilities in the data. Their proposed method is also robust to outlying years (due to pandemics, wars etc.) and allows for nonparametric smoothing. Shang et al. (2016) proposed to use a multilevel functional data method, coupled with the use of parametric bootstrapping, to model and forecast mortality in their population projection model (see also Bell and Monsell, 1991; Hyndman and Booth, 2008).

Some scholars acknowledge the significance of producing coherent mortality forecasts (in the sense of having some convergence properties). Li and Lee (2005) proposed a coherent forecasting method to ensure convergence of forecasted rates of the sub-populations in the long run, where they illustrated their methodology in addressing the narrowing of gender differences. Convergence is deemed significant because it ensures plausible forecasts in the long run; without convergence, forecasted rates tend to diverge to generate implausible discrepancies among the sub-populations. This can be achieved by stipulating that mortality improvement of each sub-groups is dictated by an overall improvement due to the entire population, subject to group-specific adjustments which converges to constants in the long run. Yang et al. (2011) extended this methodology and applied it on the mortality data of the US and Canada for both genders. On the other hand, Hyndman et al. (2012) proposed to model the square root of product and ratio of mortality rates using functional time series models (with up to six principal components), where coherency is ensured when the coefficients of the functional time series of the ratio are required to be stationary.

1.5 The Poisson LC (PLC) Model

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

$$D_{xt} \sim \text{Poisson}(e_{xt}\mu_{xt}) \quad \text{with} \quad \log \mu_{xt} = \alpha_x + \beta_x \kappa_t. \tag{1.5}$$

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.

- Summing Equation (1.5) across time and applying the constraint $\sum_t \kappa_t = 0$, we have
$$\alpha_x = \frac{\sum_t \log \mu_{xt}}{T} \equiv \overline{\log \mu_{xt}}.$$

This implies that α_x is the average of the log mortality rates over time. Hence, it mimics the general shape of the mortality age profile described in Section 1.2.

- A first order differentiation of Equation (1.5) yields

$$\frac{\mathrm{d}\log\mu_{xt}}{\mathrm{d}t} = \beta_x \frac{\mathrm{d}\kappa_t}{\mathrm{d}t}, \tag{1.6}$$

  where $\frac{\mathrm{d}\log\mu_{xt}}{\mathrm{d}t}$ represents the rate of change of log mortality rate for age x and $\frac{\mathrm{d}\kappa_t}{\mathrm{d}t}$ represents the rate of change of the time variant parameter, κ_t. The identity above indicates that β_x governs the sensitivity of how the rate at each age responds to overall changes in mortality. In other words, β_x is the age-specific pattern of mortality improvement, measuring the extent to which each age's mortality is improved. As consistent with the empirical finding that mortality (on the log scale) tends to decline more rapidly at younger age groups, β_x is in general a decreasing function of age, with an exception at the "accident hump", where mortality improvement is close to being negligible. In addition, since mortality has been improving over time (i.e. $\frac{\mathrm{d}\log\mu_{xt}}{\mathrm{d}t} < 0$), it can be deduced from Equation (1.6) that β_x are positive values as κ_t is generally a decreasing function of time as well (i.e. $\frac{\mathrm{d}\kappa_t}{\mathrm{d}t} < 0$, see the next point).

- Summing Equation (1.5) across age and using the constraint $\sum_x \beta_x = 1$, we obtain

$$\kappa_t = \sum_x(\log\mu_{xt} - \alpha_x) = \sum_x(\log\mu_{xt} - \overline{\log\mu_{xt}}). \tag{1.7}$$

  This suggests that κ_t is the sum of the individual deviations of the log mortality rates from their temporal mean, $\overline{\log\mu_{xt}}$, across age. Clearly, it captures the overall time trend of mortality change (after being appropriately modulated by β_x). Differentiating Equation (1.7) with respect to t, we have

$$\frac{\mathrm{d}\kappa_t}{\mathrm{d}t} = \sum_x \frac{\mathrm{d}\log\mu_{xt}}{\mathrm{d}t}.$$

  With mortality improving over time for most countries (i.e. $\frac{\mathrm{d}\log\mu_{xt}}{\mathrm{d}t} < 0$), we infer that $\frac{\mathrm{d}\kappa_t}{\mathrm{d}t} < 0$ since it involves a summation of negative values. Hence, κ_t is generally a decreasing function of time.

In order to fit this model, weighted least squares (with D_{xt} as the weights, see Wilmoth, 1993) or Newton's iterative updating scheme can be used to obtain the maximum likelihood estimates (MLE) $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$ (see Renshaw and Haberman, 2005 for details). The ordinary generalised regression method does not work here due to the bilinear terms in Equation (1.5). One can, however, fit this model within the generalised linear model (GLM) framework by iteratively conditioning on one of β or κ parameters (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}.

As an illustration, the PLC model is fitted on the EW female mortality data within the frequentist framework. The MLE of the model parameters, α_x, β_x and κ_t are shown in Figure 1.4. As expected, the fitted α_x takes the shape of the general mortality age profile. The fitted β_x are positive for all ages, with a relatively more erratic age pattern than α_x. Remarkably, the fitted κ_t appears to be a **linearly** decreasing function of t for the EW female mortality data (this is useful to know when specifying the model for κ_t).

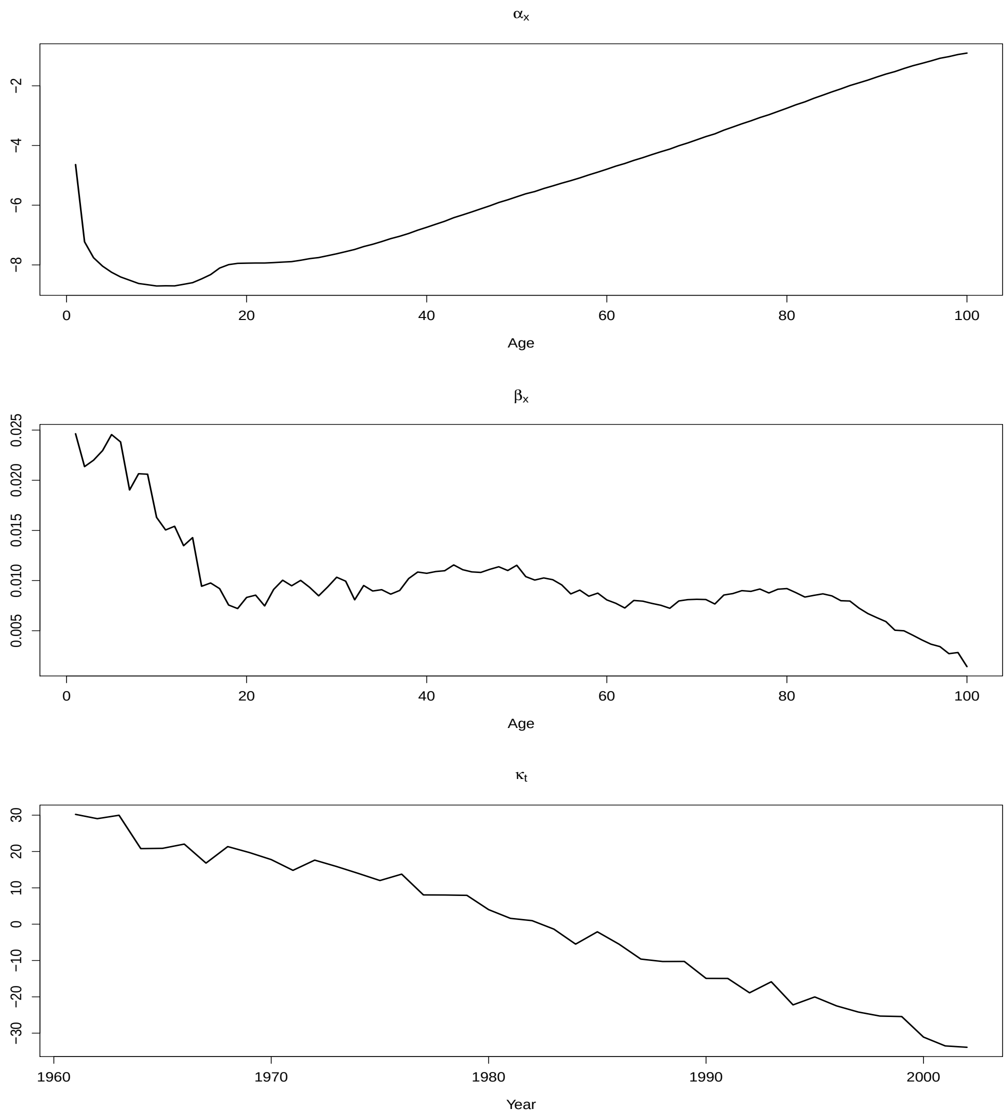


Figure 1.4: Plots of the MLE of α_x, β_x and κ_t under the PLC model.

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).

Some of the disadvantages of the classical LC approach are as follows:

- Potential inconsistencies may arise due to their two-stage model fitting procedures (Czado et al., 2005). In particular, 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.
- Any prior knowledge from the subject-matter experts cannot be incorporated.
- Parameter uncertainty and model uncertainty are typically ignored.
- Long term projected mortality age-profile will not be smooth. This is a direct consequence of the assumption that the age-specific pattern of mortality improvement, β_x, is constant over time, causing the projected age-specific log mortality rates to eventually "fan out" after a point when projected too far out (see Girosi and King, 2007 for a detailed description).
- The cohort effect is neglected.
- Specific to the Poisson LC model, the mean and variance of D_{xt} are restricted to be the same (see Chapter 3.1 for a detailed explanation).

1.6 Agenda

The rest of the thesis is organised in the following way. Chapter 2 provides several statistical preliminaries necessary to understand the materials in this thesis. In Chapter 3, we illustrate the evidence of extra variabilities in the England and Wales female mortality data, and present two mixed Poisson LC models as possible solutions. We also focus on the construction of well-calibrated prediction intervals in a coherent manner using Bayesian methods. A detailed description of the Markov Chain Monte Carlo algorithms for posterior sample generation is provided. Finally, the mixed Poisson LC models are compared with that of Czado et al. (2005) (to highlight the importance of accounting for overdispersion) and with each other. In Chapter 4, four simulation studies are conducted to study the behaviour of the bridge sampling estimator, which we then use to improve the accuracy of the marginal likelihood estimation for model comparison. In Chapter 5, we propose a structurally simpler mortality model that postulates a log-linear relationship between the mortality rate and time. A rigorous investigation is performed to ensure the consistency of prior information specified for this new model and one of the mixed Poisson LC models from Chapter 3. A comparison is then carried out between the two models in terms of the goodness of fit and predictive ability. Incorporation of the cohort effect is investigated in Chapter 6, where the log-linear model presented in Chapter 5 is extended to include cohort components. In Chapter 7, a conclusion reviewing our contributions in the thesis and several potential areas for future work are presented.

Chapter 2

Statistical Preliminaries

2.1 Bayesian Methodology

We perform the model fitting within a Bayesian paradigm. Essentially, analysts start with their own set of prior beliefs (or inspired by subject-matter experts) about a particular problem, which are then updated in light of the observed data to come up with posterior beliefs. The fundamental usage of Bayesian inference is based on Bayes' Theorem. In particular, suppose that $\boldsymbol{d}$ is a data matrix, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^\top$ is a vector of parameters to be estimated, $f(\boldsymbol{d}|\boldsymbol{\theta})$ is the corresponding likelihood function representing the data generating mechanism, and prior beliefs on $\boldsymbol{\theta}$ are converted into the probability distribution, $f(\boldsymbol{\theta})$. Then Bayes' theorem states that

$$f(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{f(\boldsymbol{d}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{d})} = \frac{f(\boldsymbol{d}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{d}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.1}$$

Mathematically, Bayes' Theorem is merely a straightforward consequence of the definition of conditional probability. When applied in the context of data analysis, this yields Bayesian inference. A crucial concept here is that the vector of unknown parameter, $\boldsymbol{\theta}$, is now treated as random variables as compared to classical statistics, where parameters are regarded as being unknown constant. As such, the denominator of Equation 2.1 (which involves integral that can be difficult to evaluate) is often not of interest, leading to the following expression:

$$f(\boldsymbol{\theta}|\boldsymbol{d}) \propto f(\boldsymbol{d}|\boldsymbol{\theta}) \times f(\boldsymbol{\theta}). \tag{2.2}$$

In words, this is

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

2.1.1 Prior Distributions

Clearly, the prior distribution plays an important role in determining the posterior distribution. Ideally, the prior distributions formulated should correctly reflect our uncertainty/prior knowledge about a particular statistical problem. Elicitation of the knowledge from subject matter experts should therefore be carried out wherever applicable. In the absence of prior knowledge, the corresponding prior distributions should possess a lot of uncertainties to reflect ignorance, commonly referred to as non-informative priors. Other default and reference priors also appear to be sensible choices in this situation and are extensively discussed in Kass and Wasserman (1994).

Secondly, recall that a family of priors is said to demonstrate conjugacy if the resulting posterior distribution also belongs to the same family of distributions (Diaconis and Ylvisaker, 1979). Here, we reintroduce the concept of conditional conjugacy as defined in Gelman (2006). In particular, suppose that our parameters, $\boldsymbol{\theta}$, are partitioned into two subsets $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$. Then a prior distribution $f(\boldsymbol{\theta}_1)$ is said to be conditionally conjugate for $\boldsymbol{\theta}_1$ if the conditional posterior distribution, $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{d})$, also belongs to the same family of distributions. This is a useful idea in a hierarchical modelling framework because it simplifies the derivation of conditional posterior distributions, thereby permitting the use of Gibbs algorithm in the sampling scheme (provided it is possible to generate directly from this class of prior distributions). Moreover, the conjugacy relationship is preserved even when the model is expanded hierarchically.

Furthermore, we incorporated the identifiability constraints directly through the prior distributions. Effectively, this can be undertaken within a Bayesian paradigm by simply treating the intended constraints as prior knowledge without uncertainty (in the form of conditioning, see Section 3.4.1.1). This is in contrast to Czado et al. (2005), who modified their Markov Chain Monte Carlo (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 (the detailed balance equation for MCMC to work is potentially violated). They merely provided some visual evidence of convergence from the resulting trajectories of the constrained parameters, which is an inadequate proof for MCMC convergence.

2.2 Markov Chain Monte Carlo (MCMC) Simulations

The classic difficulty of adopting Bayesian methodology is that Bayesian computation involves evaluating the integral in the denominator of Equation 2.1 which is often analytically intractable, thereby hindering subsequent inferences. Several methods have been developed to deal with this issue. For example, Laplace approximation corresponds to a non-sample based approach that directly approximates the required integral, allowing posterior distributions to be acquired in closed form, at least approximately (for more information, see O'Hagan and Forster, 2004).

In this thesis, we focus mainly on sample based approach to perform Bayesian inference. Essentially, a sample based approach to Bayesian inference involves generating samples from the posterior distribution. Features of the posterior distribution (such as marginal/joint distributions and other quantities of interest) can then be estimated from the samples generated (using kernel smoothing for example). However, it is typically the case that independent realizations from the posterior distributions are unavailable, especially in complex hierarchical model set up. Nevertheless, MCMC methods provide a promising solution in this regard by enabling generation of dependent samples, rendering Bayesian inference feasible. Their availability is the main reason for the substantive increase in the application of Bayesian methods over recent years. Basically, an ergodic Markov Chain which is straightforward to sample from and possesses $f(\boldsymbol{\theta}|\boldsymbol{d})$ as the target distribution is constructed (by ensuring it satisfies the detailed balance equation, see O'Hagan and Forster, 2004). The samples generated can then be used to summarise our posterior distribution (at least in an asymptotic sense). One crucial criterion to bear in mind when constructing such algorithm is that the speed of convergence (in terms of the sample size required) should be within the practical constraints of time and computational power.

The number of iterations required for convergence varies from applications to applications. Current literature suggests that there is no formal proof to show that the constructed Markov Chain has converged to the intended equilibrium distribution. Nevertheless, there exists some negative evidence as an indication of insufficient iterations. In particular, Gelman and Rubin (1992) suggest that, for several chains of iterations, the within-chain variance should not differ markedly from across-chain variance and proceed to test this by introducing the potential scale reduction, which is the ratio of across-chain variance to within-sequence variance. A value that differs substantially from 1 violates convergence properties. Several other methods are also available to test for convergence. The approach undertaken in this thesis to monitor convergence is by running several parallel chains with different initial values as long as possible, then undergo comparisons to ensure there is no considerable difference between the chains (mostly visually).

Usually, the first portion of the simulated samples is discarded to diminish the effect of initial values chosen, a process known as **burn-in**. On the other hand, **k-thinning** (retaining samples every k iteration) can be applied to obtain a set of approximately independent MCMC samples, where k is selected on the basis that the sample autocorrelations are close to zero. To implement the above strategy, we merely need algorithms to construct the Markov Chain with the intended properties. The MCMC methods considered in this thesis consist of mainly the **Gibbs sampler** and **Metropolis-Hastings algorithm**.

2.2.1 The Metropolis-Hastings (MH) Algorithm

The MH algorithm is one of the most useful and general method of constructing an MCMC sampler with the required theoretical properties. It was described by Hastings (1970), who generalised the original Metropolis algorithm of Metropolis et al. (1953). In particular, the MH algorithm can be summarised with the procedures below:

1. Select an initial value $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \dots, \theta_k^0)$.
2. Set $i = 1$, so the current value is $\boldsymbol{\theta}^{i-1}$.
3. Now propose $\boldsymbol{\theta}^*$ from a proposal density, $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{i-1})$.
4. Define
$$a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = \min\left\{1, \frac{f(\boldsymbol{\theta}^*|\boldsymbol{d})q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^{i-1}|\boldsymbol{d})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})}\right\}.$$
Then with probability $a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})$, accept the proposal and set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$; otherwise, reject the proposal and assign $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$.
$$\left[\text{Generate } U \sim \text{Uniform}(0,1), \text{ then } \boldsymbol{\theta}^i = \begin{cases} \boldsymbol{\theta}^* & \text{if } U \le a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}), \\ \boldsymbol{\theta}^{i-1} & \text{if } U > a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}). \end{cases}\right]$$
5. Set $i = i + 1$ and repeat steps 3-4.
6. Stop when $i = n$, where n is the number of iterations required.

From the constructed Markov Chain, it is straightforward to check that it satisfies the detailed balance criteria, which is an essential condition. Moreover, provided that the thus far arbitrary $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{i-1})$ is irreducible and aperiodic on an appropriate state space, it can be shown that the equilibrium distribution of the constructed Markov Chain is indeed the target distribution, $f(\boldsymbol{\theta}|\boldsymbol{d})$ (see Gelman et al. (1995) for more theoretical details).

Crucially, notice that the posterior distribution of interest only appears in the algorithm as a ratio,
$$\frac{f(\boldsymbol{\theta}^*|\boldsymbol{d})}{f(\boldsymbol{\theta}^{i-1}|\boldsymbol{d})},$$
meaning that any normalising constant of our posterior distribution will vanish. This implies that we only need to be able to evaluate the unnormalised posterior density and to sample from the proposal density $q(\cdot|\cdot)$ in order to implement the MH algorithm.

Thus far, the choice of the proposal distribution, $q(\cdot|\cdot)$ remains arbitrary. In fact, the choice of $q(\cdot|\cdot)$ defines the specific algorithm created. For instance, if $q(\cdot|\cdot)$ is symmetric, i.e. $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)$, then the transition probability collapses to

$$a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = \min\left\{1, \frac{f(\boldsymbol{\theta}^*|\boldsymbol{d})}{f(\boldsymbol{\theta}^{i-1}|\boldsymbol{d})}\right\}. \tag{2.3}$$

The resulting algorithm is referred to as the **Metropolis algorithm**. On the other hand, if a random walk proposal is used, i.e.

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{i-1} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ has a symmetric distribution around $\mathbf{0}$ (e.g. multivariate normal or t-distribution). The constructed algorithm corresponds to the well known **random walk MH algorithm**. Typically, a multivariate normal distribution is used:

$$\boldsymbol{\epsilon} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the proposal variance matrix. Then we have $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = q(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{i-1})$ and the $a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})$ reduces to Equation (2.3). It is immediately obvious that any proposal that possesses higher posterior density than the current iteration will be accepted with probability 1; while proposals that have smaller posterior density than the current iteration will only be accepted according to a probability that depends on the ratio of the posterior densities. Intuitively, the algorithm allocates a higher chance for the Markov Chain to move "uphill", in the meantime also enables the chain to move "downhill", facilitating the exploration of the whole parameter space of the posterior distribution.

Alternatively, the prior distribution can be used as the proposal, that is

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = f(\boldsymbol{\theta}^*),$$

which is independent of the current iteration. Then the acceptance probability simplifies to the ratio of likelihoods:

$$a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = \frac{f(\boldsymbol{\theta}^*|\boldsymbol{d})f(\boldsymbol{\theta}^{i-1})}{f(\boldsymbol{\theta}^{i-1}|\boldsymbol{d})f(\boldsymbol{\theta}^*)} = \frac{f(\boldsymbol{d}|\boldsymbol{\theta}^*)}{f(\boldsymbol{d}|\boldsymbol{\theta}^{i-1})},$$

resulting in another variant of the MH algorithm, the **independence sampler**. Note that this MCMC sampler is only efficient if the prior distribution is a good approximation to the posterior, which is rarely the case. Finally, if any other distribution with no special characteristics is used, then this is called the general **Hastings algorithm**.

2.2.2 Gibbs Sampler

Geman and Geman (1984) was the first paper to formally describe the Gibbs sampling algorithm, and has been widely applied across different fields primarily due to its ease of implementation. In brief, the Gibbs sampler takes advantage of the tractability of the conditional posterior distributions. Sequential sampling from each successive conditional posterior then yields a Markov Chain, which under some mild regular conditions, converges in distribution to the target distribution.

Using similar notation to Section 2.1, suppose we need to simulate from the joint posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{d}) = f(\theta_1, \theta_2, \ldots, \theta_k|\boldsymbol{d})$. Additionally, let $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$, be the set of $\boldsymbol{\theta}$ excluding the i^{th} component, and let $f(\theta_i|, \boldsymbol{d}, \boldsymbol{\theta}_{-i})$ denote the induced full conditional posterior distribution (tractable). A generic Gibbs sampler algorithm proceeds as follows:

1. Initiate the chain with arbitrary values, $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_k^0)$.

2. Set $i = 1$.

3. Perform a sequential draw from the full conditional distributions, i.e.

$$\begin{aligned}
&\text{Draw } \theta_1^i \text{ from } f(\theta_1|\boldsymbol{d}, \boldsymbol{\theta}_{-1}^{i-1});\\
&\text{Draw } \theta_2^i \text{ from } f(\theta_2|\boldsymbol{d}, \theta_1^i, \theta_3^{i-1}, \ldots, \theta_k^{i-1});\\
&\text{Draw } \theta_3^i \text{ from } f(\theta_3|\boldsymbol{d}, \theta_1^i, \theta_2^i, \theta_4^{i-1}, \ldots, \theta_k^{i-1});\\
&\qquad\quad\vdots\\
&\text{Draw } \theta_k^i \text{ from } f(\theta_k|\boldsymbol{d}, \boldsymbol{\theta}_{-k}^i).
\end{aligned}$$

   These complete a newly generated iteration of the algorithm, $\boldsymbol{\theta}^i = (\theta_1^i, \ldots, \theta_k^i)$.

4. Set $i = i + 1$ and repeat step 3 until a desirable amount of iterations, say n, is obtained.

After burning in, the resulting samples can subsequently be regarded as realizations from the joint posterior, $f(\boldsymbol{\theta}|\boldsymbol{d})$. Note that Gibbs sampling is a special case of the MH algorithm with acceptance probability equals to one (by choosing the conditional posterior distributions as the corresponding proposal distributions).

2.3 Our MCMC Strategy

Smith and Roberts (1993) states that a range of hybrid strategies are viable by combining several different chains in various ways. The MCMC method we propose to use is the variable-at-a-time MH algorithm as described by 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. Hence, this MCMC scheme is sometimes referred to as the Metropolis-within-Gibbs algorithm. Moreover, we will mostly be using the random walk MH algorithm (with multivariate normal increments), with the proposal variance matrix properly tuned/chosen to optimise the exploration.

In addition, we will be adopting the idea of blocking wherever possible within our MCMC updating scheme. This is because univariate updating scheme can be rather inefficient when the parameters exhibit high degree of dependence, prohibiting large transitions to parts of the parameter space with high posterior probability (see Roberts and Sahu, 1997 for details). On the contrary, blocking enables the MCMC algorithm to acknowledge the correlation structure of the parameters. Crucially, this allows the algorithm 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. Furthermore, blocking prevents excessive looping in the algorithm, reducing the computational time substantially (especially if the sampling is performed in R). 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. In the extreme case where all the parameters are allocated in a single block, the resulting computations relating to the posterior density is clearly more computationally demanding, particularly in complex problems like ours. Finding efficient proposal distributions for a MH algorithm to simultaneously update every single component of the parameter can also be very difficult. Therefore, our general strategy of blocking is to allocate highly-correlated parameters in a single block such that the correlations between blocks are reduced. Note that efficient block updates using the random walk MH algorithm necessitates clever specification of the proposal variance matrix, which will be discussed in Section 3.6.3.

On a related matter, Roberts and Sahu (1997) commented that the particular sequential order in which the updating schemes within an MCMC algorithm is performed influences its rate of convergence. Specifically, they carried out rigorous examination on the effect of different updating strategies on the efficiency of the Gibbs Sampler in achieving convergence theoretically. Two popular updating orders that are commonly used in practice are the Deterministic Updating Gibbs Sampling (DUGS), where the updating order is fixed throughout the iterations, and the Random Sweep Gibbs Sampling, where the sequence of updating is randomised throughout the sampling algorithm. They demonstrated that each of the updating schemes prevails on different occasions. Nevertheless, this is not our main concern here so we shall adopt the DUGS for simplicity, where the updating sequence is deterministically chosen for our constructed MCMC algorithm.

Chapter 3

Bayesian Mortality Forecasting with Overdispersion

3.1 Overdispersion

The Poisson Lee-Carter (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) 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 backgrounds 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:

$$r_{xt}^2 = \frac{(d_{xt} - \mathbb{E}[D_{xt}])^2}{\mathrm{Var}[D_{xt}]}\bigg|_{\mu_{xt}=\hat{\mu}_{xt}} = \frac{(d_{xt} - e_{xt}\hat{\mu}_{xt})^2}{e_{xt}\hat{\mu}_{xt}}, \tag{3.1}$$

where $\hat{\mu}_{xt} = \exp(\hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t)$ 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 3.1.

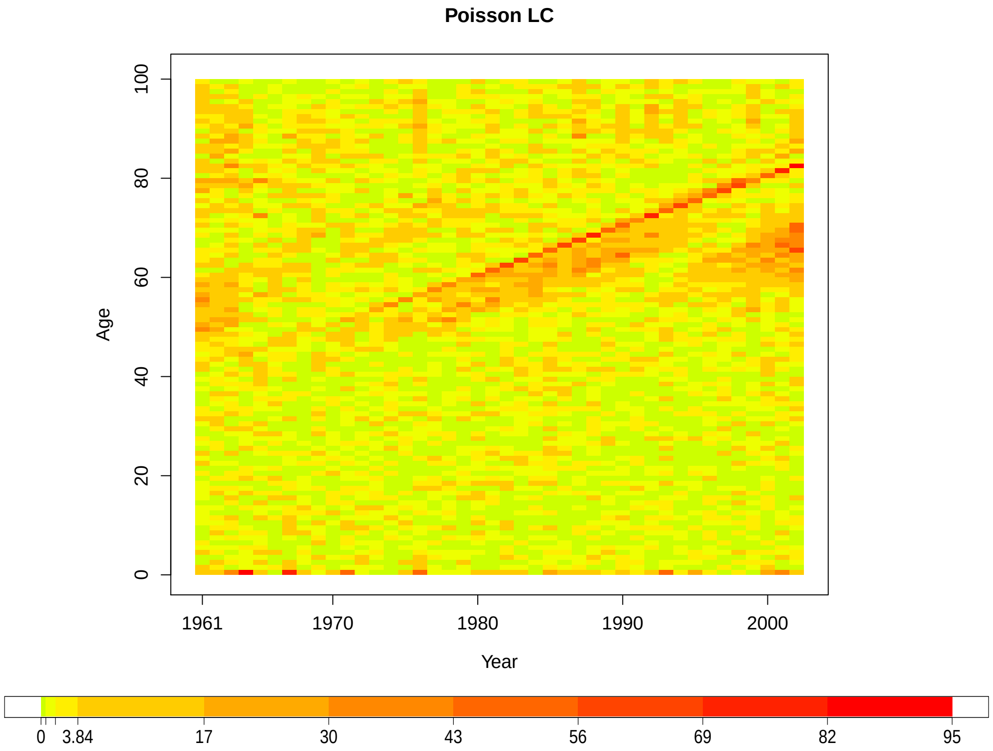


Figure 3.1: Heat map of the squared Pearson residuals of the PLC model, accompanied by the corresponding colour code. Green/yellow rectangular cells indicate areas with good fit, while orage/red coloured cells indicate areas with significantly poor fit.

Under the null hypothesis that the PLC is the true underlying model, and certain mild conditions (e.g. provided that the expected deaths in each age group and year is greater than 5), each r^2_{xt} has an approximate chi-squared distribution with degrees of freedom 1 (χ^2_1) asymptotically. Ideally, we should expect only about 5% of the r^2_{xt} ($AT \times 0.05 = 210$) to be larger than 3.84 (95^{th} percentile of χ^2_1). However, a quick skim through Figure 3.1 shows that there is still a large proportion of orange/red regions, and is especially obvious for the infants and ages above 40. In fact, there is a total of 1044 (about 25%) r^2_{xt} having values greater than 3.84, way exceeding the expected number of 210, suggesting model inadequancy in accounting for extra variations in the data.

In addition, we can also perform the Pearson's chi-squared overall goodness of fit test. In particular, model deviance computed as the sum of r^2_{xt},

$$r^2 = \sum_{x,t} r^2_{xt} = \sum_{x,t} \frac{(d_{xt} - e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x \hat{\kappa}_t))^2}{e_{xt}\exp(\hat{\alpha}_x + \hat{\beta}_x \hat{\kappa}_t)} \tag{3.2}$$

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) derived as

$$\begin{aligned}\text{df} &= \text{Number of observations}-\text{Number of parameters estimated}+\text{Number of constraints}\\ &= AT - A - A - T + 2\\ &= (A-1)(T-2).\end{aligned}$$

In this case, we have $A = 100$ and $T = 42$, so $r^2 \sim \chi^2_{3960}$ under the null hypothesis. Since the model deviance of 15378.73 is substantially larger than the critical value of the conventional chi-squared statistics (i.e. the 95th percentile of χ^2_{3960} is 4107.51), this clearly suggests that PLC model does not provide a satisfactory fit to the data.

Moreover, closer inspection reveals obvious orange/red diagonal lines in the heat map, which corresponds to a cohort effect. Specifically, the red-coloured diagonal line corresponds to the 1919 cohort, and those underneath it corresponds to cohorts born from 1920-1927. Our first sensible guess to the effect of these cohorts is their linkage to the 1918 influenza and the post effects of First World War. Indeed, these have already been thoroughly discussed in Willets (2004), who performed an extensive research on the existence as well as the associated insights of the cohort effect in England and Wales. Our main interest in the first part of the thesis is on overdispersion so, as we shall see in the next few sections, we made no attempt to account for the cohort effect in our proposed models. Nevertheless, incorporation of the cohort effect is investigated in Chapter 6. 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.

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.

3.2 Advantages of Bayesian Mortality Forecasting

The rationale for considering Bayesian methodology is it provides a natural framework in which prior knowledge in the relevant field can be incorporated and various sources of uncertainty can be coherently incorporated to produce a more representative prediction interval. The sources of uncertainty under consideration, as described in Keilman (1990) (see also de Beer, 2000), include:

1. Natural uncertainty: Error due to inherent random variation.

2. Parameter uncertainty: Error due to parameters misestimation.
3. Model uncertainty: Error due to model misspecification.
4. Forecast uncertainty: Uncertainty due to projections.
5. Uncertainty due to expert opinion: uncertain judgements involved in the elicitation of expert knowledge (see O'Hagan et al., 2006).

The classical Lee-Carter (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 the 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 (as described in Section 1.5). 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 3.7).

Furthermore, carefully calibrated percentiles of the posterior predictive distribution carry valuable information necessary to characterise the uncertainties we encounter during forecasting. This is in contrast to the rather ad-hoc construction of scenario forecasting within the classical settings, which involves unnecessary assumptions. In practice, any percentile can be used as a point estimate instead of the posterior mean or median. For instance, the 75th percentile of posterior predictive distribution should be used if we deem underestimation of parameters three times worse than overestimation (see Appendix F). In general, a well-defined loss function should be formulated to penalise misestimation of parameters based on the user's preferences. The appropriate point estimate can then be derived by minimizing the expected posterior loss according to the chosen loss function. The resulting point estimate then accounts for the uncertainties in light of the posterior distribution. Alternatively, probabilistic statement such as, "there is a 10% chance that the projected 10-years ahead mortality rate for age 0 exceeds the 90th percentile of the posterior predictive distribution", is equally reliable (see for example Azose et al., 2016). In short, probabilistic forecasts provide more flexibility to the users in their decision making.

3.3 Mixed PLC 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, forming the mixed PLC models (known interchangeably as overdispersion models).

3.4 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 is added on $\log \mu_{xt}$ for an extra layer of variability in the model:

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

Or equivalently,

$$\begin{aligned}
D_{xt}|\mu_{xt} &\sim \text{Poisson}(e_{xt}\mu_{xt}) \\
\log \mu_{xt}|\alpha_x, \beta_x, \kappa_t, \sigma^2_\mu &\sim N(\alpha_x + \beta_x\kappa_t, \sigma^2_\mu).
\end{aligned}$$

In essence, σ^2_μ 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.

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

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 the 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

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

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.

3.4.1 Priors

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-driven inference. In addition, we also attempt to be indifferent in terms of prior specification under both overdispersion models to facilitate model comparison later on. It is also worth mentioning our prior specification differs from those suggested by Czado et al. (2005), who used the empirical Bayes approach in formulating their non-informative priors.

3.4.1.1 Prior Distributions for α_x, β_x, σ_β^2, σ_μ^2, and ϕ

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 on the unidentified β parameters, which automatically yields proper posterior inference, as stated by Gelfand and Sahu (1999). 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 expected more erratic behaviour as compared to α_x.

As pointed out in Section 3.4, σ_μ^2 serves as the dispersion parameter. 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. It is worth noting that we investigated the dependence of the posterior inferences on the priors informally by a sensitivity analysis (see for example Gelman et al., 1995). In particular, we found that a choice of different constants for the hyperpriors, say the use of $a_\mu = b_\mu = 0.01$ instead, does not result in a substantial change in the posterior inferences (primarily because of the size of our mortality data, where the priors are dominated by the likelihood).

3.4.1.2 Prior Distribution for κ_t

κ_t represents the overall mortality level at time t, which forms the crucial element for stochastic forecasts. For reasons mentioned in Section 2.1, 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,

$$\begin{cases} \kappa_t - \eta_t = \rho(\kappa_{t-1} - \eta_{t-1}) + \epsilon_t, & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \eta_1 + \epsilon_1 \end{cases}, \tag{3.3}$$

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 (3.3) 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. Note that the choice of which ARIMA model to use is essentially a model selection problem. Thus, it is possible to consider specification of various ARIMA models on κ_t. Posterior model probabilities can then be computed to serve as a criterion for model determination. However, this is not our main focus here. 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

$$\begin{cases} \boldsymbol{\kappa}_{-1} \sim N(\boldsymbol{Y}_{-1}\boldsymbol{\psi} - \rho \boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi}, \sigma_\kappa^2 \boldsymbol{Q}^{-1}) \\ \kappa_1 = 0 \end{cases}, \tag{3.4}$$

where $\boldsymbol{\kappa}_{-1} = (\kappa_2, \kappa_3, \ldots, \kappa_T)^\top$, $\boldsymbol{\psi} = (\psi_1, \psi_2)^\top$, $\boldsymbol{Q} = \boldsymbol{R}^\top \boldsymbol{R}$, $\boldsymbol{R} = \boldsymbol{I}_{T-1} - \boldsymbol{P}$,

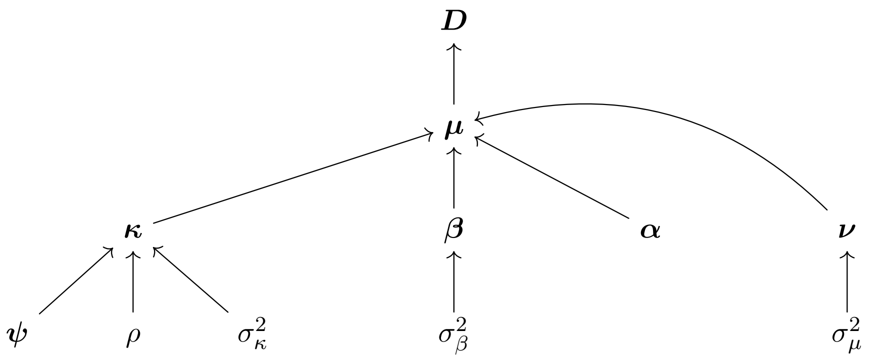
$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \rho & 0 & & & \vdots \\ 0 & \rho & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho & 0 \end{pmatrix}_{(T-1)\times(T-1)}, \boldsymbol{Y}_{-1} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ \vdots & \vdots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

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 3.4.2.

The resulting model can be represented by the following directed graphical diagram (Lunn et al., 2013).



3.4.2 MCMC Scheme for the PLNLC Model

Due to the model structure and the prior distributions specified, the conditional posterior distributions of all of the parameters can be conveniently recognised as standard distributions, except for the μ_{xt}. This is because the log-normal prior on the mortality rates, μ_{xt}, is non-conjugate with respect to the Poisson likelihood. Streftaris and Worton (2008) proposed to apply an approximation in the form of a mixture of log-normal and gamma densities on the conditional posterior distributions of μ_{xt}, where they then demonstrated that this leads to more accurate and efficient inferences than the exact methods (for a given sample size) as measured by the Bayes risk. However, given that the approximation has to be applied on each and every μ_{xt}, it is rather unlikely that their method would improve the computational efficiency in our case as the number of μ_{xt} involved is huge (4200 in total). Hence, we choose to implement the exact method using MH algorithm for updating the μ_{xt}. The MCMC updating scheme under the PLNLC model can then be easily implemented by iterating through a series of Gibbs steps, together with some MH steps for the remaining μ_{xt}.

We describe in detail the MH steps for the μ_{xt} in the next subsection. The conditional posterior distribution of the rest of the parameters then follows. For a full derivation of the conditional posterior distributions/densities for this model, please refer to Appendix A.

3.4.2.1 MH Step for μ_{xt}

For technical reasons, it is more convenient to work with μ_{xt} on the logarithmic scale. Each $\log \mu_{xt}$ is updated univariately using a random walk MH algorithm. 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

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

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,

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

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 $\sigma_{\mu_{xt}}^2$ can be numerically determined and are depicted in Figure 3.2.

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).

In this case, notice that the MH sampling has to be performed for each and every age group and time successively, which can be computationally inefficient. One possible solution to this is to allocate all $\log \mu_{xt}$ in a single block. The MH update of $\log \mu_{xt}$ then proceeds simultaneously in a single step using a multivariate proposal distribution. However, we dismiss blocking here due to the immense dimensionality involved.

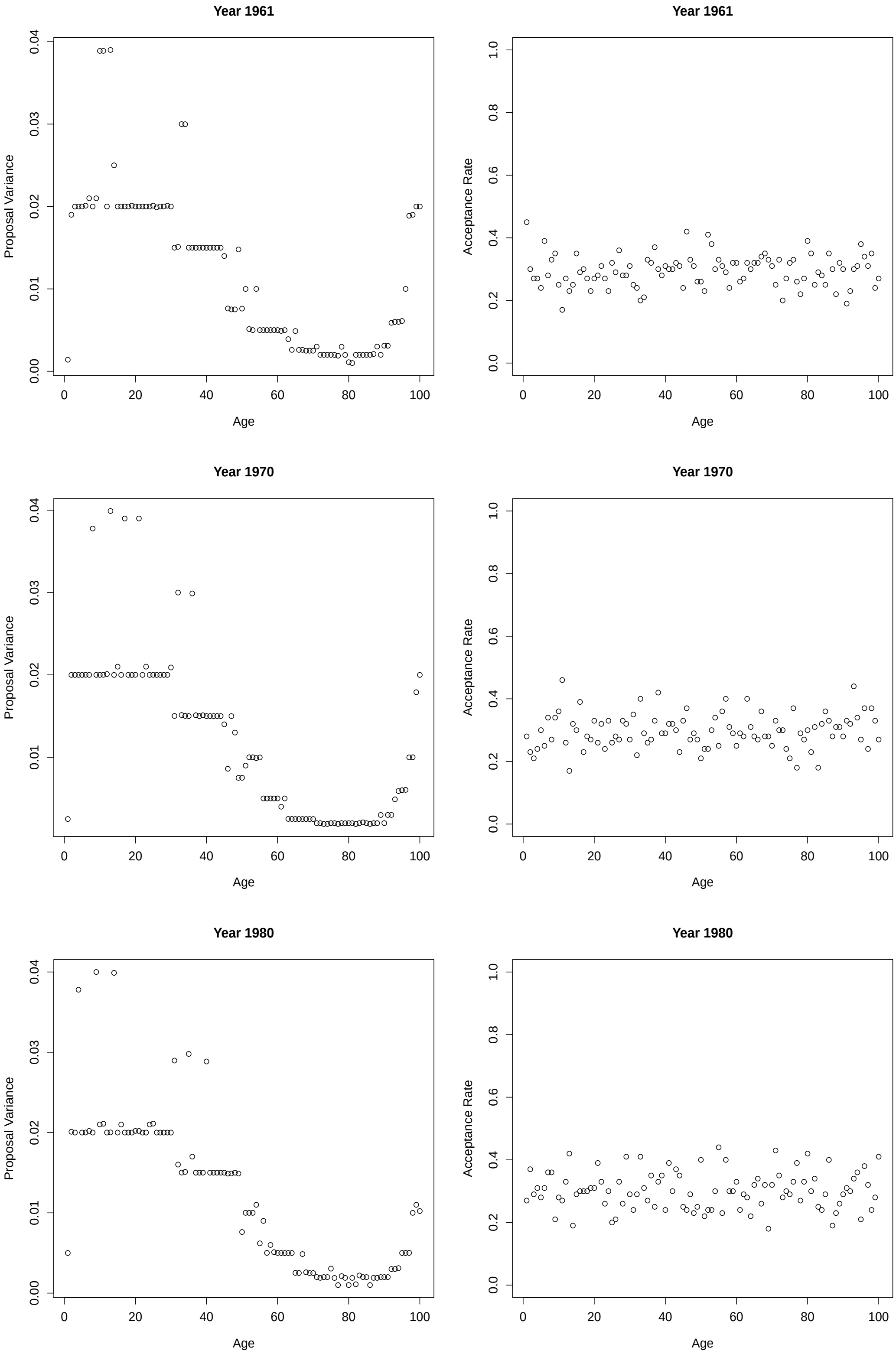


Figure 3.2: Plots of proposal variances, $\sigma^2_{\mu_{xt}}$ (left panels) and the corresponding acceptance rates of μ_{xt} (right panels) for years 1961, 1970 and 1980.

3.4.2.2 Gibbs Step for κ_t

Denote $\boldsymbol{\mu}_t = (\mu_{1t}, \mu_{2t}, \ldots, \mu_{At})^\top$ as the mortality rates corresponding to year t and $\boldsymbol{\kappa}_{-1,t} = (\kappa_2, \ldots, \kappa_{t-1}, \kappa_{t+1}, \ldots, \kappa_T)^\top$. The conditional posterior distribution of κ_t is then given as

$$\kappa_t | \boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2 \sim \begin{cases} N(\mu_\kappa^*, (\sigma_\kappa^*)^2) & \text{for } 1 < t < T, \\ N(\mu_\kappa', (\sigma_\kappa')^2) & \text{for } t = T, \end{cases}$$

where

$$\begin{aligned} (\sigma_\kappa^*)^2 &= \left(\frac{\sum_x \beta_x^2}{\sigma_\mu^2} + \frac{1+\rho^2}{\sigma_\kappa^2} \right)^{-1}, \\ \mu_\kappa^* &= (\sigma_\kappa^*)^2 \times \left[\frac{\sum_x \beta_x (\log \mu_{xt} - \alpha_x)}{\sigma_\mu^2} + \frac{\eta_t + \rho(\kappa_{t-1} - \eta_{t-1}) + \rho(\kappa_{t+1} - \eta_{t+1} + \rho\eta_t)}{\sigma_\kappa^2} \right], \\ (\sigma_\kappa')^2 &= \left(\frac{\sum_x \beta_x^2}{\sigma_\mu^2} + \frac{1}{\sigma_\kappa^2} \right)^{-1}, \\ \mu_\kappa' &= (\sigma_\kappa')^2 \times \left[\frac{\sum_x \beta_x (\log \mu_{xt} - \alpha_x)}{\sigma_\mu^2} + \frac{\eta_t + \rho(\kappa_{t-1} - \eta_{t-1})}{\sigma_\kappa^2} \right]. \end{aligned}$$

3.4.2.3 Gibbs Step for β_x

Denoting $\boldsymbol{\beta}_{-1,x} = (\beta_2, \ldots, \beta_{x-1}, \beta_{x+1}, \ldots, \beta_A)^\top$ as a vector of β excluding the 1st and $x - th$ components, we have

$$\begin{cases} \beta_x | \boldsymbol{\beta}_{-1,x}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \rho, \boldsymbol{\psi}, \sigma_\mu^2 \sim N(\mu_\beta^*, (\sigma_\beta^*)^2) & \text{for } 2 \leq x \leq A, \\ \beta_1 = 1 - \beta_2 - \ldots - \beta_A, \end{cases}$$

where

$$\begin{aligned} (\sigma_\beta^*)^2 &= \left(\frac{2\sum_t \kappa_t^2}{\sigma_\mu^2} + \frac{2}{\sigma_\beta^2} \right)^{-1}, \\ \mu_\beta^* &= (\sigma_\beta^*)^2 \times \left[\frac{\sum_t \kappa_t (\log \mu_{xt} - \alpha_x)}{\sigma_\mu^2} + \frac{\sum_t \kappa_t (-\log \mu_{1t} + \alpha_1 + \kappa_t (1 - \sum_{i \neq 1,x} \beta_i))}{\sigma_\mu^2} \right. \\ &\quad \left. + \frac{1 - \sum_{i \neq 1,x} \beta_i}{\sigma_\beta^2} \right]. \end{aligned}$$

3.4.2.4 Gibbs Step for α_x

Define $\boldsymbol{\alpha}_{-x} = (\alpha_1, \ldots, \alpha_{x-1}, \alpha_{x+1}, \ldots, \alpha_A)^\top$ as a vector of α without its $x - th$ component, then

$$\alpha_x | \boldsymbol{\alpha}_{-x}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2 \sim N(\alpha_x^*, (\sigma_\alpha^*)^2),$$

where

$$\begin{aligned}
(\sigma_\alpha^*)^2 &= \left(\frac{T}{\sigma_\mu^2} + \frac{1}{\sigma_\alpha^2}\right)^{-1}, \\
\alpha_x^* &= (\sigma_\alpha^*)^2 \times \left(\frac{\sum_t \log \mu_{xt} - \beta_x \sum_t \kappa_t}{\sigma_\mu^2} + \frac{\alpha_0}{\sigma_\alpha^2}\right) \\
&= \frac{\sigma_\mu^2 \alpha_0 + \sigma_\alpha^2 (\sum_t \log \mu_{xt} - \beta_x \sum_t \kappa_t)}{\sigma_\mu^2 + T\sigma_\alpha^2}.
\end{aligned}$$

3.4.2.5 Gibbs Step for σ_μ^2

The distribution of σ_μ^2 given the rest of the parameters is given by

$$\begin{aligned}
&\sigma_\mu^{-2} | \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \rho, \boldsymbol{\psi} \\
\sim\ &\text{Gamma}\left(a_\mu + \frac{AT}{2}, b_\mu + \frac{1}{2}\sum_{x,t}(\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2\right).
\end{aligned}$$

3.4.2.6 Gibbs Step for σ_κ^2

The conditional posterior distribution of σ_κ^2 can be written as

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

3.4.2.7 Gibbs Step for σ_β^2

The distribution of σ_β^2 given the rest of the parameters is similarly given by

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

3.4.2.8 Gibbs Step for ρ

The conditional posterior distribution of ρ is

$$\rho | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \sigma_\mu^2 \sim N(\mu_\rho^*, (\sigma_\rho^*)^2),$$

where

$$\begin{aligned}
(\sigma_\rho^*)^2 &= \frac{\sigma_\kappa^2}{a_\rho + \frac{\sigma_\kappa^2}{\sigma_\rho^2}}, \\
\mu_\rho^* &= \frac{b_\rho}{a_\rho + \frac{\sigma_\kappa^2}{\sigma_\rho^2}}, \\
a_\rho &= \sum_{t=2}^{T} (\kappa_{t-1} - \eta_{t-1})^2, \\
b_\rho &= \sum_{t=2}^{T} (\kappa_t - \eta_t)(\kappa_{t-1} - \eta_{t-1}).
\end{aligned}$$

3.4.2.9 Gibbs Step for $\boldsymbol{\psi}$

Similar derivation yields

$$\boldsymbol{\psi} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \rho, \sigma_\mu^2 \sim N_2(\boldsymbol{\psi}^*, \boldsymbol{\Sigma}_\psi^*), \tag{3.5}$$

where

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

3.4.3 Poisson Log-Normal LC Model with Blocking

Since the conditional posterior distributions of α_x, β_x and κ_t are univariate normal distributions, this further motivates blocking of parameters as multivariate normal distributions are straightforward to work with. A sensible blocking strategy here is to allocate the α_x, β_x and κ_t each in a separate block, and the rest of the parameters updated univariately. It is worth mentioning that allocating all of α_x, β_x and κ_t in a single block is disadvantageous because the resulting conditional posterior distribution will no longer be a normal distribution due to the multiplicative bilinear term, $\beta_x\kappa_t$. The conditional posterior distribution of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{-1}$ and $\boldsymbol{\kappa}_{-1}$ are presented below, with the associated derivations provided in Appendix B.

3.4.3.1 Gibbs Step for $\boldsymbol{\alpha}$

The conditional posterior distribution of $\boldsymbol{\alpha}$ is given by

$$(\boldsymbol{\alpha} | \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log \boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \sim N_A(\boldsymbol{\mu}_\alpha^*, \boldsymbol{\Sigma}_\alpha^*),$$

where

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

3.4.3.2 Gibbs Step for $\boldsymbol{\beta}_{-1}$

First, denote $\boldsymbol{\mu}_x=(\mu_{x1},\mu_{x2},\ldots,\mu_{xT})^\top$ as the vector of mortality rates corresponding to age group x, and $\boldsymbol{\mu}_{-x,t}=(\mu_{1t},\ldots,\mu_{x-1\,t},\mu_{x+1\,t},\ldots,\mu_{At})^\top$ as the vector of mortality rates corresponding to year t, but excluding the $x-th$ component. Then, we have

$$\begin{cases}\boldsymbol{\beta}_{-1}|\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma^2_\kappa,\sigma^2_\beta,\boldsymbol{\psi},\rho,\sigma^2_\mu \sim N_{A-1}(\boldsymbol{\mu}^*_\beta,\boldsymbol{\Sigma}^*_\beta),\\ \beta_1=1-\beta_2-\ldots-\beta_A,\end{cases}$$

where

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

3.4.3.3 Gibbs Step for $\boldsymbol{\kappa}_{-1}$

Similarly, we denote $\boldsymbol{\mu}_{x,-t}=(\mu_{x1},\ldots,\mu_{x\,t-1},\mu_{x\,t+1},\ldots,\mu_{xT})^\top$ as the vector of mortality rates corresponding to age group x, excluding year t. The resulting conditional posterior distribution is given by

$$\boldsymbol{\kappa}_{-1}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma^2_\kappa,\sigma^2_\beta,\boldsymbol{\psi},\rho,\sigma^2_\mu\sim N_{T-1}(\boldsymbol{\mu}^*_\kappa,\boldsymbol{\Sigma}^*_\kappa),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}^*_\kappa &= \left[\frac{\sum_x\beta_x^2}{\sigma^2_\mu}\boldsymbol{I}_{T-1}+\frac{1}{\sigma^2_\kappa}\boldsymbol{Q}\right]^{-1},\\ \boldsymbol{\mu}^*_\kappa &= \boldsymbol{\Sigma}^*_\kappa\times\left[\frac{1}{\sigma^2_\mu}\sum_x\beta_x(\log\boldsymbol{\mu}_{x,-1}-\alpha_x\mathbf{1}_{T-1})+\frac{1}{\sigma^2_\kappa}\boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\psi}+\boldsymbol{R}^{-1}\boldsymbol{a})\right],\\ \boldsymbol{a} &= (\rho\eta_1,0,\ldots,0)^\top.\end{aligned}$$

3.5 Poisson-Gamma LC (PGLC) Model

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

$$\begin{aligned} D_{xt}|\mu_{xt} &\overset{\text{ind}}{\sim} \text{Poisson}(e_{xt}\mu_{xt}), \\ \mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi &\sim \text{Gamma}\left(\phi, \frac{\phi}{\exp(\alpha_x + \beta_x\kappa_t)}\right). \end{aligned}$$

Or equivalently,

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

where $\phi > 0$ 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.

The likelihood function also consists of two parts, that is

$$\text{Likelihood} = f(\boldsymbol{d}|\boldsymbol{\mu}) \times f(\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \phi),$$

with,

i.

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

ii.

$$
\begin{aligned}
& f(\boldsymbol{\mu}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa},\phi) \\
= & \prod_{x,t} \frac{\left[\frac{\phi}{\exp(\alpha_x+\beta_x\kappa_t)}\right]^{\phi}}{\Gamma(\phi)} \mu_{xt}^{\phi-1} \exp\left[-\frac{\phi\mu_{xt}}{\exp(\alpha_x+\beta_x\kappa_t)}\right] \\
= & \frac{\phi^{AT\phi}}{\Gamma(\phi)^{AT}\exp(T\phi\sum_x \alpha_x + \phi\sum_t \kappa_t)} \left(\prod_{x,t}\mu_{xt}\right)^{\phi-1} \exp\left[-\phi\sum_{x,t}\frac{\mu_{xt}}{\exp(\alpha_x+\beta_x\kappa_t)}\right]. \\
& (\text{because } \sum_x \beta_x = 1)
\end{aligned}
$$

3.5.1 Prior Distributions for the PGLC Model

To facilitate model comparison later, we impose the same prior distributions as before for all the parameters of this model, except for ϕ (refer to Section 3.4.1). 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 PGLC model, and ignoring the variabilities due to α_x, β_x, and κ_t, we have

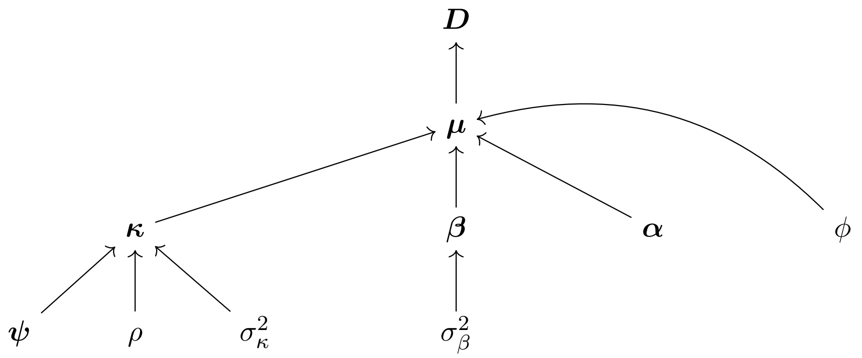
$$
\mathrm{Var}(\log\mu_{xt}) = \mathrm{Var}(\log\nu_{xt}) \approx \left(\frac{d\log z}{dz}\right)^2\Bigg|_{z=\mathbb{E}(\nu_{xt})} \times \mathrm{Var}(\nu_{xt}) = \frac{1}{\phi}.
$$

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

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

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

The Bayesian hierarchical model specified above can be summarised by the directed graphical diagram as follows.



3.5.2 MCMC Scheme for the PGLC Model

Due to the conjugacy relationship between a Poisson likelihood and a gamma distribution, the conditional posterior distribution of the mortality rate, μ_{xt}, is tractable and hence can be updated using the Gibbs algorithm. On the other hand, we apply the random walk MH algorithm on α_x, β_x and κ_t instead because the normal prior distributions are no longer conditionally conjugate. The random walk MH updating is performed univariately without blocking. Nevertheless, the Gibbs steps for $\rho, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}$ are unaffected (refer to Section 3.4.2) because they belong to the lower part of the hierarchical model, and hence, their conditional posterior distributions remain the same conditional upon α_x, β_x and κ_t.

3.5.2.1 Gibbs Step for μ_{xt}

Denoting $\boldsymbol{\mu}_{-xt}$ as the vector of mortality rates excluding the $xt-th$ component (see Section 3.4.2.1 for an expression of it), the conditional posterior distribution of μ_{xt} is simply

$$\mu_{xt}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\boldsymbol{\mu}_{-xt},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\rho,\phi \sim \text{Gamma}(d_{xt}+\phi, e_{xt}+\frac{\phi}{\exp(\alpha_x+\beta_x\kappa_t)}).$$

3.5.2.2 MH Step for κ_t

Following similar procedures, the conditional posterior density of κ_t up to a proportionality constant is

i. For $t=T$,

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

ii. For $2<t<T$,

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

A candidate is proposed from a normal distribution centered at the current iteration,

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

with appropriately tuned proposal variance, $\sigma^2_{\kappa_t}$. The acceptance probability is given by the following:

i. For $t = T$,

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

ii. For $1 < t < T$,

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

Using the search algorithm described in Section 3.4.2, the selected set of values of $\sigma^2_{\kappa_t}$ and the consequent acceptance rates are given conveniently in Figure 3.3.

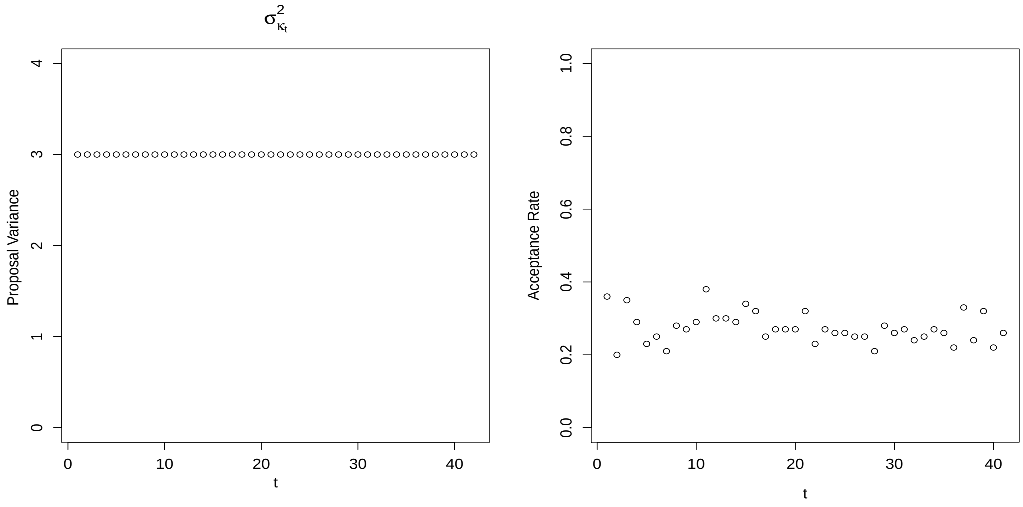


Figure 3.3: Plots of proposal variances, $\sigma^2_{\kappa_t}$ and the corresponding acceptance rates of κ_t.

As evident from Figure 3.3, proposal variances of κ_t, $\sigma^2_{\kappa_t}$, are strikingly identical across the years, signifying that the marginal posterior variances of κ_t under this model are rather similar. This is in contrast to $\log\mu_{xt}$, where their proposal variances vary substantially across age and time.

3.5.2.3 MH Step for α_x

The conditional posterior density of α_x can be expressed as

$$\begin{aligned} & f(\alpha_x|\boldsymbol{\alpha}_{-x}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \boldsymbol{\mu}, \rho, \sigma^2_\kappa, \sigma^2_\beta, \boldsymbol{\psi}, \phi) \\ \propto \quad & \exp\left\{ -T\phi\alpha_x - \phi\sum_t \frac{\mu_{xt}}{\exp(\alpha_x + \beta_x\kappa_t)} - \frac{1}{2\sigma^2_\alpha}(\alpha_x - \alpha_0)^2 \right\}. \end{aligned}$$

With a proposal for $\alpha^*_x \sim N(\alpha^{i-1}_x, \sigma^2_{\alpha_x})$ in each iteration, the acceptance probability of α_x for each age groups is given by

$$\begin{aligned} a(\alpha^*_x|\alpha^{i-1}_x) \quad = \quad & \min\left\{1, \exp\left[-T\phi(\alpha^*_x - \alpha^{i-1}_x) - \phi\left(\sum_t \frac{\mu_{xt}}{\exp(\beta_x\kappa_t)}(e^{-\alpha^*_x} - e^{\alpha^{i-1}_x}) \right) \right.\right. \\ & \left.\left. - \frac{1}{2\sigma^2_\alpha}[(\alpha^*_x - \alpha_0)^2 - (\alpha^{i-1}_x - \alpha_0)^2] \right]\right\}. \end{aligned}$$

The numerically determined proposal variances, $\sigma^2_{\alpha_x}$, and the corresponding acceptance rates for a 100-iterations pilot run are provided in Figure 3.4.

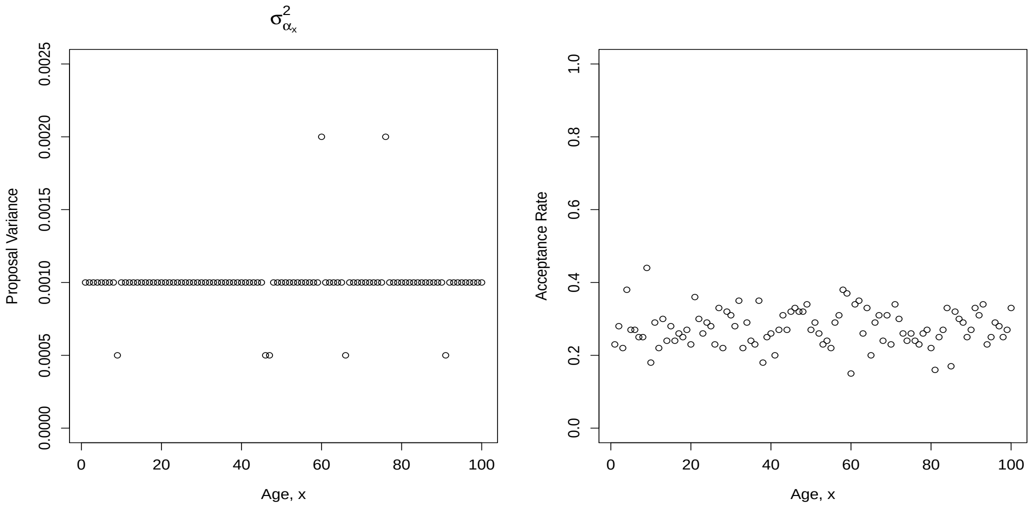


Figure 3.4: Plots of proposal variances, $\sigma^2_{\alpha_x}$ and the corresponding acceptance rates of α_x.

Surprisingly, the age pattern displayed by $\sigma^2_{\alpha_x}$ under the PGLC model is insensitive to age, indicating that the random walk MH for α_x is not very responsive to the choice of proposal variances.

3.5.2.4 MH Step for β_x

The conditional posterior density of β_x can be expressed as

$$f(\beta_x|\boldsymbol{\beta}_{-1,x},\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\boldsymbol{\mu},\rho,\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\phi) \propto \exp\left\{-\phi\sum_t \frac{\mu_{xt}}{\exp(\alpha_x+\beta_x\kappa_t)}\right. \\ \left. -\phi\sum_t\frac{\mu_{1t}}{\exp\left[\alpha_1+\kappa_t(1-\sum_{j\neq 1}\beta_j)\right]} - \frac{1}{2\sigma_\beta^2}\left(1-\sum_{j\neq1}\beta_j\right)^2 - \frac{1}{2\sigma_\beta^2}\beta_x^2\right\}.$$

With a proposal for $\beta_x^* \sim N(\beta_x^{i-1},\sigma_{\beta_x}^2)$ in each iteration, the acceptance probability of β_x for each age groups is given by

$$a(\beta_x^*|\beta_x^{i-1}) \;=\; \min\left\{1, \frac{\exp\left[-\phi\sum_t\frac{\mu_{xt}}{\exp(\alpha_x+\beta_x^*\kappa_t)} - \phi\sum_t\frac{\mu_{1t}}{\exp[\alpha_1+\kappa_t(1-\beta_x^*-\sum_{j\neq1,x}\beta_j)]}\right]}{\exp\left[-\phi\sum_t\frac{\mu_{xt}}{\exp(\alpha_x+\beta_x^{i-1}\kappa_t)} - \phi\sum_t\frac{\mu_{1t}}{\exp[\alpha_1+\kappa_t(1-\beta_x^{i-1}-\sum_{j\neq1,x}\beta_j)]}\right]}\right. \\ \left. \times\frac{\exp\left[-\frac{1}{2\sigma_\beta^2}\left(1-\beta_x^*-\sum_{j\neq1,x}\beta_j\right)^2-\frac{1}{2\sigma_\beta^2}(\beta_x^*)^2\right]}{\exp\left[-\frac{1}{2\sigma_\beta^2}\left(1-\beta_x^{i-1}-\sum_{j\neq1,x}\beta_j\right)^2-\frac{1}{2\sigma_\beta^2}(\beta_x^{i-1})^2\right]}\right\}.$$

The numerically determined proposal variances, $\sigma_{\beta_x}^2$ and the corresponding acceptance rates for a 100-iterations pilot run are provided in Figure 3.5.

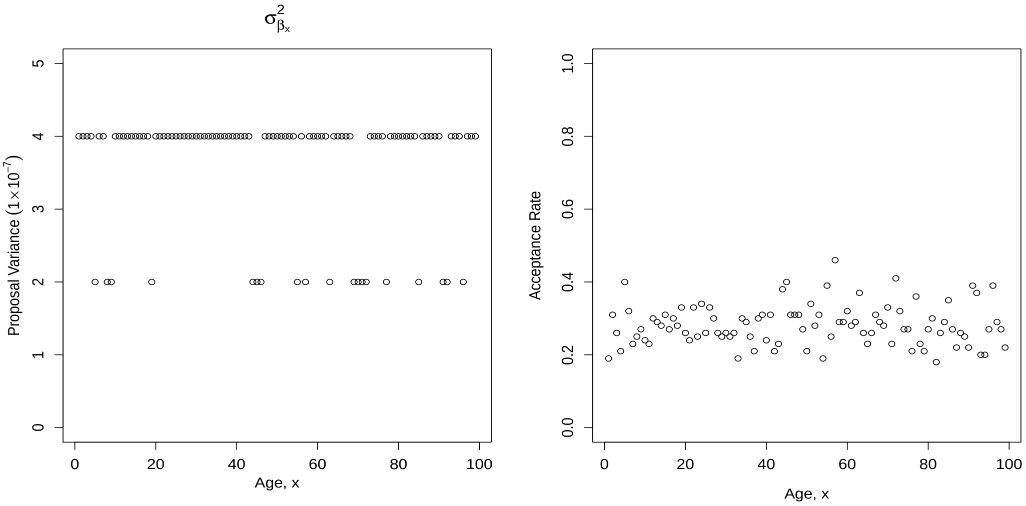


Figure 3.5: Plots of proposal variances, $\sigma_{\beta_x}^2$ and the corresponding acceptance rates of β_x.

3.5.2.5 MH Step for the Dispersion Parameter, ϕ

Using the prior distribution $\phi \sim \text{Gamma}(a_\phi, b_\phi)$, we find that

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

Once again, under the random walk MH algorithm, a candidate ϕ^* is proposed from

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

in each iteration, and is accepted with probability

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

where ϕ^{i-1} is the current iteration of ϕ. A proposal variance tuned at $\sigma_\phi^2 = 0.02$ yields an acceptance rate of approximately 0.27 for a 100-iteration pilot run.

3.6 PGLC Model with μ_{xt} Integrated Out (Negative Binomial LC Model)

One attractive feature about the PGLC model is that the latent variables, μ_{xt}, can be conveniently integrated out, producing its equivalent version, which we refer to as the Negative Binomial LC (NBLC) model. That is,

$$D_{xt}|\alpha_x, \beta_x, \kappa_t, \phi \sim \text{ Neg-Bin } \left(\phi, \frac{\phi}{e_{xt}\exp(\alpha_x + \beta_x\kappa_t) + \phi}\right), \tag{3.6}$$

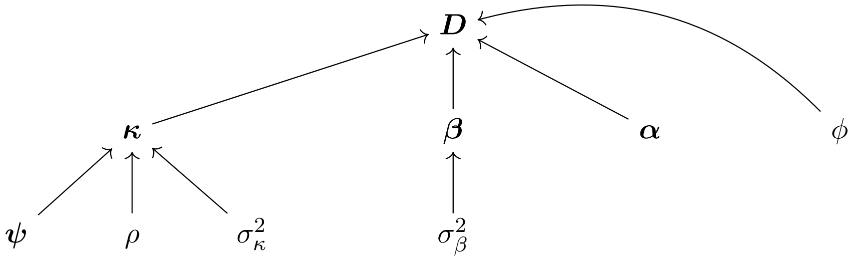
where the likelihood function now consists of only one part:

$$\begin{aligned}
& f(d_{xt}|\alpha_x, \beta_x, \kappa_t, \phi) \\
= \quad & \int_0^\infty f(d_{xt}|\mu_{xt}, \alpha_x, \beta_x, \kappa_t, \phi) f(\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi) d\mu_{xt} \\
= \quad & \frac{\Gamma(d_{xt} + \phi)}{\Gamma(\phi)\Gamma(d_{xt} + 1)} \left[\frac{e_{xt}\exp(\alpha_x + \beta_x\kappa_t)}{e_{xt}\exp(\alpha_x + \beta_x\kappa_t) + \phi}\right]^{d_{xt}} \left[\frac{\phi}{e_{xt}\exp(\alpha_x + \beta_x\kappa_t) + \phi}\right]^\phi.
\end{aligned}$$

The joint likelihood is just the product across age and time due to independence:

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

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 to deal with a slightly more complicated likelihood function. Note that this model has already been considered by Delwarde et al. (2007), but within the classical framework. The directed acyclic graphical diagram for this model is illustrated below.



3.6.1 MCMC Scheme for the NBLC Model

Due to the fact that the likelihood function is now a negative binomial distribution, the conditional posterior distribution of all of α_x, β_x, and κ_t will no longer show (conditional) conjugacy with the prior distributions imposed. This necessitates the use of random walk MH algorithms for the updating of these parameters. We will be considering both univariate updating and block updating of α_x, β_x and κ_t for this model. Note again that the conditional posterior distributions of σ^2_κ, σ^2_β, $\boldsymbol{\psi}$, and ρ are unaffected.

3.6.1.1 MH Step for κ_t

The conditional posterior densities of κ_t are given by:

i. For $t = T$,

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

ii. For $1 < t < T$,

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

The acceptance probability can be straightforwardly calculated as

$$
a(\kappa_t^*|\kappa_t^{i-1}) = \min\left\{ 1, \frac{f(\kappa_t^*|\boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi)}{f(\kappa_t^{i-1}|\boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi)} \right\},
$$

where $\kappa_t^* \sim N(\kappa_t^{i-1}, \sigma_{\kappa_t}^2)$ is the random walk proposal. Our chosen set of values for the proposal variances, $\sigma_{\kappa_t}^2$, and the corresponding acceptance rates are illustrated in Figure 3.6.

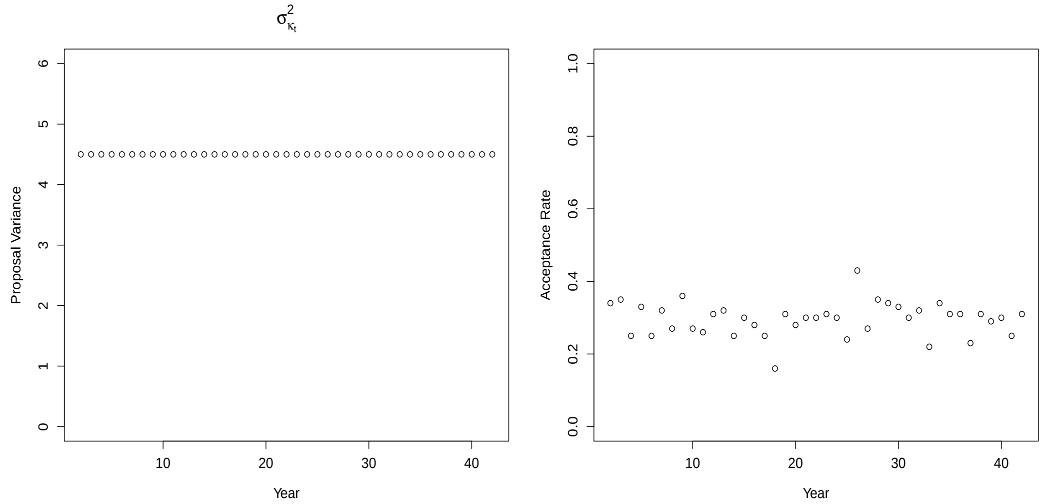


Figure 3.6: Plots of proposal variances, $\sigma_{\kappa t}^2$ and the corresponding acceptance rates of κ_t.

As before, the proposal variances of κ_t are all the same, indicating that their marginal posterior variances are rather similar.

3.6.1.2 MH Step for α_x

The conditional posterior densities of α_x can be written as

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

with acceptance probabilities given by

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

According to Figure 3.7, the proposal variances of α_x, $\sigma_{\alpha_x}^2$, demonstrate a rather similar age pattern to $\sigma_{\mu_{xt}}^2$ at any given time (see Figure 3.2). This is perhaps not so surprising since α_x represent log mortality rates in the base year.

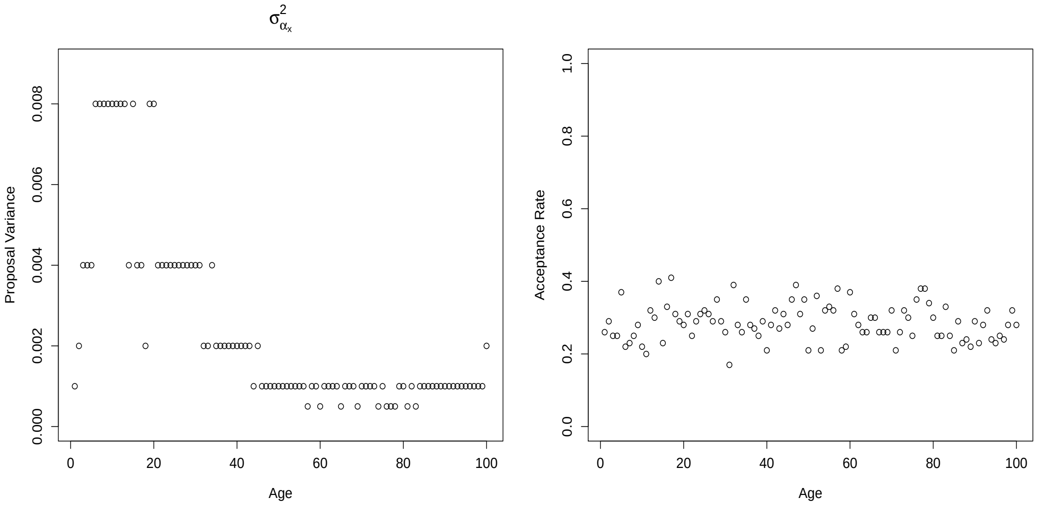


Figure 3.7: Plots of proposal variances, $\sigma_{\alpha_x}^2$ and the corresponding acceptance rates of α_x.

3.6.1.3 MH Step for β_x

The conditional posterior densities of β_x for $x = 2, \ldots, A$ are provided below:

$$\begin{aligned} & f(\beta_x|\boldsymbol{\beta}_{-1,x}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi) \\ \propto\ & \prod_t \left\{\frac{\exp(d_{xt}\beta_x\kappa_t)}{[e_{xt}\exp(\alpha_x + \beta_x\kappa_t) + \phi]^{d_{xt}+\phi}} \times \frac{\exp(d_{1t}\beta_1\kappa_t)}{[e_{1t}\exp(\alpha_1 + \beta_1\kappa_t) + \phi]^{d_{1t}+\phi}}\right\} \\ & \times \exp\left[-\frac{1}{2\sigma_\beta^2}\beta_1^2\right]\exp\left(-\frac{1}{2\sigma_\beta^2}\beta_x^2\right), \end{aligned}$$

where we recall that $\beta_1 = 1 - \beta_2 - \ldots - \beta_A$. The corresponding acceptance probabilities of a random walk proposal $\beta_x^* \sim N(\beta_x^{i-1}, \sigma_{\beta_x}^2)$ for our MH steps are then

$$\begin{aligned} & a(\beta_x^*|\beta_x^{i-1}) \\ = \; & \min\left\{1, \exp\left[(\beta_x^* - \beta_x^{i-1})\sum_t (d_{xt}\kappa_t) + (\beta_1^* - \beta_1^{i-1})\sum_t (d_{1t}\kappa_t) - \frac{(\beta_x^*)^2 - (\beta_x^{i-1})^2}{2\sigma_\beta^2}\right.\right. \\ & \left. - \frac{(\beta_1^*)^2 - (\beta_1^{i-1})^2}{2\sigma_\beta^2}\right] \times \prod_t \left[\frac{e_{xt}\exp(\alpha_x + \beta_x^{i-1}\kappa_t) + \phi}{e_{xt}\exp(\alpha_x + \beta_x^*\kappa_t) + \phi}\right]^{d_{xt}+\phi} \\ & \left. \times \prod_t \left[\frac{e_{1t}\exp(\alpha_1 + \beta_1^{i-1}\kappa_t) + \phi}{e_{1t}\exp(\alpha_1 + \beta_1^*\kappa_t) + \phi}\right]^{d_{1t}+\phi}\right\}. \end{aligned}$$

The numerically determined proposal variances, $\sigma_{\beta_x}^2$, for β_x to achieve acceptance rates that are within the benchmark range are shown in Figure 3.8.

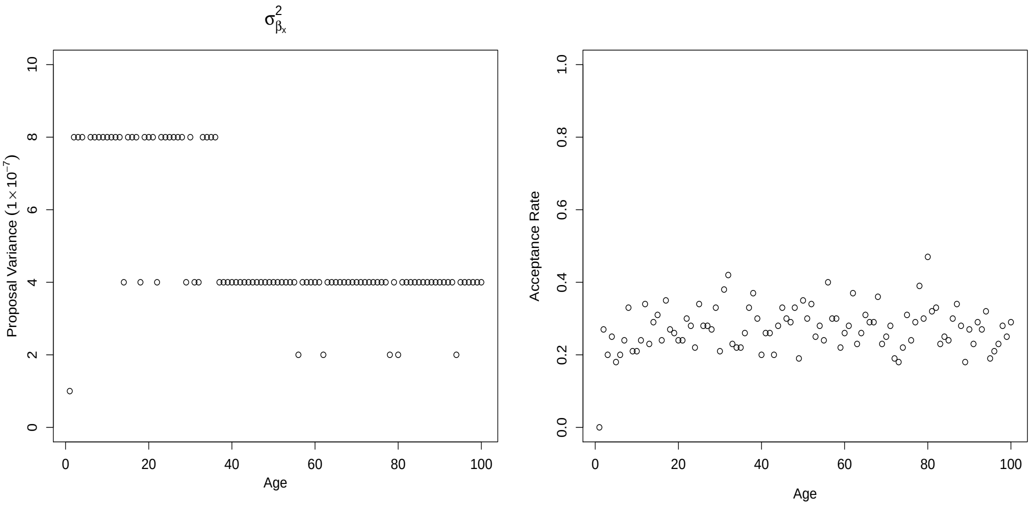


Figure 3.8: Plots of proposal variances, $\sigma_{\beta_x}^2$ and the corresponding acceptance rates of β_x.

Despite being rather similar, the age pattern exhibited by $\sigma_{\beta_x}^2$ is less sensitive to age than those of the $\log \mu_{xt}$ as well as $\sigma_{\alpha_x}^2$.

3.6.1.4 MH Step for ϕ

The conditional posterior density of ϕ is given as

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

The resulting acceptance probability of our proposal, $\phi^* \sim N(\phi^{i-1}, \sigma^2_\phi)$, is

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

A value of 0.08 for the proposal variance, σ^2_ϕ, will return an acceptance rate of approximately 0.30 for ϕ.

3.6.2 Generating Posterior Samples of 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 the density $f(\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d})$ can be derived as

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

implying that

$$\mu_{xt}|\alpha_x, \beta_x, \kappa_t, \phi, \boldsymbol{d} \sim \text{Gamma}\left(d_{xt}+\phi, e_{xt} + \frac{\phi}{\exp(\alpha_x+\beta_x\kappa_t)}\right). \qquad (3.7)$$

Therefore, the posterior samples of μ_{xt} can be generated by simulating from the expression in (3.7), 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. Note that Equation (3.7) is exactly the same as the Gibbs step for updating μ_{xt} within the MCMC algorithm under the PGLC model, which is to be expected since the PGLC and NBLC models are essentially equivalent.

3.6.3 NBLC Model with Blocking

Here, we consider allocating all of α_x, β_x and κ_t in a single block to account for the correlation structure of the relevant parameters. This has the potential to improve the overall efficiency of our MCMC algorithm by combining each individual MH updating steps into a single MH step, reducing the computational time of the algorithm by avoiding excessive looping, provided the corresponding proposal variance matrix is appropriately specified.

Suppose $\boldsymbol{\delta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_{-1}^\top, \boldsymbol{\kappa}_{-1}^\top)^\top$, the resulting conditional posterior density can be expressed as

$$\begin{aligned}
&f(\boldsymbol{\delta}|\boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi)\\
\propto &\prod_{x,t}\left\{\frac{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)]^{d_{xt}}}{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi]^{d_{xt}+\phi}}\right\}\times\exp\left\{-\frac{1}{2\sigma_\kappa^2}\sum_{t=2}^{T}[\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})]^2\right\}\\
&\times\exp\left\{-\frac{(1-\beta_2-\ldots-\beta_A)^2}{2\sigma_\beta^2}-\frac{\sum_{x=2}^{A}\beta_x^2}{2\sigma_\beta^2}-\frac{\sum_{x=1}^{A}(\alpha_x-\alpha_0)^2}{2\sigma_\alpha^2}\right\}.
\end{aligned}$$

The acceptance probability is then

$$a(\boldsymbol{\delta}^*|\boldsymbol{\delta}^{i-1}) = \min\left\{1, \frac{f(\boldsymbol{\delta}^*|\boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi)}{f(\boldsymbol{\delta}^{i-1}|\boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi)}\right\},$$

where $\boldsymbol{\delta}^* \sim N_{2A+T-2}(\boldsymbol{\delta}^{i-1}, \boldsymbol{\Sigma}_\delta)$ is the multivariate random walk proposal, $\boldsymbol{\delta}^{i-1}$ is the current iterate and $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$ is the proposal variance matrix. Despite the quicker speed of implementation that this strategy offers, the choice of the proposal variance matrix, $\boldsymbol{\Sigma}_\delta$, is crucial because naive specification of it can be detrimental to the overall efficiency of our algorithm. For example, if the proposal is performed independently ($\boldsymbol{\Sigma}_\delta$ is diagonal), then the resulting rejection rate for our MH algorithm will be inevitably high due to the naively specified proposal.

Rosenthal (2014) proposed that under certain assumptions, the optimal proposal variance matrix for a block updating MH algorithm with a multivariate normal distribution as its target is formulated as

$$\frac{2.38^2}{p}\times\boldsymbol{\Sigma}, \tag{3.8}$$

where p is the dimension of the parameters, and $\boldsymbol{\Sigma}$ is the variance matrix of the target distribution. Nevertheless, it is often not directly applicable because, typically, the variance matrix of the posterior distribution is unknown. Fortunately, the negative inverse of the Hessian matrix (evaluated at the mode) serves as a reasonable approximation of the posterior variance matrix (seen frequently in normal approximation). Specifically, denoting $\boldsymbol{\theta}_{\text{NBLC}} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_{-1}^\top, \boldsymbol{\kappa}_{-1}^\top, \sigma_\kappa^2, \sigma_\beta^2, \rho, \psi_1, \psi_2, \phi)^\top$ as the full set of parameters under the NBLC model with dimensionality $p_{\text{NBLC}} = 2A+T-2$, we consider the proposal variance matrix

$$\boldsymbol{\Sigma}_\delta = \frac{2.38^2}{p_{\text{NBLC}}}\times\boldsymbol{G}(\boldsymbol{\theta}_{\text{NBLC}}^{\text{mode}}),$$

where $\boldsymbol{G}(\boldsymbol{\theta}_{\text{NBLC}}^{\text{mode}})$ is the sub-matrix of $[-\boldsymbol{H}(\boldsymbol{\theta}_{\text{NBLC}})]^{-1}$ corresponding to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{-1}$, and $\boldsymbol{\kappa}_{-1}$ evaluated at the joint posterior mode $\boldsymbol{\theta}_{\text{NBLC}}^{\text{mode}}$, while $\boldsymbol{H}(\boldsymbol{\theta}_{\text{NBLC}})$ is the Hessian matrix of the joint posterior distribution with ij^{th} element

$$[\boldsymbol{H}(\boldsymbol{\theta}_{\text{NBLC}})]_{ij} = \frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{NBLC}}|\boldsymbol{d})}{\partial\theta_j \partial\theta_i};$$

see Appendix D for a complete expression of the Hessian matrix.

Note that the multiplicative constant, 2.38^2, that appears in (3.8), is only regarded as an optimal scaling factor when the target distribution is multivariate normal and its posterior variance is known. In practice, it should be tuned accordingly using similar search algorithm described before in Section 3.4.2.1. In our application, it turns out that the recommended 2.38^2 works perfectly fine by yielding an acceptance rate of around 0.26.

In order to undertake the above approach, the joint posterior mode is required, which can be troublesome as numerical multi-dimensional optimisation is rather difficult to perform. Therefore, the iterative conditional modes search algorithm is implemented. This algorithm searches for the joint mode of a distribution by iteratively maximizing the density function with respect to each parameter, conditioned on the rest. The underlying principle of this algorithm is rather similar to Gibbs sampling, where basically, the optimisation is performed routinely using the conditional posterior distributions. Hence, this optimisation procedure is particularly useful when some of the conditional posterior distributions are tractable, as in our case. Where the conditional posterior density is intractable, numerical optimisation (the Newton-Raphson method for example) is applied (see Appendix E for the specified algorithm).

3.7 Methods for Mortality Forecasting

Projection within the Bayesian framework is particularly natural through the derivation of posterior predictive distribution. To be precise, the posterior predictive distribution of the 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

$$\begin{aligned} f(\log \mu_{xT+1}|\boldsymbol{d}) &= \int f(\log \mu_{xT+1}|\alpha_x, \beta_x, \kappa_{T+1}, \sigma_\mu^2) f(\alpha_x, \beta_x, \sigma_\mu^2|\boldsymbol{d}) f(\kappa_{T+1}|\kappa_T, \rho, \sigma_\kappa^2, \boldsymbol{\psi}) \\ &\quad \times f(\kappa_T, \rho, \sigma_\kappa^2, \boldsymbol{\psi}|\boldsymbol{d}) \mathrm{d}\alpha_x \mathrm{d}\beta_x \mathrm{d}\kappa_T \mathrm{d}\kappa_{T+1} \mathrm{d}\rho \mathrm{d}\sigma_\kappa^2 \mathrm{d}\boldsymbol{\psi} \mathrm{d}\sigma_\mu^2, \end{aligned} \tag{3.9}$$

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, in light of the model likelihood, prior distributions and projection, are fully integrated in the posterior predictive distribution. The density in (3.9) is analytically intractable, but can be empirically estimated using our MCMC samples as follows. Essentially, generation of the posterior samples of $\log \mu_{x\,T+1}$ under the PLNLC model 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-years ahead projections can be obtained by recursive implementation of the above generation procedures. Specifically, the joint posterior predictive distribution of the future log mortality rates $(\log \mu_{x\,T+1}, \log \mu_{x\,T+2}, \ldots, \log \mu_{x\,T+h})$ is given by

$$\begin{aligned}
& f(\log \mu_{x\,T+1}, \log \mu_{x\,T+2}, \ldots, \log \mu_{x\,T+h} | \boldsymbol{d}) \\
= & \int \prod_{t=T+1}^{T+h} [f(\log \mu_{xt} | \alpha_x, \beta_x, \kappa_t, \sigma_\mu^2)] f(\alpha_x, \beta_x, \sigma_\mu^2 | \boldsymbol{d}) \prod_{t=T+1}^{T+h} [f(\kappa_t | \kappa_{t-1}, \sigma_\kappa^2, \rho, \boldsymbol{\psi}, \boldsymbol{d})] \\
& f(\kappa_T, \sigma_\kappa^2, \rho, \boldsymbol{\psi} | \boldsymbol{d}) \mathrm{d}\alpha_x \mathrm{d}\beta_x \mathrm{d}\kappa_{T+h} \ldots \mathrm{d}\kappa_T \mathrm{d}\sigma_\mu^2 \mathrm{d}\sigma_\kappa^2 \mathrm{d}\rho \mathrm{d}\boldsymbol{\psi}.
\end{aligned}$$

Hence, the joint posterior predictive samples can be obtained by:

1. Generate $(\kappa_{T+1}, \ldots, \kappa_{T+h})$ repeatedly through the AR(1) model, i.e.

   $$\begin{aligned}
   \kappa_{T+1} &\sim N(\psi_1 + \psi_2(T+1) + \rho(\kappa_T - \psi_1 - \psi_2 T), \sigma_\kappa^2), \\
   \kappa_{T+2} &\sim N(\psi_1 + \psi_2(T+2) + \rho(\kappa_{T+1} - \psi_1 - \psi_2(T+1)), \sigma_\kappa^2), \\
   &\vdots \\
   \kappa_{T+h} &\sim N(\psi_1 + \psi_2(T+h) + \rho(\kappa_{T+h-1} - \psi_1 - \psi_2(T+h-1)), \sigma_\kappa^2),
   \end{aligned}$$

   where $(\kappa_T, \sigma_\kappa^2, \psi_1, \psi_2, \rho)$ are the joint posterior samples from the MCMC output.

2. Generate the future log mortality rates $(\log \mu_{x\,T+1}, \ldots, \log \mu_{x\,T+h})$ by

   $$\begin{aligned}
   \log \mu_{x\,T+1} &\sim N(\alpha_x + \beta_x \kappa_{T+1}, \sigma_\mu^2), \\
   &\vdots \\
   \log \mu_{x\,T+h} &\sim N(\alpha_x + \beta_x \kappa_{T+h}, \sigma_\mu^2),
   \end{aligned}$$

where $(\kappa_{T+1}, \ldots, \kappa_{T+h})$ are from step 1 and the joint posterior samples of $(\alpha_x, \beta_x, \sigma^2_\mu)$ are substituted accordingly.

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 user's preference, whether or not they prefer to base their policy making on the underlying rates (unobservable), or the crude rates (observable). 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.

3.8 Results

3.8.1 Comparison of the Efficiency of Different MCMC Schemes

Thus far, two main models to account for overdispersion have been presented, the Poisson Log-Normal LC model and the Poisson Gamma (Negative Binomial) LC model. We also described several possible MCMC schemes to carry out posterior sampling for each model. Specifically, there are two schemes available for the Poisson Log-Normal LC model:

- PLNLC with with univariate updating (Section 3.4),
- PLNLC with block updating (Section 3.4.3);

while three other schemes for the Poisson Gamma LC model:

- PGLC with univariate updating (Section 3.5),
- NBLC (PGLC with μ_{xt} marginalised) with univariate updating (Section 3.6),
- NBLC with block updating (Section 3.6.3).

All the sampling algorithms for generating posterior samples are produced from custom codes written in R. For a posterior sample of size 100, it takes roughly 8.8 seconds and 10.4 seconds respectively for the PLNLC with univariate and block updating schemes (without devoting too much efforts to speed up the generation processes). On the other hand, it takes approximately 4.7 seconds for the PGLC with univariate updating scheme, 2.5 seconds and 1.5 seconds respectively for the NBLC with univariate and block updating schemes. In general, posterior sample generation is computationally faster for the PGLC and NBLC models as compared to the PLNLC model, but produce samples with larger auto-correlations due to the use of MH algorithms. Note that the above sampling algorithms can be rather time consuming to execute in R for large sample sizes (especially for the PLNLC model), efforts can be dedicated to improve the codes in terms of the speed of sample generation, or simply choose a more efficient programming software such as C++. However, this is acceptable for the purpose of mortality forecasting as computational time is not highly regarded as the limiting factor in terms of its usage, as opposed to, for example, weather forecasting.

Next, we carry out a crude computational comparison between the MCMC schemes under each model to select the most efficient scheme in terms of the posterior samples generation. Clearly, provided that each MCMC scheme manages to attain convergence, they should in principle, be indifferent in terms of producing posterior samples. To compare the the efficiency of each MCMC scheme, effective posterior sample size per unit time (hour) is computed for each parameter to act as a crude measure of speed, where effective sample size is defined as the equivalent number of independent samples that contain as much information as the generated dependent posterior samples. Or mathematically (Gelman et al., 1995),

$$\text{Effective Sample Size} = \frac{\text{Original Sample Size}}{1 + 2\sum_{s=1}^{\infty} \rho_s},$$

where ρ_s is the auto-correlation at lag s. Notice that the expression above involved infinite summation of the auto-correlations. In response to this issue, Gelman et al. (1995) suggested to perform a positive partial sum, where the summation is taken only until lag S, such that the following sum of autocorrelation estimates for two successive lags, $\hat{\rho}_{S+1} + \hat{\rho}_{S+2}$, is negative.

The effective posterior sample sizes per unit time of each parameter for both MCMC schemes of PLNLC model are depicted in Figure 3.9 and 3.10. Despite being rather similar for the rest of the parameters, the PLNLC with block updating (in red) outperforms its counterpart by a considerable margin, especially for the parameters $\beta_2, \ldots, \beta_A$. In particular, the total effective posterior sample size per unit time formed by summing across the parameters indicate that the PLNLC with block updating is more efficient than that with univariate updating (given as $240,029$ and $171,204$ respectively). Therefore, PLNLC model with block updating is recommended as it is computationally more efficient in overall. Note that the β parameters benefit the most from the blocking strategy, mainly because of the fact that they are highly correlated due to the correlations induced by the constraint, $\sum_x \beta_x = 1$. One can then consider only blocking the β parameters (leaving α and κ parameters univariately updated) to avoid the theoretical derivations of the conditional posterior of $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$. However, given that the two schemes require roughly the same computational time and that only multivariate normal distributions (which are easily tractable) are involved for the derivations, not much of an advantage is to be anticipated.

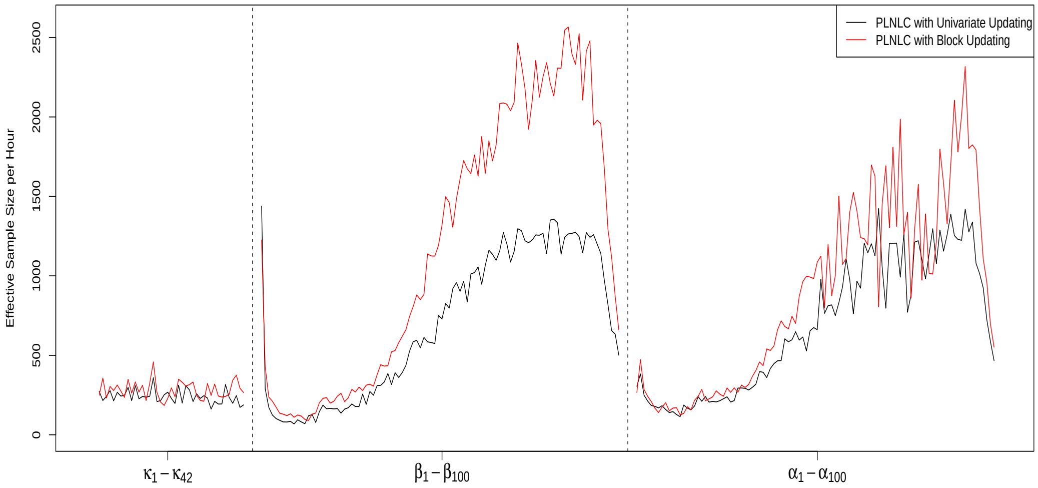


Figure 3.9: Plot of effective posterior sample sizes per hour for κ_t, β_x, and α_x under the PLNLC model.

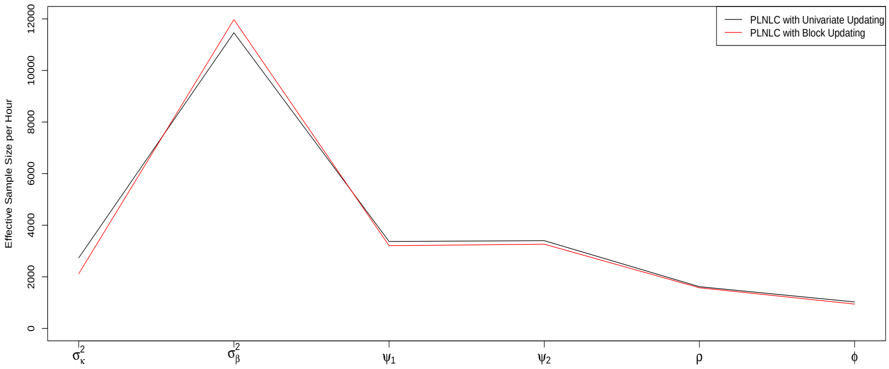


Figure 3.10: Plot of effective posterior sample sizes per hour for the rest of the parameters under the PLNLC model.

Next, we compare the three MCMC schemes for the Poisson Gamma LC model. As illustrated in Figure 3.11, the NBLC model with univariate updating clearly dominates the other two MCMC schemes in terms of the parameters α_x, β_x and κ_t. The performance of the NBLC model with block updating is rather disappointing by consistently having less efficiency than its counterparts. This finding is consistent to the observation of Roberts and Sahu (1997), who suggested that blocking can slow down the rate of convergence of a MCMC scheme in certain situations. In our case, the suboptimal performance of this MCMC scheme may be due to the poor approximation of the posterior variance matrix by the negative inverse of Hessian matrix (evaluated at the joint posterior mode). It could also be caused by the rather inefficient blocking combination, where all of α_x, β_x and κ_t are allocated in a single block. Considering the trade-off of benefits between correlations and the computational complexity of high-dimensional parameters, allocating α_x, β_x, κ_t each in a separate block may well be more efficient since the correlations between them are expected to be small.

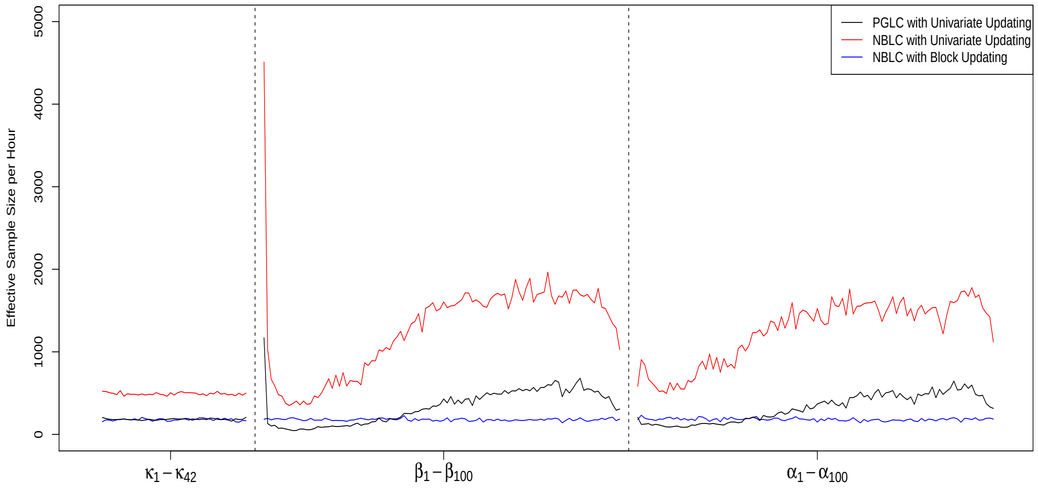


Figure 3.11: Plot of effective posterior sample sizes per hour for κ_t, β_x, and α_x under the Poisson Gamma LC model.

The NBLC model with blocking is arguably most efficient in generating the rest of the parameters (see Figure 3.12), but it is not sufficient to outweigh those of α_x, β_x and κ_t. In terms of the total effective posterior sample sizes per hour, the NBLC model with univariate updating outclasses the rest by having a value of $372,868$, compared to $107,456$ and $154,757$ of the PGLC without blocking and NBLC with blocking respectively. In conclusion, the rest of the results is based on the posterior samples generated from the PLNLC with blocking and NBLC without blocking.

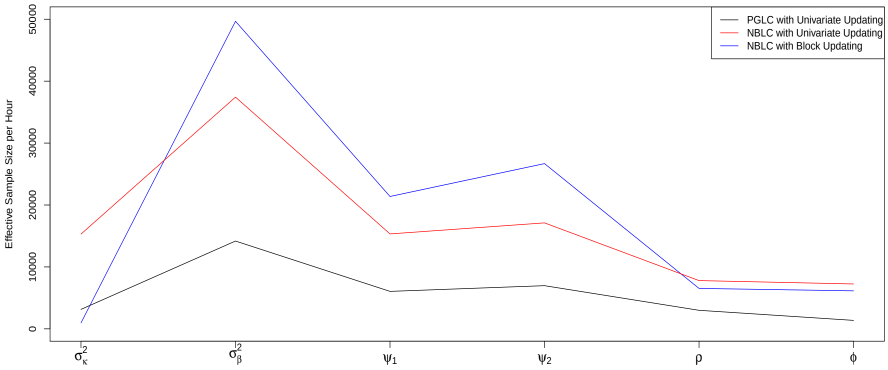


Figure 3.12: Plot of effective posterior sample sizes per hour for rest of the parameters under the Poisson Gamma LC model.

3.8.2 Summary of the MCMC Generated Posterior Samples

For initialization of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{-1}$ and $\boldsymbol{\kappa}_{-1}$, we use the maximum likelihood estimates (MLE) obtained using Goodman's method (see Renshaw and Haberman, 2005). On the other hand, the initial values of σ^2_κ and ρ are obtained by fitting an AR(1) with linear drift model on $\boldsymbol{\kappa}$, while σ^2_β is initialised by the empirical variance of the MLE of $\boldsymbol{\beta}$. Finally, $\boldsymbol{\psi}$ is initialised as $(0,0)^\top$, while the overdispersion parameters, σ^2_μ and ϕ, are initialised by 0.001 and 100 respectively. 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. We do not pursue this matter here, instead we assume that the burn-in phase (10000 iterations) is sufficiently long to diminish the effect of initialization.

Before making any inferential comparisons, trace plots and auto-correlation plots are presented as diagnostic tools for detecting anomalies in the MCMC generated posterior samples. A trace plot is a plot of the posterior samples against iteration number. For a converged MCMC trajectory, the trace plot should demonstrate proper mixing (that looks like a "fat hairy caterpillar") in which the transitions occur within a well-defined region. On the contrary, "snake"-shaped chains (see for example Lunn et al., 2013) or chains that experience random shifts in their mean levels portrayed in a trace plot indicate poorly behaved MCMC algorithms and that, require appropriate tuning.

An auto-correlation plot is a plot of the sample autocorrelations against lag, where lag-k sample auto-correlation is defined as the correlation of a parameter with itself from the same series of samples, separated by an interval of k (see Chatfield, 1984). It is used to assess the degree of dependence of the MCMC generated samples. Ideally, a fast (exponential) decay of the sample auto-correlations is desirable. Auto-correlation plots can also be used to assist the thinning process, where the amount to thin depends on the sample auto-correlations displayed. For instance, an approximately independent posterior sample can be obtained by applying a k-thinning, where k is chosen such that the lag-k sample auto-correlation first approaches a negligible value with respect to the standard error.

By examining the trace plots depicted in Figure 3.13, 3.14, 3.15 and 3.16, all of them exhibit the shape of a “fat hairy caterpillar”, with no apparent anomaly. The trajectories emerge as if convergence has been attained. Additionally, 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.

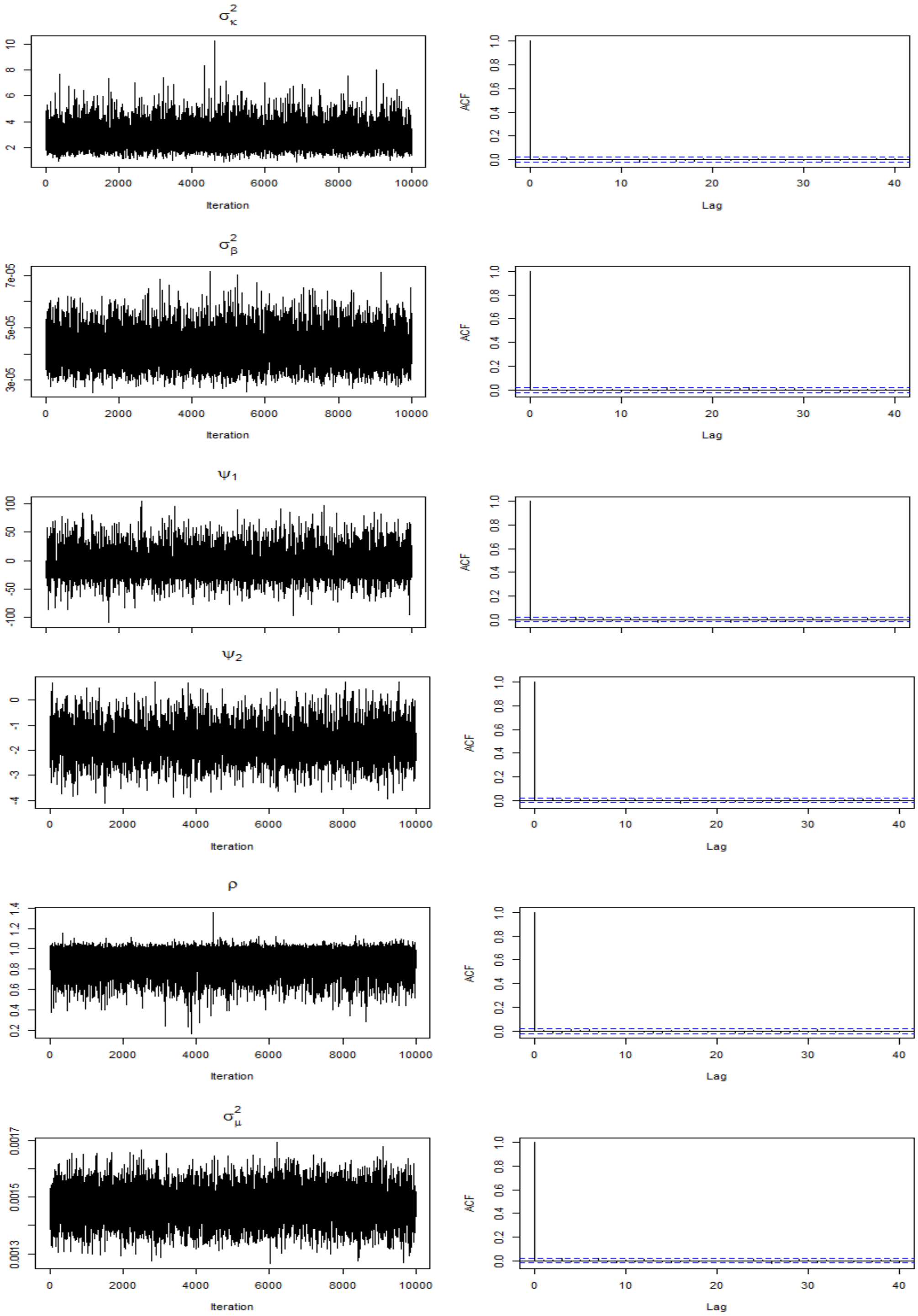


Figure 3.13: Trace plots (left panels) and auto-correlation plots (right panels) of a portion of the posterior samples for σ^2_κ, σ^2_β, ψ_1, ψ_2, ρ and σ^2_μ under the PLNLC model.

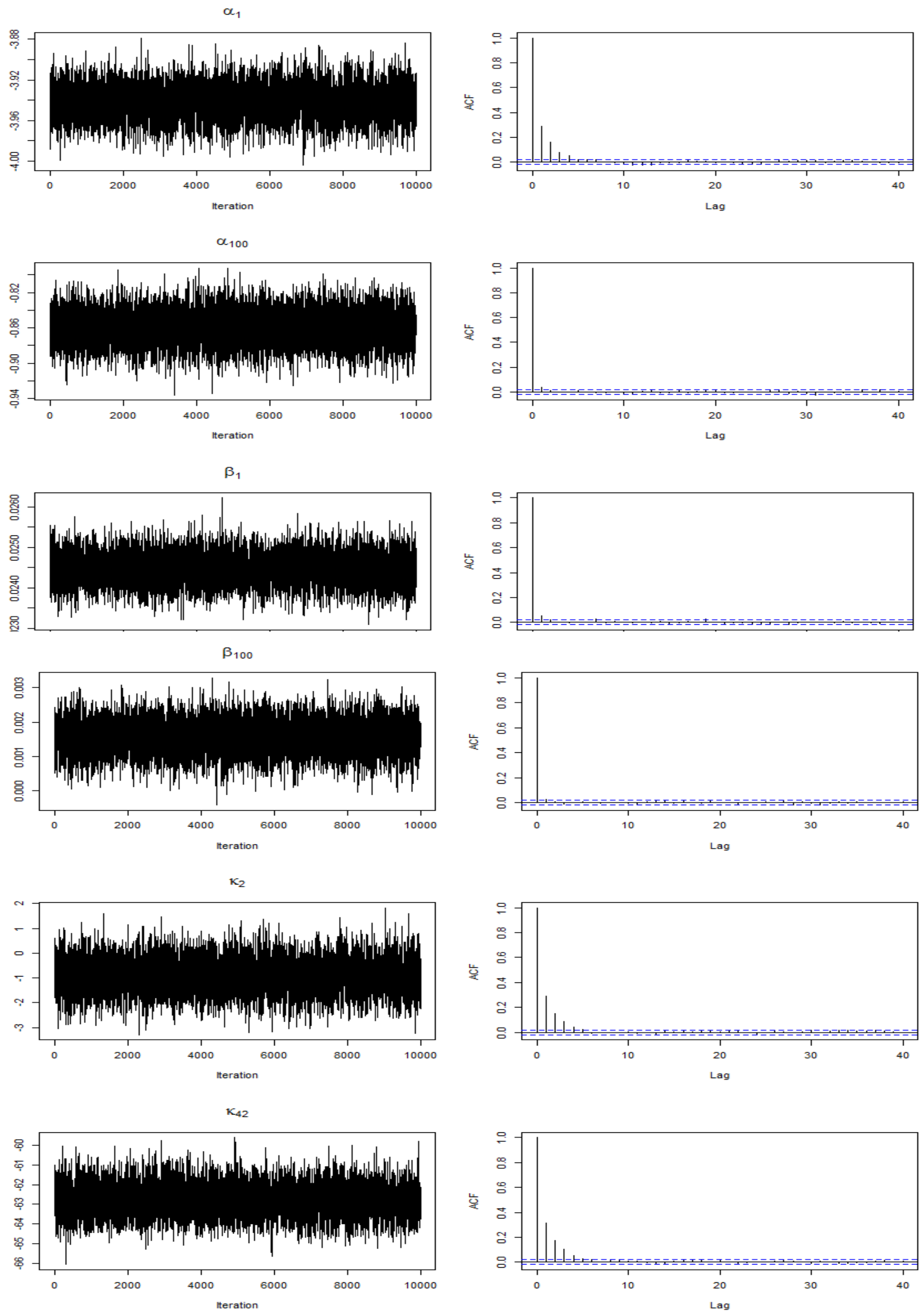


Figure 3.14: Trace plots (left panel) and auto-correlation plots of a portion of the posterior samples for several chosen α_x, β_x and κ_t under the PLNLC model.

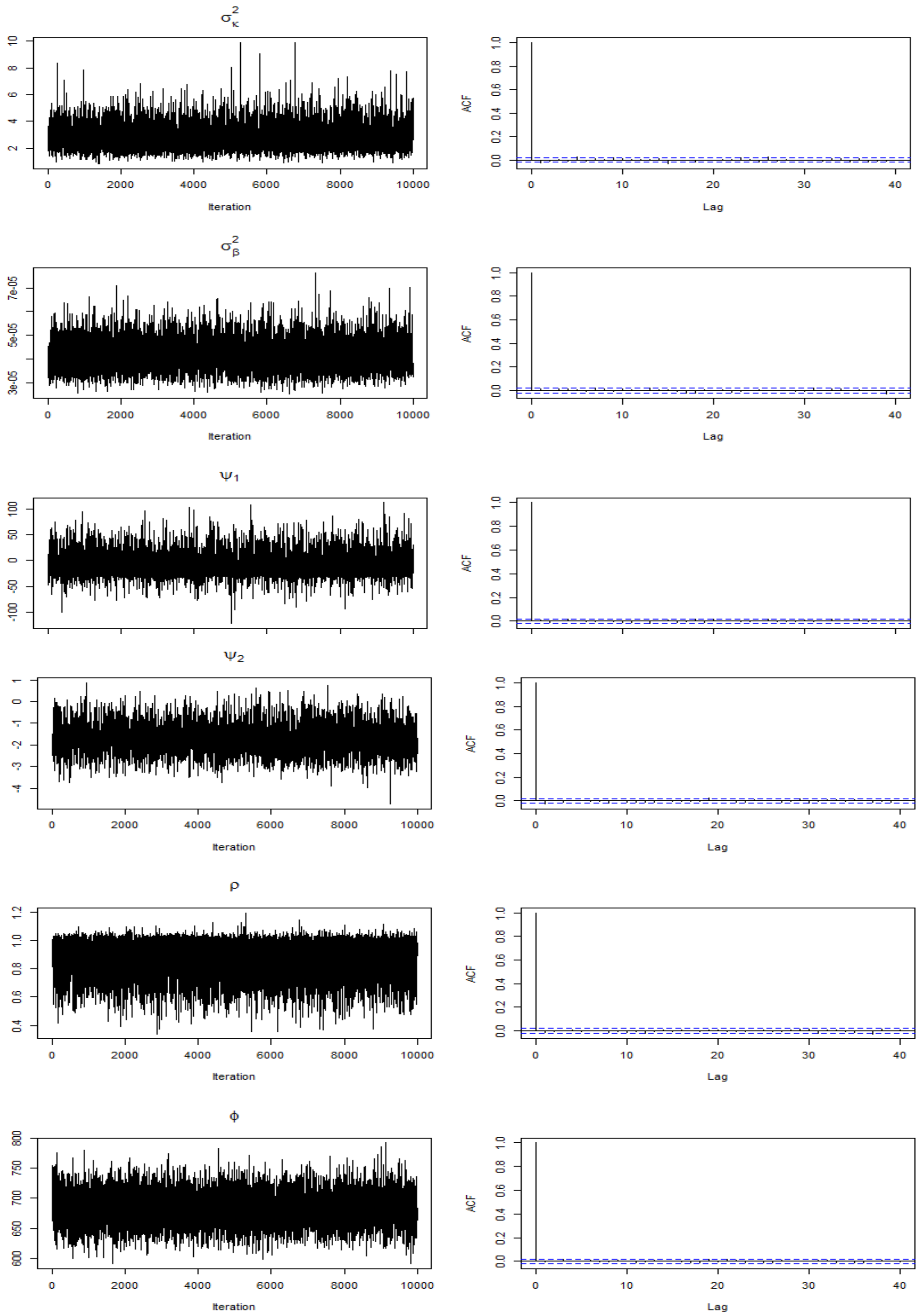


Figure 3.15: Trace plots (left panels) and auto-correlation plots (right panels) of a portion of the posterior samples for σ_κ^2, σ_β^2, ψ_1, ψ_2, ρ and σ_μ^2 under the NBLC model.

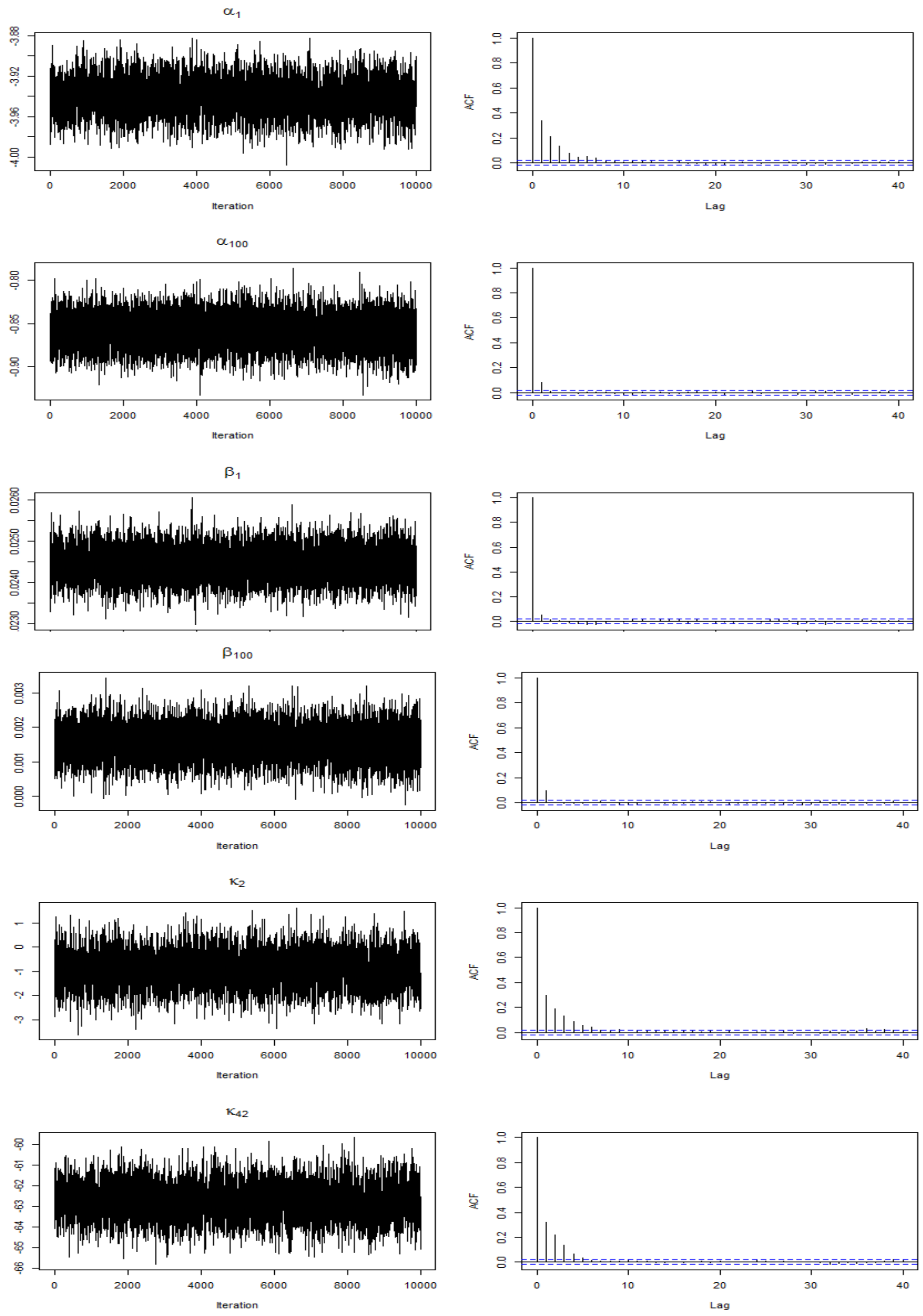


Figure 3.16: Trace plots (left panel) and auto-correlation plots of a portion of the posterior samples for several chosen α_x, β_x and κ_t under the NBLC model.

3.8.3 Estimated Parameters

Throughout the result section, we compare our proposed models with the Bayesian PLC model (Czado et al., 2005) to highlight the importance of accounting for overdispersion.

The Bayesian PLC model (i.e. the PLNLC model or PGLC/NBLC model without the overdispersion component, ν_{xt}) is fitted using Czado's methodology, except we adopt the same prior specification as in Section 3.4.1 to facilitate model comparison later on. We also provide a comparison of our proposed models with each other.

Figures 3.17 and 3.18 depict the fitted values (posterior medians) of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, accompanied by the associated 95% credible interval (computed from the sample quantiles) 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 Figures 3.17 and 3.18, 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.

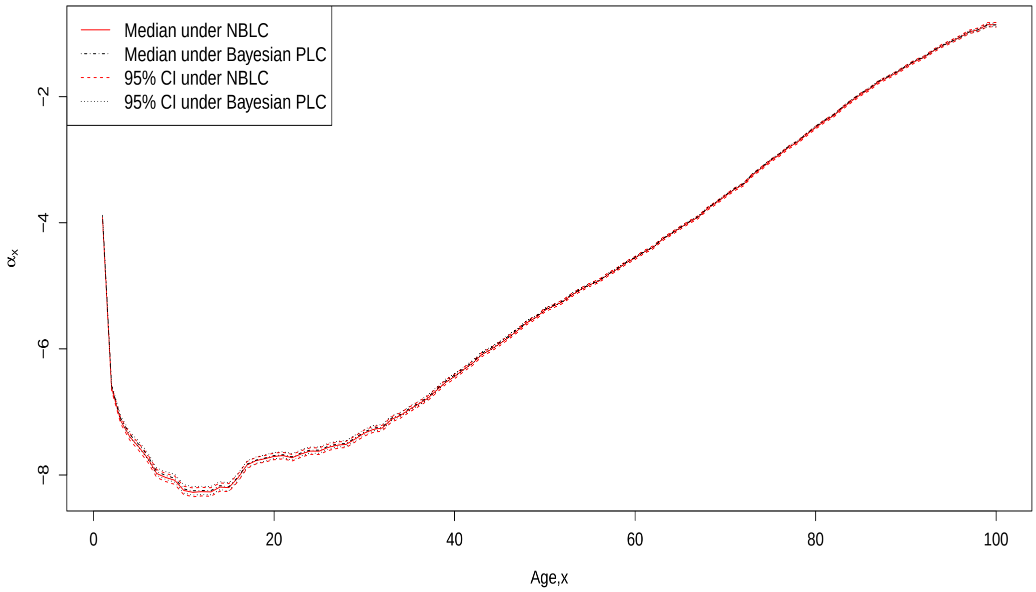


Figure 3.17: Plot of estimated α_x against age with their 95% credible intervals under the Bayesian Poisson LC model and the NBLC model.

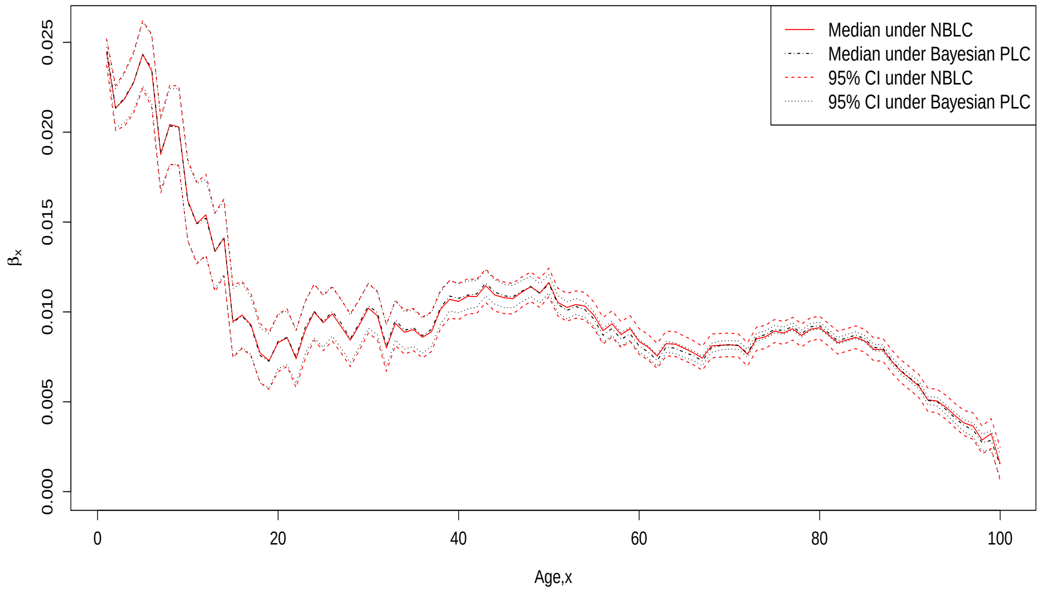


Figure 3.18: Plot of estimated β_x against age with their 95% credible intervals under the Bayesian PLC model and the NBLC model.

The main difference arises from the parameter $\boldsymbol{\kappa}$. As evident from Figure 3.19, 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, 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 flexibility for the model to describe the data, allowing more priority to be put on fitting the AR(1) prior, hence the smoother fitted values. On the other hand, with less variabilities offered 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.

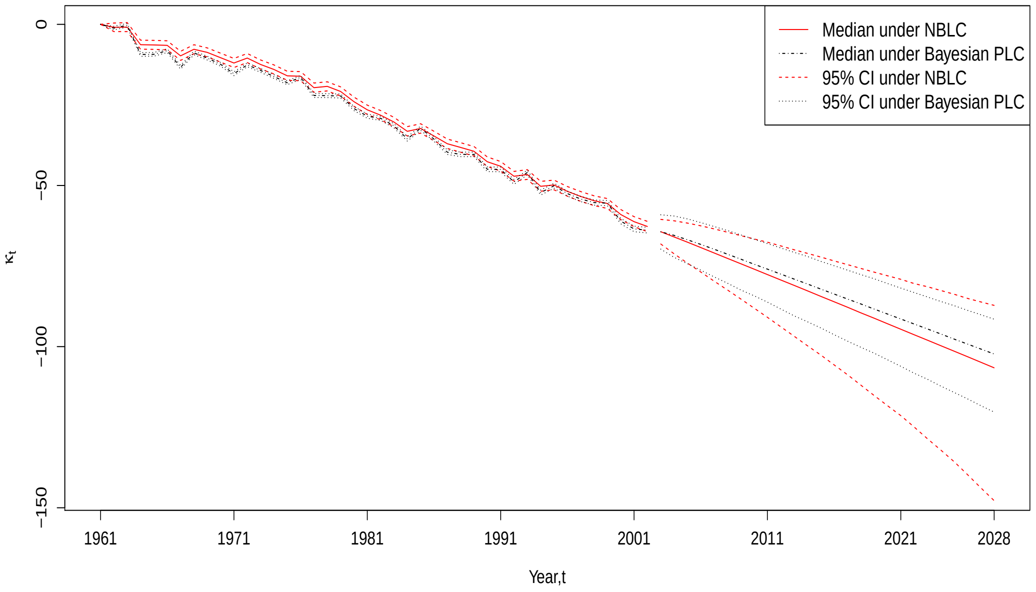


Figure 3.19: Plot of estimated κ_t and their 26-years ahead projection against years, accompanied by the corresponding 95% intervals under the Bayesian PLC model and the NBLC model.

Table 3.1 compares the posterior medians and 95% credible intervals of the rest of the hyperparameters. First of all, the fitted σ_β^2 are exactly the same across all three models (the variance of β are unaffected by the incorporation of overdispersion). By contrast, there are sizeable differences for hyperparameters relating to the time series model of κ_t (σ_κ^2, ψ_1, ψ_2 and ρ) between the Bayesian PLC and overdispersion models. Specifically, the residual variance, σ_κ^2, is larger with more uncertainty for the Bayesian PLC model. The posterior medians of ψ_1 and ψ_2 are rather similar across the models, but the 95% credible intervals are substantially wider for the overdispersion models. There is also a huge discrepancy for the auto-regressive coefficient ρ, where it can be deduced that the Bayesian PLC model favours a stationary AR(1) model on κ_t (ρ mostly smaller than one), while the overdispersion models prefer the random walk model (ρ close to one). These discrepancies will be further investigated in the next paragraph, when their kernel densities are being examined. On the other hand, the dispersion parameters (σ_μ^2 and ϕ) have virtually identical posterior medians and 95% credible intervals. Notice also that the posterior medians and credible intervals under the PLNLC and the NBLC model are remarkably similar across the hyperparameters, indicating model similarity.

A better visualization of the marginal posterior distributions is given by the kernel densities estimated from the MCMC generated samples. Kernel density estimator of the marginal posterior density of the j^{th} component of the parameter, θ_j, is a non-parametric estimator of the form

$$\hat{f}_{\text{kernel}}(\theta_j^*|\boldsymbol{d}) = \frac{1}{Nh_N}\sum_{i=1}^{N} K\left(\frac{\theta_j^* - \theta_j^i}{h_N}\right),$$

where θ_j^* is the point of evaluation, θ_j^i $(i = 1, \ldots, N)$ is a sample of size N from the MCMC samples, and h_N is the smoothing parameter that typically depends on the sample size N (see for example Chen et al., 2000). For simplicity, the kernel function, $K()$, is chosen to be a Gaussian kernel. Under some regularity conditions, Silverman (1986) showed that the kernel density estimator converges asymptotically to the marginal posterior density.

Parameter	Posterior Median		
	Bayesian PLC	PLNLC	NBLC
σ^2_κ	6.07	2.68	2.65
σ^2_β	4.1×10^{-5}	4.1×10^{-5}	4.1×10^{-5}
ψ_1	-30.8	-26.1	-26.2
ψ_2	-1.55	-1.66	-1.66
ρ	0.46	0.94	0.94
σ^2_μ and $1/\phi$		0.001468	0.001467

Parameter	95% Credible Interval		
	Bayesian PLC	PLNLC	NBLC
σ^2_κ	$(3.91, 10.3)$	$(1.51, 4.83)$	$(1.48, 4.84)$
σ^2_β	$(3.1 \times 10^{-5}, 5.5 \times 10^{-5})$	$(3.1 \times 10^{-5}, 5.5 \times 10^{-5})$	$(3.1 \times 10^{-5}, 5.5 \times 10^{-5})$
ψ_1	$(-33.1, -6.9)$	$(-42.7, 43.1)$	$(-41.4, 42.3)$
ψ_2	$(-1.97, -1.33)$	$(-2.83, -0.50)$	$(-2.85, -0.49)$
ρ	$(0.12, 0.99)$	$(0.59, 1.04)$	$(0.58, 1.04)$
σ^2_μ and $1/\phi$		$(0.00136, 0.00158)$	$(0.00136, 0.00158)$

Table 3.1: Posterior medians and 95% credible intervals of σ^2_κ, σ^2_β, ψ_1, ψ_2, ρ, and σ^2_μ under the Bayesian PLC, PLNLC and NBLC model.

As a result, kernel estimates of the marginal posterior density of the rest of the parameters, derived from the posterior samples, are presented in Figure 3.20. The kernel densities of σ^2_β are almost identical. 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 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.

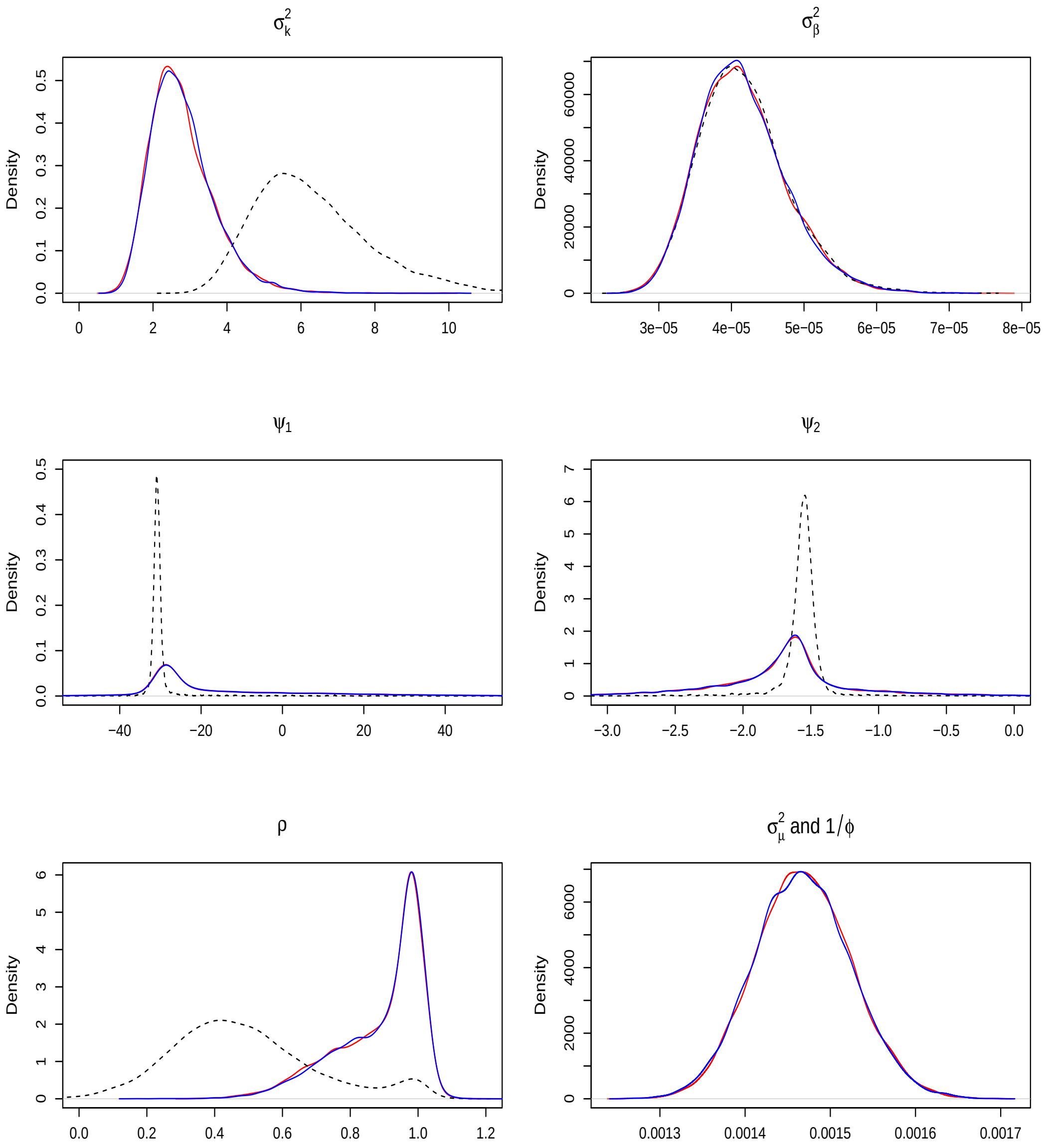


Figure 3.20: Kernel density plots of σ_κ^2, σ_β^2, ψ_1, ψ_2, ρ, and ϕ under the Bayesian PLC (black dotted), PLNLC (blue solid) and NBLC model (red solid).

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 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, when $\rho = 1$, the Gibbs step for $\boldsymbol{\psi}$ given in (3.5) simplifies to

$$\boldsymbol{\psi}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \kappa_{-1}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \rho, \sigma_\mu^2 \sim N(\boldsymbol{\psi}^*, \boldsymbol{\Sigma}_\psi^*),$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_\psi^* &= \begin{pmatrix} 1000 & 0 \\ 0 & \frac{1}{1/10+(T-1)/\sigma_\kappa^2} \end{pmatrix}, \\ \boldsymbol{\psi}^* &= \left(0, \frac{\kappa_T/\sigma_\kappa^2}{1/10 + (T-1)/\sigma_\kappa^2}\right)^\top. \end{aligned}$$

Since the conditional posterior distribution of ψ_1 does not depend on the data and other parameters, this implies that its marginal posterior distribution is $N(0, 1000)$, 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 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 $[0.00136, 0.00158]$) from the arbitrarily diffuse prior distributions (which have close to negligible densities for the region of values presented in Figure 3.20). 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 sample 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}, given as

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

under the PGLC/NBLC model. 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 PGLC/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. Recall that under the PLNLC model,

$$\mathrm{Var}[D_{xt}] = \mathbb{E}[D_{xt}] \times [1 + \mathbb{E}[D_{xt}](\exp(\sigma_\mu^2) - 1)],$$

where now $\mathbb{E}[D_{xt}](\exp(\sigma_\mu^2) - 1)$ represents the relative increase in the variance of D_{xt} with respect to 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.

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 3.21. 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 3.6.1, supporting our conjecture about the relationship between optimal proposal variances and posterior variances.

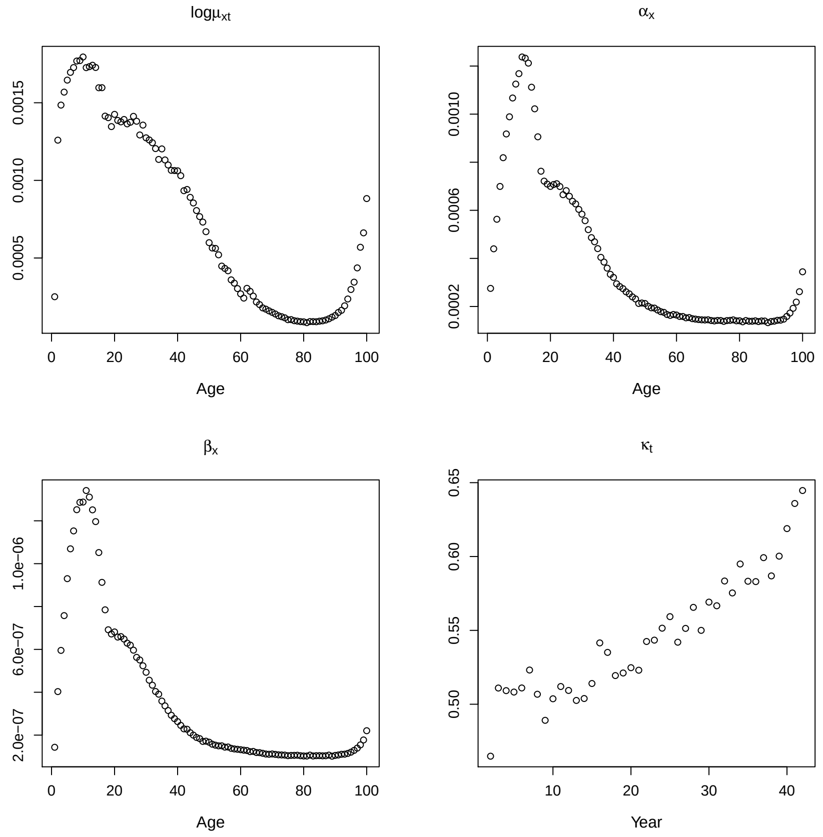


Figure 3.21: Plots of posterior variances of $\log \mu_{xt}$ for year 1980 (upper-left), α_x (upper-right), β_x (bottom-left) and κ_t (bottom-right) under the NBLC model.

3.8.4 Fitted and Projected Crude Mortality Rates

Figure 3.22 shows the fitted and projected log mortality rates for newborns, age 65 and age 80 plotted against time, 26 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 low-variance model by construction. In other words, a model that ignores overdispersion has the tendency to force the fitted values to adhere more closely to the data due to the small 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 data 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 the credible intervals for the overdispersion models 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 65). Overall, the overdispersion models provide a better description of the data variabilities, even though they also appear to have failed at capturing the important mortality trend for age 65, resulting in a sizeable discrepancy between the observed and fitted rates for the most recent year (see next paragraph).

In terms of projections, 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 (1992b). 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 rates as well as better calibrated probabilistic intervals in terms of the projection. On a side note, the sizeable jump-off discontinuity (between the most recent observed rate and the first projected rate) at the forecast origin observed for age 65 corresponds to the model failure in capturing relevant trend components of the mortality that will be mitigated by incorporating a cohort component (see Section 6.3).

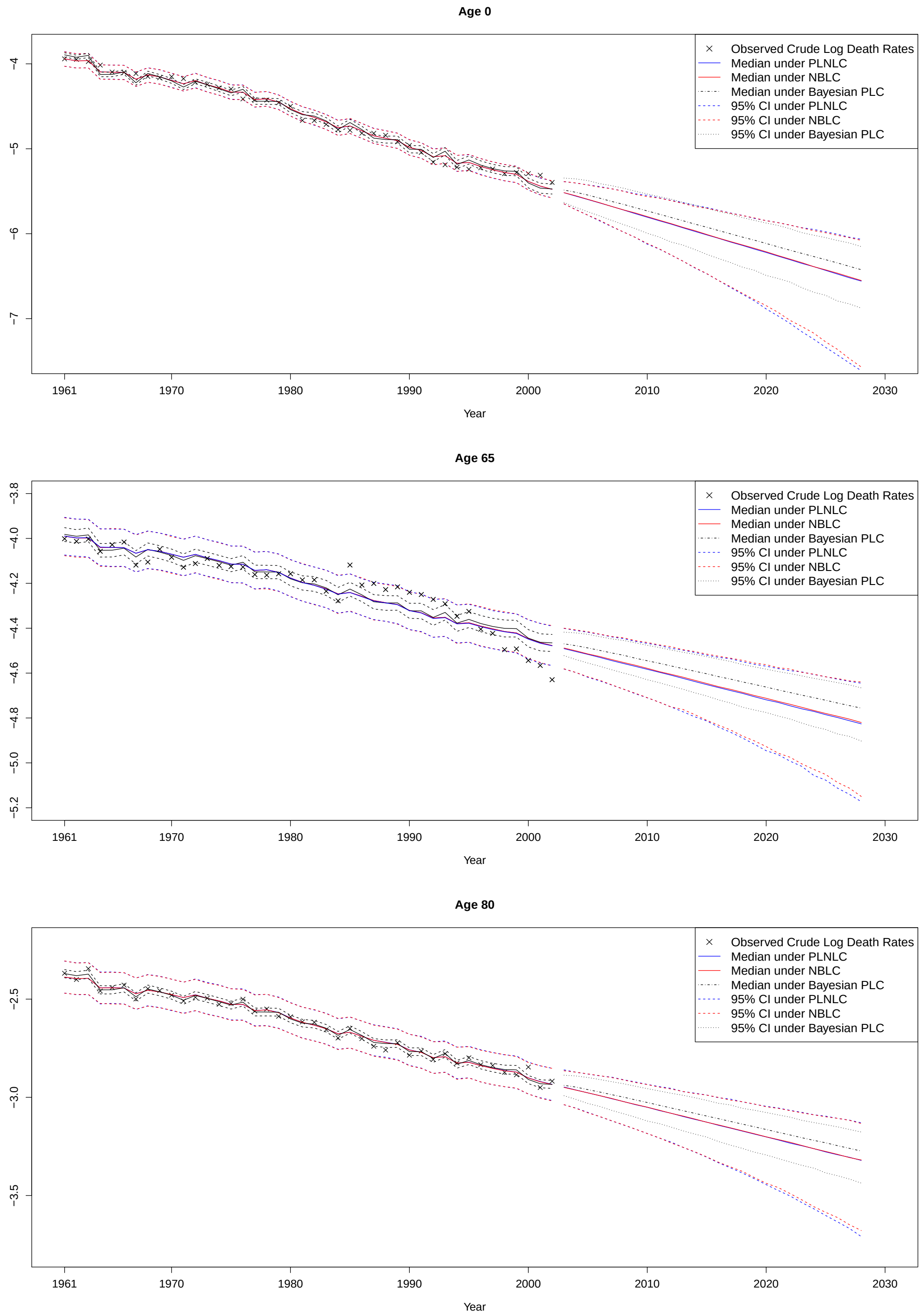


Figure 3.22: Plots of the observed log crude death rates, $\log(d_{xt}/e_{xt})$, fitted log mortality rates and the associated 26-years ahead projection of the crude log mortality rates for age 0 (upper panel), age 65 (middle panel) and 80 (lower panel) under the Bayesian PLC model and the overdispersion models, accompanied by 95% credible intervals.

The projected mortality age profile for year 2028 is provided in Figure 3.23 as an illustration. The same phenomena are observed, where the overdispersion models yield larger mortality improvement and substantially wider prediction intervals for the projection. The lack of smoothness of the projected log mortality rates across age is due to the fact that the smoothing model was postulated on the time component κ_t, but not for the age components α_x and β_x.

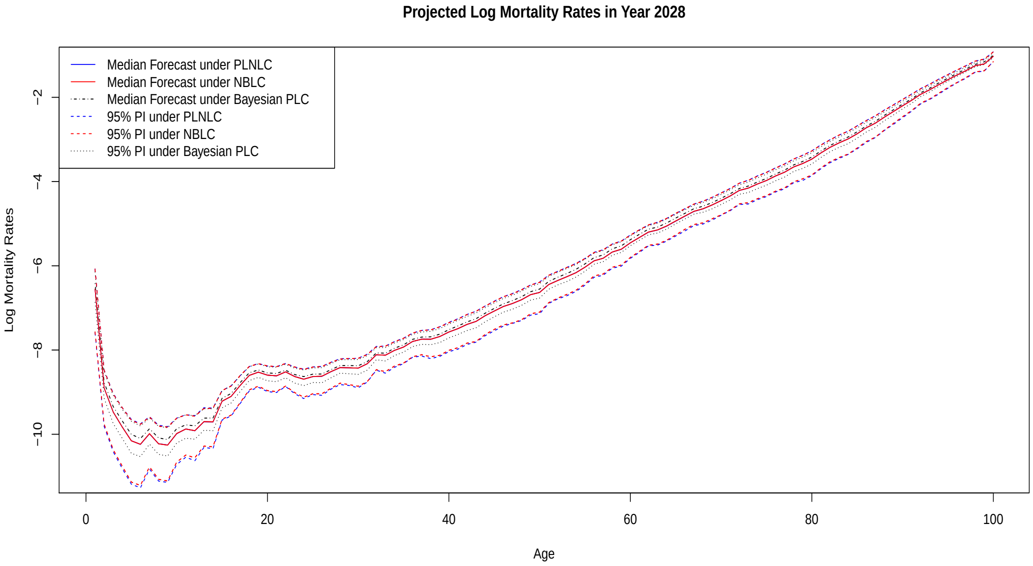


Figure 3.23: Plot of the projected log mortality rates in year 2028 against age, accompanied by 95% prediction intervals.

3.8.5 Model Assessment

We can similarly construct a heat map of the squared Pearson residuals, r_{xt}^2 for the overdispersion models. An expression of the squared Pearson residuals for the PLNLC and NBLC model is 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, and ϕ are substituted into the expression for an estimate.

As illustrated in Figure 3.24, the heat maps of the overdispersion models are much “greener” than before (Figure 3.1), indicating an overall improvement in goodness of fit.

The sum of r_{xt}^2 (r^2) 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 3.24. Nevertheless, it is rather obvious that the overdispersion models outperformed both the original PLC and Bayesian PLC model by a considerable margin.

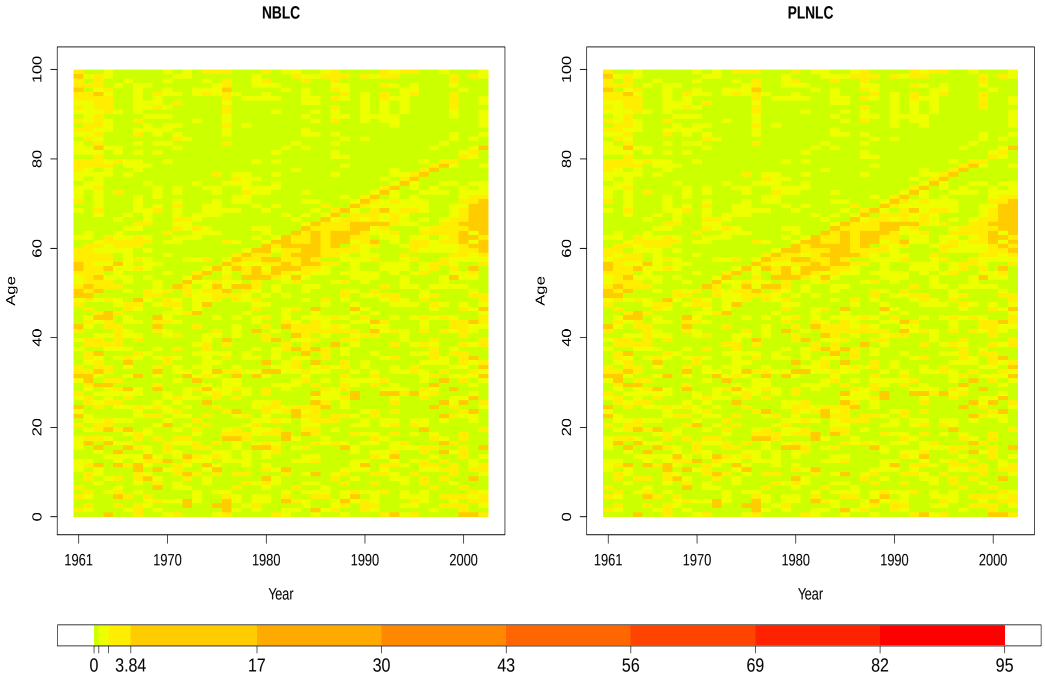


Figure 3.24: Heat map of squared Pearson residuals, r_{xt}^2, under the PLNLC (left panel) and the NBLC model (right panel), accompanied by the corresponding colour code.

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

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

from the posterior samples of $\boldsymbol{\theta}_M$ under each model. Denoting $\boldsymbol{\theta}_M^i$ $(i = 1, \ldots, N)$ as a set of posterior samples under model M, a sample of replicated data, $\boldsymbol{d}^{\text{rep}\,i}$ $(i = 1, \ldots, N)$ can be generated from the likelihood function, $f(\boldsymbol{d}^{\text{rep}}|\boldsymbol{\theta}_M^i, M)$, at each value of $\boldsymbol{\theta}_M^i$. For example, under the NBLC model, this can be accomplished by generating from

$$\text{Neg-Bin}\left(\phi^i, \frac{\phi^i}{e_{xt}\exp(\alpha_x^i + \beta_x^i\kappa_t^i) + \phi^i}\right),$$

where $\{\alpha_x^i, \beta_x^i, \kappa_t^i, \phi^i\}$ are samples from the joint posterior distribution.

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). In practice, $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$ is chosen to reflect the purpose of the analysis, which is model assessment in our case, achieved through the use of χ^2 discrepancy. An expression of $T(\boldsymbol{d}, \boldsymbol{\theta_M})$ for each of the models under consideration is presented in Appendix H. 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 3.25).

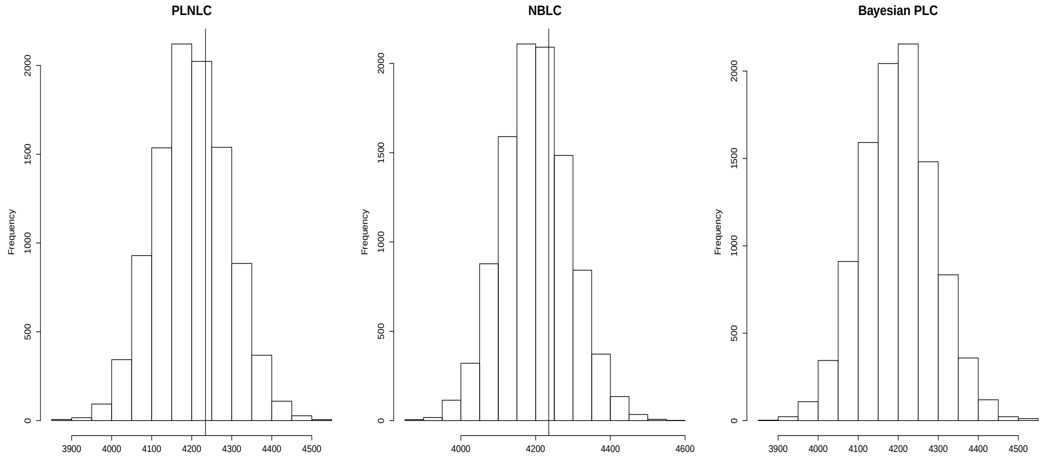


Figure 3.25: Histograms of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta_M})$ for the PLNLC, NBLC, and Bayesian PLC model, with their corresponding sum of squared Pearson residuals, r^2 included as the vertical solid lines.

The value of $T(\boldsymbol{d}, \bar{\boldsymbol{\theta}}_M)$ (denoted previously as r^2) for each model, where $\bar{\boldsymbol{\theta}}_M$ is the posterior mean under model M, is displayed in Figure 3.25 to highlight the magnitude of its discrepancy with the $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$. It can be seen that the $T(\boldsymbol{d}, \bar{\boldsymbol{\theta}}_M)$ for the overdispersion models lies somewhere in the middle of the histograms, while the $T(\boldsymbol{d}, \bar{\boldsymbol{\theta}}_M)$ for the Bayesian PLC model (15379.91) is completely off the chart. Moreover, the posterior predictive p-value, defined as

$$p_B = \Pr(T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M) \geq T(\boldsymbol{d}, \boldsymbol{\theta}_M)|\boldsymbol{d}), \tag{3.10}$$

where the probability is taken with respect to the joint distribution, $f(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M|\boldsymbol{d}, M)$, can be used to assess statistical significance formally. Note that the equation in (3.10) can be re-expressed as

$$p_B = \int I_{T(\boldsymbol{d}^{\text{rep}},\boldsymbol{\theta}_M)\geq T(\boldsymbol{d},\boldsymbol{\theta}_M)}(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M) f(\boldsymbol{d}^{\text{rep}}|\boldsymbol{\theta}_M, M) f(\boldsymbol{\theta}_M|\boldsymbol{d}, M) \mathrm{d}\boldsymbol{\theta}_M \mathrm{d}\boldsymbol{d}^{\text{rep}},$$

where

$$I_{T(\boldsymbol{d}^{\text{rep}},\boldsymbol{\theta}_M)\geq T(\boldsymbol{d},\boldsymbol{\theta}_M)}(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M) = \begin{cases} 1 & \text{if } T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M) \geq T(\boldsymbol{d}, \boldsymbol{\theta}_M), \\ 0 & \text{otherwise.} \end{cases}$$

In practice, it is easily computed as

$$p_B \approx \frac{\sum_{i=1}^N I_{T(\boldsymbol{d}^{\text{rep}},\boldsymbol{\theta}_M)\geq T(\boldsymbol{d},\boldsymbol{\theta}_M)}(\boldsymbol{d}^{\text{rep}\,i}, \boldsymbol{\theta}_M^i)}{N},$$

which is the proportion of the predictive test quantity, $T(\boldsymbol{d}^{\text{rep}\,i}, \boldsymbol{\theta}_M^i)$, which equals or exceeds the realised test quantity, $T(\boldsymbol{d}, \boldsymbol{\theta}_M^i)$ for $i = 1, \ldots, N$ (note that this is not $T(\boldsymbol{d}, \bar{\boldsymbol{\theta}}_M)$). If the model under assessment fits the data well, we should expect the values of $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$ to be close to $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$, resulting in a Bayesian p-value of around 0.5. An extreme Bayesian p-value, either too large or too small, is an indication of a lack of goodness of fit, signifying potential failure of the model under assessed. 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.

An alternative visualization is to examine the scatter plot of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$ against $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$. If the model being assessed is of good fit, those points would scatter evenly on each side of the line denoting equality. Note that the posterior predictive p-value is just the proportion of points above the equality line. As displayed in Figure 3.26, the cloud of points lies slightly towards the right side of the equality line for both the overdispersion models, suggesting that they have slightly larger values of $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$ overall, but not significantly large enough to indicate model inadequacy. For the Bayesian PLC model, the cloud of points is situated at the far right of the equality line (with $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$ overwhelmingly larger than $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$), signifying again that it is rather implausible for the data to be generated from this model.

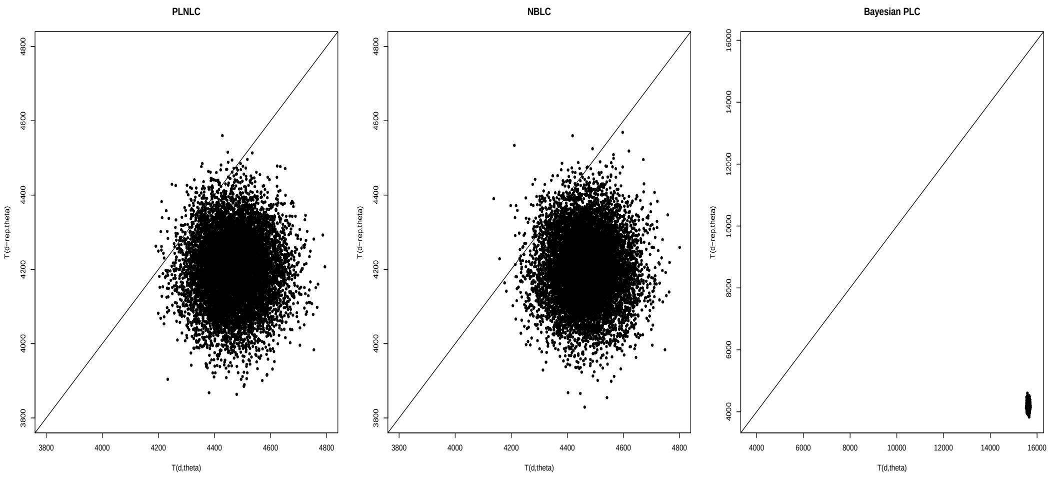


Figure 3.26: Scatter plots of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta_M})$ against $T(\boldsymbol{d}, \boldsymbol{\theta_M})$ for the PLNLC, NBLC, and Bayesian PLC model, with the solid lines denoting equality.

3.8.6 Out-of-Sample Validation

In this section, 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, derived using the projected underlying mortality rates and the holdout exposure data (see Section 3.7). The 11-years ahead forecast of crude mortality rates under the competing models and the holdout data for ages 0, 65 and 80 are depicted in Figure 3.27. The performances of the models in terms of their coverages vary across ages. In particular, the projections of mortality improvement for infants are over-optimistic by all of the candidate models. On the contrary, the projected mortality improvement for age 65 are over-pessimistic, with underwhelming coverages due to the jump-off discontinuity. This is caused by the absence of the cohort components as we shall demonstrate in Section 6.3. By contrast, 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 slightly outperform the Bayesian PLC model in terms of predictive ability. However, the relatively low coverages for some ages (e.g. age 0 and 65) are slightly worrying.

It is perhaps more useful to perform the validation based on an aggregate mortality quantity, the life expectancy at birth, derived from the projected crude mortality rates (instead of focusing on a specific age). As illustrated in Figure 3.28, 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.

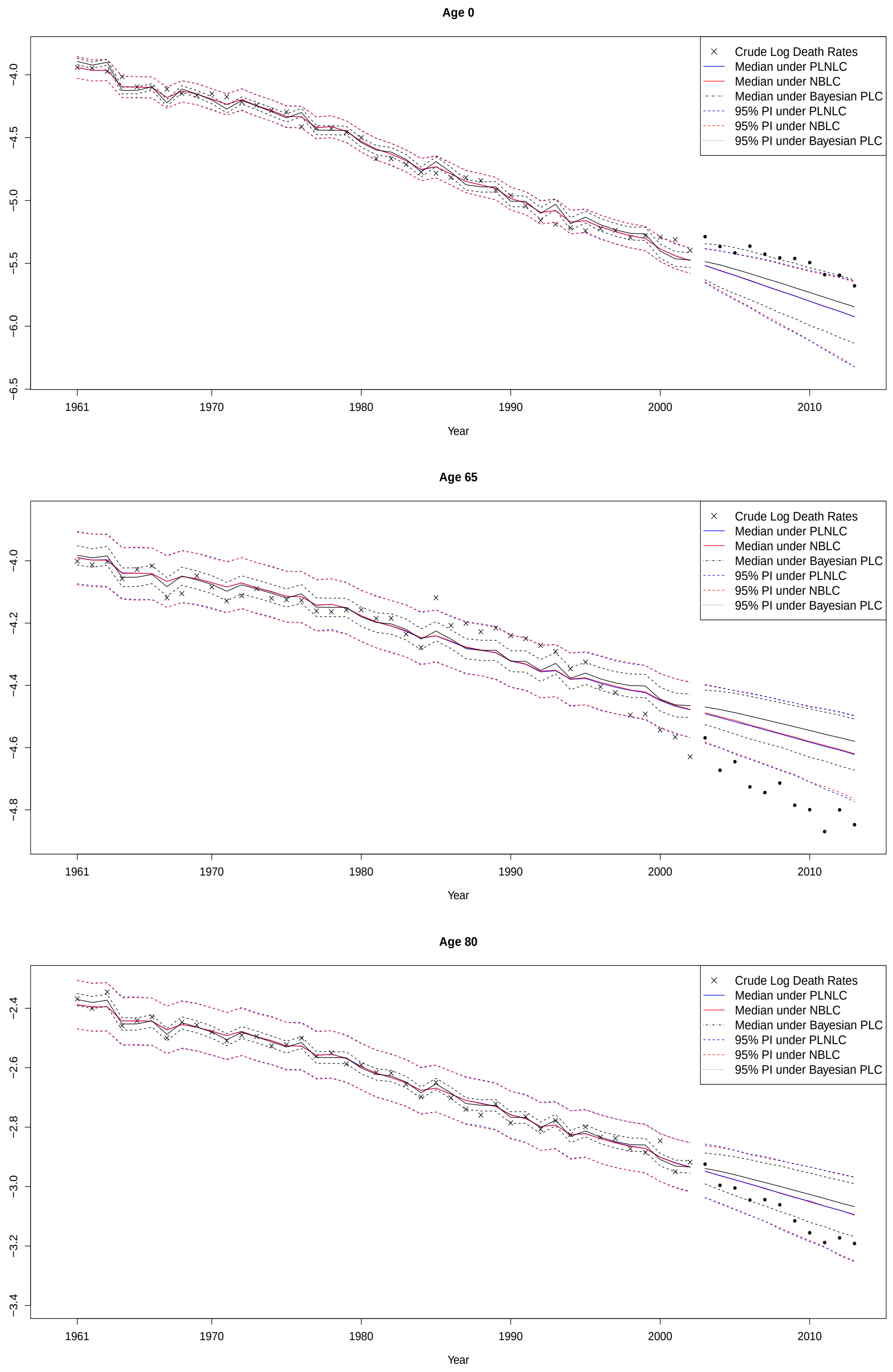


Figure 3.27: Plot of the observed, fitted crude log mortality rates and the associated 11-years ahead median forecasts of crude mortality rates for age 0, 65 and 80 under the Bayesian PLC and overdispersion models, accompanied by the 95% prediction intervals.

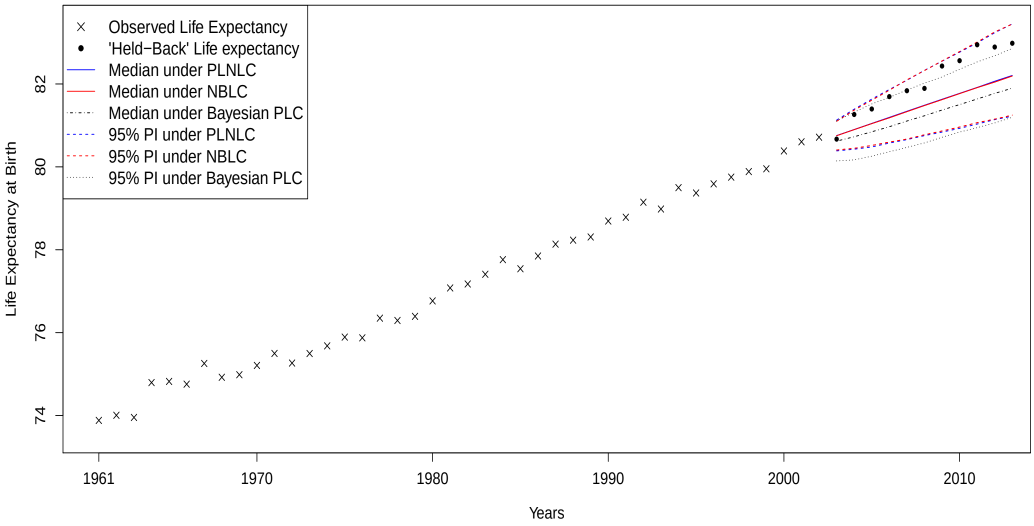


Figure 3.28: Plots of the observed life expectancy at birth and the associated 11-years ahead forecast under the Bayesian PLC and the overdispersion models, accompanied by the 95% prediction intervals.

3.8.7 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 (since the two models are essentially the same conditional on μ_{xt}). Recall that if the two distributions are identical, the sample QQ plot of the log mortality rates should coincide with the equality line. A U-shaped and S-shaped sample QQ plot indicate that one of the models possesses larger skewness and heavier tail respectively for the mortality rates.

The sample QQ plots of $\log \mu_{xt}$ for several selected ages and years are presented in Figure 3.29.

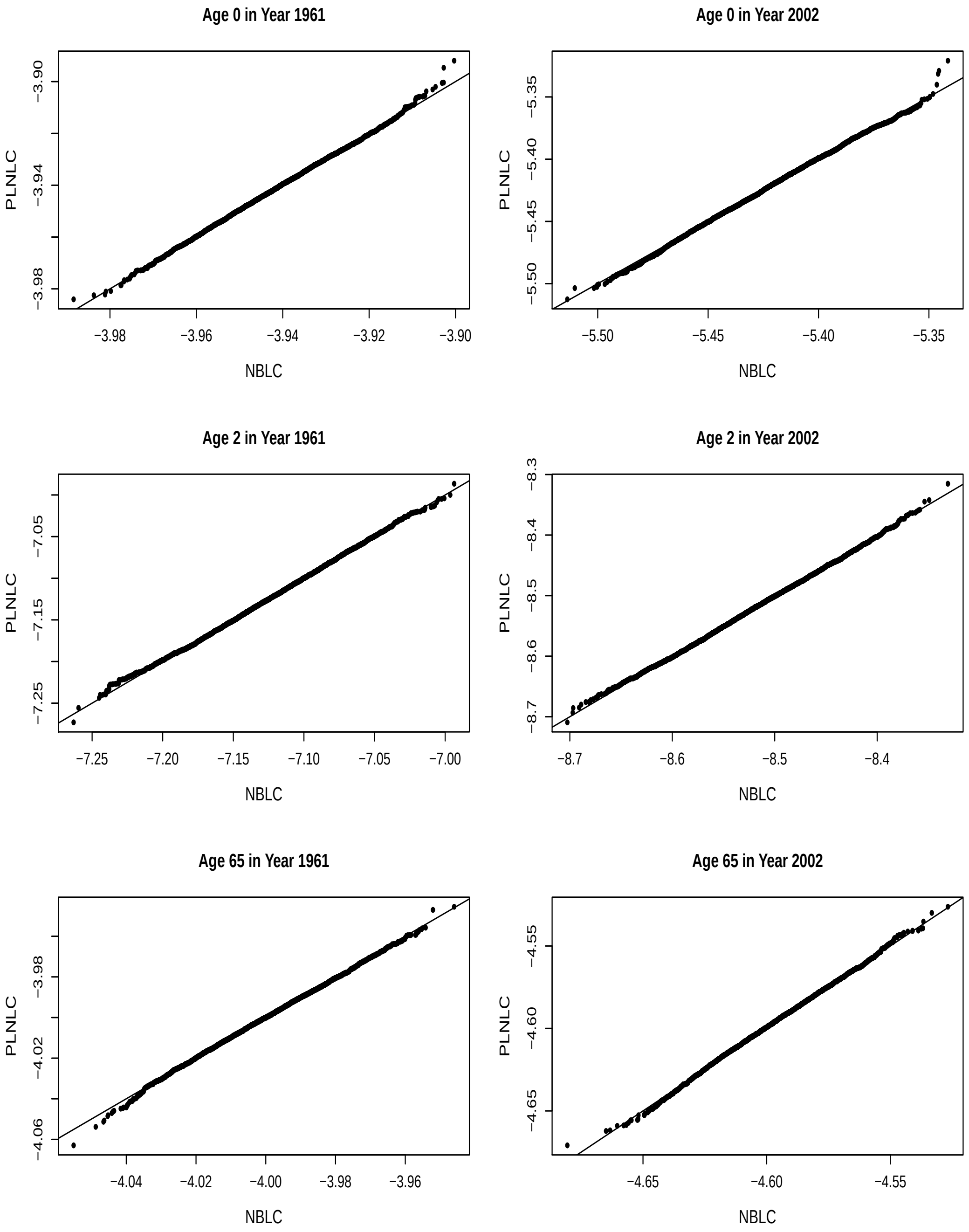


Figure 3.29: QQ Plots of the posterior samples of several chosen fitted log mortality rates, $\log\mu_{1\,1}$, $\log\mu_{1\,42}$, $\log\mu_{3\,1}$, $\log\mu_{3\,42}$, $\log\mu_{66\,1}$, $\log\mu_{66\,42}$ for the overdispersion models, with solid lines denoting equality.

From Figure 3.29, 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. Equivalently, instead of using the sample QQ plots, the empirical cumulative distribution function (ECDF) can be constructed to illustrate the similarities of the posterior distributions of $\log\mu_{xt}$ between the models. According to Figure 3.30, the ECDF constructed under both overdispersion models are virtually identical, agreeing with the result from the sample QQ plots.

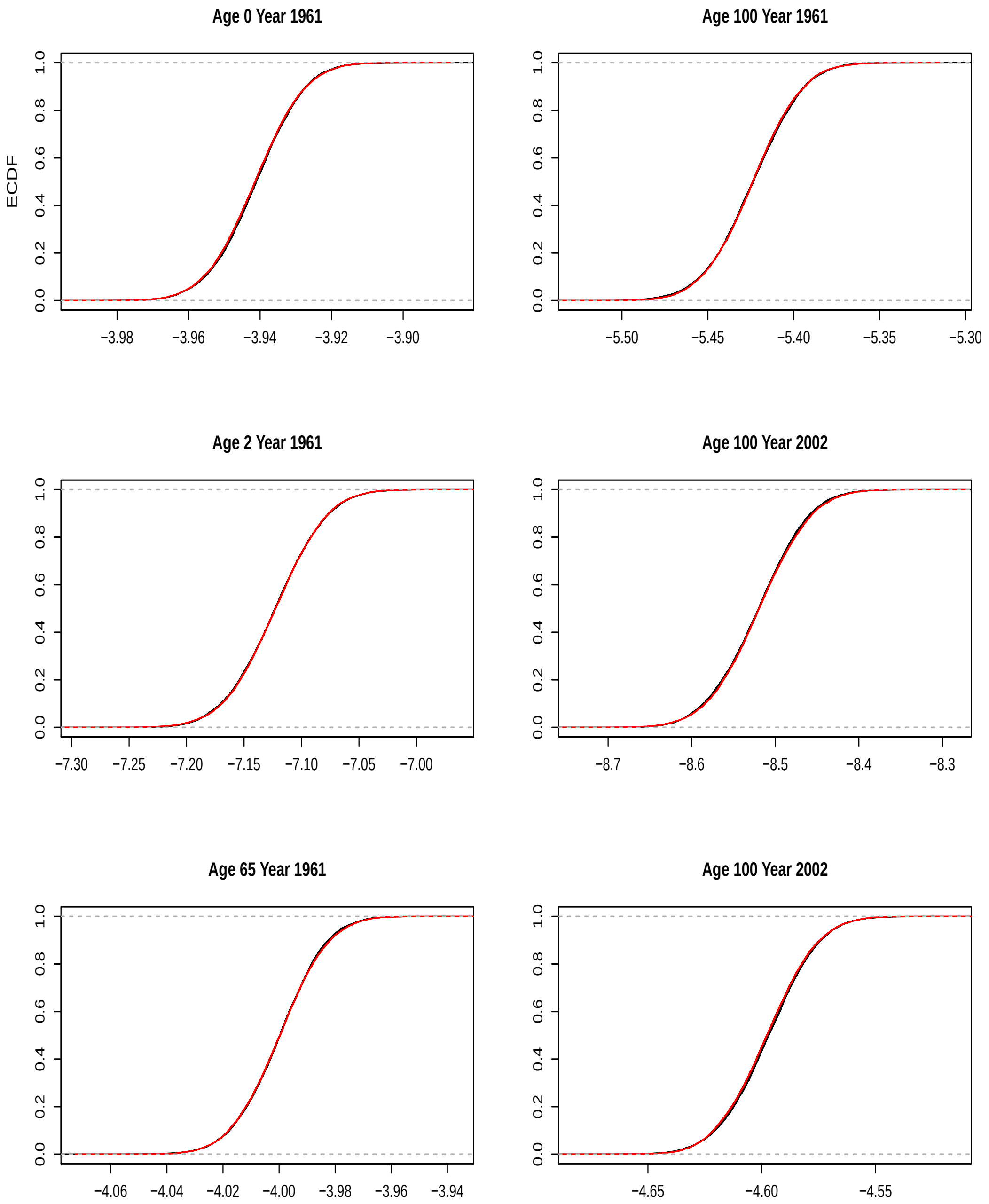


Figure 3.30: ECDF plots of posterior distributions of several chosen fitted log mortality rates, $\log \mu_{1\,1}$, $\log \mu_{1\,42}$, $\log \mu_{3\,1}$, $\log \mu_{3\,42}$, $\log \mu_{66\,1}$, $\log \mu_{66\,42}$ under the PLNLC (black) and NBLC model (red).

Recall from Section 3.8.4 that the projections of log mortality rates are visually identical for both overdispersion models. Figure 3.31 depicts the sample QQ plots of several chosen future log mortality rates. Again, most of the points lie on top of the equality line, indicating that the projection under the overdispersion models are indeed very similar.

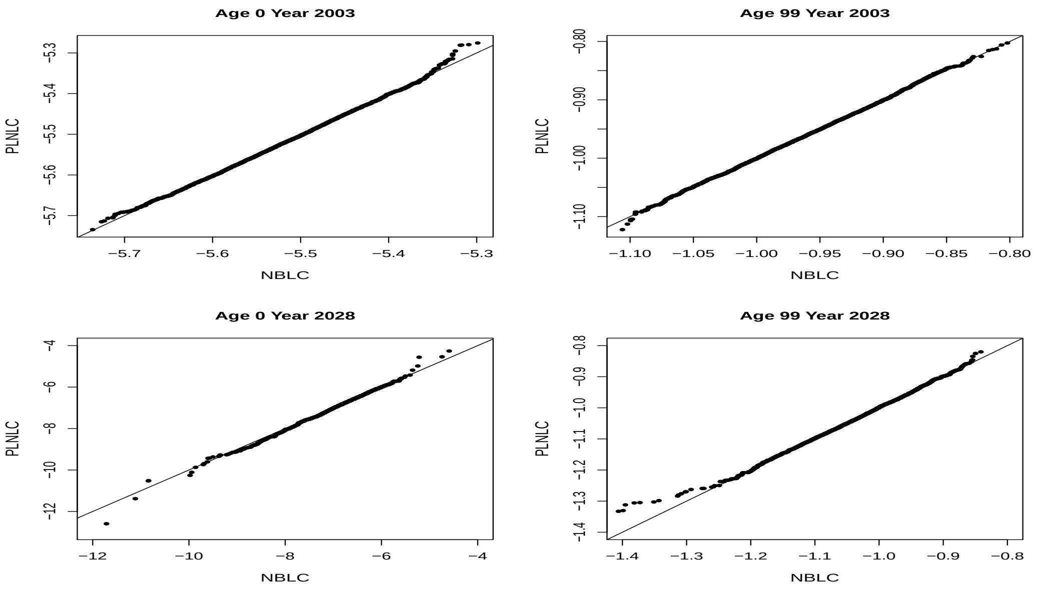


Figure 3.31: QQ plots of the posterior distributions of several selected projected log mortality rates, $\log \mu_{1\,44}$, $\log \mu_{100\,44}$, $\log \mu_{1\,68}$, $\log \mu_{100\,68}$.

Furthermore, the QQ plot of the dispersion parameters σ_μ^2 against $1/\phi$ is remarkably close to the reference line as depicted Figure 3.32, 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 3.5.1). Again, this signifies model similarity. In summary, the exploratory analysis above provides plenty of informal evidence that the two models are similar.

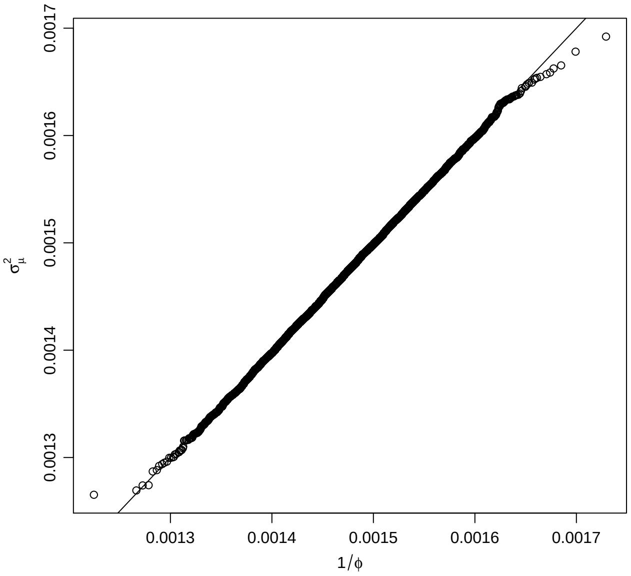


Figure 3.32: QQ plot of posterior sample of σ^2_μ against $1/\phi$, with black solid line denoting equality.

3.9 Conclusion

In this chapter, we focused on the importance of accounting for overdispersion in modelling mortality data. In particular, we presented two mixed Poisson LC models, the PLNLC and the PGLC model (or equivalently the NBLC model if the latent mortality rates are marginalised), 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. Several MCMC schemes for posterior samples generation were also considered. By comparing their speeds of generating effective posterior samples, we deduced that the PLNLC model with blocking and the NBLC model without blocking are the most efficient MCMC schemes under each overdispersion model. The subsequent inferences made were then based on posterior samples generated from these two schemes. 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 larger 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. The two proposed models provide rather similar qualitative fit, with the NBLC model producing slightly heavier-tailed posterior distributions. Formal Bayesian model comparison using posterior model probabilities will be presented in the next chapter to verify this similarity. Finally, the overdispersion models provide pronounced improvement in fit, but can be further refined by including cohort components. Until then, the dispersion parameters do not represent heterogeneity entirely in the sense that it is contaminated with the cohort effect (this point will be further elaborated in Chapter 6).

Chapter 4

Bayesian Model Determination

In the previous chapter, we witnessed some informal evidence of the similarities between our proposed overdispersion models both through the heat maps and sample QQ-plots. Here, we present a formal model comparison procedure. Model comparison within a Bayesian paradigm is particularly straightforward. Only a brief summary of Bayesian model comparison procedures is provided in this section (please see Carlin and Louis, 2000 for more details). Essentially, the model index, $M \in M^S$, is treated as a parameter and is similarly updated through the use of Bayes theorem (see Hoeting et al., 1999):

$$\begin{aligned} f(M|\boldsymbol{d}) &= \frac{f_M(\boldsymbol{d})f(M)}{f(\boldsymbol{d})} \\ &= \frac{f_M(\boldsymbol{d})f(M)}{\sum_{M \in M^S} f_M(\boldsymbol{d})f(M)}, \end{aligned} \tag{4.1}$$

where $f(M)$ is the prior model probability representing our prior belief concerning the "true" underlying model, $f(M|\boldsymbol{d})$ denotes the posterior model probability, and $f_M(\boldsymbol{d})$ denotes the marginal likelihood (ML) of model M. Notice that the denominator of Equation (4.1) does not depend on the model index, and hence can be ignored when comparing the models. Specifically, for two competing models,

$$\frac{f(M_1|\boldsymbol{d})}{f(M_2|\boldsymbol{d})} = \frac{\frac{f_{M_1}(\boldsymbol{d})f(M_1)}{f(\boldsymbol{d})}}{\frac{f_{M_2}(\boldsymbol{d})f(M_2)}{f(\boldsymbol{d})}} = \frac{f_{M_1}(\boldsymbol{d})}{f_{M_2}(\boldsymbol{d})} \times \frac{f(M_1)}{f(M_2)}.$$

In words, the posterior odds in favour of model M_1 is equal to the corresponding Bayes factor (BF) multiplied by its prior odds, where the BF in favour of M_1 is defined as the ratio of MLs,

$$BF_{12} = \frac{f_{M_1}(\boldsymbol{d})}{f_{M_2}(\boldsymbol{d})}.$$

The BF can be interpreted as the evidence of model preference given by the data, and therefore is regarded as the key quantity in Bayesian model selection. Typically, the prior model probabilities are chosen to be uniform (implying that we do not favour any model a priori), especially when the number of parameters in the competing models are the same, as in our overdispersion models. Consequently, the posterior odds equates the BF.

Computation of the BF requires the values of MLs, which can be expressed as

$$\begin{aligned} f_M(\boldsymbol{d}) &= \int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{d}, \boldsymbol{\theta}_M)\mathrm{d}\boldsymbol{\theta}_M \\ &= \int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{d}|\boldsymbol{\theta}_M) f_M(\boldsymbol{\theta}_M)\mathrm{d}\boldsymbol{\theta}_M. \end{aligned} \tag{4.2}$$

The integral form in (4.2) suggests that the ML is the marginalization of the data likelihood with respect to the prior distributions. In other words, the ML is a form of likelihood penalization according to the prior distributions specified, thus serves as data indication of model preference. In addition, ML also penalises over-parametrization because having extra unnecessary parameters at regions where the likelihood has negligible density will decrease the value of ML and, hence, implies a smaller posterior model probability.

One of the main advantages of using BF for comparing models is that it is not essential for the competing models to be nested, as the computation of the BF involves integrated likelihoods. This is in contrast to the likelihood ratio test for model comparison which involves maximised likelihoods, and thus, requires competing models to be nested. On the other hand, the choice of prior distributions has a crucial impact on the value of ML. First of all, it is ideal for all the prior distributions to be proper because improper priors lead to undefined ML by having infinite normalizing constants, which complicates the resulting posterior inferences. In some cases, specification of improper priors (e.g. Jeffrey's prior in simple hierarchical normal set up) still yields proper posterior inferences. However, in complex problems such as ours, it is rather difficult to even guarantee the propriety of our posterior unless we employ proper prior distributions throughout. Secondly, **Bartlett's Paradox** (Bartlett, 1957) dictates that arbitrarily diffuse or improper uniform prior distributions should be avoided wherever possible since a fully Bayesian model selection procedure using Bayes factor will incline towards automatically favouring the more parsimonious model (see Appendix G for a description and an example of the paradox).

4.1 Literature Review on the Computation of BFs/MLs

A brief review of some of the computational methods of BFs and MLs is presented here (see Kass and Raftery, 1995 or Carlin and Louis, 2000 for a comprehensive review). The derivation of posterior model probabilities is conceptually straightforward using Bayes' Theorem (by computing the BF), but is rather computationally intensive in most situations. Except in scenarios where point hypotheses are tested (when BF can be found by simply plugging in the point values), computation of the BF is a problem of evaluating the integration shown in (4.2). In some simple cases, analytical expressions of BFs are available if the prior distributions exhibit conjugacy. However, this is rarely the case since the BFs are often intractable, and hence, necessitates numerical methods. Numerical integration techniques such as quadrature methods can be used when the dimensionality involved is small, but rarely find successes in high-dimensional problems, especially when sample sizes are large. This is because the integrand in (4.2) is peaked at a particular region when the sample size is large, so quadrature methods (whose efficiencies rely on the partitioning of parameter spaces) generally have difficulty approximating the integration efficiently as knowledge of the location of the integrand's mass is not accounted for. One way of tackling this issue is to use the Laplace's method of approximation (see for example Tierney and Kadane, 1986), which makes use of the information about the likely location of the integrand mass. Laplace's method is essentially the use of a normal approximation (by matching the posterior mode and curvature) to estimate the ML. Again, the accuracy of the Laplace's method is dictated by the sample size relative to the dimension of the problem, because large sample size is essential for the normality assumption of the posterior to work well (Kass and Raftery, 1995 stated that a sample size of less than $5d$ is troublesome, where d is the dimension of the parameter). For other variants of the Laplace's method, please refer to Kass and Vaidyanathan (1992).

Alternatively, the Monte Carlo (MC) method (a sampling based approach) can be used to obtain an approximation of the ML. From Equation (4.2), it is straightforward to see that the ML is the expectation of the likelihood with respect to the prior distribution, $f_M(\boldsymbol{\theta}_M)$, that is

$$f_M(\boldsymbol{d}) = \mathbb{E}_{\text{prior}}[f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)].$$

Hence, the simplest MC approximation of the ML is

$$\frac{\sum_{i=1}^{N} f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i)}{N},$$

where $\{\boldsymbol{\theta}_M^i\}$, for $i = 1, \ldots, N$, is a sample of size N from the prior distribution, $f_M(\boldsymbol{\theta}_M)$. A major pitfall of this method, resulting in its limited usage, is that if the likelihood is concentrated relative to the prior distribution (which is typically the case), then the estimator will be very inefficient. Specifically, an estimate of this form consists of mostly $f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i)$ with negligible values, dominated by a few large values (that occurs occasionally when the simulated prior samples are at the location where likelihood is peaked), resulting in a large variance. Another approximation of the ML is the importance sampling method, given as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i)f_M(\boldsymbol{\theta}_M^i)}{h(\boldsymbol{\theta}_M^i)},$$

where $h()$ is the importance sampling distribution to be specified and $\{\boldsymbol{\theta}_M^i\}$, for $i = 1, \ldots, N$, is a sample of size N from $h()$. For the importance sampling estimator to be well-behaved in the sense of having a finite variance, $h()$ is required to possess tails as heavy as the posterior distribution. Otherwise, a few realizations from the tail of $h()$ will dominate the value of the estimator.

If posterior samples are available (e.g. from Markov chain Monte Carlo (MCMC) methods), the following methods can be considered. Newton and Raftery (1994) proposed the harmonic mean estimator of the ML, given as

$$\left[\frac{\sum_{i=1}^{N}(f_M(\boldsymbol{d}|\boldsymbol{\theta}_M^i))^{-1}}{N}\right]^{-1},$$

where now $\{\boldsymbol{\theta}_M^i\}$, for $i = 1, \ldots, N$, is a sample of size N from the posterior distribution. By noting that

$$\begin{aligned} 1 &= \int f_M(\boldsymbol{\theta}_M)\mathrm{d}\boldsymbol{\theta}_M \\ &= \int \frac{f_M(\boldsymbol{\theta}_M)}{f_M(\boldsymbol{\theta}_M|\boldsymbol{d})} f_M(\boldsymbol{\theta}_M|\boldsymbol{d})\mathrm{d}\boldsymbol{\theta}_M \\ &= \int \frac{f_M(\boldsymbol{d})}{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)} f_M(\boldsymbol{\theta}_M|\boldsymbol{d})\mathrm{d}\boldsymbol{\theta}_M \text{ (using Bayes Theorem)}, \end{aligned}$$

the ML can be expressed alternatively as

$$\begin{aligned} f_M(\boldsymbol{d}) &= \left[\int \frac{1}{f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)} f_M(\boldsymbol{\theta}_M|\boldsymbol{d})\mathrm{d}\boldsymbol{\theta}_M\right]^{-1} \\ &= \frac{1}{\mathbb{E}_{\text{posterior}}[(f_M(\boldsymbol{d}|\boldsymbol{\theta}_M))^{-1}]}, \end{aligned}$$

suggesting the harmonic mean estimator as the MC average of the ML. This implies that the harmonic mean estimator is essentially an application of the importance sampling method to evaluate the normalizing constant of the prior distribution, $f_M(\boldsymbol{\theta}_M)$, using the posterior distribution, $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$, as the importance sampling distribution. Since the posterior distribution is typically more concentrated (less heavy-tailed) than the prior distribution, this estimator is inefficient in practice because it is usually dominated by a few realizations at the tail of the posterior distribution, resulting in a large variance. This is regarded as the main drawback of this estimator, as noted by Newton and Raftery (1994), who proceeded to provide several modifications of the harmonic mean estimator. One of them is to use a mixture of the prior and posterior densities as the importance sampling function instead, which improves the efficiency by having a heavier-tailed importance sampling function, but requires samples from the prior distribution. On the other hand, Raftery et al. (1995) proposed the Laplace-Metropolis estimator, which uses a combination of the Laplace's method and Metropolis algorithm (for estimating the posterior mode and curvature at the mode) to approximate the ML.

The Chib's method, as proposed in Chib (1995), approximates the ML using Gibbs outputs, when the conditional posterior densities are tractable. Chib and Jeliazkov (2001) extended the original Chib's method to allow for cases where conditional posterior densities are not necessarily tractable using samples from Metropolis-Hastings outputs. The main disadvantage of both these approaches is that they are computationally demanding because additional samples (on top of posterior samples) are required. Meng and Wong (1996) proposed the bridge sampler, which is a way to estimate ratio of normalizing constants based on a simple identity using the "bridge" function (see next section), provided draws are available from the distributions involved. Meng and Wong (1996) also mentioned that bridge sampling is a generalization of several algorithms that encompass a wide range of sampling-based normalizing constants estimation methods such as the importance sampling, harmonic mean estimator, Chib's method and so forth. This technique will be the main focus of the thesis in this chapter, and so will be described in detail in Section 4.2. Gelman and Meng (1998) generalised the idea of bridge sampling by proposing to construct a continuous "path" linking the distributions instead, hence the name path sampling.

Some methods that approximate posterior model probabilities (which can be used to derive BFs) by searching directly in the model space are also available. The Product Space Search method presented by Carlin and Chib (1995) estimates posterior model probabilities using pseudo priors, treating the model indicator, M, explicitly as a parameter. Dellaportas et al. (2002) extended the method by Carlin and Chib (1995) to allow for Metropolis updating of the model selection step, forming a hybrid version known as the "Metropolised" Carlin and Chib. The reversible jump Markov chain Monte Carlo (MCMC) method by Green (1995) is a general strategy which creates a Markov chain that allows transitions between the model space and parameter spaces (of varying dimensionalities).

Thus far, most of the approaches described involve scenarios where subjective specification of proper prior distributions (with certain prior information embedded) is undertaken. On occasions where no such information is available for such specifications or automatic methods of model selection are intended, then the objective Bayesian model selection methods (mostly involve construction of default priors) can be considered. The Bayesian Information Criterion (BIC) developed by Schwarz (1978) provides an asymptotic approximation of the ML directly. It was also pointed out by Raftery (1999) that the BIC is effectively an approximate BF in an asymptotic sense when a unit information prior is used. If one wishes to use improper priors, then the fractional BF and intrinsic BF approaches developed by O'Hagan (1995) and Berger and Pericchi (1996) respectively can be used. The fractional BF approach works by converting the improper priors into proper priors using a fraction of the likelihood, while the intrinsic BF approach does so by using a part of the data. For more details on objective Bayesian model selection methods, see for example Berger and Pericchi (2001).

4.2 Bridge Sampling

Bridge sampling is a sampling-based technique originally developed by Meng and Wong (1996) to estimate the ratio of two normalizing constants. Suppose that $p_i(\boldsymbol{\theta})$ $(i = 1, 2)$ are two densities with parameter spaces $\boldsymbol{\Theta}_i \subset \mathbb{R}^d$ respectively, d is the dimension of $\boldsymbol{\theta}$, and are known up to a normalizing constant, i.e.

$$p_i(\boldsymbol{\theta}) = \frac{q_i(\boldsymbol{\theta})}{c_i},$$

where c_i are the corresponding normalizing constants of the unnormalised densities, $q_i(\boldsymbol{\theta})$. The fundamental usage of bridge sampling is based on the following simple key identity,

$$r \equiv \frac{c_1}{c_2} = \frac{\mathbb{E}_2[q_1(\boldsymbol{\theta})\omega(\boldsymbol{\theta})]}{\mathbb{E}_1[q_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})]}, \tag{4.3}$$

which can be derived by realizing that

$$1 \;=\; \frac{\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} \frac{q_1(\boldsymbol{\theta})}{c_1}p_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} p_1(\boldsymbol{\theta})\frac{q_2(\boldsymbol{\theta})}{c_2}\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{c_2}{c_1}\cdot\frac{\int_{\boldsymbol{\Theta}_2}[q_1(\boldsymbol{\theta})\omega(\boldsymbol{\theta})]p_2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}_1}[q_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})]p_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},$$

since $p_1(\boldsymbol{\theta})\cdot p_2(\boldsymbol{\theta}) = 0$ for regions $\boldsymbol{\Theta}_1\backslash(\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2)$ and $\boldsymbol{\Theta}_2\backslash(\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2)$, where $\omega(\boldsymbol{\theta})$ is the so called bridge function (defined on the common support $\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2$) satisfying

$$0 < \left|\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right| < \infty, \tag{4.4}$$

so that the ratio in Equation (4.3) is well defined. The condition in (4.4) is easily guaranteed by having

$$\int_{\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2} p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} > 0,$$

which simply means that the common support, $\boldsymbol{\Theta}_1\cap\boldsymbol{\Theta}_2$, is not an empty set (Meng and Wong 1996). In other words, the existence of $\omega()$ for Equation (4.3) (and hence the bridge sampler to be valid) is ensured as long as the two densities “overlap”. Given that the above condition is satisfied, the Monte Carlo estimate of r is provided as

$$\hat{r} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} q_1(\boldsymbol{\theta}_2^i)\omega(\boldsymbol{\theta}_2^i)}{\frac{1}{N_1}\sum_{i=1}^{N_1} q_2(\boldsymbol{\theta}_1^i)\omega(\boldsymbol{\theta}_1^i)},$$

where $\{\boldsymbol{\theta}_1^i\}$, for $i = 1, 2, \ldots, N_1$, and $\{\boldsymbol{\theta}_2^i\}$, for $i = 1, 2, \ldots, N_2$, are random (possibly dependent) realizations from $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ respectively. Under certain regularity conditions, $\hat{r}$ converges asymptotically to the true value, r (i.e. the sample averages in Equation 4.3 converge to their respective population averages).

Meng and Wong (1996) proposed that an optimal choice of $\omega()$, in the sense of minimizing the asymptotic relative mean square error, is given by the reciprocal of a mixture between the two densities,

$$\omega^*(\boldsymbol{\theta}) \propto \frac{1}{N_1 q_1 + r N_2 q_2}, \tag{4.5}$$

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 (or N_2). Alternatively, k-thinning can be applied to obtain a set of approximately independent MCMC samples, where k is chosen such that the sample auto-correlations are close to zero. Note that $\omega^*()$ still involves the unknown r, which then requires iterative computations to evaluate $\hat{r}$ (see below).

Bridge sampling can be applied in the context of approximation of ML if we construct the algorithm such that the second normalising constant is known. In particular, this can be undertaken by setting $q_1(\boldsymbol{\theta}_M) = f_M(\boldsymbol{d}|\boldsymbol{\theta}_M)f_M(\boldsymbol{\theta}_M)$ and $q_2(\boldsymbol{\theta}_M) = g_M(\boldsymbol{\theta}_M)$, where $g_M(\boldsymbol{\theta}_M)$ is the density of a normalised distribution. Then, the bridge sampling estimator of ML of model M 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 $\{\boldsymbol{\theta}^i_M\}_{i=1}^{N_1}$ is a sample of size N_1 from the posterior distribution with density $f_M(\boldsymbol{\theta}_M|\boldsymbol{d})$, $\{\tilde{\boldsymbol{\theta}}^i_M\}_{i=1}^{N_2}$ is a sample of size N_2 from a normalised distribution with density $g_M()$, and $\omega_M()$ satisfies $0 < |\int f_M(\boldsymbol{\theta}_M|\boldsymbol{d})g_M(\boldsymbol{\theta}_M)\omega_M(\boldsymbol{\theta}_M)\mathrm{d}\boldsymbol{\theta}_M| < \infty$. The asymptotically optimal bridge function, $\omega^*_M()$, is

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

in this case. However, $\omega^*_M()$ still depends on the unknown ML, $f_M(\boldsymbol{d})$, so Meng and Wong (1996) suggest an iterative procedure for estimating $f_M(\boldsymbol{d})$:

$$\hat{f}^*_M(\boldsymbol{d})^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{\tilde{l}_i}{N_1\tilde{l}_i + N_2\hat{f}^*_M(\boldsymbol{d})^{(t)}}\right]}{\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\frac{1}{N_1 l_i + N_2\hat{f}^*_M(\boldsymbol{d})^{(t)}}\right]}, \tag{4.6}$$

where $l_i = \frac{f_M(\boldsymbol{d}|\boldsymbol{\theta}^i_M)f_M(\boldsymbol{\theta}^i_M)}{g_M(\boldsymbol{\theta}^i_M)}$ and $\tilde{l}_i = \frac{f_M(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}^i_M)f_M(\tilde{\boldsymbol{\theta}}^i_M)}{g_M(\tilde{\boldsymbol{\theta}}^i_M)}$. Starting with an initial guess, $\hat{f}^*_M(\boldsymbol{d})^{(0)}$, the bridge sampling estimate, $\hat{f}^*_M(\boldsymbol{d})$, of the ML can be obtained by iterating (4.6) until convergence.

The choice of the density $g_M()$ is entirely arbitrary, but ill-specification can be detrimental to the accuracy of the bridge sampling estimate. In practice, 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. The 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, as we shall demonstrate through Simulation Study 2 in Section 4.4. Therefore, a better alternative is to use sample mean and variance computed directly from the posterior sample for 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 ML (see Overstall and Forster, 2010). We further investigate this matter and provide some recommendations through Simulation Study 3 in Section 4.5.

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 $g_M()$. Although Chen et al. (2000, p. 129) stated that the optimal choice of $\omega()$ itself is often more crucial than the optimal allocation of sample sizes, we conducted a simulation study to investigate the effect of various sample size allocations (see Simulation Study 4 in Section 4.6).

4.2.1 Marginal Likelihoods of the PLNLC and NBLC Model

Without considering in detail the potential misestimation of the bridge sampling estimator, the marginal likelihoods of the overdispersion models are computed here for the purpose of illustration. The general algorithm for computing the marginal likelihood using bridge sampling is

1. Generate a sample, $\{\boldsymbol{\theta}_M^1, \ldots, \boldsymbol{\theta}_M^N\}$, of size N 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 in model M.
3. Generate a sample, $\{\tilde{\boldsymbol{\theta}}_M^1, \ldots, \tilde{\boldsymbol{\theta}}_M^N\}$, of size N from the density $g_M()$.
4. Obtain the bridge sampling estimate of marginal likelihood, $\hat{f}_M^*(\boldsymbol{d})$, using (4.6), evaluated at the samples $\{\boldsymbol{\theta}_M^1, \ldots, \boldsymbol{\theta}_M^N\}$ and $\{\tilde{\boldsymbol{\theta}}_M^1, \ldots, \tilde{\boldsymbol{\theta}}_M^N\}$, so that $N_1 = N_2 = N$.

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

$$\begin{aligned} f_{\text{PLNLC}}(\boldsymbol{d}) &= \int f(\boldsymbol{d}|\log\boldsymbol{\mu})f(\log\boldsymbol{\mu}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\log\sigma_\mu^2)f(\boldsymbol{\alpha})f(\boldsymbol{\beta}_{-1}|\log\sigma_\beta^2)f(\boldsymbol{\kappa}_{-1}|\rho,\log\sigma_\kappa^2,\boldsymbol{\psi}) \\ &\quad\times f(\rho)f(\log\sigma_\kappa^2)f(\log\sigma_\beta^2)f(\boldsymbol{\psi})f(\log\sigma_\mu^2)\mathrm{d}\boldsymbol{\theta}_{\text{PLNLC}}, \end{aligned} \tag{4.7}$$

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

$$\begin{aligned} f_{\text{NBLC}}(\boldsymbol{d}) &= \int f(\boldsymbol{d}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\log\phi)f(\boldsymbol{\alpha})f(\boldsymbol{\beta}_{-1}|\log\sigma_\beta^2)f(\boldsymbol{\kappa}_{-1}|\rho,\log\sigma_\kappa^2,\boldsymbol{\psi}) \\ &\quad\times f(\rho)f(\log\sigma_\kappa^2)f(\log\sigma_\beta^2)f(\boldsymbol{\psi})f(\log\phi)\mathrm{d}\boldsymbol{\theta}_{\text{NBLC}}, \end{aligned} \tag{4.8}$$

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 $p_{\text{NBLC}} = 246$.

With a sample size of $N = 10000$ (after thinning by 50) from the posterior distribution, the log marginal likelihood of the NBLC model is estimated to be around -23727.48 using the above algorithm. We experienced major difficulty during the computation of the bridge sampling estimate of 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 $p_{\text{PLNLC}} = 4446$. In particular, the bridge sampling estimate appears to vary according to sample size, as illustrated in Table 4.1. While this may be an indication of non-convergence of the MCMC generated posterior samples, this is not entirely the case. As we shall demonstrate in Section 4.5, this is merely an artefact of using a normal distribution, with its moments computed from the posterior samples, as $g_M()$. Additionally, a series of simulation studies is also conducted and presented in the next few sections to gain more insights to the potential failure and to explore other possibilities for improving the accuracy of the bridge sampling estimator.

Table 4.1: The marginal likelihoods (on logarithmic scale) of the PLNLC model approximated from bridge sampling for various sample sizes.

Sample Size, N	Estimated Marginal likelihood of PLNLC Model, $\hat{f}^*_{\text{PLNLC}}(\boldsymbol{d})$
50000	-23833.26
100000	-23779.53
250000	-23747.43

4.3 Simulation Study 1: The Effect of MCMC Samples Dependence on The Accuracy of Bridge Sampling Estimator

Generally, the use of serially (positively) correlated samples in estimating the expectation of any function with respect to a distribution (using Monte Carlo average) results in an unbiased estimator, but with the standard error inflated in comparison to using independent samples (see for example Ripley, 1987). However, it is slightly more complicated for the bridge sampling estimator because the optimal bridge function, $\omega_M^*()$, also depends on the effective sample sizes and is derived based on the assumption that the samples are independent. Therefore, in Simulation Study 1, we investigate the effect of using serially correlated samples on the accuracy of the bridge sampling estimator, given our huge reliance on dependent MCMC samples for the computations.

Suppose that p_1 is the density of a p-dimensional normal distribution with mean $\mathbf{0}$ and variance matrix $\boldsymbol{I}_p$: $N(\mathbf{0}, \boldsymbol{I}_p)$. Suppose also that q_1 has the same density so that the normalising constant is one. Let p_2 and q_2 be the densities of a slightly displaced normal distribution with the same variance matrix: $N(0.1 \times \mathbf{1}_p, \boldsymbol{I}_p)$, so that the Mahalanobis distance between p_1 and p_2 is $0.1^2 \times p$. The ratio of normalising constants r is thus known to be one. Consider the following sampling mechanism:

1. The random walk Metropolis-Hastings algorithm is used to generate a set of dependent sample $\{\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^N\}$ from $N(\mathbf{0}, \boldsymbol{I}_p)$. The updating is performed univariately using the proposal variance c^2, where c is chosen such that the resulting trajectories of the MCMC exploration are fairly correlated (we chose $c = 0.38$).

2. Same as 1, but with the updating performed in a single block using the optimal proposal variance matrix $2.38^2 \times \boldsymbol{I}_p$.

The usual sampling method ("rmvnorm" in R) is then used to generate a set of pseudo independent sample $\{\boldsymbol{\theta}_2^1, \ldots, \boldsymbol{\theta}_2^{N_2}\}$ from the density p_2. We consider the sample sizes, N, from the set $\{100, 200, \ldots, 20000\}$, and let $N_1 = N_2 = N$ so that the bridge sampler is evaluated at the whole samples from p_1 and p_2. Each computation at each sample size is then replicated $R = 1000$ times. The interest here is to study the impact of using correlated samples on the accuracy of the bridge sampling estimator with respect to various sample sizes. We also examine two different dimensionalities $p = 10, 100$. Trace plots and autocorrelation plots of the resulting trajectories are illustrated in Figure 4.1. The samples for all four scenarios are fairly correlated, with the last case (block updating with $p = 100$) being the most correlated, which also appears to have poor mixing.

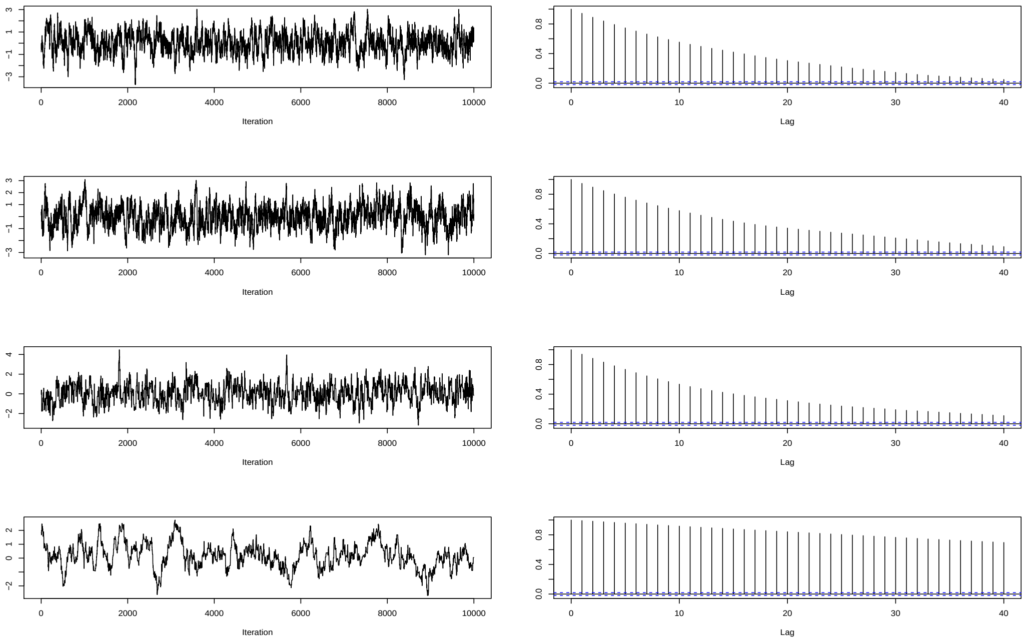


Figure 4.1: Trace plots and autocorrelation plots of the resulting trajectories of the univariate updating random walk MH with $p = 10$ (1st row), and $p = 100$ (2nd row), as well as block updating with $p = 10$ (3rd row) and $p = 100$ (4th row).

For the resulting bridge estimator $\hat{r}$, we assess its efficiency by monitoring the relative mean-square error (RMSE), defined as

$$\mathrm{RMSE}(\hat{r}) = \frac{\mathbb{E}[(\hat{r} - r)^2]}{r^2}. \tag{4.9}$$

For a given sample of size N in our case, it can be estimated as

$$\mathrm{R\hat{M}SE}(\hat{r}) \approx \frac{\frac{1}{R}\sum_{i=1}^{R}(\hat{r}^i - r)^2}{r^2} = \frac{\sum_{i=1}^{R}(\hat{r}^i - 1)^2}{R},$$

where $\hat{r}^i$ is the bridge sampling estimate computed at each replication. Note that Equation (4.9) can be re-expressed as

$$\mathrm{RMSE}(\hat{r}) = \frac{\mathrm{Var}(\hat{r}) + \mathrm{Bias}(\hat{r})^2}{r^2}, \tag{4.10}$$

where $\mathrm{Bias}(\hat{r}) = \mathbb{E}(\hat{r}) - r$. It is desirable for $\hat{r}$ to have a minimal RMSE (either by having a small variance or a small bias or both).

Plots of the median bridge sampling estimates and the associated 95% intervals against various sample sizes for different scenarios is illustrated in Figure 4.2 and Figure 4.3 respectively. Generally, both bias and standard error are decreasing functions of sample size, N, which is to be expected since larger sample sizes imply more information learned from the distributions, improving the bridge sampling estimates. For the cases where the MCMC samples are univariately updated, the bridge sampling estimator appears to perform rather efficiently by having biases that vanish quickly with respect to the sample size. Interestingly, the one with higher-dimensionality, $p = 100$ possesses smaller standard errors than that of $p = 10$. On the other hand, performing block updating results in estimates with relatively larger bias as compared to univariate updating. Moreover, the performance of bridge sampling worsens as the dimension increases in this case. We hypothesise that the relatively poor performance of block updating is possibly due to the poor mixing of the MCMC algorithm (as depicted in Figure 4.1), prohibiting the exploration of the distribution, p_1, hence the bias. Ultimately, if the sample is sufficiently long to represent the distribution, then the bridge sampling estimator will be unbiased. In other words, the main drawback of using auto-correlated samples is that it increases the standard error of the resulting bridge sampling estimator (but not the bias), unless the high correlation is directly related to the fact that the chain did not explore the distributions well enough.

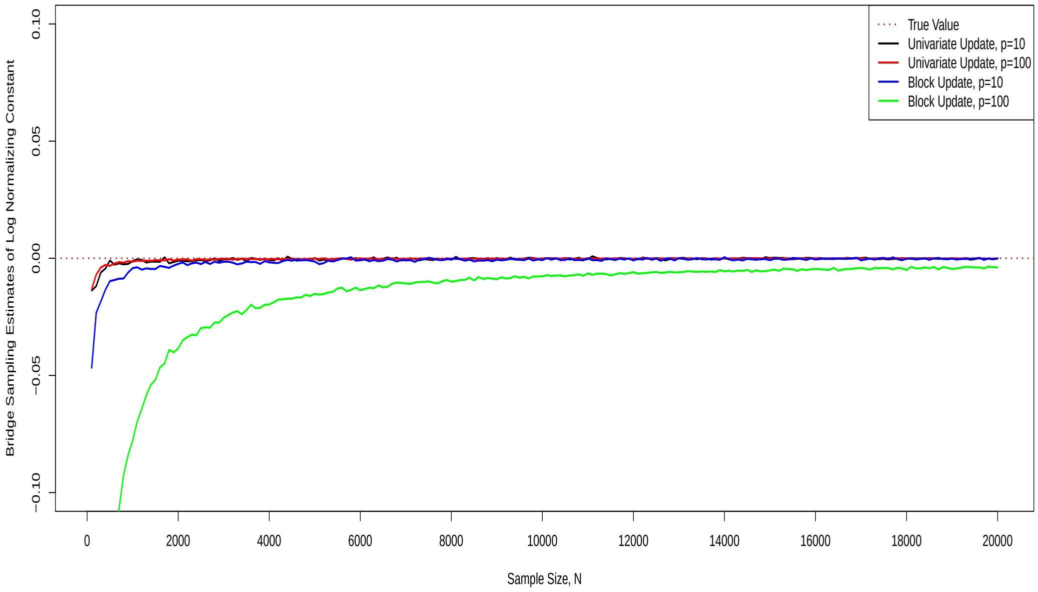


Figure 4.2: Plot of the median bridge sampling estimates of the log ratio of normalizing constant against sample size, N, in the case of using serially correlated samples.

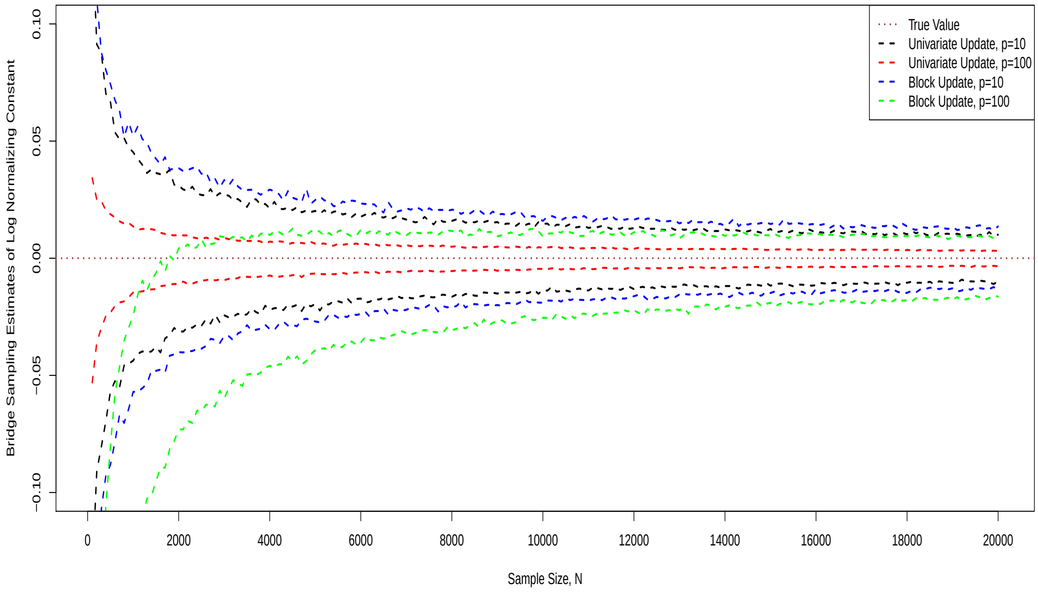


Figure 4.3: Plot of the 95% intervals of the bridge sampling estimators of the log ratio of normalizing constant against sample size, N, in the case of using serially correlated samples.

Figure 4.4 shows the log RMSE of the bridge sampling estimator under the four scenarios. They are all decreasing functions of N because the larger the N, the smaller the bias and standard error. Again, it is rather interesting that higher dimensionality for univariately updated MCMC samples produces bridge sampling estimates with lower RMSE (the red line outperforms the black line).

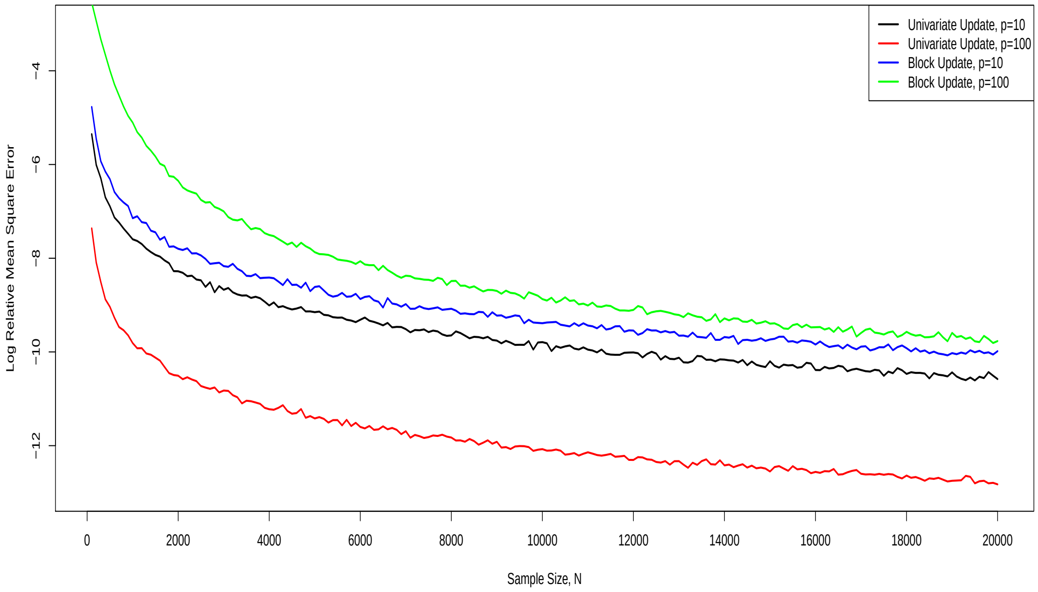


Figure 4.4: Plot of log relative mean square error against sample size, N, in the case of using serially correlated samples.

The results in this part also indicate that the misestimation due to sample dependence does not scale considerably with the dimension of parameter. For instance, setting $p = 4300$ and $N = 100000$ in the above example yield a bridge sampling estimate of approximately 0.8245 (without replications), which is not terrible considering the high-dimensionality involved. Therefore, we can ascertain that using serially correlated samples is not the primary source of the major misestimation of bridge sampling estimators.

4.4 Simulation Study 2: The Use of Mode-Curvature Matched Normal Approximation in Bridge Sampling Estimation

As mentioned before, though arbitrary, the choice of $g()$ is critical to ensure efficiency of the bridge sampling estimator. In the context of marginal likelihood estimation, this occurs when $g()$ is a good approximation of the joint posterior distribution. In various Bayesian applications, approximating the posterior distribution using a normal distribution by matching mode and curvature (negative inverse of the Hessian matrix) has been found to be rather effective, especially when the data sample size is large (so that the likelihood is peaked). It is also easily implemented since mode and curvature are relatively convenient to derive as their computation mostly involve derivatives. However, for the computation of bridge sampling estimate, mode and curvature are inadequate summary statistics in certain scenarios, particularly when the posterior distribution is heavy-tailed. To illustrate this point, the following case study is conducted.

Suppose p_1 is the density of a 100-dimensional Student's t-distribution with degrees of freedom 3, location parameter $\boldsymbol{\mu} = \mathbf{0}$, and the scale matrix $\Sigma = \boldsymbol{I}_{100}$, so that

$$p_1(\boldsymbol{\theta}) = \frac{\Gamma(103/2)}{(3\pi)^{50}\Gamma(3/2)}\left[1 + \frac{1}{3}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right]^{-103/2}.$$

It can be shown that the mode and curvature of this distribution is given by $\mathbf{0}$ and $\frac{3}{103}\boldsymbol{I}_{100}$. Hence, we set p_2 and q_2 be the densities of a $N(\mathbf{0}, \frac{3}{103}\boldsymbol{I}_{100})$. The ratio of normalising constants is again known to be one. A set of independent samples of size 100000 is then generated from each of the densities p_1 and p_2. The bridge sampling estimate of the ratio of the normalising constants came out to be about 30.49, which is substantially larger than the true value. This occurs even when the sample sizes are very large, and hence has nothing to do with having insufficient samples to learn about the distributions. The main reason is because the bridge sampling relies on the area of "overlap" to work efficiently (the larger the area of "overlap" between the densities, the more efficient the bridge sampler is). Matching mode and curvature implies that the "overlapping" between p_1 and p_2 will be small when one of them has a relatively heavier tail. Furthermore, the bias will be amplified as the dimension of the parameter increases. In particular, for a 250-dimensional Student's t-distribution, similar procedures as above yield a bridge sampling estimate of around exp(749)!

4.5 Simulation Study 3: The Use of Sample Mean-Variance Matched Normal Approximation in Bridge Sampling Estimation

It was demonstrated in the previous simulation study that the use of mode-curvature matched normal distribution can be suboptimal in situation where it involves heavy-tailed distributions. An immediate alternative is to use a normal distribution that matches the mean and variance matrix instead. The theoretical mean and variance matrix are difficult to derive since they involve integration. Therefore, sample moments are the obvious candidates. Unfortunately, as we shall show in the following simulation study, this leads to a systematic underestimation of the ratio of normalising constants.

Suppose p_1 and q_1 are the densities of a univariate standard normal distribution: $N(0, 1)$. A sample of size N is generated from this distribution to form $\{\theta_1^1, \ldots, \theta_1^N\}$. Then let p_2 and q_2 be the densities of $N(\mu_1, \sigma_1^2)$, where μ_1 and σ_1^2 are sample moments derived from $\{\theta_1^1, \ldots, \theta_1^N\}$. We assess two approaches of computing the sample moments:

1. Approach 1 (Naive): μ_1 and σ_1^2 are the sample mean and variance computed from the entire sample generated from p_1. A sample of size $N_2 = N$, $\{\theta_2^1, \ldots, \theta_2^{N_2}\}$, is generated from $N(\mu_1, \sigma_1^2)$. Hence, the bridge sampler is evaluated at the entire samples, $\{\theta_1^1, \ldots, \theta_1^{N_1}\}$ and $\{\theta_2^1, \ldots, \theta_2^{N_2}\}$, with $N_1 = N_2 = N$.

2. Approach 2 (Splitting): μ_1 and σ_1^2 are the sample mean and variance computed from a proportion, k, of the sample generated from p_1, $\{\theta_1^1, \ldots, \theta_1^{kN}\}$. A sample of size $N_2 = N$ is generated from $N(\mu_1, \sigma_1^2)$. The bridge sampler is then evaluated at the remaining samples from p_1, $\{\theta_1^{kN+1}, \ldots, \theta_1^{N}\}$, and the entire sample from p_2, $\{\theta_2^1, \ldots, \theta_2^{N}\}$. Whence, $N_1 = (1-k)N$ and $N_2 = N$.

Both approaches require approximately the same amount of computational effort (with approach 2 being slightly faster since the bridge sampler is only evaluated at a proportion of the posterior samples). We consider the sample size, N, from the set $\{100, 200, \ldots, 10000\}$, with each computation replicated $R = 10000$ times. We also examine five different splitting proportions, $k = 0.10, 0.25, 0.50, 0.75, 0.90$. Note that a similar simulation study has been performed by Overstall and Forster (2010), but only $k = 0.50$ was considered and they focused mainly on the bias correction. Their splitting approach is also slightly different by only simulating a shorter sample size from p_2 ($N_2 = 0.5N$), which can be improved with no substantial additional computational cost (relative to approach 1) in our opinion.

As depicted in Figure 4.5, approach 1 produces bridge sampling estimates with a systematic underestimation, but with comparatively narrower (and asymmetric) intervals. The bias also appears to be a decreasing function of sample size, N, which slowly converges to zero as N becomes large. By contrast, estimates from approach 2 are generally unbiased, but have wider (and symmetric) intervals. In other words, there is a trade-off between bias and variance for the two approaches. Approach 2 manages to alleviate the bias by removing the correlation between p_1 and p_2, but at the same time introduces more variations in the estimates (by having a smaller sample size, N_1, to work with). Therefore, a much more useful summary statistics for comparison in this scenario is the RMSE.

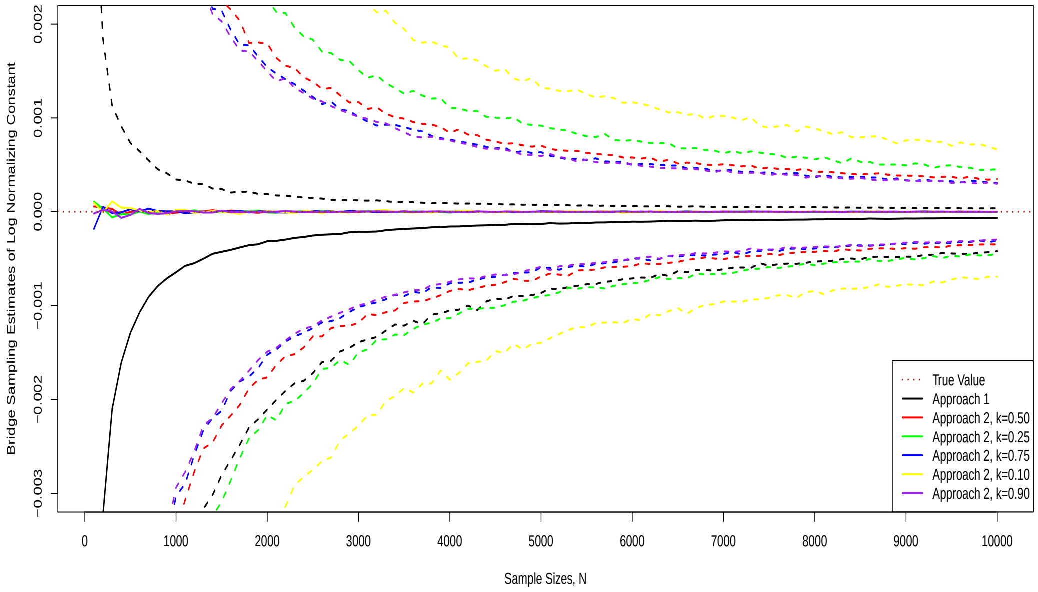


Figure 4.5: Plot of the median bridge sampling estimates (solid lines) of the log ratio of normalizing constant against sample size, N, accompanied by the associated 95% intervals (dotted lines), for various approaches in Simulation Study 3.

According to Figure 4.6, approach 2 with $k = 0.75$ and $k = 0.90$ possess the lowest RMSE, followed by approach 1 and approach 2 with $k = 0.50$. Approach 2 with $k = 0.25$ and $k = 0.10$ still have higher RMSE than approach 1. We can then deduce that for approach 2, the larger the allocated proportion, k, for moments estimation, the smaller the RMSE and hence the better the resulting estimator. For this particular toy example, approach 2 with $k = 0.75$ and $k = 0.90$ is clearly favoured, since it corrects for the bias due to correlation with relatively little loss in consistency. Approach 2 with $k = 0.50$ is arguably better than approach 1 since it alleviates the bias, despite having slightly larger RMSE.

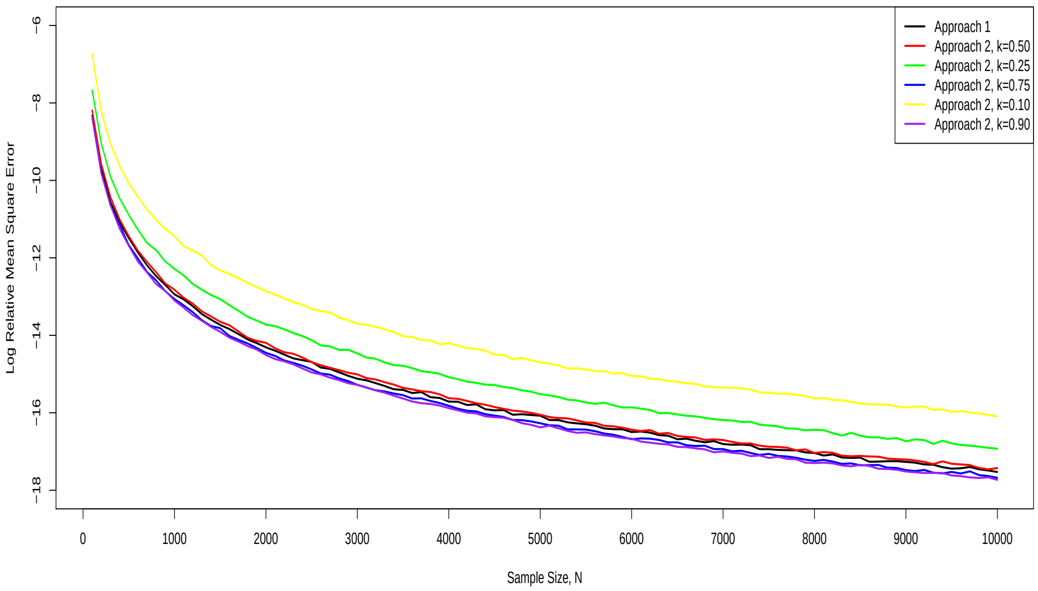


Figure 4.6: Plot of log relative mean square error against sample size, N, for various approaches in Simulation Study 3.

Additionally, the negative bias of approach 1 worsens as dimensionality increases. As an illustration, performing bridge sampling on a 4200-dimensional standard normal distribution with a sample size of 100000 yields a bridge sampling estimate of around $\exp(-44)$. To further explore the behaviour of the bridge sampling estimator with respect to dimensionality in the above set up, we consider the case when p_1 and q_1 are densities of a 10-dimensional standard normal distribution: $N_{10}(\mathbf{0}, \boldsymbol{I}_{10})$. A sample of size N is generated from this distribution to form $\{\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^N\}$. Then let p_2 and q_2 be the densities of $N_{10}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are sample mean and covariance matrix estimated from $\{\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^N\}$. We consider a similar set up as described above.

From Figure 4.7, similar patterns are realised: approach 1 systematically underestimates the ratio of normalizing constants, while approach 2 yields unbiased estimates at the expense of having larger variances. Upon closer inspection, the underestimation of approach 1 is much more apparent as compared to the unidimensional case, suggesting that the bias of approach 1 is amplified by the dimensionality of the problem. Among the different k of approach 2, the ranking is preserved, in that the larger the proportion, k, used for moments estimation, the better the resulting bridge sampling estimator. Notice now that approach 2 outperforms approach 1 for all the values of k considered in terms of RMSE, as illustrated in Figure 4.8. This is primarily because approach 2 still manages to alleviate the bias despite the increase in dimensionality, whereas the bias produced by approach 1 increases with increasing dimensionality. In brief, the problem of underestimation of approach 1 scales up as dimensionality increases, making the idea of performing splitting more valuable, since it corrects for the correlation-induced bias irrespective of the dimensionality involved.

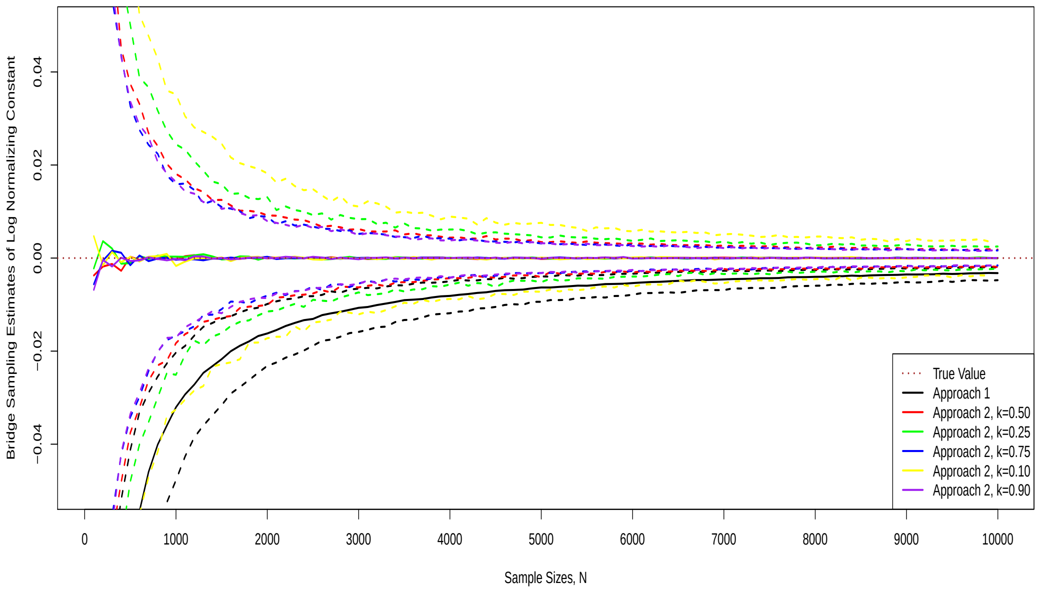


Figure 4.7: Plot of the median bridge sampling estimates (solid lines) of the log ratio of normalizing constant against sample size, N, for the 10-dimensional case, accompanied by the associated 95% intervals (dotted lines) in Simulation Study 3.

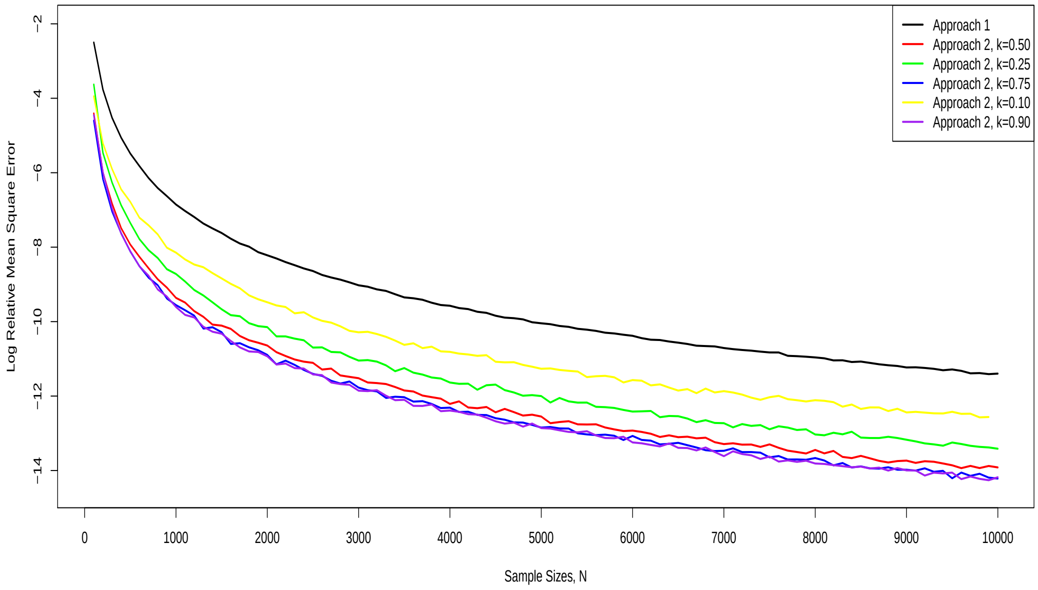


Figure 4.8: Plot of log relative mean square error against sample size, N, for the 10-dimensional case in Simulation Study 3.

4.6 Simulation Study 4: The Effect of Sample Size Allocation on The Efficiency of Bridge Sampling Estimator

The sample sizes N_1 and N_2 appear in the optimal bridge function, $\omega^*()$, as the mixture proportions of the two densities p_1 and p_2 respectively (see Equation (4.5)). Given that the bridge function plays the role to provide a linkage between the two densities, it is logical that the allocation of sample sizes directly influences the efficiency of the resulting bridge sampler. In the extreme case, when $N_1 \to \infty$ relative to N_2, we have

$$\omega^* \propto \frac{1}{q_1},$$

as the optimal bridge function, which then implies that

$$\frac{c_1}{c_2} = \frac{1}{\mathbb{E}_1[q_2/q_1]}.$$

In other words, the bridge sampler simplifies to a special case of the generalised harmonic rule when $N_1 \to \infty$. Specifically, choosing $q_1 = f_M(D|\boldsymbol{\theta}_M)f_M(\boldsymbol{\theta}_M)$ and $q_2 = f_M(\boldsymbol{\theta}_M)$ yields the harmonic mean estimator for marginal likelihood approximation, as proposed by Newton and Raftery (1994). On the other hand, when $N_2 \to \infty$ relative to N_1 (essentially all the samples are from p_2), the bridge sampler reduces to the usual non-iterative importance sampling estimator. Although these would not happen in practice as infinity is merely conceptual, but they do indicate that the behaviour of the bridge sampling estimator is affected by the relative sample sizes.

To the best of our knowledge, the effect of relative sample sizes on the efficiency of bridge sampling estimates has yet to be investigated. Here, we focus on a specific scenario that it is computationally expensive to simulate from p_1 and to evaluate the corresponding density, while it is relatively cheaper to simulate from p_2. This occurs frequently in the computation of marginal likelihood, where samples from the posterior distribution, generated from the MCMC algorithm, are generally much more time consuming to obtain, while samples from the moment-matched normal distribution are more straightforward to obtain and to evaluate their densities. Thus, in this case study, we investigate the possibility of using $N_2 > N_1$ to improve the efficiency of the bridge sampler with little increase in the computational effort.

Suppose that p_1 and q_1 are the densities of a unidimensional normal distribution with mean 0 and variance matrix 1: $N(0, 1)$. Denote μ_1 and σ_1^2 as the sample mean and variance matrix of the samples generated from p_1, $\{\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^N\}$. Then let p_2 and q_2 be the densities of the $N(\mu_1, \sigma_1^2)$ distribution. We consider the sample size, N, from the set $\{100, 200, \ldots, 10000\}$. Three different sample size allocations are examined:

i. Naive approach: $N_2 = N$.

ii. A constant multiple of N: $N_2 = 10N$.

iii. Some relatively large number: $N_2 = 50000$.

The bridge sampler is then evaluated at the entire samples from both p_1 and p_2, so that $N_1 = N$ and N_2 is as above.

As shown in Figure 4.9, a larger N_2 generally leads to better estimates by reducing both the bias and standard error of the bridge sampling estimator. For $N_2 = 50000$ (blue line), the bias shrinks to almost zero, with a minute underestimation of magnitude 0.00001 after closer inspection. For $N_2 = 10 \times N_1$ (red line), the improvement of the bridge sampling estimator with respect to sample size is consistently better relative to using $N_2 = N_1$, overtaking that of using $N_2 = 50000$ at $N = 5000$ (when the blue and red lines cross each other), where the red line outperforms the blue by having a larger N_2. Moreover, the RMSE of the red line also appears to decrease indefinitely as sample size, N, increases, while the improvement of blue line slowly decelerates with increasing N (mainly because its standard error does not reduce considerably towards the end). This suggests that this method can be further improved by using an even larger N_2. Overall, larger N_2 leads to better bridge sampling estimates (with lower bias and standard error, hence lower RMSE) in this particular case.

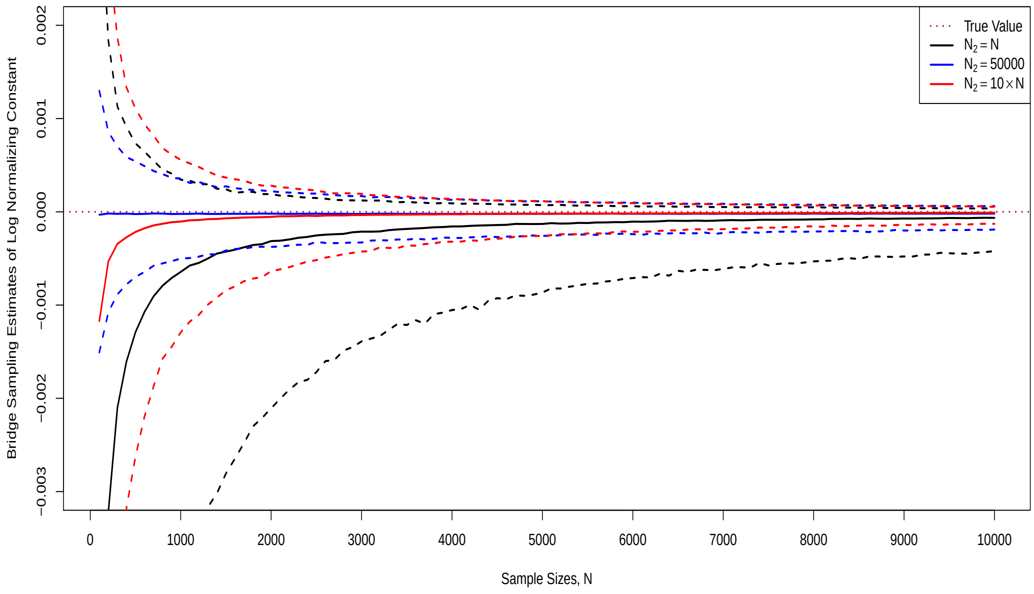


Figure 4.9: Plot of the median bridge sampling estimates (solid lines) of the log ratio of normalizing constant against sample size, N, accompanied by the associated 95% intervals (dotted lines), for various relative sample sizes in Simulation Study 4.

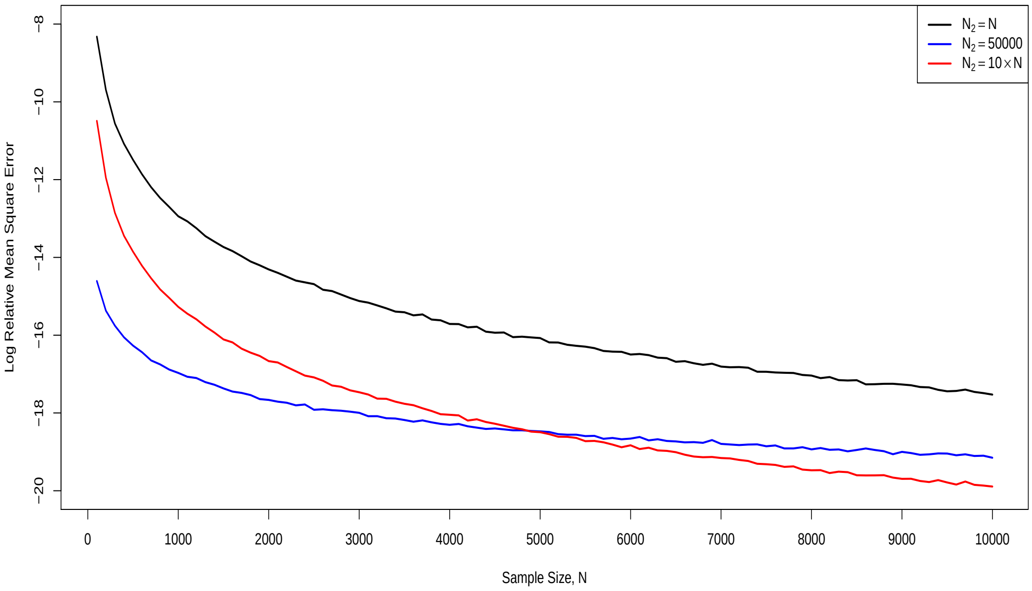


Figure 4.10: Plot of log relative mean square error against sample size, N, for various allocation of sample sizes in Simulation Study 4.

The previous result is to be expected since using a relatively large N_2 in some sense resembles performing importance sampling using a normal distribution. Also, using normal distribution as an importance sampling distribution to compute the normalizing constant of another normal distribution is certainly going to behave well as they possess similar tail heaviness. It is perhaps more interesting to consider a distribution with heavier tail, where importance sampling is known to be less efficient. Suppose now that p_1 and q_1 are the densities of a Student's t-distribution with degrees of freedom three (t_3). Using a similar set up as above, the behaviour of the bridge sampler in response to various sample sizes is examined, as displayed in Figure 4.11 and 4.12.

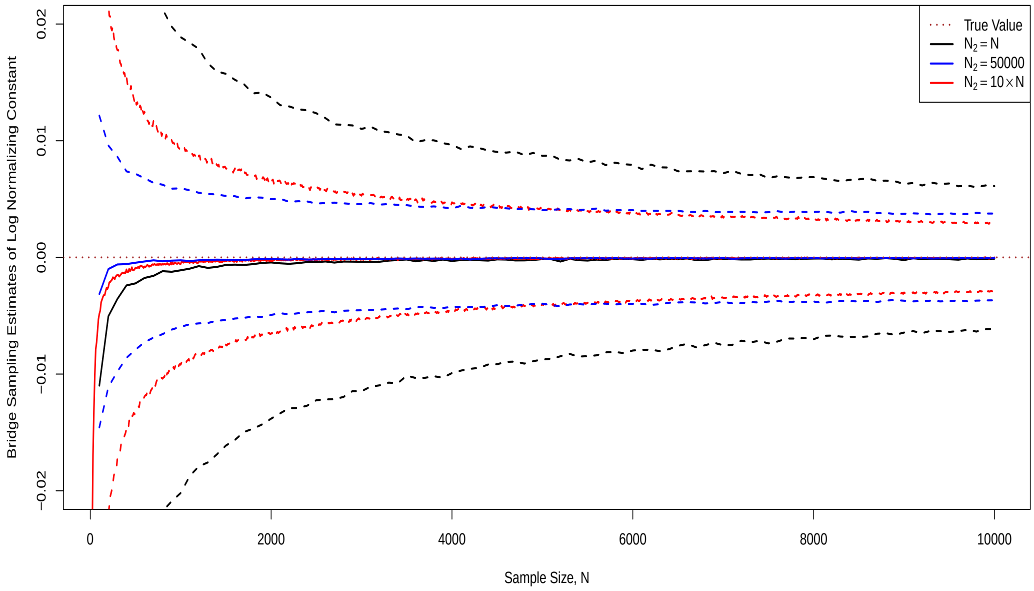


Figure 4.11: Plot of the median bridge sampling estimates (solid lines) of the log ratio of normalizing constant against sample size, N, accompanied by the associated 95% intervals (dotted lines), for Simulation Study 4 in the case where p_1 is the density of a t_3 distribution.

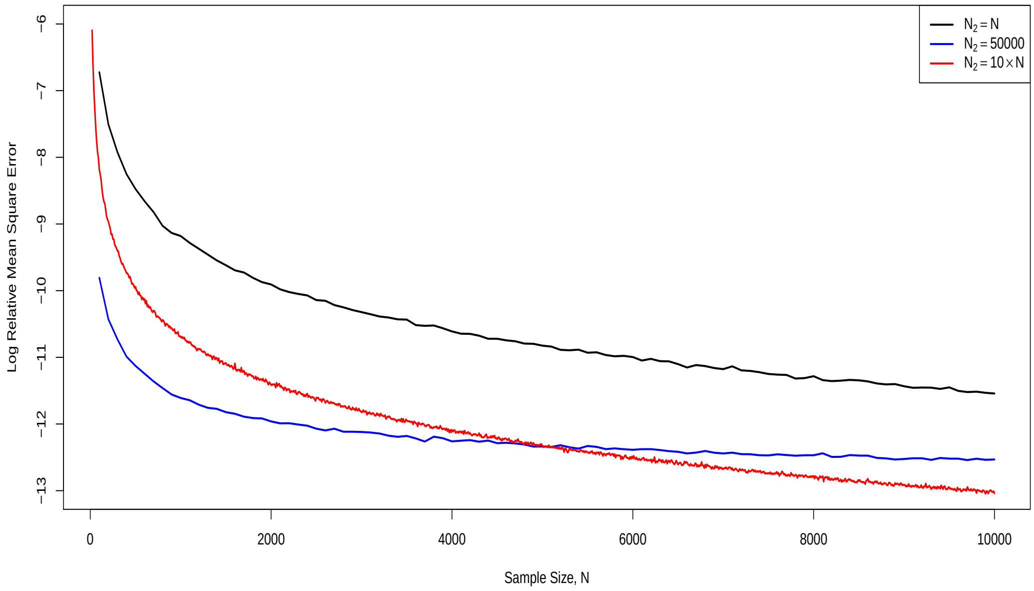


Figure 4.12: Plot of log relative mean square error against sample size, N, for Simulation Study 4 in the case where p_1 is the density of a t_3 distribution.

Remarkably, even with the use of a heavier-tailed p_1, similar pattern of behaviours are observed. One exception is that the bias and width of the uncertainty bands are now larger due to the tail heaviness, resulting in larger overall RMSE than before. To reiterate the main points again, the blue and red lines cross at $N = 5000$ as expected. The red line appears to improve indefinitely in terms of the RMSE as sample size increases and is consistently better than the black line, while the blue line has decelerating improvement with almost no reduction on the width of uncertainty bands eventually. In conclusion, it can be deduced from this simulation study that it is favourable to always use a larger sample from p_2 if $g()$ is chosen to be the (sample) moment-matched normal distribution, since this reduces both the bias and standard error of the resulting bridge sampling estimate with little increase in computational effort.

4.7 Discussions and Model Comparison

These simulation studies can potentially be used to explain why the bridge sampling estimator faltered during the computation of the bridge sampling estimate of marginal likelihood for the PLNLC model. Basically, our previous observation that the bridge sampling estimate of the PLNLC model increases as sample size increases is the consequence of using approach 1 of Simulation Study 3 (where the entire posterior sample is used to derive the moments of $g_M()$ and evaluate the bridge sampler), which produces bias that is a decreasing function of sample size, amplified by the immense dimensionality of the problem. Our hypothesis on the failure of bridge sampling in accurately estimating the marginal likelihood in this case is a combination of the lack of sufficiently long samples to obtain a good approximation of the posterior moments especially the variance matrix (due to sample autocorrelations), and the correlation induced through the use of approach 1 of Simulation Study 3. Knowing that the bias of approach 1 tends to zero as sample size increases, one possible solution is to generate more posterior samples for this purpose. However, with the MCMC algorithm only generating dependent posterior samples, it implies that an enormously large samples is essential to yield an estimate with acceptable error margin. This is practically infeasible as far as the available computational resources are concerned. Therefore, a better alternative is to apply the idea of splitting, coupled with the use of a relatively large sample, N_2, from the moment-matched normal distribution. In particular, we generated a sample of size 25000000 from the posterior distribution of the PLNLC model using the MCMC algorithm described before, applied a thinning of 100 so that we have $N = 250000$ samples from the posterior, $\{\boldsymbol{\theta}^1_{\text{PLNLC}}, \ldots, \boldsymbol{\theta}^N_{\text{PLNLC}}\}$. Next, a sample of size $N_2 = 500000$, $\{\boldsymbol{\theta}^1_2, \ldots, \boldsymbol{\theta}^{N_2}_2\}$, is generated from a 4446-dimensional normal distribution, with moments set as the sample mean and covariance matrix estimated from the first half of the posterior samples (the splitting approach with $k = 0.50$). The bridge sampler in (4.6) is then evaluated at the remaining half of the posterior samples, $\{\boldsymbol{\theta}^{125001}_{\text{PLNLC}}, \ldots, \boldsymbol{\theta}^{250000}_{\text{PLNLC}}\}$, as well as the sample, $\{\boldsymbol{\theta}_2^1, \ldots, \boldsymbol{\theta}_2^{500000}\}$, so that $N_1 = 125000$ and $N_2 = 500000$. This results in an estimate of -23723.65 for the marginal likelihood of the PLNLC model.

For the NBLC model, we apply similar idea, where a sample of size $N_2 = 50000$ is simulated from a 248-dimensional normal distribution with moments equated to the sample mean and variance of the first half of the posterior samples, $\{\boldsymbol{\theta}_{\text{NBLC}}^1, \ldots, \boldsymbol{\theta}_{\text{NBLC}}^{25000}\}$. The bridge sampler is then evaluated using the second half of the posterior samples, $\{\boldsymbol{\theta}_{\text{NBLC}}^{25001}, \ldots, \boldsymbol{\theta}_{\text{NBLC}}^{50000}\}$ and the entire sample from the moment-matched normal distribution, so that $N_1 = 25000$ and $N_2 = 50000$. This yields a marginal likelihood estimate of -23727.01 for the NBLC model. The improvement is not so pronounced comparatively because the previous estimate was already considered good given the relatively easy problem.

The marginal likelihood of each model under consideration, approximated using bridge sampling are presented in Table 4.2. 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 section 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. Even though both the overdispersion models provide similar fit qualitatively (for this particular dataset), the NBLC model is to be recommended due to its computational advantage over its counterpart by having a lower dimensionality after integrating out the latent variables, μ_{xt}.

Bayesian Poisson LC	Poisson Log-normal LC	Negative Binomial LC
−26684.10	−23723.65	−23727.01

Table 4.2: The marginal likelihoods (on logarithmic scale) of each model approximated by bridge sampling.

4.8 Conclusion

In this chapter, we perform a formal Bayesian model comparison of the candidate models from the previous chapter using posterior model probabilities. The computation of posterior model probabilities relies on an accurate estimation of the marginal likelihoods, which can be achieved by using bridge sampling. However, this proved to be a rather challenging task in a high-dimensional problem, as in the PLNLC model. Hence, four simulation studies are conducted to study the behaviours of bridge sampling. Simulation Study 1 examines the effect of using correlated samples for the evaluation of the bridge sampler, which was shown to increase the standard errors of the resulting bridge sampling estimate but not the bias, provided the samples are sufficiently long. In Simulation Study 2, we demonstrated that the mode-curvature matched normal approximation does not work well for the estimation of marginal likelihood, particularly in the case of heavily-tailed posterior distribution. The sample moment-matched normal approximation appears to be more promising but induces correlation between the distributions, causing systematic underestimation of the marginal likelihoods. Therefore, we propose the use of the splitting approach in Simulation Study 3, which has the potential to alleviate the bias, at the expense of a slight increase in standard errors. Simulation Study 4 investigates the possibility of improving the bridge sampling estimates using different allocations of sample sizes, and we proceeded to illustrate that this can be accomplished by allocating larger samples to the sample moment-matched normal distribution. The conclusions deduced from these simulation studies allow us to improve the bridge sampling estimates of the marginal likelihoods. With that, the posterior model probabilities computed indicate that the overdispersion models outperform the Bayesian PLC model considerably. The PLNLC and NBLC models also possess exceptionally close marginal likelihoods (verifying their similarities), thus we recommend the NBLC model over its counterpart for computational reasons.

Chapter 5

Poisson-Gamma (Negative Binomial) Log-Linear Model

Thus far, we have been considering Lee-Carter (LC) based mortality models. One undesirable feature of LC based models is the presence of the multiplicative bilinear term, $\beta_x \kappa_t$, which can prove difficult to handle. For example, the search for the joint posterior mode or MLE typically necessitates iterative conditioning in a log-bilinear set up (Section 3.6.3). On the other hand, the likelihood function of a log-linear model is log-concave, simplifying parameter estimation (using optimization) and various related computations. Hence, a simpler model is proposed here, where the bilinear term is replaced by a linear component:

$$\begin{aligned} D_{xt}|\mu_{xt} &\sim \text{Poisson}(e_{xt}\mu_{xt}), \\ \log \mu_{xt} &= \alpha_x + \beta_x t + \kappa_t + \log \nu_{xt}, \\ \nu_{xt}|\phi &\sim \text{Gamma}(\phi, \phi), \end{aligned} \tag{5.1}$$

where $t = -\frac{T-1}{2}, \ldots, \frac{T-1}{2}$ is centered at zero for computational stability. We refer to this model as the Poisson-Gamma Log-Linear model. Note that similar rate model has been considered by Renshaw and Haberman (2003) within a classical framework, who explicitly included the calendar time, t, as a known covariate in their generalized linear regression approach for mortality projections (where overdispersion was dealt with using the quasi-likelihood approach).

As indicated by our findings earlier, it is more computationally efficient to marginalise the $\log \mu_{xt}$, forming the Negative Binomial Log-Linear (NBLL) model. Note that this model is not identifiable by being invariant to the following transformations:

$$\begin{aligned}\alpha_x &\mapsto \alpha_x + a,\\ \beta_x &\mapsto \beta_x + b,\\ \kappa_t &\mapsto \kappa_t - bt - a,\end{aligned}$$

for any $a \in \mathbb{R}$ and $b \in \mathbb{R}$. Therefore, we impose the constraints

$$\sum_t \kappa_t = \sum_t t\kappa_t = 0,$$

which have the effect of centering the κ_t at zero and restricting its linear growth.

Throughout this chapter, we use superscripts $^{\mathrm{LC}}$ and $^{\mathrm{LL}}$ to indicate LC type models and the log-linear model respectively. Technically, t plays the role of extracting the linear drift of κ_t^{LC}, making the κ_t^{LL} appearing as random noise (driftless). In other words, $\beta_x^{\mathrm{LL}} t + \kappa_t^{\mathrm{LL}}$ collectively behave rather similarly to the $\beta_x^{\mathrm{LC}}\kappa_t^{\mathrm{LC}}$ (see Section 5.3 for details). This phenomenon is also pointed out by Wong-Fupuy and Haberman (2004), who stated that the use of the LC type models with κ_t modelled by a random walk is essentially assuming a log-linear relation between mortality rates and time.

Despite being very similar, the NBLC and NBLL models have slightly different parameter interpretation. To see this, multiply both sides of Equation (5.1) by t and summing across time (ignoring the residuals, $\log \nu_{xt}$), we see that

$$\beta_x = \frac{\sum_t t \log \mu_{xt}}{\sum_t t^2}, \tag{5.2}$$

meaning that β_x is the regression coefficient if we regress $\log \mu_{xt}$ against time, t. Moreover, κ_t no longer represents the overall change in mortality. To be more specific, by differentiating Equation (5.1) with respect to t, we have

$$\frac{\mathrm{d}\log \mu_{xt}}{\mathrm{d}t} = \beta_x + \frac{\mathrm{d}\kappa_t}{\mathrm{d}t}.$$

Hence, β_x and $\frac{\mathrm{d}\kappa_t}{\mathrm{d}t}$ still govern the rate of change of mortality rate, but do so in a different functional form. Meanwhile, α_x still represents the average log mortality rates at age x (proved easily by summing Equation (5.1) across time).

5.1 An Overview of the Chapter

An outline of this chapter is provided here. In Section 5.2, the time series prior for the time-variant parameter, κ_t, is provided. Parameter correspondence relationships between the NBLL and NBLC models are presented in Section 5.3. In Section 5.5.1, we illustrate how a naive prior specification (Section 5.4) prevents us from achieving our main objective of comparing the NBLL and NBLC models (using marginal likelihoods) by being inconsistent in terms of the prior information specified for the models and being overly heavy-tailed. We remedy this issue in several stages throughout Section 5.5 by investigating in detail the prior specification. First, we aim to tune the constants of the prior distributions such that the implied prior distributions of the log mortality rates are sensible under the NBLC model (Section 5.5.2). Secondly, moment-based approach is used to specify prior distributions with matching information for the NBLL model, retaining the family of distributions (Section 5.5.3). Next, we demonstrate that retaining the family of the prior distributions for the NBLL model (particularly those concerning normal distributions) causes the inconsistency due to a mismatch of family of distributions (Section 5.5.4). Then, we propose the use of Laplace prior distributions for the NBLL model as a solution to the inconsistent prior specification in Section 5.5.5. After ensuring consistency in the prior specification, we present MCMC schemes for both the NBLL and NBLC models in Section 5.6. In Section 5.7, the method to project mortality rates under the NBLL model is described. Finally, numerical results and conclusion are given in Section 5.8 and Section 5.9 respectively.

5.2 Time Series Modelling of κ_t

As mentioned previously, κ_t appears as random noise, having its linear drift extracted. Therefore, an appropriate projection model for κ_t is an AR(1) model without drift:

$$\begin{cases} \kappa_t = \rho\kappa_{t-1} + \epsilon_t & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \epsilon_1 \end{cases},$$

where $\epsilon_t \sim N(0, \sigma_\kappa^2)$ are independent Gaussian errors. Suppose, for the moment, we do not impose strict stationarity by using a non-truncated prior on the regression coefficient, ρ. Suppose also $\boldsymbol{Q} = (\boldsymbol{I}_T - \boldsymbol{P})^\top(\boldsymbol{I}_T - \boldsymbol{P})$, where

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \rho & 0 & & & \vdots \\ 0 & \rho & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho & 0 \end{pmatrix}_{T\times T},$$

and $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^\top$, where

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & 3 & \cdots & T-1 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{T\times T}.$$

Applying the constraints $\sum_t \kappa_t = \sum_t t\kappa_t = 0$ on the AR(1) prior (see Appendix K for derivation) and using the conditional property of a multivariate normal distribution, we obtain the following prior for $\boldsymbol{\kappa}_{-1,2} = (\kappa_3, \ldots, \kappa_T)^\top$:

$$\boldsymbol{\kappa}_{-1,2} \sim N_{T-2}(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{D}),$$

where $\boldsymbol{D} = [\boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}]$ and $\boldsymbol{B}$ is partitioned such that

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11_{2\times 2}} & \boldsymbol{B}_{12_{2\times(T-2)}} \\ \boldsymbol{B}_{21_{(T-2)\times 2}} & \boldsymbol{B}_{22_{(T-2)\times(T-2)}} \end{pmatrix}.$$

The remaining κ_1 and κ_2 are then deterministically computed from

$$\begin{cases} \kappa_1 = \sum_{i=3}^T (i-2)\kappa_i = \kappa_3 + 2\kappa_4 + 3\kappa_5 + \cdots + (T-2)\kappa_T \\ \kappa_2 = -\sum_{i=3}^T (i-1)\kappa_i = -2\kappa_3 - 3\kappa_4 - 4\kappa_5 - \cdots - (T-1)\kappa_T \end{cases}. \tag{5.3}$$

5.3 Parameter Correspondence Relationships between the NBLL and NBLC Models

Here, we establish parameter correspondence relationships between the NBLL and NBLC models by performing a term-by-term comparison. For simplicity, we ignore the constraints of the ARIMA models for κ_t^{LC} and κ_t^{LL} momentarily, and consider their marginal distributions, given respectively by

$$\begin{cases} \kappa_t^{\text{LC}} = \psi_1^{\text{LC}} + \psi_2^{\text{LC}} t + \epsilon_t'^{\text{LC}} \\ \kappa_t^{\text{LL}} = \epsilon_t'^{\text{LL}} \end{cases}, \tag{5.4}$$

where

$$\epsilon_t'^{\text{LC}} \sim N\left(0, \frac{(\sigma_\kappa^{\text{LC}})^2}{1-(\rho^{\text{LC}})^2}\right) \text{ and } \epsilon_t'^{\text{LL}} \sim N\left(0, \frac{(\sigma_\kappa^{\text{LL}})^2}{1-(\rho^{\text{LL}})^2}\right)$$

are independent Gaussian errors. Hence, the log mortality rates under the NBLC and NBLL models can be expressed as follows:

$$\begin{cases} \log \mu_{xt}^{\mathrm{LC}} = \alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}}t + \beta_x^{\mathrm{LC}}\epsilon_t'^{\mathrm{LC}} + \nu_{xt}^{\mathrm{LC}} \\ \\ \log \mu_{xt}^{\mathrm{LL}} = \qquad \alpha_x^{\mathrm{LL}} \qquad + \quad \beta_x^{\mathrm{LL}}t \quad + \quad \epsilon_t'^{\mathrm{LL}} \quad + \nu_{xt}^{\mathrm{LL}} \end{cases} .$$

Clearly, the following parameter correspondence can be established:

$$\alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}} \longleftrightarrow \alpha_x^{\mathrm{LL}}, \quad \beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}} \longleftrightarrow \beta_x^{\mathrm{LL}}, \quad \beta_x^{\mathrm{LC}}\epsilon_t'^{\mathrm{LC}} \longleftrightarrow \epsilon_t'^{\mathrm{LL}}, \quad \nu_{xt}^{\mathrm{LC}} \longleftrightarrow \nu_{xt}^{\mathrm{LL}}.$$

5.4 Naive Prior Specification

First, we consider a naive specification of prior distributions for the parameters under the NBLL model. In particular, the prior distributions described for the PLNLC and NBLC models in Section 3.4.1.1 are re-used here for the relevant parameters (except κ_t). As we shall demonstrate later, this leads to an inconsistent model comparison procedure by inherently favouring one of the models through the prior specification.

Without going much into the details, MCMC methods are used to generate a set of posterior sample from the NBLL model using the naive prior specification above. We applied a burn-in phase of 10000 iterations and the resulting chain was thinned by a 100, yielding a sample of size 10000. Bridge sampling is then applied to this sample to obtain a marginal likelihood estimate of about $\exp(-23769.50)$, which is considerably smaller than that of the NBLC model, $\exp(-23727.01)$. This result is in contrast to that suggested by BIC, where the NBLL model (BIC: 47169.46) is indicated to be more superior than the NBLC model (BIC: 47217.47). This is rather counter-intuitive considering that the BIC is effectively an approximate Bayes factor in an asymptotic sense when unit information prior is used (Raftery, 1999). Fundamentally, this is because the Bayes factor is known to be very sensitive to the prior distributions used, as pointed out by Weakliem (1999). While it is acceptable for the conclusion from Bayes factor to disagree with that of BIC, we believe that consistency in terms of the prior information provided for the competing models should be ensured to some extent if data dominated inference is required, as in our case here. Hence, in the next few sections, we shall investigate the prior specification to illustrate that the inconsistency of prior information imposed tilted the marginal likelihood into favouring the NBLC model.

5.5 Investigating The Exact Impact of Our Prior Specification

Previously vague priors were chosen for illustrative purposes. Here, we examine in detail the implication of these priors in terms of the information indicated for the mortality rates, as they are generally the determining factor for the subsecquent mortality projection (also because they are present in most mortality models and typically have the same interpretation across different mortality models). This can be undertaken by monitoring the implied prior distribution on $\log \mu_{xt}$. Since the prior distributions are specified for the rest of the (hyper)parameters and not on $\log \mu_{xt}$ itself, it is rather challenging to derive its implied prior distribution theoretically. However, we can generate a sample of $\log \mu_{xt}$ using the prior distributions, from which the implied prior distribution can be estimated through kernel density estimation. For example, under the NBLL model, the generation of prior $\log \mu_{xt}$ proceeds in two steps:

1. Generate κ_t from the double-constrained AR(1) model,

$$\begin{cases} \boldsymbol{\kappa}_{-1,2} \sim N_{T-2}(\mathbf{0}, \sigma_\kappa^2 \boldsymbol{D}) \\ \kappa_1 = \sum_{i=3}^{T} (i-2)\kappa_i = \kappa_3 + 2\kappa_4 + 3\kappa_5 + \cdots + (T-2)\kappa_T \\ \kappa_2 = -\sum_{i=3}^{T} (i-1)\kappa_i = -2\kappa_3 - 3\kappa_4 - 4\kappa_5 - \cdots - (T-1)\kappa_T \end{cases},$$

   where samples from the prior distributions of ρ and σ_κ^2 are substituted correspondingly.

2. Generate μ_{xt} from

$$\mu_{xt} \sim \text{Gamma}\left(\phi, \frac{\phi}{\exp(\alpha_x + \beta_x t + \kappa_t)}\right),$$

   where κ_t is from step 1 and (α_x, β_x, and ϕ) are samples from the corresponding prior distributions.

5.5.1 Implied Priors on $\log \mu_{xt}$ for the Naive Prior Specification

Figure 5.1 illustrates several chosen kernel estimates of the prior distribution of $\log \mu_{xt}$ under the naive prior specification. This figure is not particularly informative because we can hardly visualise the resulting distributions. Fundamentally, this is because the implied prior distributions of $\log \mu_{xt}$ are so diffused that the resulting kernel density estimates appear as straight lines when a particular region is focused (also because it is practically infeasible to generate a sample of sufficient length to obtain good kernel density estimates of these extremely heavy-tailed distributions). Therefore, we first aim to tune the constants for the prior distributions for both models such that the implied prior distributions on $\log \mu_{xt}$ have most of their density within a reasonable range in the sense of realistic mortality rate, typically around $[-10, 0)$ on the log scale. This is critical because using overly heavy-tailed prior distributions implies non-sensible prior information and has a higher tendency to induce Bartlett's paradox (see Section 4).

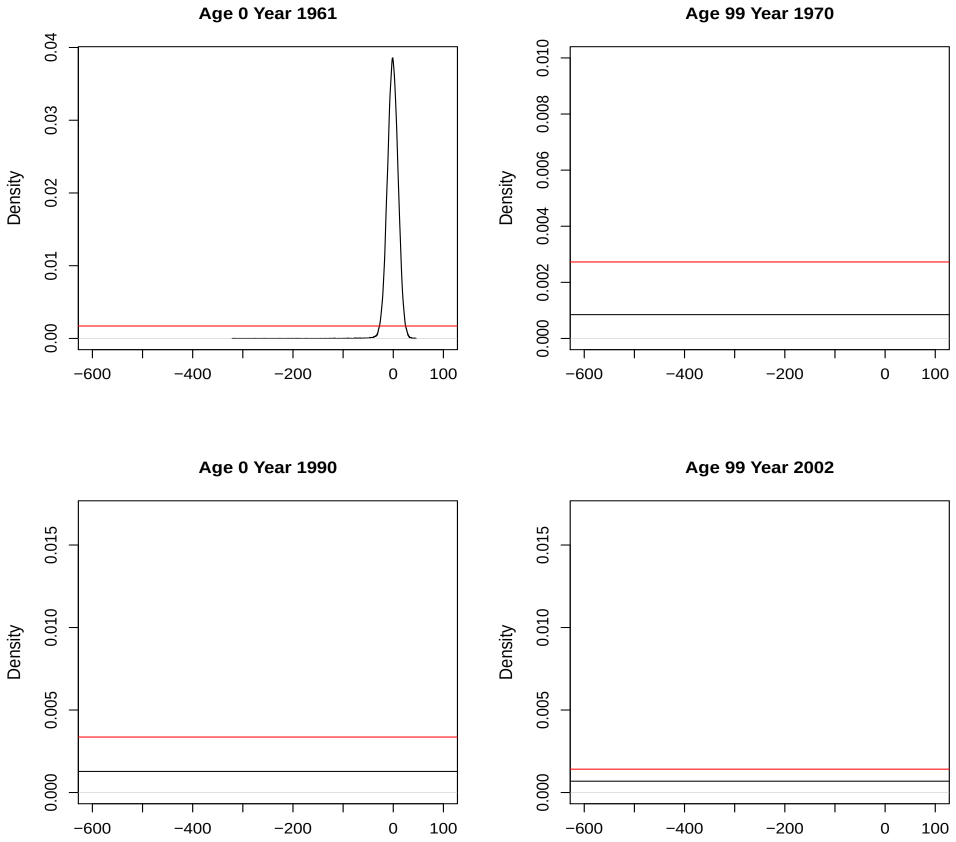


Figure 5.1: Kernel density plots of several chosen prior distributions of $\log \mu_{xt}$ as labelled, under the NBLC (black) and NBLL (red) models.

5.5.2 Prior Specification with Sensible Implied Priors on $\log \mu_{xt}$ under the NBLC model

After some preliminary investigations, the prior distributions chosen for the NBLC model are as follows:

$$\begin{aligned}
\alpha_x^{\mathrm{LC}} &\overset{\mathrm{ind}}{\sim} N(-5, 4),\\
\boldsymbol{\beta}_{-1}^{\mathrm{LC}} &\sim \left(\frac{1}{A}\mathbf{1}_{A-1}, (\sigma_\beta^{\mathrm{LC}})^2\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)\right),\\
(\sigma_\beta^{\mathrm{LC}})^2 &= 0.005,\\
\frac{\rho^{\mathrm{LC}}+1}{2} &\sim \mathrm{Beta}(3,2),\ \text{ where } \rho \in (-1,1),\\
(\sigma_\kappa^{\mathrm{LC}})^2 &\sim \text{Inverse Gamma}(0.1, 0.001) \text{ truncated to } (0, 10],\\
\boldsymbol{\psi}^{\mathrm{LC}} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2000 & 0 \\ 0 & 2 \end{pmatrix}\right),\\
\phi^{\mathrm{LC}} &\sim \mathrm{Gamma}(0.0001, 0.0001) \text{ truncated to } [1, 50000].
\end{aligned}$$

There are a few remarks to be noted here. First, the $(\sigma_\beta^{\mathrm{LC}})^2$ is given a point mass value at 0.005 rather than a distribution a priori for simplicity. Recall also that the posterior distribution of ρ^{LC} is a mixture of a stationary AR(1) model and a random walk with drift model. Here, we explicitly separate the two cases through our prior specification. In particular, this can be accomplished by using a transformed beta prior (as shown above), with density function given as

$$f(\rho^{\mathrm{LC}}) = \frac{3}{4}(\rho^{\mathrm{LC}}+1)^2(1-\rho^{\mathrm{LC}})^1, \qquad \text{for } \rho^{\mathrm{LC}} \in (-1,1).$$

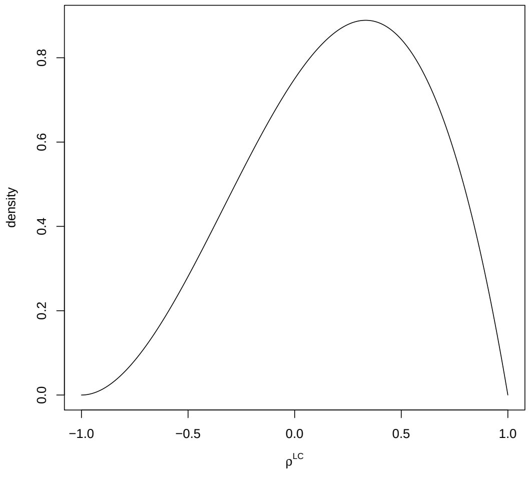


Figure 5.2: Plot of the density function of the transformed beta prior on ρ^{LC}.

This distribution has most of its density concentrated at the region $(-1, 1)$, slightly favouring positive values and declines rapidly as it approaches its end points (see Figure 5.2). Because this prior distribution assigns zero density at $\rho^{\mathrm{LC}} = 1$, this results in a stationary AR(1) model as the posterior distribution of ρ^{LC}. The random walk model can easily be obtained by simply setting $\rho^{\mathrm{LC}} = 1$ deterministically within the corresponding MCMC algorithm. Posterior model probabilities computed (Section 5.8) can subsequently be used to compare the different specification explicitly (hopefully they reflect the mixture proportion as displayed in Figure 3.20). For the purposes of the investigation here, we focus on using the transformed beta prior for comparison, since the case with random walk models then follows trivially. On the other hand, the rationale of applying truncation on some of the prior distributions is that the possibility of the parameter values lying within the corresponding truncated region are deemed implausible. For example, it is highly unlikely that ϕ^{LC} exceeds the value of 50000 (which corresponds to a dispersion measure of $= 0.00002$), implying a close to negligible level of overdispersion which would be contradictory to our previous study on the overall overdispersion present in our mortality data (the previous fitted value was around $\phi^{\mathrm{LC}} = 681$).

For consistency, we revert to the conventional constraint on κ_t^{LC}, $\sum_t \kappa_t^{\mathrm{LC}} = 0$ (so that α_x^{LC} represents the average log mortality rates). Suppose that $\boldsymbol{Q}^{\mathrm{LC}} = (\boldsymbol{I}_{T-1} - \boldsymbol{P}^{\mathrm{LC}})^\top (\boldsymbol{I}_{T-1} - \boldsymbol{P}^{\mathrm{LC}})$, where

$$\boldsymbol{P}^{\mathrm{LC}} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \rho & 0 & & & \vdots \\ 0 & \rho & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho & 0 \end{pmatrix}_{T\times T},$$

$\boldsymbol{B}^{\mathrm{LC}} = \boldsymbol{A}^{\mathrm{LC}} (\boldsymbol{Q}^{\mathrm{LC}})^{-1} (\boldsymbol{A}^{\mathrm{LC}})^\top$ with

$$\boldsymbol{A}^{\mathrm{LC}} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{T\times T},$$

and

$$\boldsymbol{Y}_{-1}^{\mathrm{LC}} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & T \end{pmatrix}_{(T-1)\times 2}.$$

The change in constraint leads to a change in the prior distribution of κ_t^{LC}, where now

$$\boldsymbol{\kappa}_{-1}^{\text{LC}} \sim N_{T-1}\left(\boldsymbol{\mu}_\kappa^{\text{LC}}, (\sigma_\kappa^{\text{LC}})^2 \boldsymbol{D}^{\text{LC}}\right),$$

with $\boldsymbol{\mu}_\kappa^{\text{LC}} = Y_{-1}^{\text{LC}}\boldsymbol{\psi} - \boldsymbol{B}_{21}^{\text{LC}}(\boldsymbol{B}_{11}^{\text{LC}})^{-1} \cdot \sum_{t=1}^{T}(\psi_1^{\text{LC}} + \psi_2^{\text{LC}} t)$, $\boldsymbol{D}^{\text{LC}} = [\boldsymbol{B}_{22}^{\text{LC}} - \boldsymbol{B}_{21}^{\text{LC}}(\boldsymbol{B}_{11}^{\text{LC}})^{-1}\boldsymbol{B}_{12}^{\text{LC}}]$, and the matrix $\boldsymbol{B}^{\text{LC}}$ is partitioned such that

$$\boldsymbol{B}^{\text{LC}} = \begin{pmatrix} \boldsymbol{B}_{11_{1\times 1}}^{\text{LC}} & \boldsymbol{B}_{12_{1\times(T-1)}}^{\text{LC}} \\ \boldsymbol{B}_{21_{(T-1)\times 1}}^{\text{LC}} & \boldsymbol{B}_{22_{(T-1)\times(T-1)}}^{\text{LC}} \end{pmatrix}.$$

Henceforth, these are the prior distributions used for the NBLC model and will not be changed. The aim from this point is to specify prior distributions for the NBLL model such that the prior information embedded within is consistent between the two models.

5.5.3 Moment-Based Approach for Matching the Prior Specification between the NBLL and NBLC Models

After setting the prior distributions for the NBLC model, we attempt to specify distributions that have similar prior information embedded within for the NBLL model using moment-matching. Suppose we retain the family of distributions for the corresponding parameters, but with their first few moments carefully matched between the models. To achieve that, a simple identity is required and can be derived as follows. Suppose that U and V are two independent random variables, then

$$\mathbb{E}(UV) = \mathbb{E}_U[\mathbb{E}_V(UV|U)] = \mathbb{E}_U[U\mathbb{E}_V(V|U)] \overset{\text{ind}}{=} \mathbb{E}_U[U\mathbb{E}_V(V)] = \mathbb{E}_V(V)\mathbb{E}_U(U)$$

and

$$\begin{aligned} \text{Var}(UV) &= \text{Var}_U[\mathbb{E}_V(UV|U)] + \mathbb{E}_U[\text{Var}_V(UV|U)] \\ &= \text{Var}[U\mathbb{E}_V(V|U)] + \mathbb{E}_U[U^2\text{Var}(V|U)] \\ &= \text{Var}_U[U\mathbb{E}_V(V)] + \mathbb{E}_U[U^2\text{Var}_V(V)] \qquad \text{(independence)} \\ &= [\mathbb{E}_V(V)]^2\text{Var}_U(U) + \text{Var}_V(V)\mathbb{E}_U(U^2). \end{aligned}$$

Using the parameter correspondence established in Section 5.3 and the formula above, we require

$$\mathbb{E}(\alpha_x^{\text{LL}}) = \mathbb{E}(\alpha_x^{\text{LC}} + \beta_x^{\text{LC}}\psi_1^{\text{LC}}) = \mathbb{E}(\alpha_x^{\text{LC}}) + \mathbb{E}(\beta_x^{\text{LC}})\mathbb{E}(\psi_1^{\text{LC}}) = -5 + 0 \times 0 = -5$$

and

$$\begin{aligned}
\mathrm{Var}(\alpha_x^{\mathrm{LL}}) &= \mathrm{Var}(\alpha_x^{\mathrm{LC}}+\beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}}) \\
&= \mathrm{Var}(\alpha_x^{\mathrm{LC}})+[\mathbb{E}(\beta_x^{\mathrm{LC}})]^2\mathrm{Var}(\psi_1^{\mathrm{LC}})+\mathbb{E}((\psi_1^{\mathrm{LC}})^2)\mathrm{Var}(\beta_x^{\mathrm{LC}}) \\
&= 4+0^2\times 2000+[2000+0^2]\times 0.005 \\
&= 14.
\end{aligned}$$

Similarly, we require

$$\mathbb{E}(\beta_x^{\mathrm{LL}})=\mathbb{E}(\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}})=\mathbb{E}(\beta_x^{\mathrm{LC}})\mathbb{E}(\psi_2^{\mathrm{LC}})=0\times 0=0$$

and

$$\begin{aligned}
\mathrm{Var}(\beta_x^{\mathrm{LL}}) &= \mathrm{Var}(\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}}) \\
&= [\mathbb{E}(\beta_x^{\mathrm{LC}})]^2\mathrm{Var}(\psi_2^{\mathrm{LC}})+\mathbb{E}((\psi_2^{\mathrm{LC}})^2)\mathrm{Var}(\beta_x^{\mathrm{LC}}) \\
&= 0^2\times 2+(2+0^2)\times 0.005 \\
&= 0.01.
\end{aligned}$$

Finally, it is slightly more complicated for the last parameter correspondence due to the model structure. Ignoring the constraints, we require

$$\mathbb{E}(\epsilon_t'^{\mathrm{LL}})=\mathbb{E}(\beta_x^{\mathrm{LC}}\epsilon_t'^{\mathrm{LC}})=\mathbb{E}(\beta_x^{\mathrm{LC}})\mathbb{E}(\epsilon_t'^{\mathrm{LC}})=0\times 0=0$$

and

$$\begin{aligned}
\mathrm{Var}(\epsilon_t'^{\mathrm{LL}}|(\sigma_\kappa^{\mathrm{LL}})^2) &= \mathrm{Var}(\beta_x^{\mathrm{LC}}\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2) \\
&= [\mathbb{E}(\beta_x^{\mathrm{LC}})]^2\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2)+\mathbb{E}((\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2)^2)\mathrm{Var}(\beta_x^{\mathrm{LC}}) \\
&= 0^2\times\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2)+[\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2) \\
&\quad +(\mathbb{E}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2))^2]\times 0.005 \\
&= 0.005\times[\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2)+(\mathbb{E}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2))^2]. \qquad (5.5)
\end{aligned}$$

Now using law of total expectation to marginalise over ρ^{LC}, and knowing that

$$\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2,\rho^{\mathrm{LC}}\sim N\left(0,\frac{(\sigma_\kappa^{\mathrm{LC}})^2}{1-(\rho^{\mathrm{LC}})^2}\right),$$

we have

$$\mathbb{E}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2)=\mathbb{E}[\mathbb{E}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2,\rho^{\mathrm{LC}})]=\mathbb{E}[0]=0$$

and

$$\begin{aligned}
\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2) &= \mathbb{E}[\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2, \rho^{\mathrm{LC}})] + \mathrm{Var}[\mathbb{E}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2, \rho^{\mathrm{LC}})] \\
&= \mathbb{E}\left[\frac{(\sigma_\kappa^{\mathrm{LC}})^2}{1-(\rho^{\mathrm{LC}})^2}\right] + \mathrm{Var}(0) \\
&= (\sigma_\kappa^{\mathrm{LC}})^2 \mathbb{E}\left(\frac{1}{1-(\rho^{\mathrm{LC}})^2}\right).
\end{aligned}$$

Since

$$\mathbb{E}\left(\frac{1}{1-(\rho^{\mathrm{LC}})^2}\right) = \int_{-1}^{1} \frac{1}{1-(\rho^{\mathrm{LC}})^2} \times \frac{3}{4}(1+\rho^{\mathrm{LC}})^2(1-\rho^{\mathrm{LC}})\mathrm{d}\rho^{\mathrm{LC}} = \frac{3}{2},$$

we obtain

$$\mathrm{Var}(\epsilon_t'^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2) = \frac{3}{2}(\sigma_\kappa^{\mathrm{LC}})^2.$$

Suppose also that we impose the same transformed beta prior distribution as ρ^{LC} on ρ^{LL}, which is sensible as they are both auto-regressive coefficients, that is

$$\frac{\rho^{\mathrm{LL}}+1}{2} \sim \mathrm{Beta}(3,2), \text{ for } \rho^{\mathrm{LL}} \in (-1,1).$$

Then, a similar calculation as above yields

$$\mathrm{Var}(\epsilon_t'^{\mathrm{LL}}|(\sigma_\kappa^{\mathrm{LL}})^2) = \frac{3}{2}(\sigma_\kappa^{\mathrm{LL}})^2.$$

Substituting all these back into Equation (5.5), this implies that we require

$$(\sigma_\kappa^{\mathrm{LL}})^2 = 0.005(\sigma_\kappa^{\mathrm{LC}})^2.$$

After matching the first two moments, the following is a summary of the prior distributions elicited for the NBLL model:

$$\begin{aligned}
\alpha_x^{\mathrm{LL}} &\overset{\mathrm{ind}}{\sim} N(-5, 14) \\
\beta_x^{\mathrm{LL}} &\overset{\mathrm{ind}}{\sim} N(0, (\sigma_\beta^{\mathrm{LL}})^2) \\
(\sigma_\beta^{\mathrm{LL}})^2 &= 0.01 \\
\frac{\rho^{\mathrm{LL}}+1}{2} &\sim \mathrm{Beta}(3,2), \text{ where } \rho^{\mathrm{LL}} \in (-1,1) \\
(\sigma_\kappa^{\mathrm{LL}})^2 &\sim \text{Inverse Gamma}(0.1, 0.000005) \text{ truncated to } (0, 0.05] \\
\phi^{\mathrm{LL}} &\sim \text{Gamma}(0.0001, 0.0001) \text{ truncated to } [1, 50000].
\end{aligned}$$

5.5.4 Investigating the Issues with the Moment-Based Approach: Laplace Distributions as A Potential Alternative

From Figure 5.3, it is evident that the implied prior distributions of $\log \mu_{xt}$ are much more plausible, with most of their density concentrated around the region $[-10, 0)$. However, the discrepancies between the kernel densities are still rather apparent. In fact, they can be used to explain why the marginal likelihood still inherently favours the NBLC model. Specifically, the prior distributions of $\log \mu_{xt}$ under the NBLC model have more density allocated to the region $[-10, 0)$ (with a sharper peak), where the likelihood is expected to dominate. Hence, being an integrated likelihood with respect to prior, the marginal likelihood clearly favours the NBLC model, since the implied prior distribution of $\log \mu_{xt}$ has larger density around the region where the likelihood is non-negligible. On the contrary, the NBLL model over penalises the likelihood function by having prior distributions that allocate an excessive weight to regions where the likelihood is essentially negligible. This is undesirable as the consequent model comparison appears to be unfair from the perspective of prior specification.

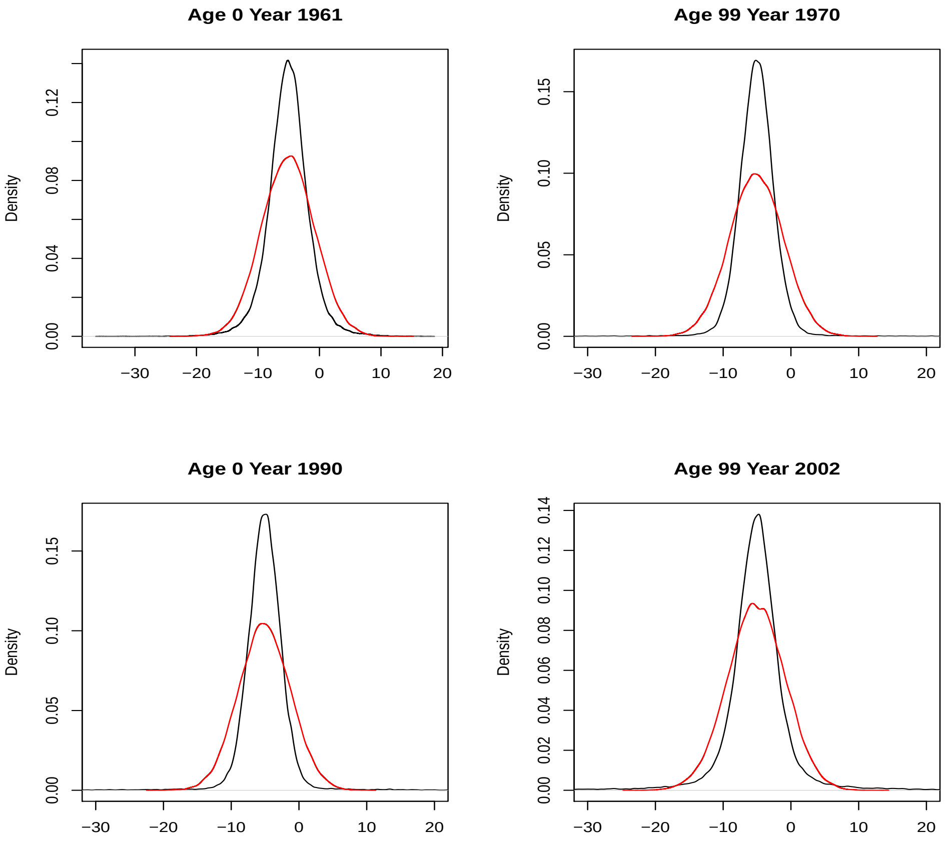


Figure 5.3: Kernel density plots of several chosen prior distributions of $\log \mu_{xt}$, as labelled, under the NBLC (black) and NBLL (red) models.

Moreover, notice also that the prior distribution of $\log \mu_{xt}$ under the NBLC model appears to possess heavier tail than its counterpart (for instance, log mortality rate at age 99 in year 2002). This can be verified by the QQ-plot depicted in Figure 5.4, where the apparent S-shape illustrated indicates a heavier tail for $\log \mu_{xt}$ under the NBLC model. The difference in the tail-heaviness arises from a mismatch in the family of prior distributions specified, which shall be further investigated promptly.

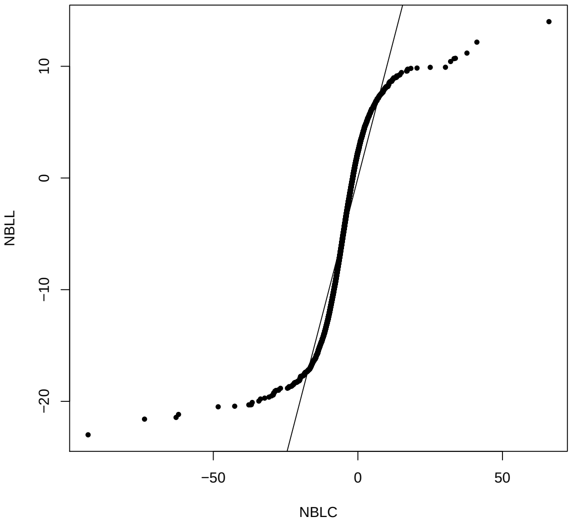


Figure 5.4: QQ-plot of $\log \mu_{xt}$ at age 99 in year 2002 for the NBLL model against the NBLC model.

For the prior distributions set up in Section 5.5, a closer inspection of the relationship of the parameter correspondence using kernel densities estimated from the simulated samples for α_x and β_x are shown in Figure 5.5. Evidently, prior densities of the parameters under the NBLC model demonstrate heavier tails in general and possess dramatically sharper peaks, especially for the β parameter (somewhat similar pattern as the log mortality rates). Basically, this occurs because the parameter correspondence relationships are violated irrespective of the moment-matching procedures. In particular, for the correspondence $\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}} \longleftrightarrow \beta_x^{\mathrm{LL}}$, left-hand-side is a product of two normally distributed random variables, while right-hand-side is a normal random variable (similarly for the α_x and κ_t). This is a mismatch of family of distributions because a random variable formed by the product of two normal random variables is known to possess a much heavier tail (which is ultimately, why the implied priors of $\log \mu_{xt}$ are much heavier-tailed under the NBLC model). Specifically, suppose U and V are zero-centered normal distribution with variance σ_u^2 and σ_v^2 respectively, i.e. $U \sim N(0, \sigma_u^2)$ and $V \sim N(0, \sigma_v^2)$, then it can be shown that their product $W = UV$ has a probability density function given by

$$f_W(w) = \frac{1}{\pi \sigma_u \sigma_v} K_0 \left(\frac{|w|}{\sigma_u \sigma_v} \right), \tag{5.6}$$

where $K_0()$ is the modified Bessel function of the second kind of order zero (see Craig, 1936 for details). Henceforth, a distribution with density function $f_W(w)$, as given in (5.6), is called the Bessel distribution with parameters σ_u and σ_v.

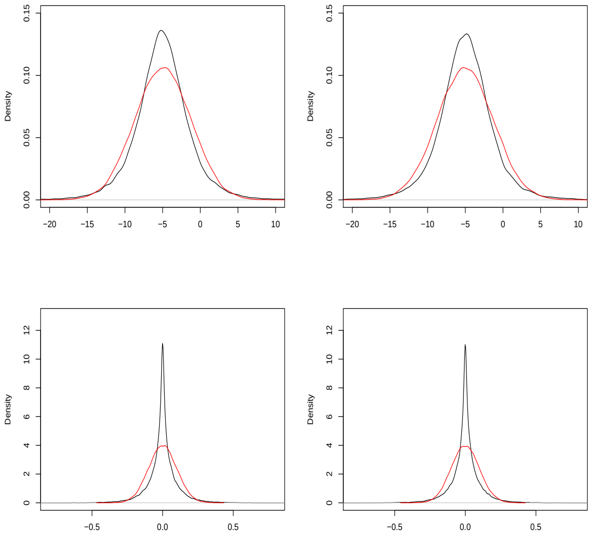


Figure 5.5: A plot of kernel density estimates of the moment-matched prior distribution of $\alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}}$ (black) against α_x^{LL} (red) in the upper panels, and $\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}}$ against β_x^{LL} in the lower panels for age 0 (left) and 99 (right).

To further visualise their differences, three distributions,

1. The Bessel distribution with parameters σ_u and σ_v,

2. $N(0, \sigma_u^2\sigma_v^2)$,

3. Laplace $\left(0, \sqrt{\frac{\sigma_u^2\sigma_v^2}{2}}\right)$ so that mean=0 and variance=$\sigma_u^2\sigma_v^2$,

where $\sigma_u^2 = 10$ and $\sigma_v^2 = 1000$ are compared, as illustrated in Figure 5.6.

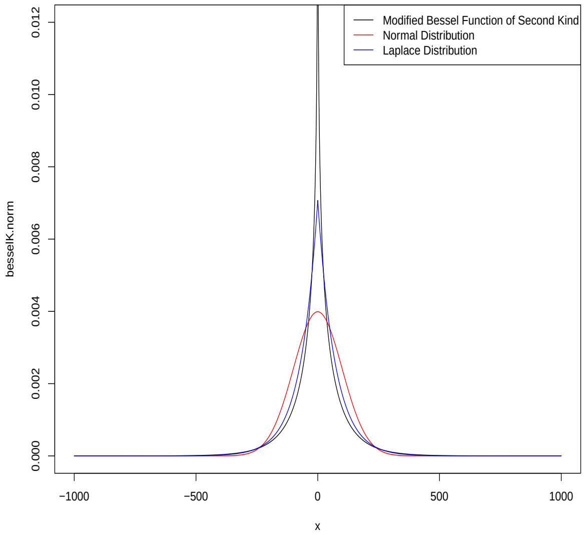


Figure 5.6: A plot of the density function of the Bessel distribution, a moment-matched normal distribution and the Laplace distribution.

According to Figure 5.6, the Bessel distribution possesses visibly heavier tail and significantly sharper peak at 0 (that asymptotes to 0 from both directions) than the moment-matched normal distribution. In fact, the kurtosis of the Bessel distribution is nine (see Appendix J for its derivation), in contrast to that of a normal distribution, which is only three. This is the primary reason why the log mortality rates under the NBLC model are observed to possess heavier tails than those of the NBLL model even after the moment-matching (Section 5.5). Therefore, if we intend to retain the prior distributions of the NBLC model, then the Bessel distribution should be considered as the prior distributions for the NBLL model instead.

Alternatively, we consider a Laplace distribution as an approximation of the modified Bessel function. A Laplace distribution (also known as the double-exponential distribution, since it is essentially formed by combining two mirrored exponential distributions) with location parameter $-\infty < a < \infty$ and scale parameter $b > 0$, Laplace(a, b), for a variable X has a density

$$\begin{aligned} f_X(x) &= \frac{1}{2b}\exp\left(-\frac{|x-a|}{b}\right) \\ &= \begin{cases} \frac{1}{2b}\exp\left(-\frac{a-x}{b}\right), & \text{for } x < a \\ \frac{1}{2b}\exp\left(-\frac{x-a}{b}\right), & \text{for } x \geq a \end{cases}. \end{aligned}$$

This distribution has a single mode at a, where two mirrored exponential distributions meet to form a cusp (see Figure 5.6). The cumulative distribution function of

Laplace(a, b) is

$$F_X(x) = \begin{cases} \frac{1}{2}\exp\left(-\frac{a-x}{b}\right), & \text{for } x < a \\ 1 - \frac{1}{2}\exp\left(-\frac{x-a}{b}\right), & \text{for } x \geq a \end{cases},$$

with upper and lower quartiles given as $a \pm b \log 2$.

As depicted in Figure 5.6, the moment-matched Laplace distribution appears to provide a descent approximation of the Bessel distribution in terms of its tail weight and the density around the peak. The approximation is not perfect, but is improved significantly when compared to a normal approximation. In fact, the kurtosis of a Laplace distribution is six, which is slightly larger than that of a normal distribution. A particularly nice feature about this distribution is that it is a compound normal distribution, formed by specifying an exponential distribution on the variance parameter (see for example, Johnson et al., 1995). That is, if

$$X|\sigma^2 \sim N(\mu, 2\sigma^2) \text{ with } \sigma^2 \sim \text{Exp}(\lambda),$$

where μ is a known constant, then the marginal distribution of X is

$$X \sim \text{Laplace}\left(a = \mu, b = \frac{1}{\sqrt{\lambda}}\right).$$

Thus, the normal prior distributions postulated previously can be easily modified into Laplace distributions by allowing the variances to be hyperparameters with exponential distributions. One interesting question here is, why is a Laplace distribution preferred over the Student's t-distribution (an inverse gamma mixture of normals), which is a classic extension of normal distributions to elevate tail weight. An explanation for this is that a Laplace distribution is believed to perform better in terms of characterizing the dramatic peak around the center of the Bessel distribution than a Student's t-distribution, which has difficulty matching sharpness of the peak and the tail weight simultaneously. To be precise, Balanda (1987) identified Laplace distribution as being more sharp peaked than a Cauchy distribution (Student's t with one degrees of freedom) despite being suggested otherwise by their kurtosis, which are respectively given by six and infinity. This is because moment-based comparison is inadequate in recognizing the dominant features of the two distributions.

5.5.5 Corrected Prior Specification for the NBLL Model: Laplace Priors

In this section, we provide a consistent version of prior distributions for the NBLL model by carefully accounting for the parameter correspondence and correctly modifying the priors (including its family of distribution). Additionally, our preliminary study (and the discussions above) indicates that moment-based comparisons have the tendency to neglect some crucial features about a distribution, particularly the dramatic peak of a modified Bessel function and its tail probabilities, which are essential elements when ensuring the similarity of those distributions. Therefore, a better alternative is to match the lower and upper quantiles instead of the variances.

The prior distributions of α_x^{LL} and β_x^{LL} we propose are

$$\begin{aligned}\alpha_x^{\mathrm{LL}} &\overset{\mathrm{ind}}{\sim} \mathrm{Laplace}\left(a_\alpha, b_\alpha\right),\\ \beta_x^{\mathrm{LL}} &\overset{\mathrm{ind}}{\sim} \mathrm{Laplace}\left(a_\beta, b_\beta\right),\end{aligned}$$

where $a_\alpha = -5$ and $a_\beta = 0$ by directly matching the modes, while b_α and b_β are chosen on the basis of quantile-matching. Specifically, b_α is such that

$$b_\alpha = -\frac{-5 - L_{\alpha;0.05}}{\log(2 \times 0.05)},$$

where $L_{\alpha;0.05}$ is the sample 5$^{\mathrm{th}}$ percentile of $\alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}}$. Similarly, b_β is such that

$$b_\beta = -\frac{0 - L_{\beta;0.05}}{\log(2 \times 0.05)},$$

where $L_{\beta;0.05}$ is the sample 5$^{\mathrm{th}}$ percentile of $\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}}$. The numerically determined b_α and b_β are roughly 2 and 0.03 respectively. As portrayed in Figure 5.7, the Laplace approximation appears to perform substantially better at capturing the relevant features (sharp peaks and heavy tails) than the previously used normal approximation (Figure 5.5).

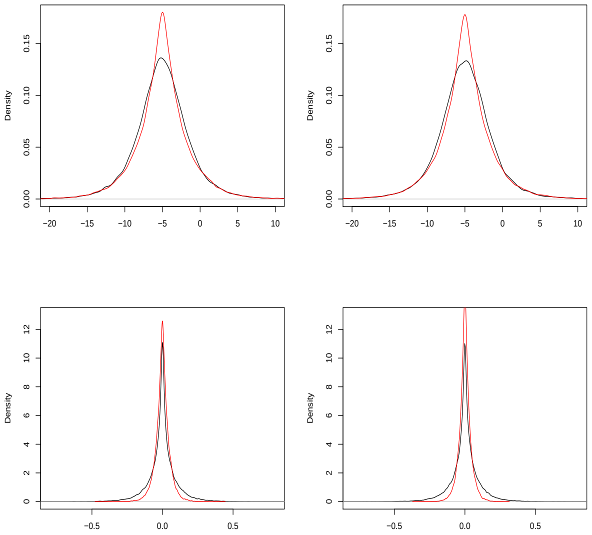


Figure 5.7: A plot of kernel density estimates of quantiles-matched prior distribution of $\alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}}$ (black) against α_x^{LL} (red) in the upper panels, and $\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}}$ against β_x^{LL} in the lower panels for age 0 (left) and 99 (right).

It is slightly trickier for the last parameter correspondence because of the interdependent autoregressive components of κ_t^{LL}. Ideally, we wish to preserve the Gaussian AR(1) model structure for κ_t^{LL} (at least conditionally) due to its nice properties, with the variance parameter, $(\sigma_\kappa^{\mathrm{LL}})^2$, modelled explicitly as a random variable for the purpose of projection. With the Laplace distribution being an exponential mixture of a normal, this can be achieved by appropriately expanding the hierarchical model. Writing in univariate form and ignoring the constraints temporarily, we propose

$$\begin{aligned} \kappa_t^{\mathrm{LL}} | \kappa_{t-1}^{\mathrm{LL}}, \rho^{\mathrm{LL}}, (\sigma_\kappa^{\mathrm{LL}})^2 &\sim N(\rho^{\mathrm{LL}}\kappa_{t-1}^{\mathrm{LL}}, 2(\sigma_\kappa^{\mathrm{LL}})^2), \\ (\sigma_\kappa^{\mathrm{LL}})^2 | \lambda^{\mathrm{LL}} &\sim \mathrm{Exp}(\lambda^{\mathrm{LL}}), \end{aligned}$$

where λ^{LL} is a newly introduced hyperparameter to be given a prior distribution. The distribution of κ_t^{LL} (integrated over $(\sigma_\kappa^{\mathrm{LL}})^2$) is then

$$\kappa_t^{\mathrm{LL}} | \kappa_{t-1}^{\mathrm{LL}}, \rho^{\mathrm{LL}}, \lambda^{\mathrm{LL}} \sim \mathrm{Laplace}\left(\rho^{\mathrm{LL}}\kappa_{t-1}^{\mathrm{LL}}, \frac{1}{\sqrt{\lambda^{\mathrm{LL}}}}\right).$$

Essentially, this is the AR(1) model with Laplace innovations as described in Wolf and Gastwirth (1967). The use of quantile-matching as previously to deduce the prior distribution of λ^{LL} is complicated by the AR(1) model structure as well as the constraints imposed. Thus, we revert to the moment-based approach, where the parameter correspondence,

$$\epsilon_t'^{\mathrm{LL}} \longleftrightarrow \beta_x^{\mathrm{LC}} \epsilon_t'^{\mathrm{LC}},$$

is revisited. Note that the distribution of κ_t^{LL} above can also be expressed marginally as

$$\kappa_t^{\mathrm{LL}}|\rho^{\mathrm{LL}},(\sigma_\kappa^{\mathrm{LL}})^2 \quad\sim\quad N\left(0,\frac{2(\sigma_\kappa^{\mathrm{LL}})^2}{1-(\rho^{\mathrm{LL}})^2}\right)$$

by integrating over κ_1^{LL}, ..., $\kappa_{t-1}^{\mathrm{LL}}$. Therefore, knowing that the innovation variance,

$$\frac{(\sigma_\kappa^{\mathrm{LL}})^2}{1-(\rho^{\mathrm{LL}})^2}|\rho^{\mathrm{LL}} \sim \mathrm{Exp}((1-(\rho^{\mathrm{LL}})^2)\lambda^{\mathrm{LL}}),$$

is still an exponential variable (conditional on ρ^{LL}), the marginal distribution of κ_t^{LL} (integrated over $(\sigma_\kappa^{\mathrm{LL}})^2$) is

$$\kappa_t^{\mathrm{LL}}|\rho^{\mathrm{LL}},\lambda^{\mathrm{LL}} \quad\sim\quad \mathrm{Laplace}\left(0,\frac{1}{\sqrt{(1-(\rho^{\mathrm{LL}})^2)\lambda^{\mathrm{LL}}}}\right).$$

Notice now that the conditional variance of the error term of κ_t^{LL} (marginalised over $(\sigma_\kappa^{\mathrm{LL}})^2$) under the NBLL is

$$\mathrm{Var}(\epsilon_t'^{\mathrm{LL}}|\rho^{\mathrm{LL}},\lambda^{\mathrm{LL}}) = \frac{2}{(1-(\rho^{\mathrm{LL}})^2)\lambda^{\mathrm{LL}}}.$$

Hence, using the exact same technique as in Section 5.5.3 to average over ρ^{LL}, we find

$$\mathrm{Var}(\epsilon_t'^{\mathrm{LL}}|\lambda^{\mathrm{LL}}) = \frac{3}{2}\times\frac{2}{\lambda^{\mathrm{LL}}}.$$

We also have from previously that

$$\mathrm{Var}(\beta_x^{\mathrm{LC}}\epsilon_t^{\mathrm{LC}}|(\sigma_\kappa^{\mathrm{LC}})^2) = \frac{3}{2}\times 0.005(\sigma_\kappa^{\mathrm{LC}})^2.$$

By matching the two variances and knowing that

$$(\sigma_\kappa^{\mathrm{LC}})^2 \sim \text{Inverse Gamma}(0.1,0.001) \text{ truncated to } (0,10],$$

these imply that a sensible prior on the new hyperparameter could be

$$\begin{aligned}\lambda^{\mathrm{LL}} &\sim \frac{2}{0.005}\times\text{Gamma}(0.1,0.001)\text{ truncated }\left[\frac{1}{10}\times\frac{2}{2000},\infty\right)\\ \Rightarrow\lambda^{\mathrm{LL}} &\sim \text{Gamma}(0.1,1)\text{ truncated }[0.0001,\infty).\end{aligned}$$

A summary of the prior distributions deduced for the NBLL model with matching prior information is as follows,

$$
\begin{aligned}
\alpha_x^{\mathrm{LL}} &\sim \text{Laplace}(-5, 2) \\
\beta_x^{\mathrm{LL}} &\sim \text{Laplace}(0, 0.03) \\
\boldsymbol{\kappa}_{-1,2}^{\mathrm{LL}} | \rho^{\mathrm{LL}}, (\sigma_\kappa^{\mathrm{LL}})^2 &\sim N_{T-2}(\mathbf{0}, 2(\sigma_\kappa^{\mathrm{LL}})^2 \boldsymbol{D}^{\mathrm{LL}}), \\
\frac{\rho^{\mathrm{LL}} + 1}{2} &\sim \text{Beta}(3, 2), \text{ where } \rho \in (-1, 1) \\
(\sigma_\kappa^{\mathrm{LL}})^2 &\sim \text{Exp}(\lambda^{\mathrm{LL}}) \\
\lambda^{\mathrm{LL}} &\sim \text{Gamma}(0.1, 1) \text{ truncated to } [0.0001, \infty) \\
\phi^{\mathrm{LL}} &\sim \text{Gamma}(0.0001, 0.0001) \text{ truncated to } [1, 50000].
\end{aligned}
$$

The resulting implied prior distributions of several chosen log mortality rates are illustrated in Figure 5.8. Remarkably, the density of the prior of log mortality rates between the two models are practically the same now, particularly for the region $[-10, 0)$ (they cannot be exactly the same because of the naturally distinct model structures and constraints between the models). After successfully matching the prior information through the comprehensive analysis above, the marginal likelihoods (and hence Bayes factor) can now be convincingly computed to serve as a model selection criterion. Note that despite the major impact on Bayes factor, the change in prior distributions is not consequential in the estimation of the parameters, given the size of our mortality data. It should be stressed that all the hard work to determine consistent prior specifications across the models is important for obtaining comparable Bayes factors, mainly because data-driven inference is intended. In principle, elicitation of prior distributions should primarily reflect prior knowledge, regardless of their impact on the model comparison procedures.

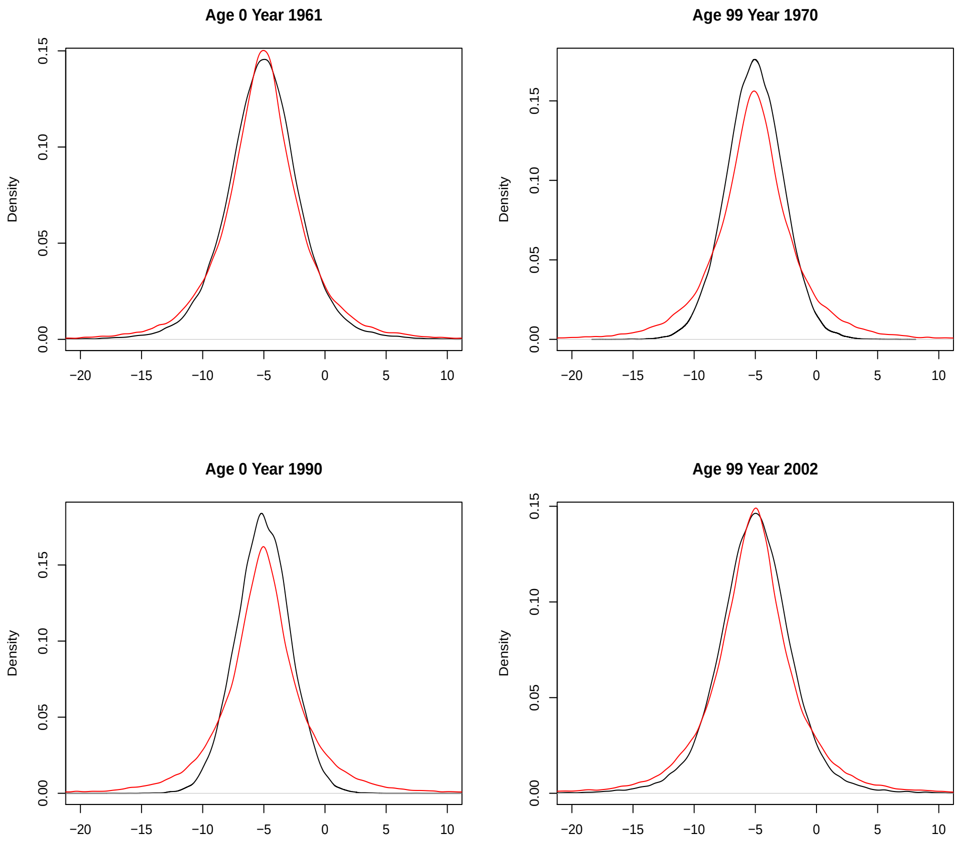


Figure 5.8: Kernel density plots of several chosen prior distributions of $\log \mu_{xt}$, as labelled, under the NBLC (black) and NBLL (red) models using Laplace approximations.

5.6 MCMC Schemes

In this section, we will be dealing primarily with the NBLL model, thus the superscripts are suppressed appropriately for clarity.

5.6.1 MCMC Scheme for the NBLL Model

For the NBLL model, we mainly applied the random walk MH algorithm for posterior sample generation because the newly appointed prior distributions are mostly not conditionally conjugate. With regards to our blocking strategy, we follow the findings on the NBLC model previously to perform univariate updating for all of the parameters, except for $\boldsymbol{\kappa}_{-1,2}$. It will be demonstrated shortly that updating the κ_t univariately leads to a poorly behaved MCMC algorithm due to the constraints.

5.6.1.1 MH Steps for α_x, β_x, σ^2_κ, λ, ρ and ϕ

Using previously defined notation, the conditional posterior densities of α and β are given respectively as

$$f(\alpha_x|\boldsymbol{\alpha}_{-x},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho,\phi) \propto \frac{\exp(\sum_t d_{xt}\alpha_x)}{\prod_t[e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t)+\phi]^{d_{xt}+\phi}}\exp\left(-\frac{|\alpha_x-a_\alpha|}{b_\alpha}\right)$$

and

$$f(\beta_x|\boldsymbol{\alpha},\boldsymbol{\beta}_{-x},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho,\phi) \propto \frac{\exp(\sum_t d_{xt}\beta_x t)}{\prod_t[e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t)+\phi]^{d_{xt}+\phi}}\exp\left(-\frac{|\beta_x-a_\beta|}{b_\alpha}\right).$$

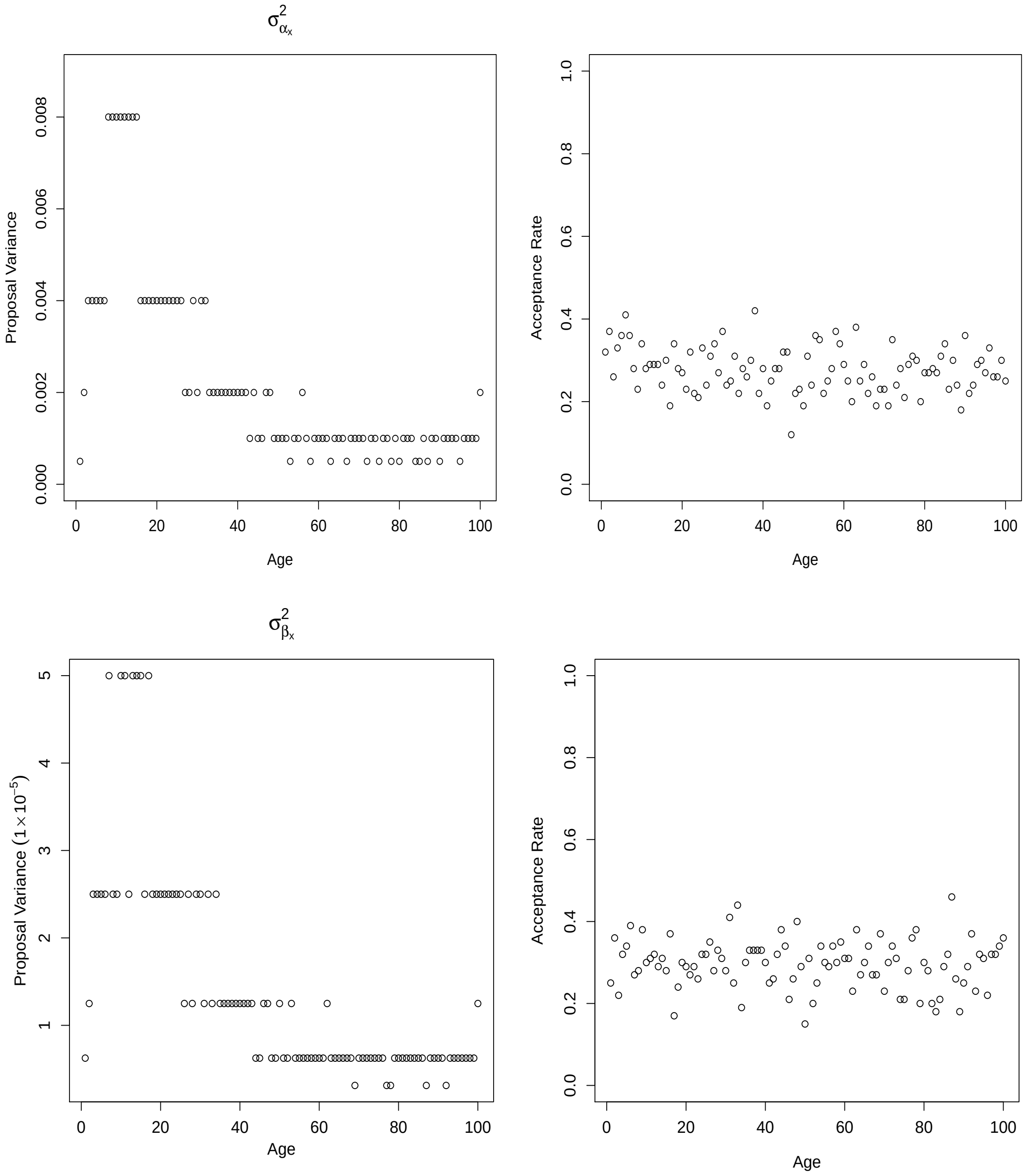


Figure 5.9: Plots of the proposal variances (left panels), $\sigma^2_{\alpha_x}$ and $\sigma^2_{\beta_x}$ and their corresponding acceptance rates (right panels).

A set of numerically determined proposal variances $\sigma^2_{\alpha_x}$ and $\sigma^2_{\beta_x}$ of the random walk MH algorithm for α_x and β_x respectively is illustrated in Figure 5.9. They demonstrate almost identical age patterns as those displayed by the NBLC model, indicating rather similar patterns of posterior variances between the models.

The conditional posterior density of σ^2_κ is

$$f(\sigma^2_\kappa|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\lambda,\rho,\phi) \propto (\sigma^2_\kappa)^{-\frac{T-1}{2}} \exp\left[-\lambda\sigma^2_\kappa - \frac{1}{4\sigma^2_\kappa}\boldsymbol{\kappa}^\top_{-1,2}\boldsymbol{D}^{-1}\boldsymbol{\kappa}_{-1,2}\right].$$

With a proposal variance of 1 for its random walk MH updating, the resulting acceptance rate returned is around 0.22. Due to the ability to preserve the Gaussian AR(1) model structure for $\boldsymbol{\kappa}_{-1,2}$ (conditional upon σ^2_κ), the conditional posterior distribution of λ is tractable and is given by

$$\lambda|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\sigma^2_\kappa,\rho,\phi \sim \text{Gamma}\left(a_\kappa+1,\frac{b_\kappa}{0.001}+\sigma^2_\kappa\right) \text{ truncated to } [0.0001,\infty),$$

where $a_\kappa = 0.1$ and $b_\kappa = 0.001$.

The conditional posterior density of ρ can be expressed as

$$\begin{aligned} f(\rho|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\sigma^2_\kappa,\lambda,\phi) \quad \propto \quad & |\boldsymbol{D}|^{-\frac{1}{2}} \exp\left[-\frac{1}{4\sigma^2_\kappa}\boldsymbol{\kappa}^\top_{-1,2}\boldsymbol{D}^{-1}\boldsymbol{\kappa}_{-1,2}\right] \\ & \times(1+\rho)^2(1-\rho)\times I_{-1<\rho<1}(\rho), \end{aligned}$$

where

$$I_{-1<\rho<1}(\rho) = \begin{cases} 1, & \text{if } -1<\rho<1 \\ 0, & \text{otherwise} \end{cases}$$

is an indicator function. Note that $\boldsymbol{D} = \boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$ is implicitly a function of ρ, i.e. $\boldsymbol{D} = \boldsymbol{D}(\rho)$. Despite the truncation range, we still use the usual non-truncated proposal,

$$\rho^* \sim N(\rho^{i-1},(\sigma^*_\rho)^2),$$

since any proposal outside $(-1,1)$ is automatically rejected (by having zero conditional posterior density). A proposal variance of $(\sigma^*_\rho)^2 = 0.5$ returns an acceptance rate of around 0.18. For the dispersion parameter, ϕ, its conditional posterior density is given by

$$\begin{aligned} f(\phi|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho) \quad \propto \quad & \prod_{x,t}\left[\frac{\Gamma(d_{xt}+\phi)}{(e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t)+\phi)^{d_{xt}+\phi}}\right]\frac{1}{[\Gamma(\phi)]^{AT}}\phi^{a_\phi-1} \\ & \times\exp(-b_\phi\phi)\times I_{1<\phi<50000}(\phi), \end{aligned}$$

where

$$I_{1<\phi<50000}(\phi) = \begin{cases} 1, & \text{if } 1<\phi<50000 \\ 0, & \text{otherwise} \end{cases}.$$

Here, we suggest the use of a log-normal proposal for ϕ to improve the computational efficiency since it is strictly positive, that is

$$\log(\phi^*) \sim N(\log(\phi^{i-1}), \sigma^2_\phi)$$

with proposal density given as

$$q(\phi^*|\phi^{i-1}) = \frac{1}{\phi^* \sqrt{2\pi\sigma^2_\phi}} \exp\left[-\frac{1}{2\sigma^2_\phi}(\log(\phi^*) - \log(\phi^{i-1}))^2\right].$$

Thus, the acceptance probability reduces to

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

A proposal variance of $\sigma^2_\phi = 0.10$ yields an acceptance rate of around 0.24.

5.6.1.2 MH Step For κ_t

The way that the constraints are handled is by expressing κ_1 and κ_2 in terms of $\boldsymbol{\kappa}_{-1,2}$, then applying the constraints on the marginal prior distribution of $\boldsymbol{\kappa}$ in the form of conditioning (see Section 5.4). κ_1 and κ_2 are thus eliminated from the parameter set (and can be deterministically retrieved later), forming a lower-dimensional hierarchical model. Once again, this corresponds to re-expressing the constraints using proper point mass priors as stated by Gelfand and Sahu (1999). One concern is that although the choice of which κ_t to remove is clearly arbitrary, it has a direct impact on the speed of convergence of the resulting MCMC algorithm. In particular, Gelfand and Sahu (1999) argued that the more informative the prior on the unidentified parameter, the slower the rate of convergence since it limits Bayesian learning from the data. In our case, this is essentially because the constraints inevitably induce correlation among the relevant parameters (through Equation 5.3), which occurs irrespective of whether the naive or the amended prior specification is undertaken. As pointed out earlier, undertaking a univariate updating MCMC scheme in the presence of highly correlated parameters is detrimental to the convergence of the algorithm. To illustrate this point, a pilot run of the univariate updating MCMC algorithm with length 1000 iterations is executed. According to Figure 5.10, the resulting trajectories of the κ_t display rather poor mixing, except for κ_1 and κ_2, which are deterministically computed from Equation (5.3) using the rest of the simulated κ_t. In fact, the lag-1 autocorrelations computed from the posterior samples of κ_t exhibit increasing trend with respect to t and are mostly very close to one as shown in Figure 5.11, indicating poor convergence.

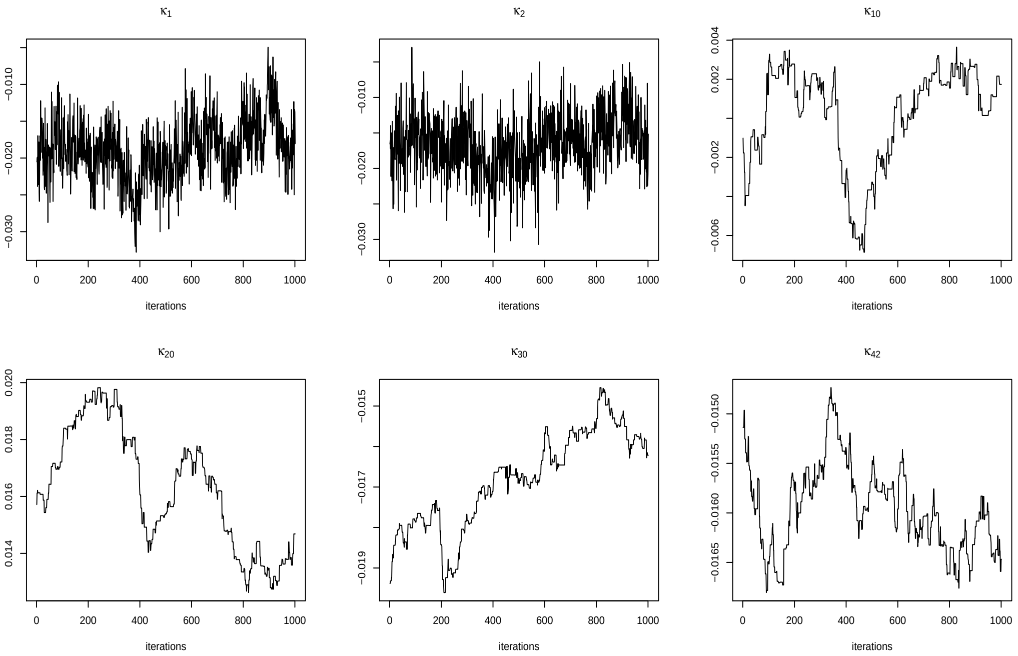


Figure 5.10: Trace plots of several chosen κ_t generated from a univariate updating MCMC scheme under the NBLL model.

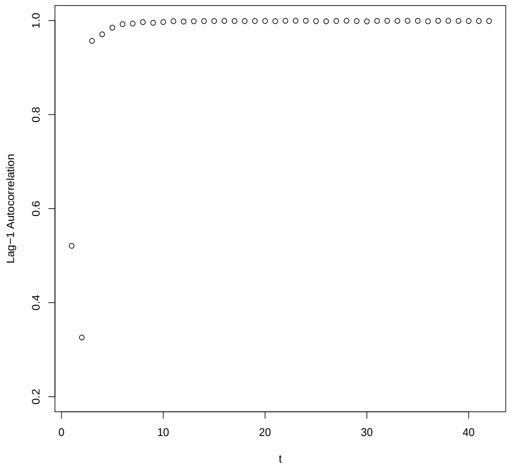


Figure 5.11: Plot of the lag-1 autocorrelations computed from the posterior samples of κ_t under the NBLL model.

Moreover, the choice of which κ_t to remove also affects the computational stability of the resulting algorithm, which shall be ignored at the moment but will be revisited in Section 6.1.

The solution to this issue is to perform block updating on the κ_t. That is, we propose a value, $\boldsymbol{\kappa}^*_{-1,2}$ multivariately at the i^{th} iteration,

$$\boldsymbol{\kappa}^*_{-1,2} \sim N_{T-2}(\boldsymbol{\kappa}^{i-1}_{-1,2}, \boldsymbol{\Sigma}_\kappa),$$

where $\boldsymbol{\kappa}^{i-1}_{-1,2}$ is the current iterate and $\boldsymbol{\Sigma}_\kappa$ is the proposal variance matrix to be specified. The proposal is then accepted with probability

$$a(\boldsymbol{\kappa}^*_{-1,2}|\boldsymbol{\kappa}^{i-1}_{-1,2}) = \min\left\{1, \frac{f(\boldsymbol{\kappa}^*_{-1,2}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho,\phi)}{f(\boldsymbol{\kappa}^{i-1}_{-1,2}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho,\phi)}\right\},$$

where

$$f(\boldsymbol{\kappa}_{-1,2}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{d},\sigma^2_\kappa,\lambda,\rho,\phi) \propto \frac{\exp(\sum_{x,t} d_{xt}\kappa_t)}{\prod_{x,t}[e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t)+\phi]^{d_{xt}+\phi}} \exp\left(-\frac{1}{4\sigma^2_\kappa}\boldsymbol{\kappa}^\top_{-1,2}\boldsymbol{D}^{-1}\boldsymbol{\kappa}_{-1,2}\right),$$

is the conditional posterior density of $\boldsymbol{\kappa}_{-1,2}$ and remembering that κ_1 and κ_2 are functions of $\boldsymbol{\kappa}_{-1,2}$ through Equation 5.3.

Now suppose that $\boldsymbol{\theta}_{\text{NBLL}} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\kappa}^\top_{-1,2}, \sigma^2_\kappa, \lambda, \rho, \phi)^\top$ is the full set of parameters for the NBLL model and $\boldsymbol{H}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$ is the corresponding Hessian matrix of the joint posterior distribution with ij^{th} element

$$[\boldsymbol{H}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})]_{ij} = \frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d})}{\partial\theta_{\text{NBLL}\,j}\partial\theta_{\text{NBLL}\,i}}.$$

The proposal variance matrix we suggest is

$$\boldsymbol{\Sigma}_\kappa = \frac{2.38^2}{39} \times [-\boldsymbol{H}^\kappa_{\text{NBLL}}(\boldsymbol{\theta}^{\text{mode}}_{\text{NBLL}})]^{-1},$$

where $\boldsymbol{H}^\kappa_{\text{NBLL}}(\boldsymbol{\theta}^{\text{mode}}_{\text{NBLL}})$ is the sub-matrix of $\boldsymbol{H}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$ corresponding to $\boldsymbol{\kappa}_{-1,2}$, evaluated at the joint posterior mode, $\boldsymbol{\theta}^{\text{mode}}_{\text{NBLL}}$ (found by applying the iterative conditional mode search algorithm). An expression of the sub Hessian matrix, $\boldsymbol{H}^\kappa_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$, is provided in Appendix I. Note that $[-\boldsymbol{H}^\kappa_{\text{NBLL}}(\boldsymbol{\theta}^{\text{mode}}_{\text{NBLL}})]^{-1}$, which is only an approximation of the conditional posterior variance of $\boldsymbol{\kappa}_{-1,2}$ instead of the marginal posterior variance, is used to determine the proposal variance matrix (in contrast to Section 3.6.3). In fact, this choice of proposal variance does not have a major implication in terms of the MCMC convergence since only an approximation of the curvature of $\boldsymbol{\kappa}_{-1,2}$ is required to facilitate efficient transitions across the parameter spaces. Hence, $\boldsymbol{H}^\kappa_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$ is used as it is relatively easier to derive compared to the full Hessian matrix, $\boldsymbol{H}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$. This proposal variance matrix yields an acceptance rate of around 0.26 for the MH algorithm of $\boldsymbol{\kappa}_{-1,2}$. The resulting block updating MCMC algorithm is much faster to execute and produces trajectories with better mixing and slightly lower (and stabilised) lag-1 sample autocorrelations across t as depicted in Figure 5.12 and 5.13 respectively.

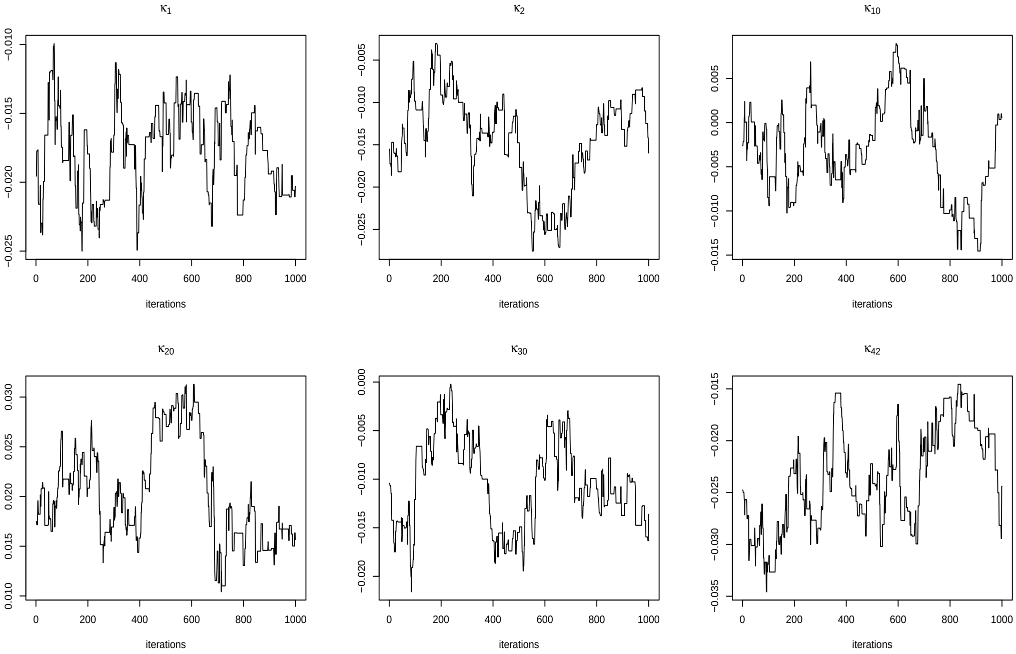


Figure 5.12: Trace plots of several chosen κ_t generated from block updating MCMC scheme under the NBLL model.

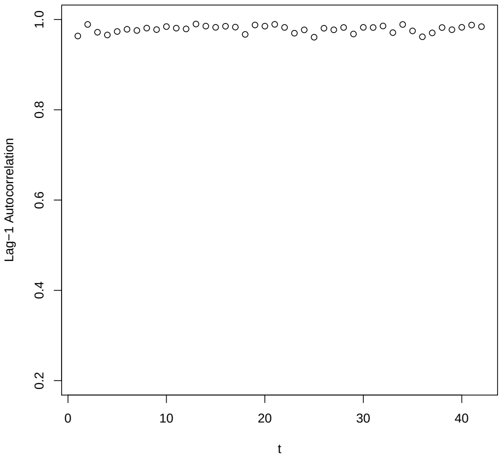


Figure 5.13: Plot of the lag-1 autocorrelations computed from the posterior samples of κ_t under the NBLL model with block updating.

5.6.2 MCMC Scheme for the NBLC Model with Amended Priors

Most of the conditional posterior distributions/densities provided in Section 3.4.2 need to be altered appropriately due to the change in prior distributions as discussed in Section 5.5. Nevertheless, the changes mainly involve different constants for the prior distributions which still belong to the same family of distributions. Hence, the previous results can be used by simply altering the constants. On the other hand, the consequence of using truncated prior distributions is effectively restricting the parameter space of the posterior distributions to the corresponding region without changing the core of the conditional posterior density. In other words, conditional posterior distributions derived in Section 3.4.2 can be used by appropriately shrinking the posterior parameter space according to the truncation. In particular, for $(\sigma_\kappa^{\mathrm{LC}})^2$, which is truncated to the region $(0,10]$, its conditional posterior distribution is now

$$\begin{aligned} & (\sigma_\kappa^{\mathrm{LC}})^2|\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\boldsymbol{\beta}_{-1},\boldsymbol{d},\rho,\boldsymbol{\psi},\phi \\ \sim\ & \text{Inverse Gamma}\left(a_\kappa+\frac{T-1}{2},b_\kappa+\frac{1}{2}(\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}-\boldsymbol{\mu}_\kappa^{\mathrm{LC}})^\top(\boldsymbol{D}^{\mathrm{LC}})^{-1}(\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}-\boldsymbol{\mu}_\kappa^{\mathrm{LC}})\right), \end{aligned}$$

truncated to $(0,10]$, where $\boldsymbol{\mu}_\kappa^{\mathrm{LC}}$ and $\boldsymbol{D}^{\mathrm{LC}}$ are as defined in Section 5.5.

The conditional posterior densities of ρ^{LC} and $\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}$ experience major changes because their prior distributions are changed entirely. The conditional posterior density of ρ^{LC} is now

$$\begin{aligned} & f(\rho^{\mathrm{LC}}|\boldsymbol{\alpha}^{\mathrm{LC}},\boldsymbol{\beta}_{-1}^{\mathrm{LC}},\boldsymbol{\kappa}_{-1}^{\mathrm{LC}},\boldsymbol{d},(\sigma_\kappa^{\mathrm{LC}})^2,\boldsymbol{\psi}^{\mathrm{LC}},\phi^{\mathrm{LC}}) \\ \propto\ & |\boldsymbol{D}^{\mathrm{LC}}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma_\kappa^2}(\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}-\boldsymbol{\mu}_\kappa^{\mathrm{LC}})^\top(\boldsymbol{D}^{\mathrm{LC}})^{-1}(\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}-\boldsymbol{\mu}_\kappa^{\mathrm{LC}})\right]\times(1+\rho^{\mathrm{LC}})^2(1-\rho^{\mathrm{LC}}) \\ & \times I_{-1<\rho^{\mathrm{LC}}<1}(\rho^{\mathrm{LC}}), \end{aligned}$$

where $I_{-1<\rho^{\mathrm{LC}}<1}(\rho^{\mathrm{LC}})$ is the indicator function and $\boldsymbol{\mu}_\kappa^{\mathrm{LC}}$ and $\boldsymbol{D}^{\mathrm{LC}}$ are as above. A proposal variance of 0.25 produces an acceptance rate of about 0.28. Again, the constraint, $\sum_t \kappa_t^{\mathrm{LC}}=0$ induces correlations among the κ_t^{LC}. Hence, we perform block updating MCMC scheme by allocating κ_t^{LC} in a single block. Its conditional posterior density can be derived as

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

After some proper tuning, the proposal variance matrix we propose is

$$\frac{6.38^2}{39}\times\boldsymbol{G}_{\mathrm{NBLC}}^{\kappa}(\boldsymbol{\delta}_{\mathrm{NBLC}}^{\mathrm{MLE}}),$$

where $\boldsymbol{G}_{\mathrm{NBLC}}^{\kappa}(\boldsymbol{\delta}_{\mathrm{NBLC}}^{\mathrm{MLE}})$ is the sub-matrix of $[\boldsymbol{H}_{\mathrm{NBLC}}]^{-1}$ corresponding to $\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}$ evaluated at the MLE of the parameters, $\boldsymbol{\delta}_{\mathrm{NBLC}}=(\boldsymbol{\alpha}^{\mathrm{LC}\top},\boldsymbol{\beta}_{-1}^{\mathrm{LC}\top},\boldsymbol{\kappa}_{-1}^{\mathrm{LC}\top},\phi^{\mathrm{LC}})^\top$, and $\boldsymbol{H}_{\mathrm{NBLC}}$ is the Hessian matrix of the likelihood function of the NBLC model (not the joint posterior distribution). In practice, $\boldsymbol{H}_{\mathrm{NBLC}}$ can be estimated by fitting the NBLC model in the frequentist framework, then extracting the corresponding covariance matrix estimated (for example, using the function "glm" in R). Again, this crude approximation of the marginal variance matrix of $\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}$ works because only an approximation of its curvature is required (notice the different proportionality constant, $6.38^2/39$).

5.7 Mortality Forecast under the NBLL Model

Mortality projection under the NBLL model is similar to that under the NBLC model. The posterior predictive density of 1-year ahead log mortality rates for each age group under this model can be written as

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

where $f(\alpha_x, \beta_x, \kappa_{T+1}, \phi|\boldsymbol{d})$ and $f(\kappa_T, \rho, \lambda, \sigma_\kappa^2|\boldsymbol{d})$ are the joint posterior distributions. This suggests the generation of posterior samples of $\log \mu_{x\,T+1}$ as follows:

1. Generate κ_{T+1} from
$$\kappa_{T+1} \sim N(\rho\kappa_T, 2\sigma_\kappa^2),$$
where joint posterior samples of $(\kappa_T, \rho, \sigma_\kappa^2)$ from the MCMC output are substituted into the expression.

2. Generate $\mu_{x\,T+1}$ from
$$\mu_{x\,T+1} \sim \mathrm{Gamma}\left(\phi, \frac{\phi}{\exp(\alpha_x + \beta_x(T+1) + \kappa_{T+1})}\right),$$
where κ_{T+1} is from step 1 and $(\alpha_x, \beta_x, \phi)$ are joint posterior samples from the MCMC output.

The h-years ahead projections can then be obtained by recursive implementation of the above generation procedures as before (see Section 3.7).

5.8 Numerical Results

In this section, we let NBLL-AR1 and NBLL-RW be abbreviations of the NBLL model with stationary AR(1) and random walk priors on κ_t^{LL} respectively. By analogy, NBLC-AR1 and NBLC-RW denote the NBLC model with stationary AR(1) and random walk with drift priors respectively.

Figure 5.14 illustrates the medians and prediction intervals of the fitted and projected α_x, β_x and κ_t under the NBLL with both stationary AR(1) and random walk models.

The fitted values of all of them almost coincide between the two time series models. The main discrepancy arises from the projected κ_t, where the median forecast of the stationary AR(1) model asymptotes towards zero; while that of random walk model levels off at the most recent value of κ_t. In addition, the 95% prediction intervals under the NBLL-AR1 model maintain roughly constant width across the years, as opposed to the NBLL-RW model, which fans out relatively quickly.

For the rest of the parameters, the kernel density estimates of λ and the dispersion parameter, $1/\phi$ are virtually identical between the two time series models (see Figure 5.15). Arguably, the variance of κ_t, σ^2_κ under the NBLL-RW model is slightly larger than that under the NBLL-AR1 model.

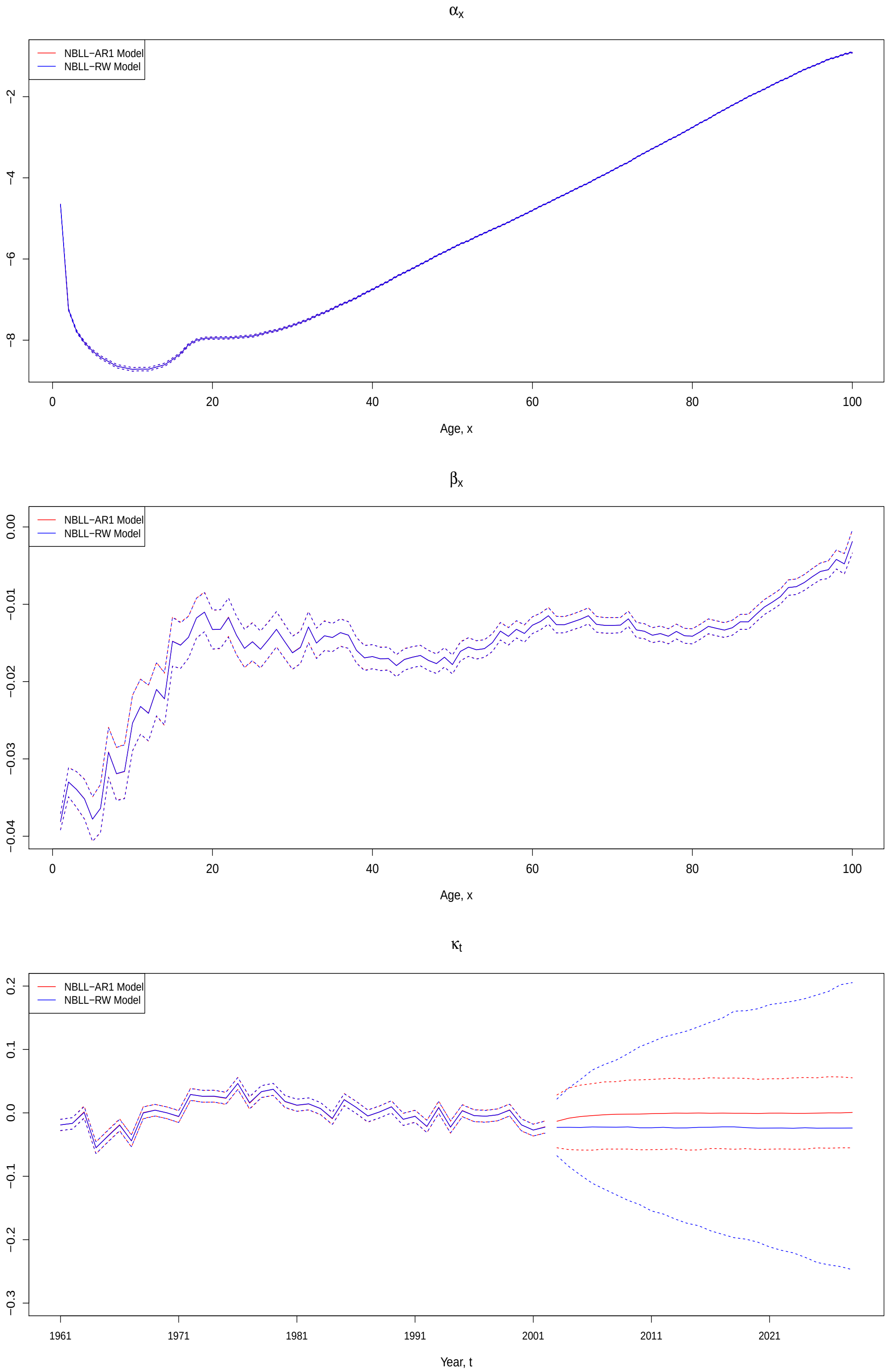


Figure 5.14: Plot of the medians (solid) of α_x, β_x, fitted and projected κ_t, accompanied by their corresponding 95% intervals (dotted) under the NBLL-AR1 and the NBLL-RW models.

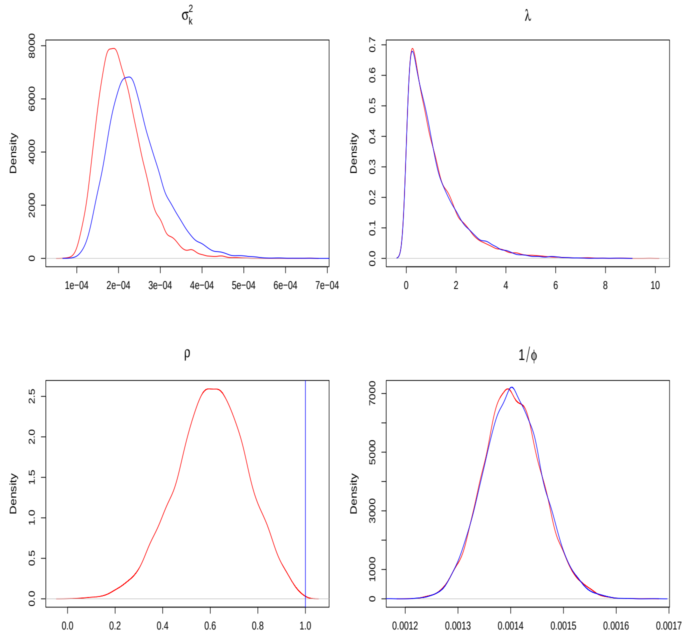


Figure 5.15: Kernel density plots for σ^2_κ, λ, ρ, and $1/\phi$ under the NBLL-AR1 (red) and the NBLL-RW (blue) models.

Under the NBLC model (with amended priors), the fitted values of α_x, β_x and κ_t are again almost identical between the two time series models (see Figure 5.16). The projected κ_t under the NBLC-AR1 model is a decreasing function of t at a decelerating rate, with prediction intervals that fan out slowly, while the NBLC-RW model not only forecasts a linearly decreasing trend, but also yields prediction intervals that fan out relatively quickly.

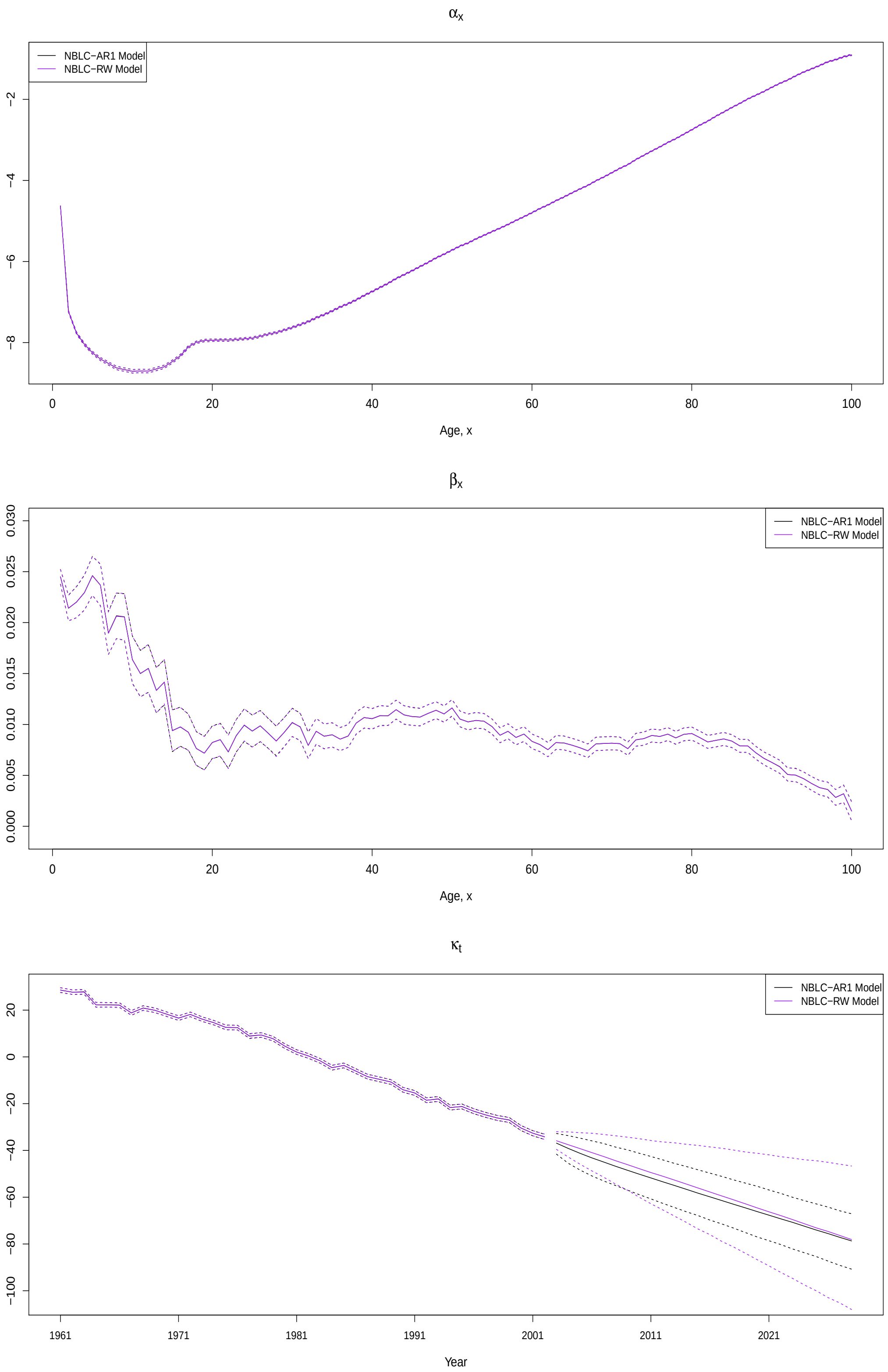


Figure 5.16: Plot of the medians (solid) of α_x, β_x, fitted and projected κ_t, accompanied by their corresponding 95% intervals (dotted) under the NBLC-AR1 and the NBLC-RW models.

From Figure 5.17, we witness similar properties as the NBLL model, where the posterior distributions of $1/\phi$ are the same, while the residual variance, σ_κ^2 is slightly larger for the NBLC-RW model. Interestingly, the kernel posterior densities of ψ_1 and ψ_2 differ substantially between the two times series models, with the random walk model producing distributions with comparably larger variabilities. This is fundamentally because when $\rho^{\text{LC}} = 1$, the time series model on κ_t^{LC} reduces to

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

meaning that ψ_1^{LC} is a redundant parameter (that is non-identifiable) and ψ_2^{LC} is the drift term. Being a redundant parameter, nothing is learned about ψ_1 by the model, hence the large variabilities.

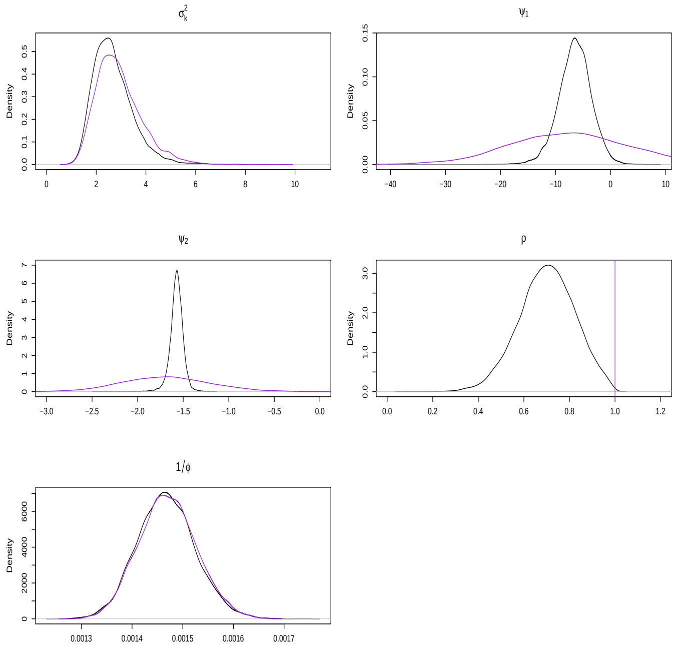


Figure 5.17: Kernel density plots for σ_κ^2, ψ_1, ψ_2, ρ, and $1/\phi$ under the NBLC-AR1 (black) and NBLC-RW (purple) models with amended prior distributions.

The 26-years ahead forecasts of log mortality rates for ages 0 and 85 under each models are depicted in Figure 5.18. First, we focus on the infant log mortality rates. For the NBLL models, the projected infant mortality rates under both time series models are rather similar, with the random walk model yielding slightly wider prediction intervals. Note that the prediction intervals under the NBLL-AR1 model appear to be implausibly narrow. On the other hand, the NBLC models forecast lower infant mortality rates in overall for both time series models, accompanied with comparatively wider prediction intervals. Again, the NBLC-RW model leads to wider prediction intervals than the NBLC-AR1 model (this time wide enough to cause explosive behaviour in long term projection).

For age 85, similar properties are observed, in that the NBLL-AR1 model underestimates mortality improvement with implausibly narrow prediction intervals, while both the NBLC-AR1 and NBLC-RW models project better mortality improvements. However, one major difference is that the NBLL and NBLC models with random walk now produce forecasts that are remarkably similar. Overall, all of the models have their own shortcomings in terms of projection, mainly because stationary AR(1) models are known to yield prediction intervals that are too narrow, while those of random walk models are too explosive.

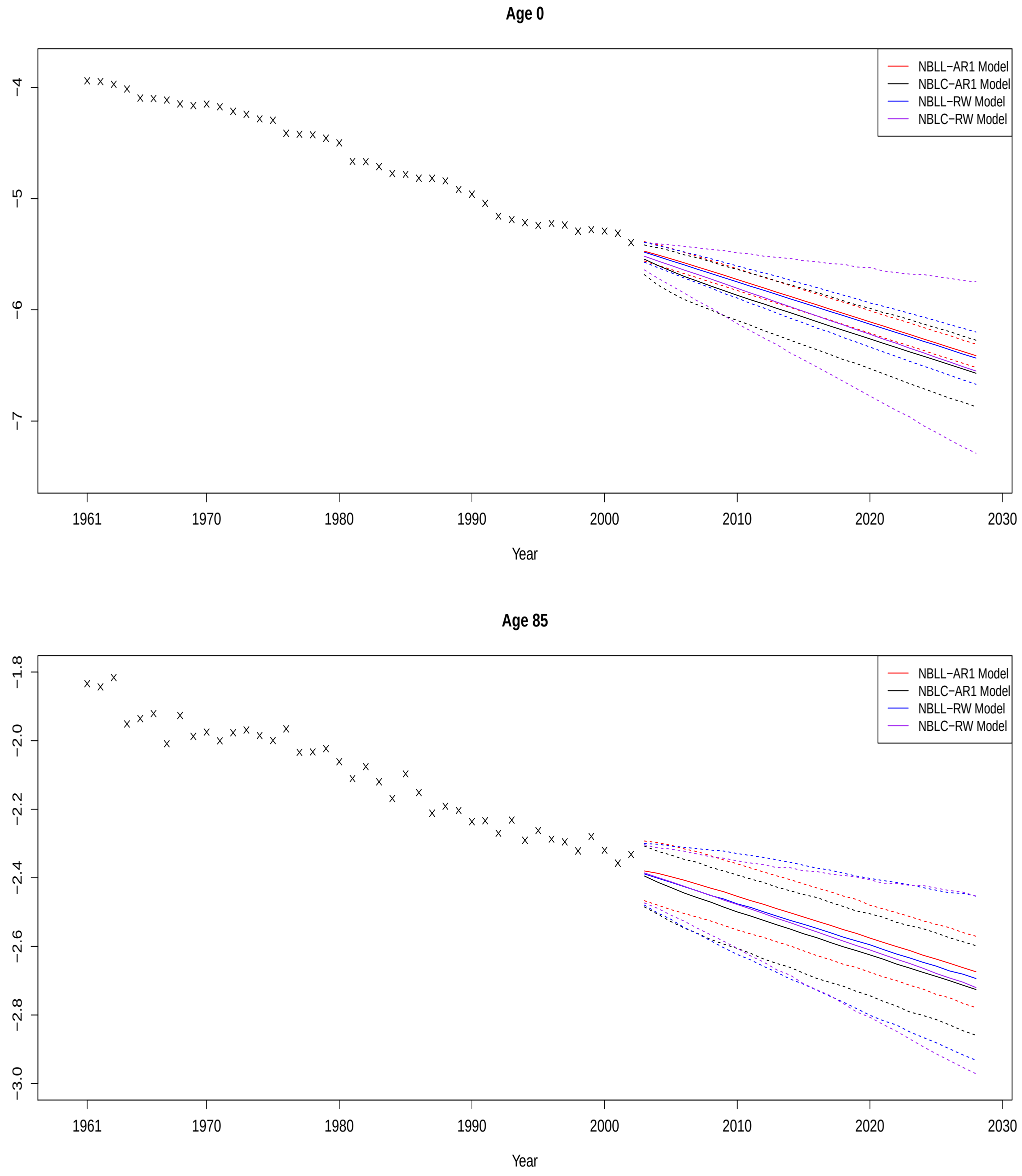


Figure 5.18: Plots of the log crude death rates, and the associated 26-years ahead projection of the underlying log mortality rates for age 0 (upper panel) and 85 (lower panel) under the NBLL and NBLC models, accompanied by the 95% credible intervals.

We can similarly perform the out-of-sample validation against the hold-out samples. According to Figure 5.19, the performance of the NBLL-AR1 model is underwhelming. Not only does it greatly underestimate the gains in life expectancy, the intervals produced are also overly narrow. In contrast, the NBLC-AR1 model leads to forecast that best describes the gains in life expectancy out of the competing models. Specifically, its median forecast is closest to the hold-out life expectancies, with prediction intervals that cover most of the hold-out samples. Both the NBLL-RW and NBLC-RW models yield rather similar projections, with the NBLL-RW model producing slightly wider intervals. They also appear to under-perform by underestimating the gains in life expectancy, which are compensated by having unnecessarily wide prediction intervals.

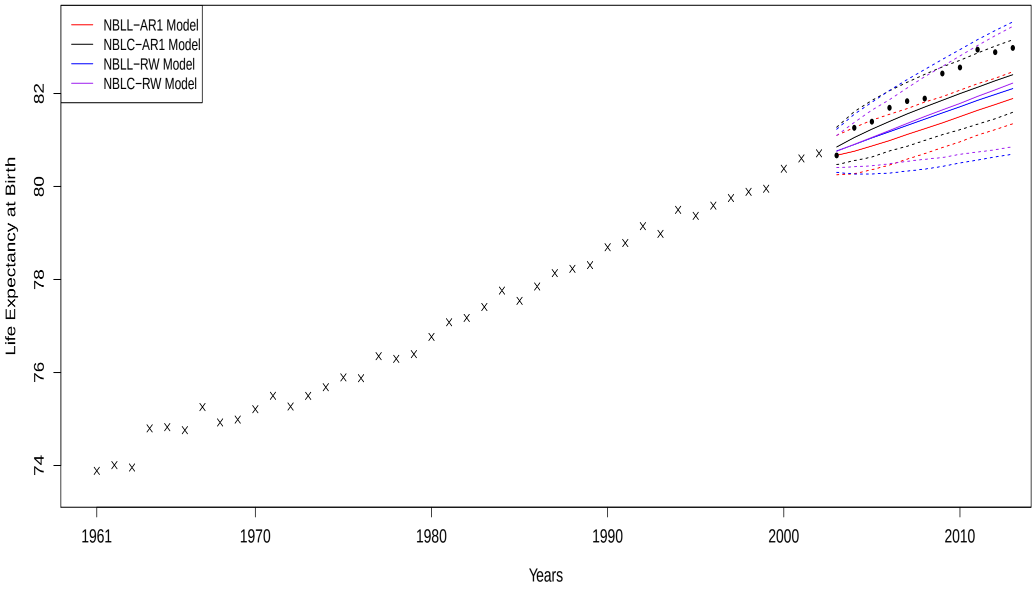


Figure 5.19: Plots of the observed life expectancy at birth and the associated 11-years ahead median forecasts (solid lines) under the NBLL and NBLC models, accompanied by the 95% prediction intervals (dotted lines).

It is evident from Table 5.1 that after amending the prior specification, the marginal likelihoods estimated agree with the conclusion drawn from the BIC. In particular, the NBLL models outperform the NBLC models by a considerable margin in the sense of posterior model probabilities (a Bayes factor of about 100 on the log scale in favour of the NBLL models). Therefore, the usual Bayesian model comparison procedure implies that the NBLL models are favoured. Between time series models, the marginal likelihoods are particularly close, with the random walk models having a slight edge over the stationary AR(1) models. This is probably not so surprising as it indicates that the regression coefficient, ρ, is again a mixture distribution under both models.

Model	Estimated Log Marginal Likelihood
NBLL-AR1	-23695.86
NBLC-AR1	-23795.45
NBLL-RW	-23695.52
NBLC-RW	-23793.75

Table 5.1: The marginal likelihoods (on logarithmic scale) of each model approximated by bridge sampling.

5.8.1 Bayesian Model Averaging

Previously, the projection of the log mortality rates under each individual model appears to be rather implausible, with the stationary AR(1) models producing prediction intervals that are too narrow, while the random walk models lead to intervals that are too wide. Hence, this motivates the idea of combining the models using the concept of Bayesian model averaging (see for example Hoeting et al., 1999). The rationale behind this idea originates from Madigan and Raftery (1994), where it was stated that model averaging results in better predictive ability than those conditional on a single selected model in the sense of logarithmic scoring rule. Specifically, this is achieved by properly mixing the posterior distributions under each model, with the posterior model probabilities as the mixture proportions, i.e.

$$f(\boldsymbol{\theta}_M|\boldsymbol{d}) = \sum_{M} f(M|\boldsymbol{d}) f(\boldsymbol{\theta}_M|\boldsymbol{d}, M). \tag{5.7}$$

Rather than model averaging across all of the models (which would be completely dominated by the NBLL models), we propose to only average between the two time series models under each of the NBLL and NBLC models. Thus, we will not encounter any issue when applying Equation (5.7), where the parameters, $\boldsymbol{\theta}_M$, to be averaged are of the same dimensionality. Model uncertainty in light of the two time series models is incorporated into the subsequent posterior distributions. To be more precise, for the NBLL models, we have

$$\begin{aligned} f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d}) &= f(M = \text{NBLL-AR1}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLL-AR1}}|\boldsymbol{d}, M = \text{NBLL-AR1}) \\ &\quad + f(M = \text{NBLL-RW}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLL-RW}}|\boldsymbol{d}, M = \text{NBLL-RW}), \end{aligned}$$

where (assuming equal prior model probabilities)

$$f(M = \text{NBLL-RW}|\boldsymbol{d}) = \frac{\exp(-23695.52)}{\exp(-23695.52) + \exp(-23695.86)} \approx 0.5842,$$

and $f(M = \text{NBLL-AR1}|\boldsymbol{d}) = 1 - f(M = \text{NBLL-RW}|\boldsymbol{d}) \approx 0.4158$. Similarly, for the NBLC models,

$$\begin{aligned} f(\boldsymbol{\theta}_{\text{NBLC}}|\boldsymbol{d}) &= f(M = \text{NBLC-AR1}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLC-AR1}}|\boldsymbol{d}, M = \text{NBLC-AR1}) \\ &\quad + f(M = \text{NBLC-RW}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLC-RW}}|\boldsymbol{d}, M = \text{NBLC-RW}), \end{aligned}$$

where

$$f(M = \text{NBLC-RW}|\boldsymbol{d}) = \frac{\exp(-23793.75)}{\exp(-23793.75) + \exp(-23795.45)} \approx 0.8455,$$

and $f(M = \text{NBLC-AR1}|\boldsymbol{d}) = 1 - f(M = \text{NBLC-RW}|\boldsymbol{d}) \approx 0.1545$. From a sampling perspective, these suggest an incredibly straightforward way to obtain samples from the model averaged posterior distributions, $f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d})$ and $f(\boldsymbol{\theta}_{\text{NBLC}}|\boldsymbol{d})$. In particular, suppose a sample of size 10000 is required from $f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d})$, this can be accomplished by combining a sample of size $10000 \times 0.5842 = 5842$ from $f(\boldsymbol{\theta}_{\text{NBLL-RW}}|\boldsymbol{d}, M = \text{NBLL-RW})$ and a sample of size 4158 from $f(\boldsymbol{\theta}_{\text{NBLL-AR1}}|\boldsymbol{d}, M = \text{NBLL-AR1})$ (both of which are readily available from our MCMC output). Hence, the mixture posterior densities of ρ and the rest of the parameters can be estimated from these samples. The resulting posterior distributions are similar to what would be obtained if a non-truncated vague prior distribution is imposed on ρ when fitting these models (except fitted values of the auto-regressive coefficient due to the random walk model scatter around 1 rather than being exactly 1). With the model averaged posterior distributions formed, we can now compare the overall predictive power of the NBLL and NBLC models.

5.8.2 Model Averaged Results

The fitted values of α_x and β_x after applying model averaging remain the same (since they were the same between the time series models) so please refer to Section 5.8. For κ_t, the fitted values remain the same, while the prediction intervals of their projections under both the model averaged models are now the compromise between the overly narrow intervals due to the stationary AR(1) models and the overly wide intervals due to the random walk models (see Figure 5.20).

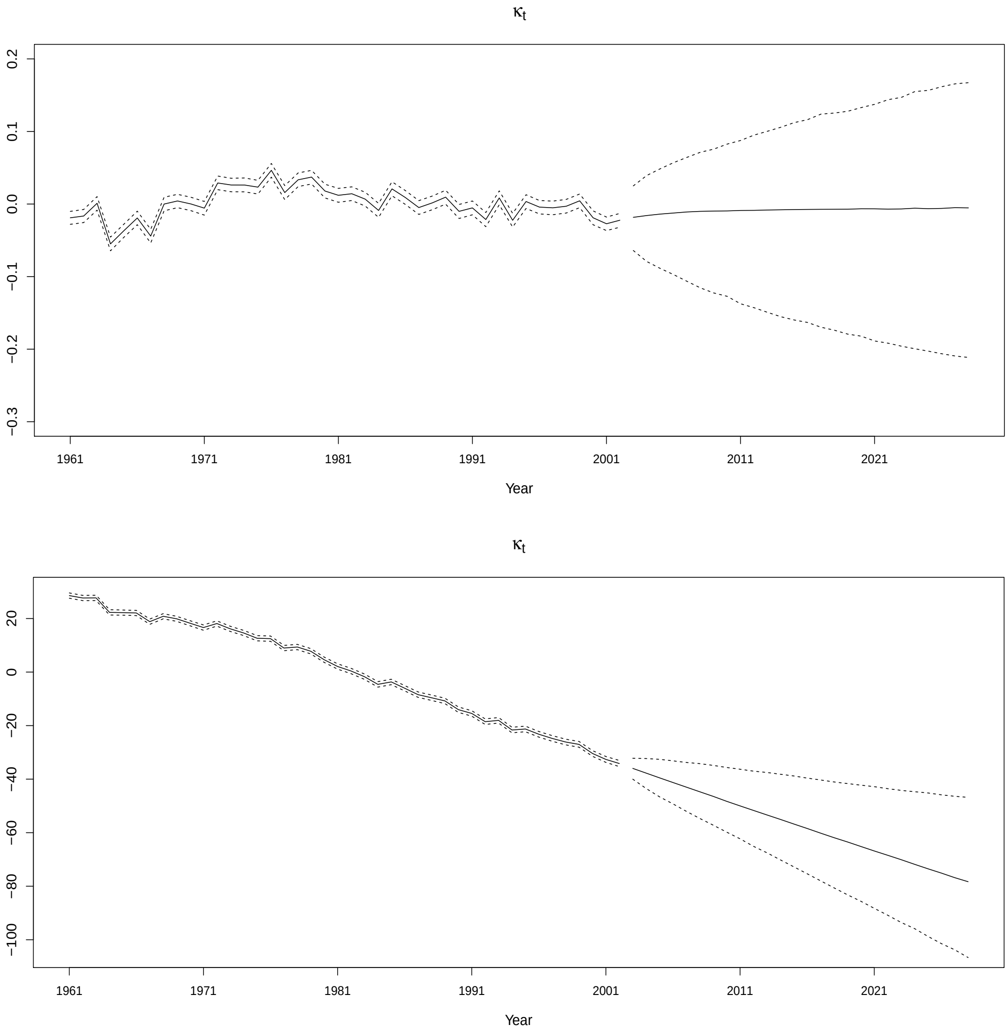


Figure 5.20: Plot of the medians of the fitted and projected κ_t, accompanied by their corresponding 95% intervals under the model averaged NBLL (upper panel) and NBLC (lower panel) models.

The estimated kernel posterior densities for the rest of the parameters under the model averaged NBLL and NBLC models are depicted respectively in Figure 5.21 and Figure 5.22. Clearly, the kernel posterior densities of the auto-regressive coefficient, ρ, are mixture distributions, and so are ψ_1 and ψ_2.

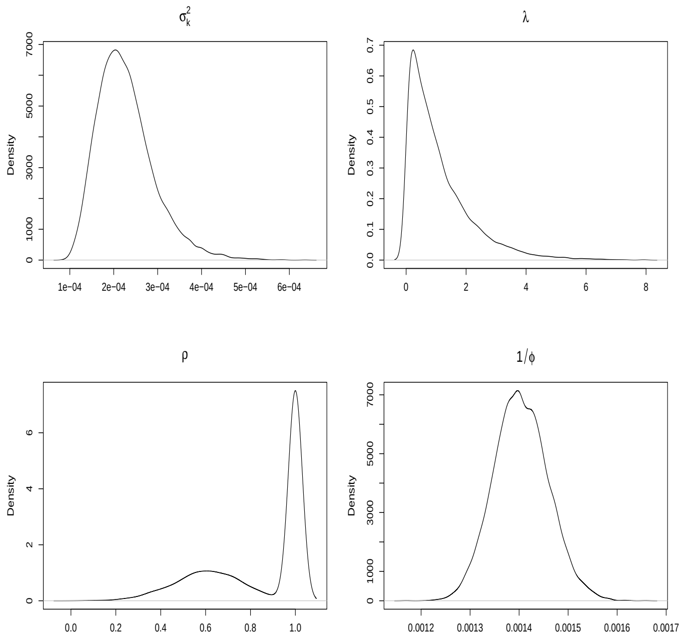


Figure 5.21: Kernel density plots for σ_κ^2, λ, ρ, and $1/\phi$ under the model averaged NBLL model, where the point mass at 1 for ρ (corresponding to the random walk model) is represented by a density concentrated around 1 with the appropriate mixture probability.

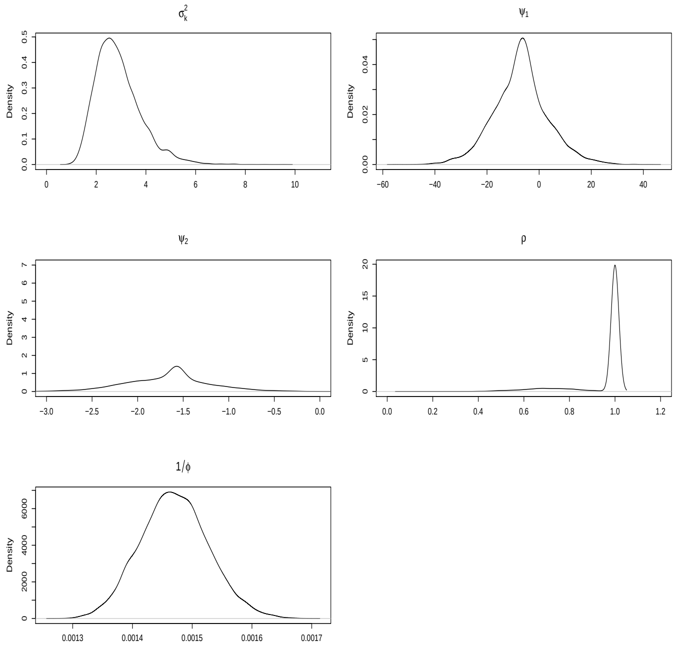


Figure 5.22: Kernel density plots for σ_κ^2, ψ_1, ψ_2, ρ, and $1/\phi$ under the model averaged NBLC model with corrected prior distributions, where the point mass at 1 for ρ (corresponding to the random walk model) is represented by a density concentrated around 1 with the appropriate mixture probability.

The fitted log mortality rates as depicted in Figure 5.23 are rather similar between the two competing models. For the projected log mortality rates, the forecasts differ considerably for age 0, where the model averaged NBLL model still yields prediction intervals that are overly narrow as compared to the model averaged NBLC model. This discrepancy diminishes as age increases. In both cases, the NBLC model forecasts a larger mortality improvement in the future. On the other hand, the fitted log mortality rates are rather similar between the NBLL and NBLC models.

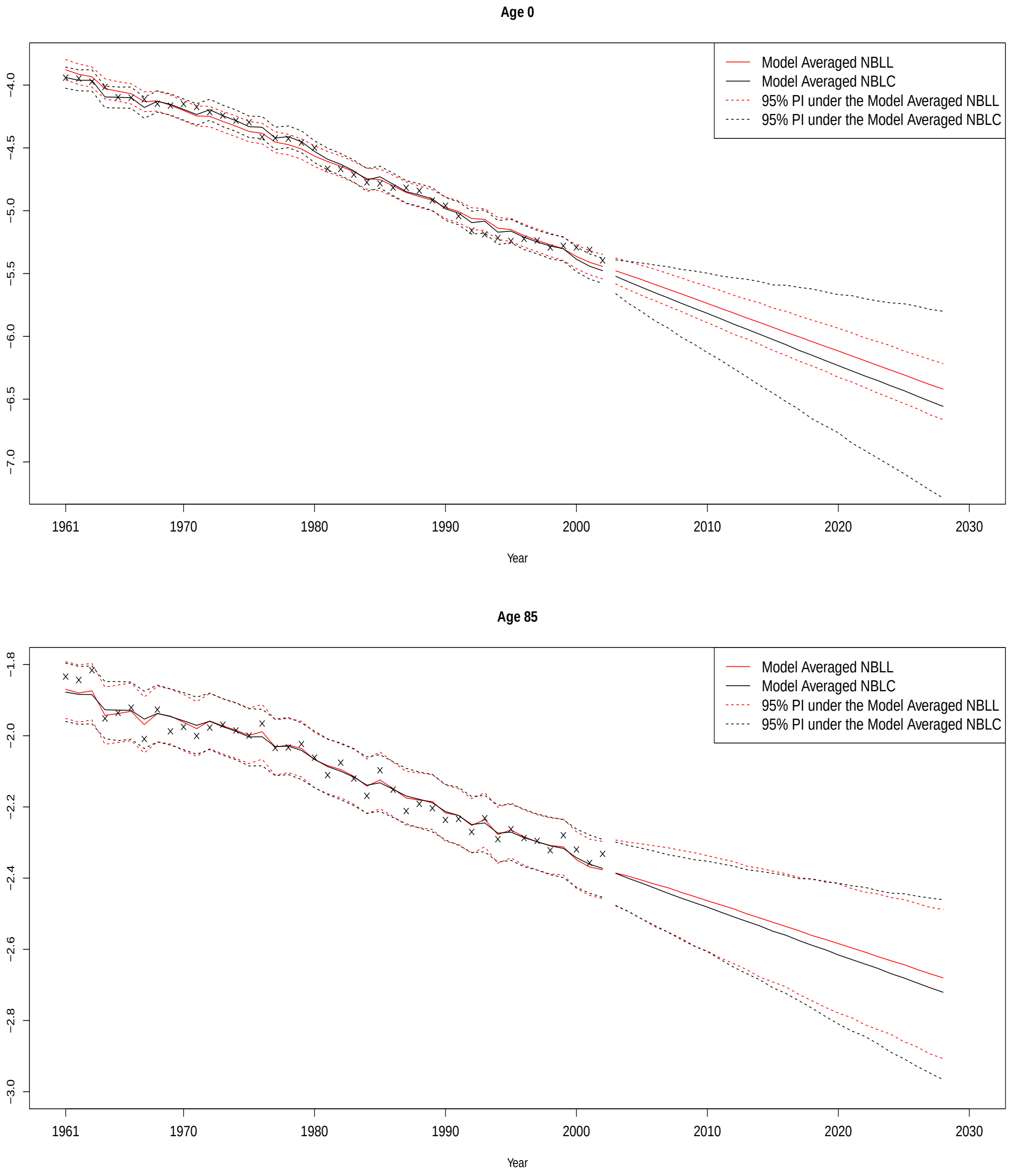


Figure 5.23: Plots of the observed and fitted log crude death rates, as well as their associated 26-years ahead forecast for age 0 (upper panel) and 85 (lower panel) under the NBLL and the NBLC models, accompanied by the 95% prediction intervals.

Albeit some minor differences, both the NBLL and NBLC models produce reasonable prediction intervals that include most of the hold-out life expectancies after performing model averaging (refer to Figure 5.24). Upon closer inspection, the NBLC model performs better in the forecast by having bias of smaller magnitude than the NBLL model, although both models underestimate the gains in life expectancy. Overall, the NBLC model predicts the next 10 years slightly more accurately than the NBLL model, despite fitting the observed data less well, as quantified by the marginal likelihoods.

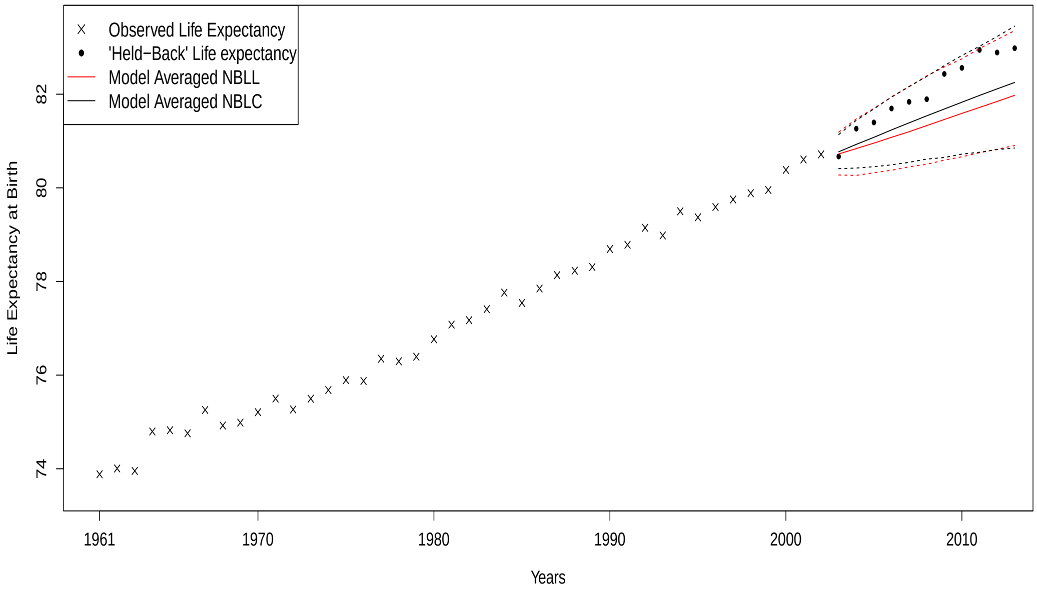


Figure 5.24: Plots of the observed life expectancy at birth and the associated 11-years ahead median forecasts (solid lines) under the model averaged NBLL and NBLC models, accompanied by the 95% prediction intervals (dotted lines).

5.8.3 Model Assessment for the Model Averaged NBLL and NBLC Models

The heat map of squared Pearson residuals as shown in Figure 5.25 looks rather promising for both models, with an exception of the orange diagonal lines (due to the cohort effect). There are approximately 6.4% of the squared Pearson residuals that exceed 3.84 under the NBLL model, and around 5.9% for the NBLC model. However, the posterior predictive p-value of the model averaged NBLL model is extremely small, with value equal to 0.0054, indicating model inadequacy at a 1% significance level (see Figure 5.26). This is also apparent from Figure 5.27, where the cloud of points for the NBLL model lies slightly more to the right, which is rather dissapointing considering that it convincingly outperforms the NBLC model in terms of the marginal likelihood.

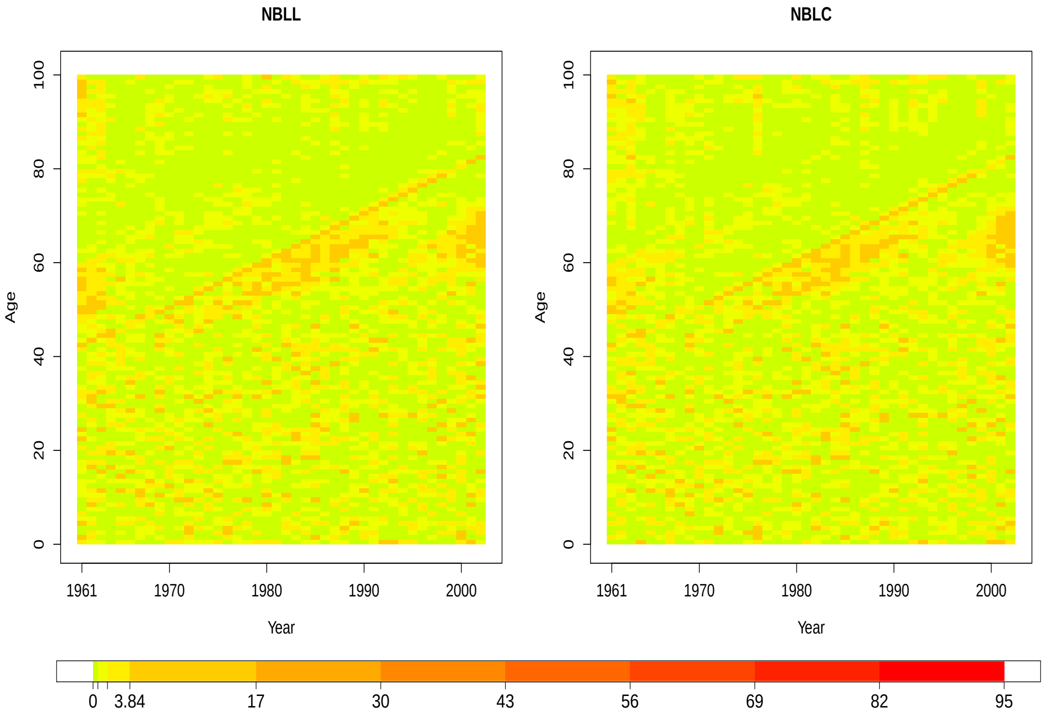


Figure 5.25: Heat map of squared Pearson residuals, r_{xt}^2, under the model averaged NBLL and NBLC models, accompanied by the corresponding colour code.

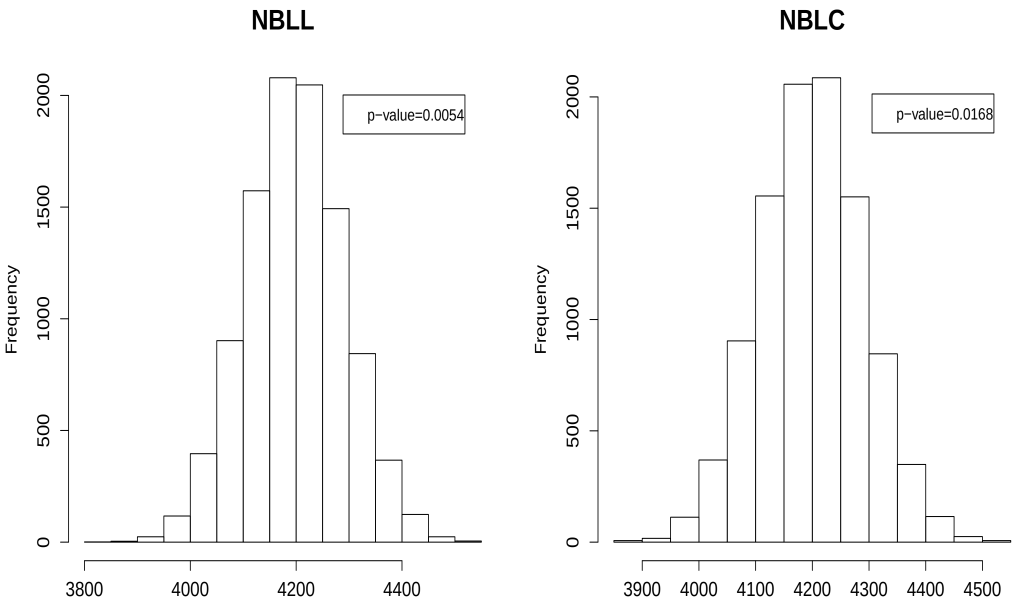


Figure 5.26: Histograms of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$ for the model averaged NBLL and NBLC models, with their corresponding sum of squared Pearson residuals included as solid lines.

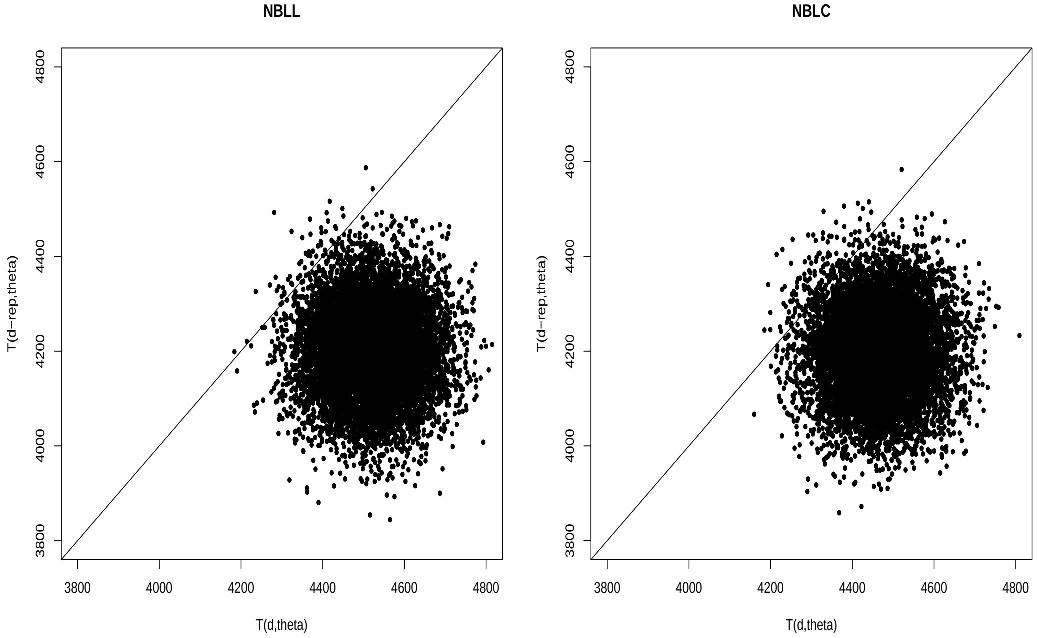


Figure 5.27: Scatter plots of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_M)$ against $T(\boldsymbol{d}, \boldsymbol{\theta}_M)$ for the model averaged NBLL and NBLC models, with solid lines denoting equality.

5.9 Conclusion

In this chapter, we present a potentially simpler model, known as the NBLL model, which is structurally more appealing than the NBLC model by avoiding the multiplicative bilinear term in the rate model. We first consider a naive prior specification for this model. We then proceed to show that this leads to unfair model comparison procedures by inherently favouring the NBLC model through the prior distributions. After some rigorous investigations on the implied prior distributions for $\log \mu_{xt}$, we manage to alleviate the prior differences by establishing parameter correspondence relationships between the models. Specifically, this can be achieved through the use of the Laplace prior distributions. The resulting model comparison based on marginal likelihoods strongly favours the NBLL models. In addition, we also explicitly distinguish between the two time series models through the use of a transformed beta prior to form four competing models, the NBLL-AR1, NBLL-RW, NBLC-AR1 and NBLC-RW models. In general, the random walk models yield forecasts with overly wide prediction intervals, while prediction intervals under the stationary AR(1) models are over-optimistic. Moreover, the NBLC models also project a larger mortality improvement in the future than the NBLL models. Nevertheless, the performances of all these models in the sense of mortality projection are rather underwhelming. Therefore, Bayesian model averaging is applied to combine the two time series models to then compare the overall predictive ability of the NBLL and NBLC models. As a result, the projections produced under both the NBLL and NBLC models are more reasonable, as consistent with the findings in Granger and Newbold (1986, p. 265-287) that weighted averages of time series forecasts are expected to be better than individual forecasts due to the component time series (irrespective of the weights used). The model averaged NBLC model was found to project larger gains in life expectancy than the model averaged NBLL model, which better described the hold-out life expectancies for this particular dataset. Finally, the Bayesian model averaging part relies upon accurate estimation of marginal likelihoods, as the posterior model probabilities are very sensitive to the values of marginal likelihoods. Therefore, methods deduced from the simulation studies in Chapter 4 could possibly be used to improve the estimates.

Chapter 6

Cohort Models

Wadsworth (1991) states that people born in the same time period experience similar health characteristics due to common exposures from the perspective of socio-economic factors, e.g. smoking behaviours, education, social welfares, diets etc. This rather crucial generational effect that has predictive possibilities on mortality is known as the cohort effect. The previous analysis of our mortality data ignored the cohort effect. However, there is clear evidence from previous studies that the cohort effect is an essential feature that is required to be properly accounted for in the UK mortality data. Specifically, Willets (2004) provided a thorough description of the cohort effect in the UK mortality data, from which he concluded that cohorts born between 1925-1945 have experienced more rapid mortality improvement than either side of this period and argued that the relevant effect is unlikely to wear off swiftly with age and time. In fact, this was also indicated by our previous results, particularly from the presence of orange/red diagonal lines in the heat maps constructed (see Figure 3.24 and 5.25). As such, it is anticipated that the cohort effect will impose an enduring impact on mortality improvements in the future decades and, thus, should be suitably incorporated.

Without accounting for the cohort effect, the dispersion parameter is contaminated in the sense that it misidentifies the uncaptured cohort components as random noises in the data, which is inappropriate as cohort is an important trend component of mortality. On the contrary, appropriately accounting for the cohort effect allows for a better calibration of the variations (and hence the overdispersion level) by properly capturing the trend components of mortality. To achieve that, the rate model of the Negative Binomial Log-Linear (NBLL) model is extended by including a cohort parameter, γ_c:

$$\log \mu_{xt} = \alpha_x + \beta_x t + \kappa_t + \gamma_c + \log \nu_{xt},$$

where $c = t - x$ is the cohort index, with a total of $C = A + T - 1 = 141$. For simplicity, we use $c = \{1, 2, \ldots, C\}$ to represent the cohorts born in $\{1861, 1862, \ldots, 2001\}$. This model, known as the NBLL-C model, is invariant to the following transformations:

$$\begin{aligned}
\alpha_x &\mapsto \alpha_x + a_x \\
\beta_x &\mapsto \beta_x + b_x \\
\kappa_t &\mapsto \kappa_t + e_t \text{ with } \sum_t e_t = \sum_t t e_t = 0 \\
\gamma_c &\mapsto \gamma_c - a_x - b_x t - e_t,
\end{aligned}$$

where a_x, b_x and e_t are age/time-specific arbitrary constants. Thus, we adopt the following three constraints, each with a unique impact on the subsequent shape of γ_c,

$$\begin{aligned}
\sum_c \gamma_c = 0 &\quad : \quad \text{removes the mean level of } \gamma_c, \\
\sum_c c\gamma_c = 0 &\quad : \quad \text{restricts its linear growth,} \\
\sum_c c^2\gamma_c = 0 &\quad : \quad \text{restricts its quadratic growth.}
\end{aligned}$$

Hence, this model has a total of five constraints (including $\sum_t \kappa_t = \sum_t t\kappa_t = 0$). Note that this particular model has also been considered by Continuous Mortality Investigation Bureau (2016) within a classical framework very recently, and was referred to as the Age-Period-Cohort Improvement (APCI) model. This is because the model is equivalent to modelling mortality improvements using the well-known Age-Period-Cohort (APC) model. To validate this statement, notice that our model can be rewritten in terms of mortality improvements as

$$\log\left(\frac{\mu_{xt}}{\mu_{x\,t-1}}\right) = \log\mu_{xt} - \log\mu_{x\,t-1} = \beta_x + (\kappa_t - \kappa_{t-1}) + (\gamma_{t-x} - \gamma_{t-x-1}) + \log\nu_{xt} - \log\nu_{x\,t-1},$$

which is essentially the APC model after reparameterization (see also Börger and Aleksic, 2014).

6.1 Prior Distribution for γ_c

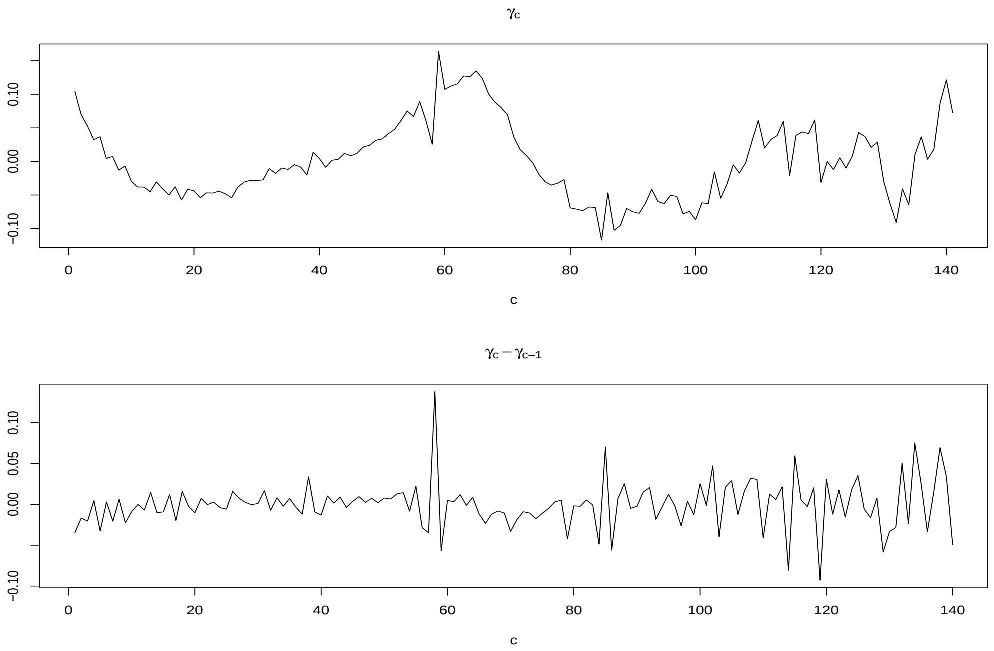


Figure 6.1: Plot of the cohort component, γ_c (top panel), fitted within the frequentist framework and its first difference (bottom panel).

Figure 6.1 illustrates the values of γ_c and its first difference, fitted within the frequentist framework. In essence, choosing an ARIMA model for the prior distribution of γ_c is a model selection problem. Ideally, several candidate ARIMA models should be fitted, and posterior model probabilities then used as a criterion to select an appropriate model. However, we do not pursue this matter here. We merely performed an ad-hoc analysis to determine the time series model for γ_c. In particular, we fitted the model in the frequentist framework, and used BIC as an indicator to search for the best model. All possible combinations of ARIMA (a, b, c), where $0 \leq a, b, c \leq 4$, were considered. The results for the first few models (ordered according to their ranking) are tabulated in Table 6.1.

Table 6.1: BIC of the first few ARIMA models, arranged in increasing order of the BIC values.

Order of the ARIMA Model	BIC
$(1, 1, 0)$	-609.21
$(0, 1, 1)$	-608.95
$(2, 0, 2)$	-606.92
$(2, 0, 0)$	-605.88
$(0, 1, 2)$	-605.59
$(1, 0, 1)$	-605.20

The best model in the sense of BIC is ARIMA$(1,1,0)$, which is consistent with Figure 6.1, where the first difference of γ_c appears to possess constant mean. Hence, an appropriate time series model for γ_c is

$$\begin{cases} (\gamma_c - \gamma_{c-1}) = \rho_\gamma(\gamma_{c-1} - \gamma_{c-2}) + \epsilon_c^\gamma, & \text{for } c = 3, \dots, C, \\ \gamma_2 - \gamma_1 = \frac{1}{\sqrt{1-\rho_\gamma^2}}\epsilon_2^\gamma, \\ \gamma_1 = 100\epsilon_1^\gamma, \end{cases}$$

where ρ_γ is the auto-regressive coefficient and $\epsilon_c^\gamma \overset{\text{ind}}{\sim} N(0, \sigma_\gamma^2)$ are random errors. Now suppose that $\boldsymbol{Q}^\gamma = (\boldsymbol{R}^\gamma)^\top \boldsymbol{R}^\gamma$, where

$$\boldsymbol{R}^\gamma = \begin{pmatrix} 1/100 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -\sqrt{1-\rho_\gamma^2} & \sqrt{1-\rho_\gamma^2} & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \rho_\gamma & -(1+\rho_\gamma) & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & \rho_\gamma & -(1+\rho_\gamma) & 1 & 0 & \cdots & 0 \end{pmatrix}_{C\times C},$$

and $\boldsymbol{B}^\gamma = \boldsymbol{A}^\gamma(\boldsymbol{Q}^\gamma)^{-1}(\boldsymbol{A}^\gamma)^\top$, where

$$\boldsymbol{A}^\gamma = \begin{array}{ll} \begin{array}{l} row1 \\ row2 \\ row3 \\ row4 \\ row5 \\ \vdots \\ row73 \\ row74 \\ \vdots \\ rowC \end{array} & \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ 1 & 2 & 3 & 4 & 5 & \cdots & \cdots & \cdots & \cdots & C \\ 1^2 & 2^2 & 3^2 & 4^2 & 5^2 & \cdots & \cdots & \cdots & \cdots & C^2 \\ 0 & 1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & & \ddots & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}_{C\times C} \end{array}.$$

Applying the constraints $\sum_c \gamma_c = \sum_c c\gamma_c = \sum_c c^2\gamma_c = 0$ on the ARIMA$(1,1,0)$ prior (see Appendix K), the model for γ_c can be equivalently expressed in its multivariate form as

$$\boldsymbol{\gamma}' \sim N_{C-3}(\mathbf{0}, \sigma_\gamma^2 \boldsymbol{D}^\gamma),$$

where $\boldsymbol{\gamma}' = (\gamma_2, \ldots, \gamma_{71}, \gamma_{73}, \ldots, \gamma_{C-1})^\top$, $\boldsymbol{D}^\gamma = [\boldsymbol{B}^\gamma_{22} - \boldsymbol{B}^\gamma_{21}(\boldsymbol{B}^\gamma_{11})^{-1}\boldsymbol{B}^\gamma_{12}]$ and the matrix $\boldsymbol{B}^\gamma$ is partitioned such that

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}^\gamma_{11_{3\times3}} & \boldsymbol{B}^\gamma_{12_{3\times(C-3)}} \\ \boldsymbol{B}^\gamma_{21_{(C-3)\times3}} & \boldsymbol{B}^\gamma_{22_{(C-3)\times(C-3)}} \end{pmatrix}.$$

In this case, the three cohort components removed from the parameter space are γ_1, γ_{72} and γ_C, which can be deterministically computed from the rest as

$$\begin{aligned} \gamma_1 &= \frac{1}{71 \times (C-1)} \sum_{c \neq 1,72,C} (c-72)(C-c)\gamma_c \ , \\ \gamma_{72} &= -\frac{1}{69 \times 71} \sum_{c \neq 1,72,C} (C-c)(c-1)\gamma_c \ , \\ \gamma_C &= \frac{1}{69 \times (C-1)} \sum_{c \neq 1,72,C} (c-1)(72-c)\gamma_c \ . \end{aligned}$$

The reason behind this choice is purely due to computational stability, where if γ_1, γ_2, and γ_3 are removed instead, the matrix $\boldsymbol{D}^\gamma$ is non-positive definite for certain values of ρ^γ, hindering the proper exploration of the posterior distribution by the subsequent MCMC algorithm. Additionally, each of rows $1-3$ of matrix $\boldsymbol{A}^\gamma$ is standardised by subtracting their corresponding row mean and dividing by their row standard deviation for similar purposes (not shown above).

For complete specification of the model for γ_c, the unknown parameters ρ^γ and σ^2_γ are treated as hyperparameters with the following prior distributions:

$$\begin{aligned} \rho^\gamma &\sim N(0, (\sigma^\gamma_\rho)^2), \\ \sigma^{-2}_\gamma &\sim \text{Gamma}(a_\gamma, b_\gamma), \end{aligned}$$

where $(\sigma^\gamma_\rho)^2 = 1$, $a_\gamma = 1000$ and $b_\gamma = 0.001$. Notice that rather than using a vague prior distribution, we adopt an informative prior on the innovation variance, σ^2_γ. This prior distribution has most of its density concentrated around region with very small values, expressing our belief of their small magnitude a priori. This has the effect of smoothing the cohort parameters, which we believe is sensible because cohorts born in nearby periods are expected to experience rather similar generational characteristics, resulting in a similar mortality experience (unless there is a major evolutionary medical breakthrough).

6.2 MH Step for γ'

Due to the constraints imposed, we perform block updating on $\boldsymbol{\gamma}'$ for computational efficiency. The conditional posterior density of $\boldsymbol{\gamma}'$ is given as

$$\begin{aligned} f(\boldsymbol{\gamma}'|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2}\boldsymbol{d},\rho,\sigma_\kappa^2,\lambda,\rho_\gamma,\sigma_\gamma^2,\phi) \quad &\propto \quad \frac{\exp(\sum_{x,t} d_{xt}\gamma_c)}{\prod_{x,t}[e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t+\gamma_c)+\phi]^{d_{xt}+\phi}} \\ &\quad \times \exp\left[-\frac{1}{2\sigma_\gamma^2}(\boldsymbol{\gamma}')^\top(\boldsymbol{D}^\gamma)^{-1}\boldsymbol{\gamma}'\right]. \end{aligned}$$

Using the proposal variance matrix

$$\frac{1.38^2}{137}\times \boldsymbol{G}^\gamma_{\text{NBLL-C}}(\boldsymbol{\delta}^{\text{MLE}}_{\text{NBLL-C}}),$$

where $\boldsymbol{G}^\gamma_{\text{NBLL-C}}(\boldsymbol{\delta}^{\text{MLE}}_{\text{NBLL-C}})$ is the sub-matrix of $[\boldsymbol{H}_{\text{NBLL-C}}]^{-1}$ corresponding to $\boldsymbol{\gamma}'$ evaluated at the MLE of the parameters, $\boldsymbol{\delta}_{\text{NBLL-C}} = (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa}_{-1,2},\boldsymbol{\gamma}',\phi)^\top$, and $\boldsymbol{H}_{\text{NBLL-C}}$ is the Hessian matrix of the likelihood function of the NBLL-C model (estimated through fitting the NBLL-C model in the frequentist framework). Note that the multiplicative constant, $1.38^2/137$, is chosen (and modified from Equation (3.8)) based on pilot runs of 100 iterations to ensure that the acceptance rate is within the recommended benchmark. An acceptance rate of around 0.33 is returned for the random walk MH updating of $\boldsymbol{\gamma}'$ using this proposal variance matrix.

6.2.1 Mortality Forecast under the NBLL-C Model

Mortality projection under the NBLL-C model is simply an extension to that of the NBLL model in Section 5.7. The posterior predictive density of 1-year ahead log mortality rates for each age group under this model can be written as

$$\begin{aligned} f(\log\mu_{x\,T+1}|\boldsymbol{d}) \quad = \quad & \int f(\log\mu_{x\,T+1}|\alpha_x,\beta_x,\kappa_{T+1},\gamma_{T-x+1},\phi)f(\alpha_x,\beta_x,\kappa_{T+1},\gamma_{T-x+1},\phi|\boldsymbol{d}) \\ & \times f(\kappa_{T+1}|\rho,\kappa_T,\sigma_\kappa^2)f(\kappa_T,\rho,\lambda,\sigma_\kappa^2|\boldsymbol{d})f(\gamma_{T-x+1}|\gamma_{T-x},\gamma_{T-x-1},\rho_\gamma,\sigma_\gamma^2) \\ & \times f(\gamma_{T-x},\gamma_{T-x-1},\rho_\gamma,\sigma_\gamma^2|\boldsymbol{d})\mathrm{d}\alpha_x\mathrm{d}\beta_x\mathrm{d}\kappa_T\mathrm{d}\kappa_{T+1}\mathrm{d}\rho\mathrm{d}\lambda\mathrm{d}\sigma_\kappa^2\mathrm{d}\phi, \end{aligned}$$

where $f(\alpha_x,\beta_x,\kappa_{T+1},\gamma_{T-x+1},\phi|\boldsymbol{d})$, $f(\kappa_T,\rho,\lambda,\sigma_\kappa^2|\boldsymbol{d})$ and $f(\gamma_{T-x},\gamma_{T-x-1},\rho_\gamma,\sigma_\gamma^2|\boldsymbol{d})$ are the joint posterior distributions. This suggests that the generation of posterior samples of $\log\mu_{x\,T+1}$ can be carried out as follows:

1. Generate κ_{T+1} from

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

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

2. For $x = 2, \ldots, A$, the fitted values of the cohort parameter from the MCMC output can be used directly. For $x = 1$, generate γ_T by projecting the ARIMA$(1, 1, 0)$ forward,
$$\gamma_T \sim N(\gamma_{T-1} + \rho_\gamma(\gamma_{T-1} - \gamma_{T-2}), \sigma^2_\gamma),$$
where joint posterior samples of $(\gamma_{T-1}, \gamma_{T-2}, \rho_\gamma, \sigma^2_\gamma)$ from the MCMC output are substituted into the expression.

3. Generate $\mu_{x\,T+1}$ from
$$\mu_{x\,T+1} \sim \text{Gamma}\left(\phi, \frac{\phi}{\exp(\alpha_x + \beta_x(T+1) + \kappa_{T+1} + \gamma_{T-x+1})}\right),$$
where κ_{T+1} is from step 1, γ_{T-x+1} is from step 2, and $(\alpha_x, \beta_x, \phi)$ are joint posterior samples from the MCMC output.

The h-years ahead projections can then be obtained by recursive implementation of the above generation procedures, appropriately projecting the cohort components wherever necessary.

6.3 Numerical Results

Again, we explicitly postulate two time series prior distributions for κ_t, where we denote the one with a stationary AR(1) model for κ_t as NBLL-C-AR1, and the one with a random walk model as NBLL-C-RW. The bridge sampling estimate of the marginal likelihoods of both the cohort models are provided in Table 6.2.

Model	Estimated Log Marginal Likelihood
NBLL-C-AR1	-22495.94
NBLL-C-RW	-22498.15

Table 6.2: The marginal likelihoods (on logarithmic scale) of the cohort models approximated using bridge sampling.

The idea of Bayesian model averaging described in Section 5.8.1 is then adopted to combine the two models to form the model averaged NBLL-C model. That is,

$$\begin{aligned} f(\boldsymbol{\theta}_{\text{NBLL-C}}|\boldsymbol{d}) &= f(M = \text{NBLL-C-AR1}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLL-C-AR1}}|\boldsymbol{d}, M = \text{NBLL-C-AR1}) \\ &\quad + f(M = \text{NBLL-C-RW}|\boldsymbol{d}) \times f(\boldsymbol{\theta}_{\text{NBLL-C-RW}}|\boldsymbol{d}, M = \text{NBLL-C-RW}), \end{aligned}$$

where (assuming equal prior model probabilities)

$$f(M = \text{NBLL-C-AR1}|\boldsymbol{d}) = \frac{\exp(-22495.94)}{\exp(-22495.94) + \exp(-22498.15)} \approx 0.9011,$$

and $f(M = \text{NBLL-C-RW}|\boldsymbol{d}) = 1 - f(M = \text{NBLL-C-AR1}|\boldsymbol{d}) \approx 0.0989$. In what follows, the results illustrated are based entirely on the model averaged NBLL-C model.

The fitted values of the rate model parameters and their associated 95% credible intervals under the model averaged NBLL-C model are presented in Figure 6.2 and 6.3, with those from the model averaged NBLL model superimposed as a comparison.

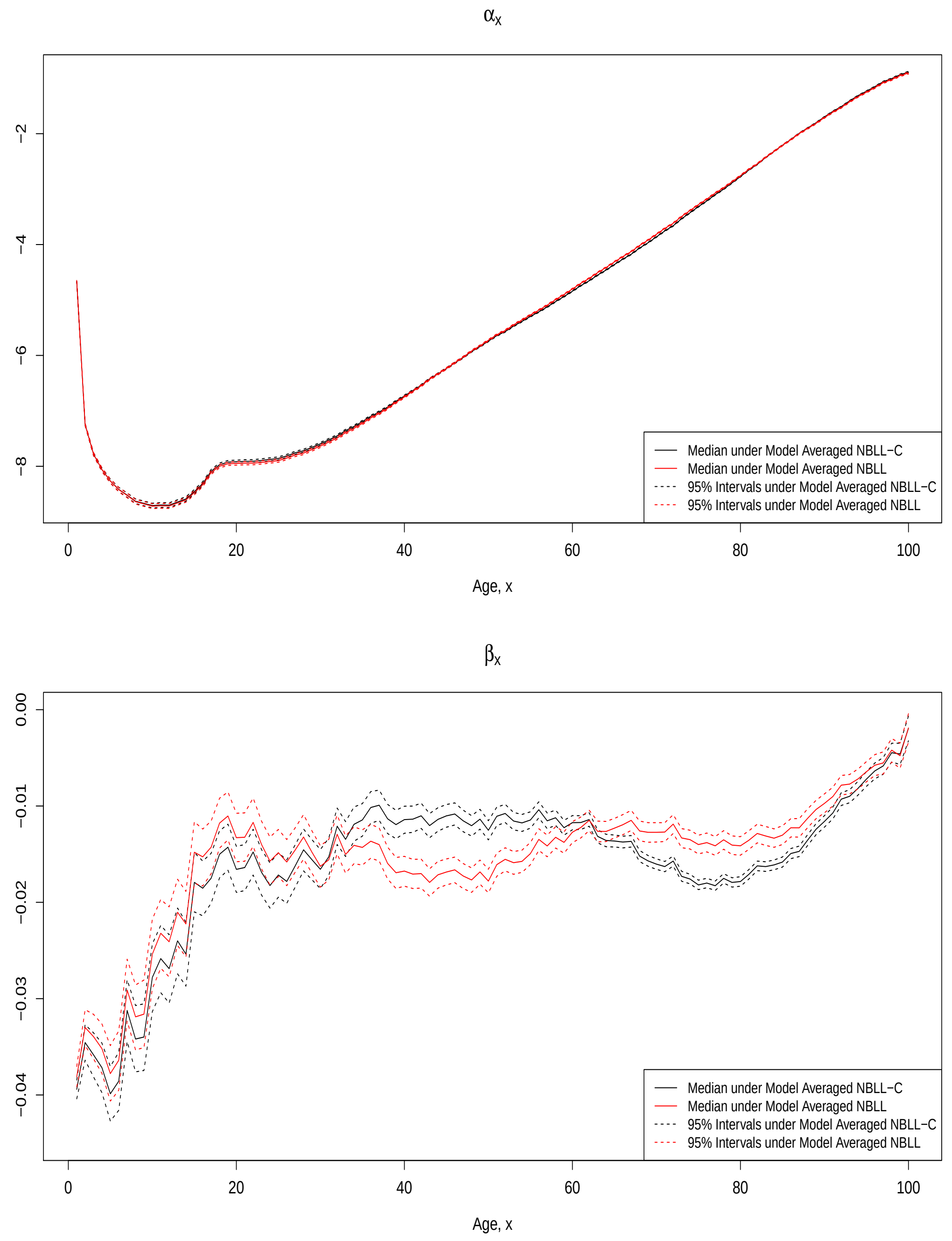


Figure 6.2: Plots of the medians of α_x and β_x, accompanied by their corresponding 95% credible intervals under the NBLL-C and NBLL models.

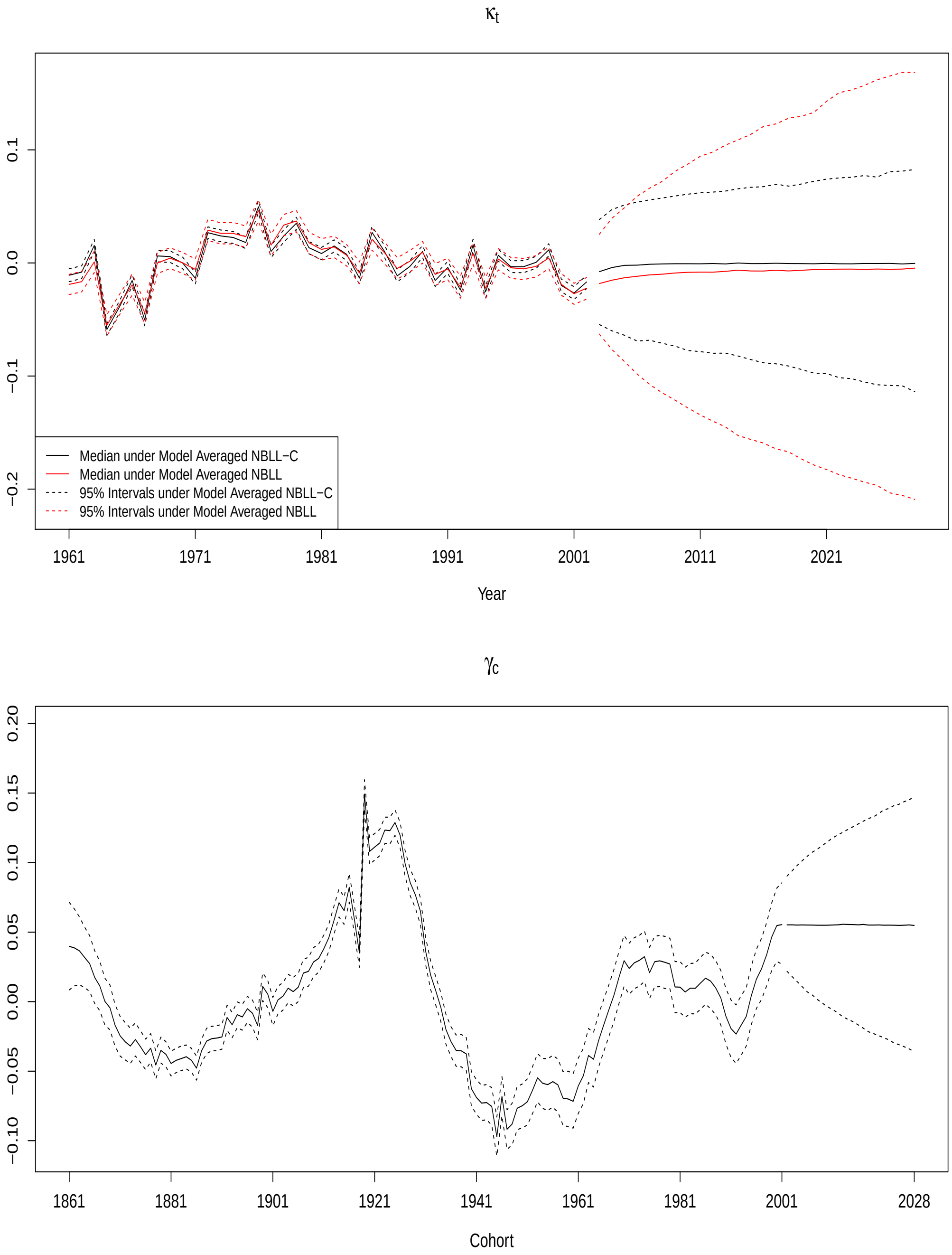


Figure 6.3: Plots of the fitted and projected κ_t as well as γ_c, accompanied by their corresponding 95% credible intervals under the NBLL-C and NBLL models.

The α_x are very similar between the two models, with only slight variations across the ages. The fitted values of β_x under the NBLL-C model exhibit rather distinctive age pattern compared to the NBLL model. This is mainly because β_x is no longer given by Equation (5.2) after the inclusion of cohort components (together with the constraints). For κ_t, the medians are fairly similar between the two models, with the NBLL-C model producing narrower intervals in general (the narrower prediction intervals are because of the smaller mixture proportion for the random walk component of the model). The general shape of the fitted γ_c is rather similar to that fitted within the classical framework, as depicted in Figure 6.1. However, posterior medians of γ_c are much smoother across c because of the smoothing effect imposed by the prior distribution on σ_γ^2. Note the obvious irregularities around the 1919 and 1945 cohort (which remain apparent after being mildly moderated by the smoothing). The first irregularity occurs at 1919, where there is a strong dip, followed by an immediate surge in 1920, while the second irregularity occurs at 1945, where a moderate dip is followed by another sudden surge in 1946. Not surprisingly, they are closely related to three of the well-known events, World War I as well as the 1918 Influenza Pandemic for the first irregularity, and World War II for the second irregularity. These two anomalies have been identified by Cairns et al. (2016) as a consequence of unreliably estimated exposures due to an uneven pattern of births. We do not attempt to address this issue here, but we note the potential impact of these anomalies in the resulting mortality projection, as discussed in Cairns et al. (2016). The projection of γ_c is relatively straightforward, where the median levels off at the value of the most recent cohort, accompanied by widening 95% prediction intervals.

Kernel density estimates of the marginal posterior distributions of the rest of the parameters are illustrated in Figure 6.4.

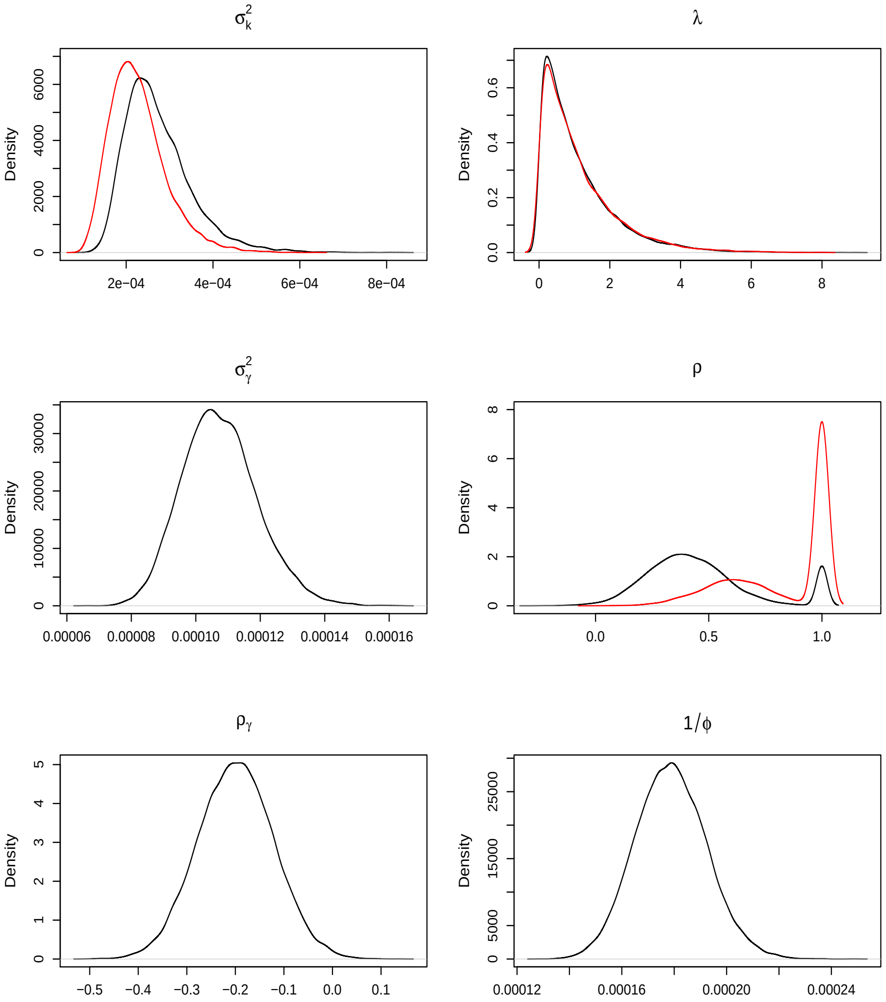


Figure 6.4: Kernel density plots of σ_κ^2, λ, σ_γ^2, ρ, ρ_γ and $1/\phi$ under the NBLL-C model (black solid lines), with those from the NBLL model superimposed for relevant parameters (as red solid lines). Note that the point mass at one for ρ (corresponding to the random walk model) is represented with a density concentrated at around one with the appropriate mixture probability.

The kernel posterior densities of λ are essentially the same under both models. On the other hand, the fitted values of σ_κ^2 are slightly larger for the NBLL-C model. Furthermore, the posterior distribution of ρ under the NBLL-C model is again a mixture distribution, with a smaller weight given to the spike at $\rho = 1$ (corresponding to the random walk model), implying that the stationary AR(1) model on κ_t is favoured instead of a random walk model. The stationary part of the model (i.e. $\rho \neq 1$) also shifts to the left, meaning that its magnitude appears to have been reduced on the inclusion of cohort components. These are the main reasons why the NBLL-C model produces intervals of narrower width in general. Regarding the level of overdispersion, the value of $1/\phi$ under the NBLL-C model is approximately 0.00016, which is 10 times smaller than that of the NBLL model. This is to be expected because without incorporating the cohort parameter, the residuals due to cohort are misidentified by the model as a form of overdispersion. Therefore, the dispersion parameter does not represent heterogeneity entirely under the NBLL model due to the contamination by the cohort effect, as pointed out in Section 3.9. Finally, the posterior density of ρ_γ is centered at -0.2 and is mostly concentrated in the region $[-0.4, 0]$, indicating that a negatively auto-correlated stationary AR(1) model is fitted on the first difference of γ_c.

6.3.1 Fitted and Projected Mortality Rates under the NBLL-C Model

As depicted in Figure 6.5, the medians of the fitted log mortality rates under the NBLL-C model adhere very closely to the observed log rates across the ages, indicating the ability of this model in capturing the underlying mortality trend components. Moreover, the associated 95% credible intervals are also much narrower than the NBLL model, primarily due to the smaller residuals under the NBLL-C model. Hence, the NBLL-C model is able to provide good coverages of the observed log rates across all ages despite having much narrower credible intervals. In essence, this model offers a better calibration of signals and errors present in the mortality data, which then lower the overall magnitude of overdispersion by reducing the residuals due to the unexplained mortality trend (cohort) components. By preventing the model from misidentifying the cohort effect as a form overdispersion, the dispersion parameter (ϕ) fitted now provides a more precise description of the mortality heterogeneity present in our mortality data.

The forecast of log mortality rates also shows major disparities between the NBLL-C and NBLL models, both for the medians and credible intervals (especially for older ages). This is primarily due to the inevitable propagation of the existing cohort effect into the future, and also the fitted stationary time series models for the NBLL-C model. Specifically, the peculiar zigzag patterns for age 65 and 80 correspond to the cohort effect propagating through the projection into the future. Particularly noteworthy are the effects of the 1919 and 1945 cohorts progressing into the future, appearing as obvious anomalies for age 80 and age 65 respectively in Figure 6.5. Moreover, the NBLL-C model appears to yield more sensible median forecasts that better describe the trend components of mortality. In particular, the obvious jump-off discontinuity between the latest log mortality rate and the 1-year ahead projection for age 65 is mitigated by the inclusion of cohort components. By contrast, the NBLL model fails to appropriately describe (and hence project) the mortality trends for certain ages due to the lack of the crucial cohort components, necessitating wider prediction intervals.

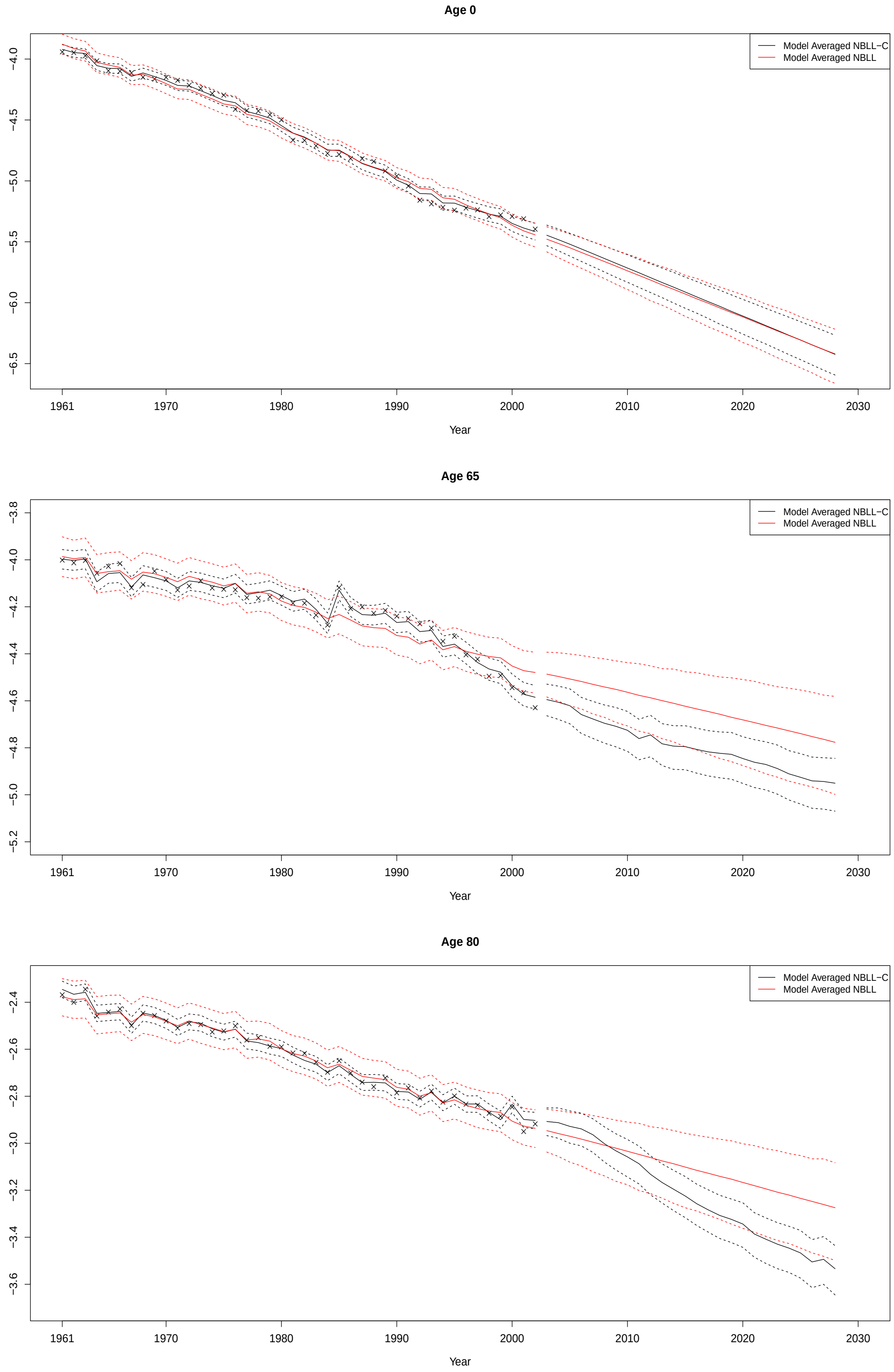


Figure 6.5: Plots of the observed and fitted log crude death rates, as well as the associated 26-years ahead projection of the underlying log mortality rates for age 0, age 65 and age 80 under the NBLL-C and NBLL models, accompanied by the 95% credible intervals.

6.3.2 Out-of-Sample Validation

The 11-years ahead forecast of crude mortality rates under the NBLL-C and NBLL models for ages 0, 60, 65, 75 and 80 are depicted in Figures 6.6 and 6.7. There are substantial differences of the projected crude log mortality rates between the models. The most apparent one being the correction for jump-off discontinuities for ages 60 and 65 (two of which are chosen for exemplification) offered by the NBLL-C model, thereby producing more sensible projections that are more consistent with the holdout data. This is an interesting observation because the issue of jump-off discontinuity appears to have been solved by fitting a model that has a better description of data signals, which in our case, is by introducing cohort trend components. Therefore, the deterministic adjustments suggested by Lee and Miller (2001) to alleviate the jump-off error can be avoided for models that incorporate cohort components when fitting the UK mortality data. Moreover, except for the infant mortality rates, the projections generated under the NBLL-C model yield considerably smaller biases and provide better coverages in general (e.g. ages 75 and 80), despite having much narrower prediction intervals. These highlight the importance of accounting for the propagation of the cohort effect into the future for mortality projections, which reduces biases and avoids unnecessarily wide prediction intervals.

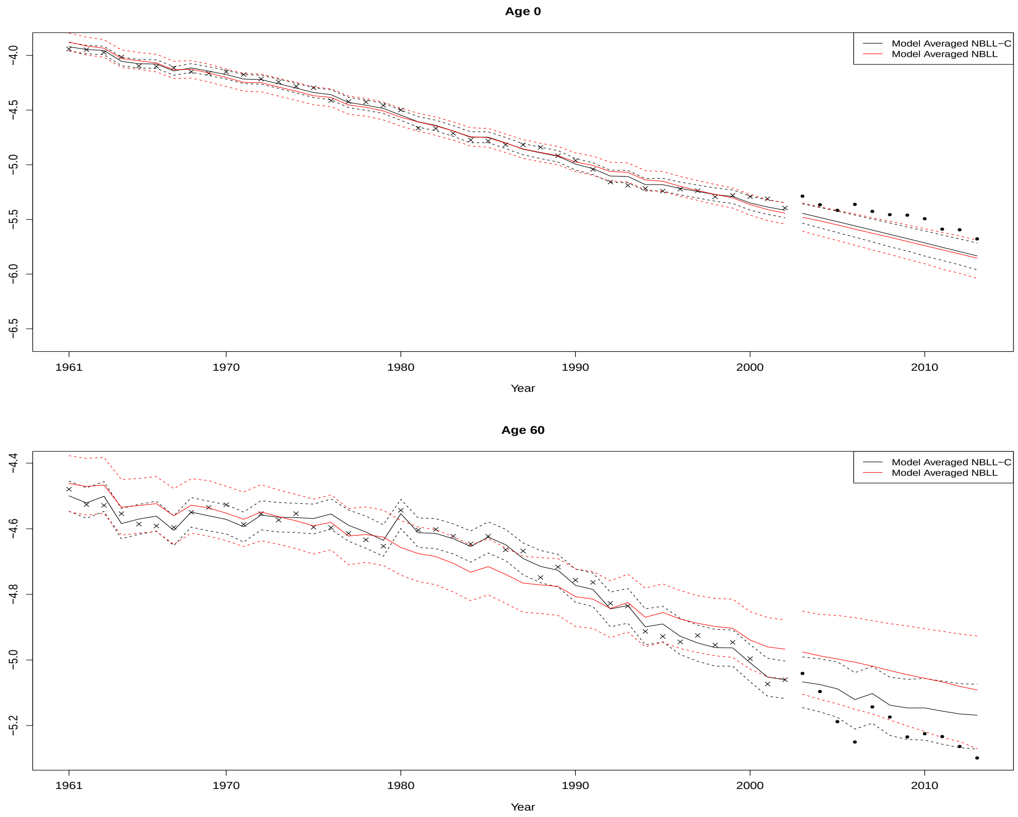


Figure 6.6: Plots of the observed and fitted log crude death rates, and the associated 11-years ahead forecast of crude mortality rates for age 0 and age 60 under the NBLL-C and NBLL models, accompanied by the 95% prediction intervals.

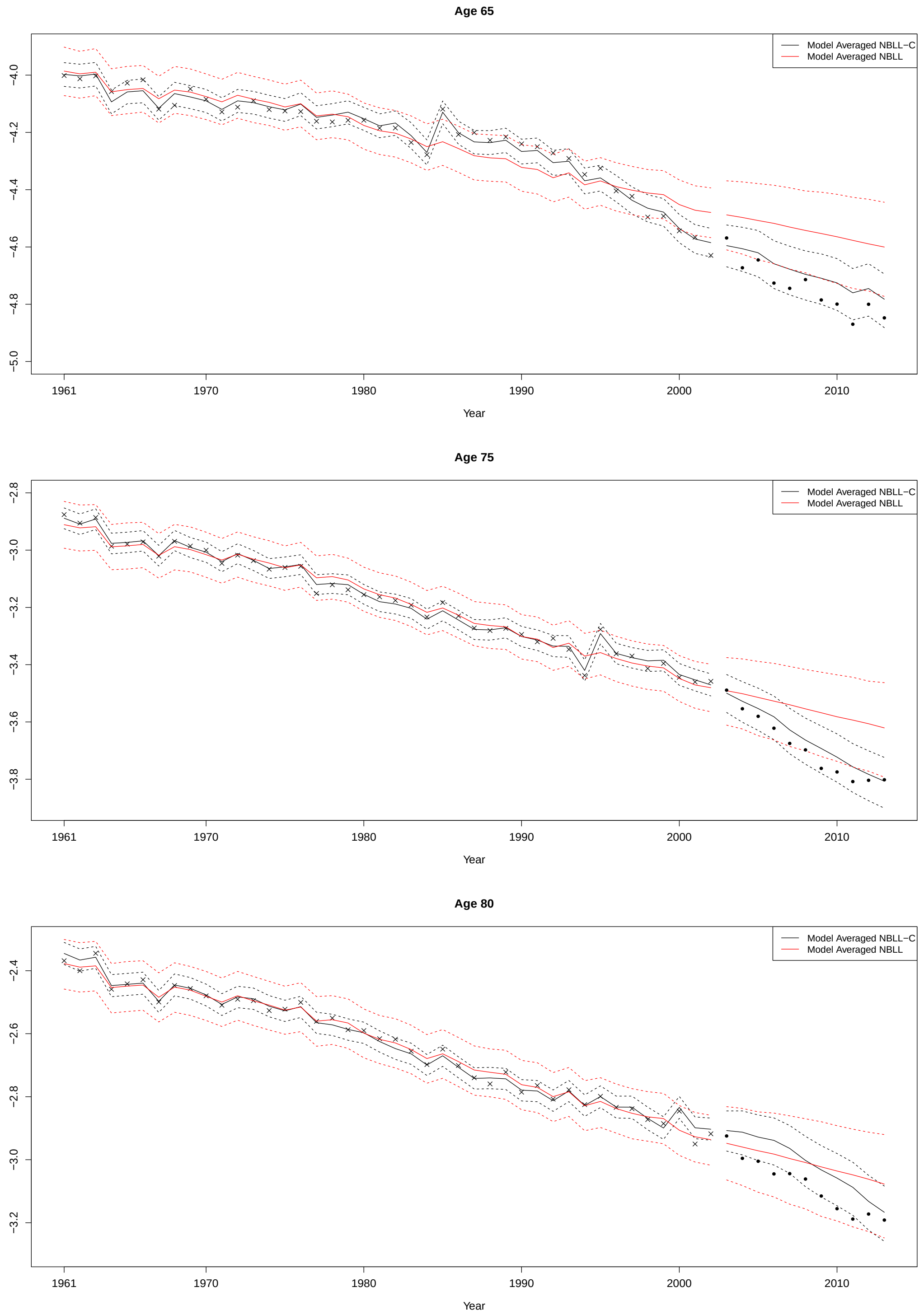


Figure 6.7: Plots of the observed and fitted log crude death rates, as well as the associated 11-years ahead forecast of crude mortality rates for age 65, age 75 and age 80 under the NBLL-C and NBLL models, accompanied by the 95% prediction intervals.

According to Figure 6.8, the NBLL-C model has an increased median forecast of life expectancy with a clear reduction of the underestimation in the life expectancy gains.

Albeit the much narrower prediction intervals, the coverage under the NBLL-C model is still reasonable, including most of the hold-out life expectancies. This is primarily because the NBLL-C model manages to capture the trend components of mortality better than the NBLL model, thus, narrower intervals are sufficient to provide a good coverage; in contrast to the NBLL model, which requires wider intervals due to the miscalibration of signals and errors. Therefore, the NBLL-C model undoubtedly outperforms the NBLL model in terms of the predictive ability by generating smaller biases for the projected life expectancies for this particular dataset.

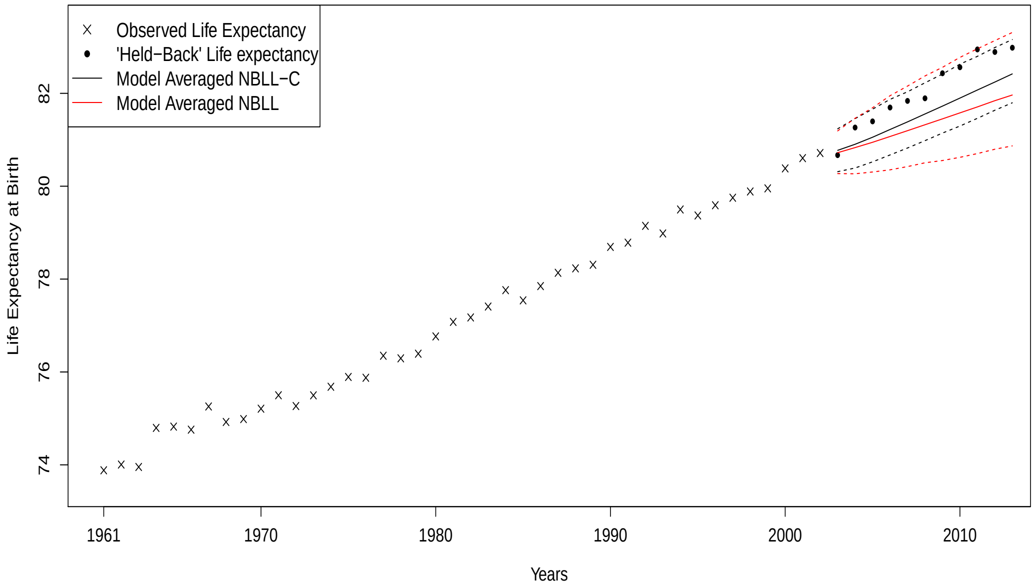


Figure 6.8: Plots of the observed life expectancy at birth and the associated 11-years ahead forecast under the NBLL-C and NBLL models, accompanied by the 95% prediction intervals.

6.3.3 Model Assessment

Remarkably, the diagonal discrepancies corresponding to the cohort effect are completely removed under the NBLL-C model, as illustrated in Figure 6.9. The 1919 cohort also appears to have been correctly modelled, despite being flagged by Cairns et al. (2016) as having unreliable exposures. The amount of rectangular cells having a value of r_{xt}^2 that is greater than the critical value 3.84 is now 233 (5.5%), which is slightly improved relative to the NBLL model (6.4%). An interesting issue here is that the NBLL-C model produces larger residuals at older ages as compared to the NBLL model, in spite of the ability to remove residuals associated with cohort effects.

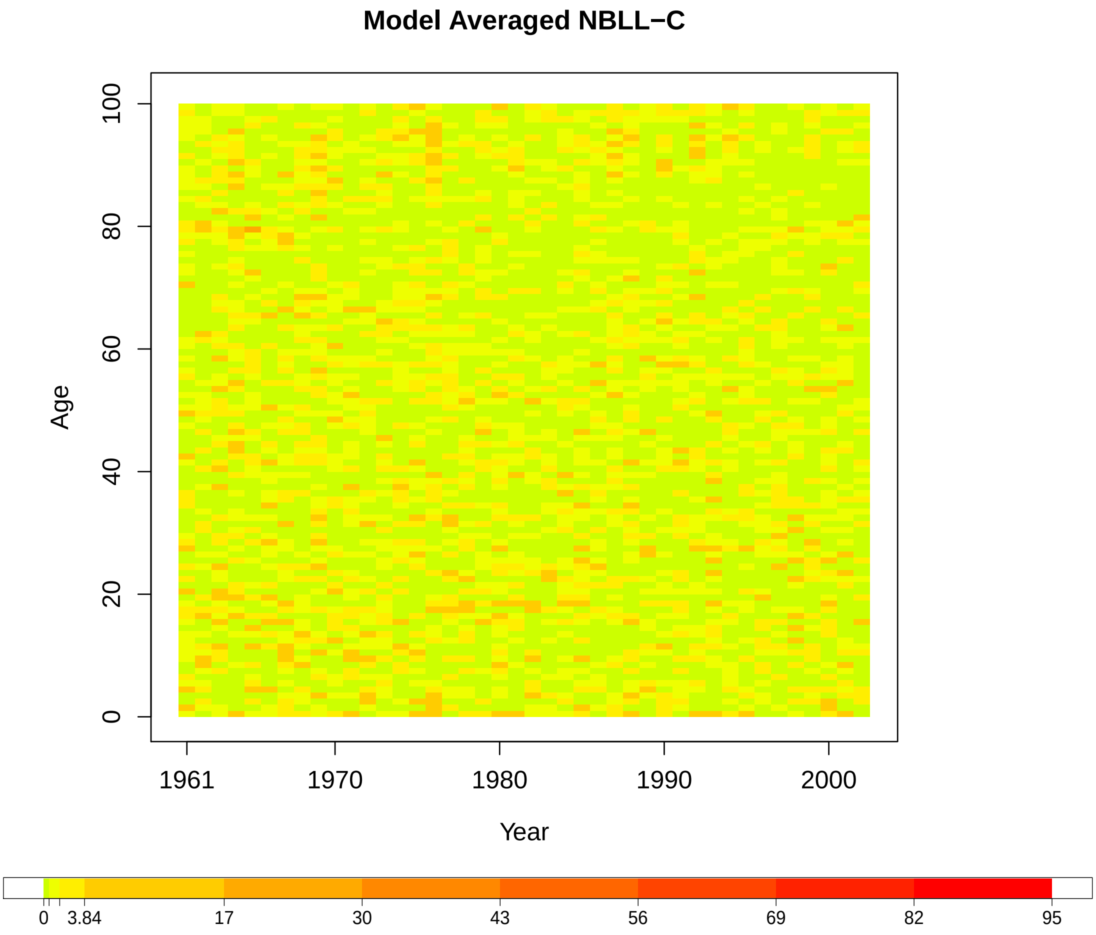


Figure 6.9: Heat map of squared Pearson residuals, r_{xt}^2, under the NBLL-C model.

The improvement in goodness of fit is also indicated by the marginal likelihoods, where the marginal likelihoods of the cohort models are overwhelmingly larger than those of the NBLL and NBLC models. Take the stationary AR(1) models for example, the marginal likelihood of the NBLL-C-AR1 model is $\exp(-22495.94)$, which is substantially larger than both $\exp(-23695.86)$ and $\exp(-23795.45)$ of the NBLL-AR1 and NBLC-AR1 respectively.

Figures 6.10 and 6.11 illustrate the histogram of the test quantity, $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_{\text{NBLL-C}})$ and the scatter plot of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_{\text{NBLL-C}})$ against $T(\boldsymbol{d}, \boldsymbol{\theta}_{\text{NBLL-C}})$. The posterior predictive p-value of the NBLL-C model is extraordinarily small, with a value of 0.00055 (even smaller than the NBLL model), strongly indicating model inadequacy. This is rather contradictory to the rest of the findings as it was suggested by our previous results that the inclusion of cohort components improves the goodness of fit substantially. While this necessitates further investigation, our conjecture is that this is an indication of the potential flaw of the posterior predictive p-value. There are several criticisms of the use of posterior predictive checking which may be related to this issue. Firstly, the posterior predictive p-value is not asymptotically uniform (see for example Robins et al., 2000). Secondly, Bayarri and Berger (2000b) pointed out that it involves double use of the data, where the data is first used to compute the posterior (predictive) distribution, and the tail region of the test statistics is then computed based on the same data. There are also concerns with the use of test statistics that depend heavily on parameters, which can complicate the interpretation of the p-value despite being a theoretically appealing feature of Bayesian model checking (Gelman, 2013).

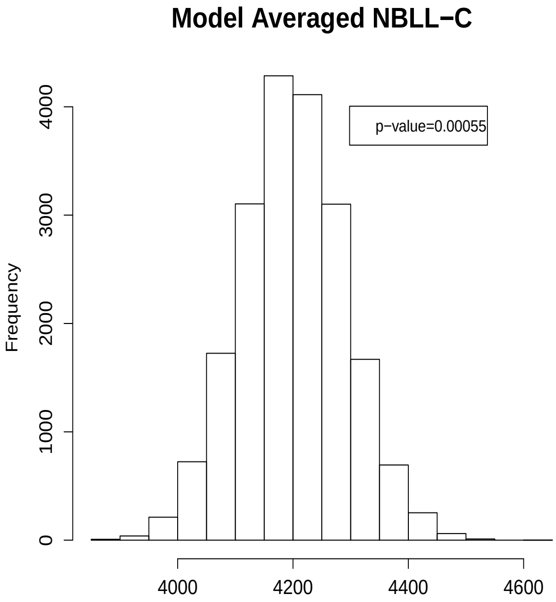


Figure 6.10: Histogram of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_{\text{NBLL-C}})$ for the NBLL-C model, with the value of $T(\boldsymbol{d}^{\text{rep}}, \bar{\boldsymbol{\theta}}_{\text{NBLL-C}})$ included as a vertical solid line.

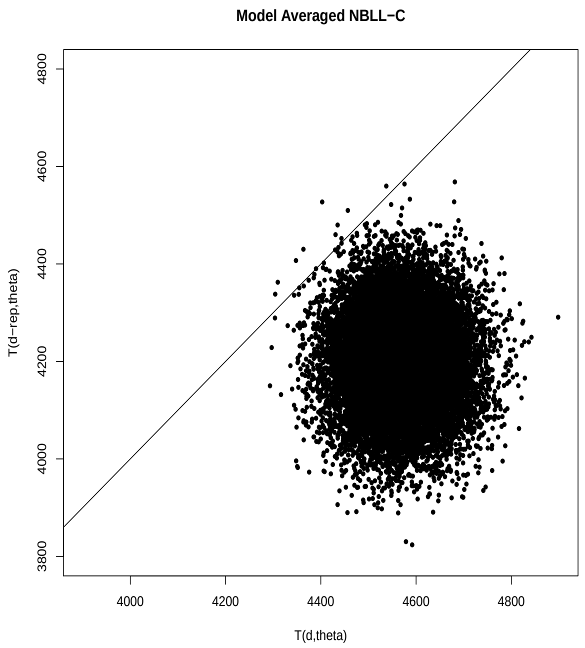


Figure 6.11: Scatter plot of $T(\boldsymbol{d}^{\text{rep}}, \boldsymbol{\theta}_{\text{NBLL-C}})$ against $T(\boldsymbol{d}, \boldsymbol{\theta}_{\text{NBLL-C}})$ for the NBLL-C model, with solid line denoting equality.

6.4 Conclusion

Knowing that the existence of the cohort effect has already been demonstrated in the UK mortality data, we investigate the inclusion of cohort components in this chapter, which is expected to improve the fit substantially. This is undertaken by introducing an extra cohort parameter, γ_c, into the rate model of the NBLL model. We also performed some smoothing on the cohort parameter, γ_c across c, through the prior specification of the innovation variance, σ_γ^2 (as we believe that nearby cohorts are expected to have similar mortality experience). As in the NBLL model, two time series models are fitted to κ_t for the NBLL-C model, and Bayesian model averaging is then used to combine the two cases. The estimated marginal likelihoods suggest that the NBLL-C model vastly outperforms the NBLL model in terms of the goodness of fit. From the perspective of mortality projection, the NBLL-C model describes the mortality trend components better, and hence, produces more sensible projection in overall (for example, the jump-off discontinuity of the projected log mortality rate for age 65 at the forecast origin was mitigated). It also offers a better calibration of the uncertainty bands by not misidentifying the unexplained signals as errors. Specifically, the prediction intervals of log mortality rates and life expectancy produced under the NBLL-C model are considerably narrower than those under the NBLL model (which is also attributable to the fact that the NBLL-C-RW model is less favoured than the NBLL-C-AR1 model), but still gives reasonable coverages of the hold-out data. The improvement in fit is also indicated in the heat map constructed, where the previous discrepancies due to the unexplained cohort effect (orange/red diagonal lines in the previous heat maps) vanished after the inclusion of cohort parameters. Nevertheless, the extreme posterior predictive p-value of the NBLL-C model is rather counter-intuitive (given the significant improvement in fit), which requires further investigation.

Chapter 7

Summary and Future Work

7.1 Summary

A brief summary of our contributions is presented here to provide a short overview of the thesis. The aim of this thesis is to develop a methodology to produce accurate mortality projections with carefully calibrated prediction intervals to characterise the underlying uncertainty associated with the projections. In order to achieve our goals, we prioritised the implementation of Bayesian methodology throughout the thesis. The primary advantage of Bayesian methods is that it allows for the possibility of incorporating various sources of uncertainty (in the form of probability distributions) in natural and coherent manners through the computation of joint posterior distributions. Bayesian mortality modelling also avoids the two-stage model fitting procedure of the original LC approach by directly specifying an ARIMA prior on the time-varying parameter, which can then be easily extrapolated for projection. Prior mortality research knowledge can also be elicited wherever applicable if desired. MCMC methods are then employed to summarise the joint posterior distribution and compute quantities of interest for the purpose of mortality projection.

The second core element that constitutes our contributions is the incorporation of the concept of overdispersion, when a Poisson distribution is specified on the death data. Mortality data often possess extra variabilities relative to their mean due to the presence of heterogeneity among individuals, leading to extra mortality variations that invalidate the assumption of the mean-variance equality of a Poisson distribution. Failure to account for overdispersion typically leads to over-fitting and miscalibration of the uncertainty involved, producing over-optimistic mortality projection due to the neglection of the extra source of variation. Two mixed Poisson models (the PLNLC and NBLC models) are presented in Chapter 3.1, both of which extended the original Poisson Lee-Carter model by introducing a single dispersion parameter. Vague prior distributions were used for illustrative purposes. Our results indicated that the overdispersion models provide a drastic improvement over the Bayesian PLC model in various perspectives (e.g. goodness of fit and out-of-sample validation). Accounting for overdispersion provides a pronounced improvement in the overall performance of the mortality projections, but is still not ideal due to the contamination by the cohort effect, which were addressed in Chapter 6.

Between the two overdispersion models presented, they offer rather similar qualitative fit, as illustrated throughout the result section of Chapter 3.1. In Chapter 4, we carried out formal Bayesian model determination procedures to ascertain their similarity. This was achieved through the computation of Bayes factors and marginal likelihoods, where we proposed to use the bridge sampling estimator to approximate these quantities. However, we experienced major difficulty during the computation of the marginal likelihood of the PLNLC model due to high-dimensionality. Therefore, four simulation studies were conducted to investigate the relevant issues of using bridge sampling and the possibility of improving its efficiency. By adopting the idea of "splitting" and altering the allocation of sample sizes, we managed to obtain an approximation of the marginal likelihood of the PLNLC model with reasonable accuracy. Based on the marginal likelihoods computed, we concluded that the PLNLC and NBLC models are essentially the same in terms of fitting the data. Therefore, the NBLC model is to be recommended over the PLNLC model for computational reasons.

In Chapter 5, we considered a structurally simpler model than Lee-Carter type models, the NBLL model, which postulates a log-linear relationship between mortality rate and time. To compare this model with the NBLC model, precautions need to be exercised in the prior specification because naively specified prior distributions leads to unfair model comparison procedures by inherently favouring the NBLC model. After some rigorous investigations on the implied prior distributions for the log mortality rates, we managed to alleviate the prior differences through the use of Laplace prior distributions. We also explicitly distinguished between the stationary AR(1) and the random walk models for both the NBLL and NBLC models, which were combined later using the idea of Bayesian model averaging (because their individual performances in the sense of mortality projection were underwhelming). The results after performing model averaging are much more reasonable in the sense that the mortality projection under each of the NBLL and NBLC models is now a compromise between the projection due to the stationary AR(1) model (known to be overly narrow) and the random walk model (known to be overly wide). For this particular dataset, the posterior model probability heavily favours the NBLL model, but is outperformed by the NBLC model in terms of the predictive ability.

Finally, incorporation of the cohort effect is presented in Chapter 6. This was undertaken by introducing a cohort parameter into the rate model of the NBLL model, producing the NBLL-C model. The cohort components were also smoothed by imposing an informative prior distribution on the innovation variance, σ_γ^2, based on our prior belief that cohorts born in nearby periods are expected to have similar mortality experience. The NBLL-C model outperformed the NBLL model by a considerable margin, both in terms of goodness of fit (as indicated by the marginal likelihoods) and predictive ability (as indicated by the out-of-sample validations). Fundamentally, this is because the NBLL-C model captures the mortality trend components better, and hence, yields a better calibration of uncertainty bands by not misidentifying the unexplained signals as errors. Additionally, the dispersion parameter ϕ gives a more precise description of the overall level of heterogeneity present in our mortality data after including the cohort effect.

Our research has the potential to lead to meaningful impacts in the actuarial industry. The incorporation of overdispersion in a mortality data has always been of considerable interest to the actuary and even though it has been undertaken within the frequentist framework (see 1.4), has yet to be widely considered within a Bayesian paradigm to the best of our knowledge. Bayesian methodology is deemed important because it offers a framework to coherently integrate over all the sources of uncertainty. Therefore, we believe that the use of Bayesian methods should be promoted in the practical industry (in the presence of appropriate training) especially when inferences are no longer hindered by the computational feasibility of posterior distributions due to the invention of MCMC methods. For instance, the coupling use of Bayesian methods and the decision theory (as discussed before in Section 3.2) for deriving point estimates that account for relevant uncertainties can prove useful for practitioners who are involved in making risk-controlled decision. Rest of the efforts that underlay the thesis then focused on further improving the calibration of uncertainties (by allowing for model uncertainties and the incorporation of cohort effects) in order to ensure the resulting mortality projections are accompanied by prediction intervals that are well representatives of the associated uncertainties. The methodology we developed here will hopefully serve as a platform for other researchers to further build on, modify or simplify to become more user-friendly for implementation within the practical industry (as all the algorithms in this thesis are all hard-coded in R).

7.2 Future Work

Our work has the potential to be extended in various ways. First and foremost, the analyses illustrated in the thesis are mostly undertaken from a statistical perspective. It would be informative to put these analyses to an actuarial context, where we will be able to shed light on the practical significance of our developed methodology in a realistic actuarial application. For instance, the projected mortality rates can be used to derive other annuity functions (or any other actuarial quantities) of interest. The resulting quantities generated under each model then serve as bases for model comparison (e.g. Brouhns et al., 2002; Renshaw and Haberman, 2005).

Secondly, the marginal likelihoods estimated in Chapter 5 and Chapter 6 should be further refined as the model averaging procedures to combine the time series models are rather sensitive to the marginal likelihood (and hence the posterior model probabilities) computed. The results of the simulation studies conducted in Chapter 4 can be used to potentially enhance the accuracies of the marginal likelihood estimates.

Next, we can continue to expand our universe of models which fit mortality data reasonably well. The simplest extension of the presented models is to consider postulating different ARIMA time series models on the time variant parameters, κ_t and γ_c, for forecasting. Bayesian model averaging (applied on the projected mortality rates for example) can be used to average across the ARIMA models. Model uncertainty in light of the time series models is then incorporated in our projections. A similar idea was already adopted in Chapter 5 and 6, where only the ARIMA$(1, 0, 0)$ and ARIMA$(0, 1, 0)$ for κ_t were considered for the purpose of separating the mixture distribution of the autoregressive coefficient, ρ. Given how close their marginal likelihoods were, it is likely that other ARIMA models are good candidates too. Hence, we can easily generalise this by including several possible candidate ARIMA models for κ_t. Additionally, recall that the ARIMA model imposed on the cohort parameter, γ_c, was inspired from an ad-hoc analysis. Judging from the BIC of the different ARIMA models fitted on γ_c, it is highly probable that other ARIMA models provide adequate fit. Therefore, model averaging across these models is informative to generate prediction intervals that account for model uncertainty in this regard.

Other than using different ARIMA models, the models presented by Cairns et al. (2007) and their variants are some plausible candidates too. Moreover, another cohort model, formed by extending the NBLC model to include the cohort effect (which we shall refer to as the NBLC-C model) can be considered. Given that the better predictive ability of the NBLC model over the NBLL model, it will be interesting to examine the relative performance of an improved version of the NBLC model. In principle, performing a Bayesian model averaging across the pool of models (including those presented in this thesis) is ideal because the prediction intervals constructed would then include model uncertainty. However, it is hardly useful because the resulting projection will typically be dominated by a single model (in our case, the cohort model).

Recall that the posterior predictive p-value of the NBLL-C model is unusually small. More investigations should be conducted to examine the behaviour of the posterior predictive p-value in order to justify this phenomenon. One possibility is to consider a different choice of the test quantity, T, for detecting model inadequacy, possibly one with less dependency on the parameters. While the original posterior predictive checking procedure allows any T in theory, Gelman (2013) recommended against the use of T that depends on unknown parameters unless the amount of imputed information contained is minimal. On the other hand, Bayarri and Berger (2000a) proposed the partial posterior predictive p-value, which is essentially a slight modification of the original posterior predictive p-value to avoid double use of the data through by properly conditioning on the relevant variables.

In this thesis, vague prior distributions are specified mostly for illustrative purposes. In practice, it may be useful for the expert knowledge related to mortality from various disciplines to be elicited in the form of probability distributions (this is straightforward in principle through the use of Bayesian methods, but can be challenging to execute). For example, our methodology only provides the possibility to smooth mortality rates across time (using the ARIMA prior), but does not allow for mortality smoothing across ages. Future research in this regard would be interesting to prevent implausible age patterns for long term mortality projections.

Moreover, the UK exposure data are assumed to be fixed (and correct) throughout the thesis. However, the reliability of the UK exposure data has been questioned by Cairns et al. (2016). Hence, the method proposed by Cairns et al. (2016) can be used to adjust the exposure data. It is also worth considering postulating a model for the exposure data. In addition, the issue of performing mortality estimation for countries where mortality data are expected to be sparse (typically developing countries), can effectively be treated as a missing data problem, where methods such as multiple imputation could potentially be used to improve the accuracy of the estimates (see Gelman et al., 1995). Finally, we put little emphasis on mortality rates at oldest ages in this thesis. Issues with inaccurately registered ages of death and relatively small exposures to risk result in highly variable data for the oldest ages. There is plenty of research that suggests treating mortality data at advanced ages separately, and imputing their values by extrapolating certain assumed model (see for example Coale and Kisker, 1990).

Appendix A

Conditional Posterior Distributions for the PLNLC Model

A.1 Conditional Posterior Density of $\log \mu_{xt}$

Denoting $\boldsymbol{\mu}_{-xt}$ as a vector of all the death rates without its $xt - th$ component, i.e.

$$\boldsymbol{\mu}_{-xt} = (\mu_{11}, \mu_{21}, \ldots, \mu_{x-1\,t}, \mu_{x+1\,t}, \ldots, \mu_{AT})^\top,$$

and using the property of the conditional probability, we have

$$\begin{aligned}
& f(\log \mu_{xt} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}_{-xt}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
= \; & \frac{f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2)}{f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}_{-xt}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2)} \text{ (denominator is independent of } \log \mu_{xt}) \\
\propto \; & f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
= \; & f(\boldsymbol{d} | \log \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
& \times f(\log \boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho) f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto \; & f(\boldsymbol{d} | \log \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \times f(\log \boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho).
\end{aligned}$$

Given the log mortality rates $\log \boldsymbol{\mu}$, the number of death, $\boldsymbol{D}$ are mutually independent elementwise. Moreover, given $\log \boldsymbol{\mu}$, the distribution of $\boldsymbol{D}$ does not depend on $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2)$. In a similar fashion, given $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \sigma_\mu^2)$, all of the $\log \mu_{xt}$ are mutually independent and the conditional distribution of $\log \boldsymbol{\mu}$ does not depend on $(\sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho)$ given $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2)$. The conditional posterior of $\log \mu_{xt}$ can then be simplified as follows:

$$\begin{aligned}
& f(\log \mu_{xt}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}_{-xt}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\boldsymbol{d}|\log \boldsymbol{\mu}) \times f(\log \boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) \\
\propto\ & \prod_{x,t} f(d_{xt}|\log \boldsymbol{\mu}) \times \prod_x \prod_t f(\log \mu_{xt}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \sigma_\mu^2) \\
=\ & \prod_{x,t} f(d_{xt}|\log \mu_{xt}) \times \prod_{x,t} f(\log \mu_{xt}|\alpha_x, \beta_x, \kappa_t, \sigma_\mu^2) \\
\propto\ & f(d_{xt}|\log \mu_{xt}) f(\log \mu_{xt}|\alpha_x, \beta_x, \kappa_t, \sigma_\mu^2) \\
\propto\ & \exp(-e_{xt}\mu_{xt})\mu_{xt}^{d_{xt}} \times \exp\left[-\frac{1}{2\sigma_\mu^2}(\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2\right].
\end{aligned}$$

A.2 Conditional Posterior Distribution of κ_t

$$\begin{aligned}
& f(\kappa_t|\boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\kappa_t, \boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log \boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) f(\boldsymbol{\kappa}_{-1}|\boldsymbol{\psi}, \rho, \sigma_\kappa^2).
\end{aligned}$$

Note that the vectors $\log \boldsymbol{\mu}_t$ are independent of each other given the set $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2)$. Therefore,

$$\begin{aligned}
& f(\kappa_t|\boldsymbol{\kappa}_{-1,t}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{D}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log \boldsymbol{\mu}_1, \ldots, \log \boldsymbol{\mu}_T|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) f(\kappa_2, \ldots, \kappa_T|\boldsymbol{\psi}, \rho, \sigma_\kappa^2) \\
=\ & f(\log \boldsymbol{\mu}_1|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) \times \ldots \times f(\log \boldsymbol{\mu}_T|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) \\
& \times f(\kappa_T|\kappa_{T-1}, \boldsymbol{\psi}, \rho, \sigma_\kappa^2) \times \ldots \times f(\kappa_2, \boldsymbol{\psi}, \rho, \sigma_\kappa^2) \\
\propto\ & f(\log \boldsymbol{\mu}_2|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \kappa_2, \sigma_\mu^2) \times \ldots \times f(\log \boldsymbol{\mu}_T|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \kappa_T, \sigma_\mu^2) \\
\propto\ & \prod_{s=2}^{T} f(\log \boldsymbol{\mu}_s|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \kappa_s, \sigma_\mu^2) f(\kappa_s|\kappa_{s-1}, \boldsymbol{\psi}, \rho, \sigma_\kappa^2).
\end{aligned}$$

There are now two cases to be considered:

i. $t = T$,

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

$$\begin{aligned}
&= \exp\left[-\frac{1}{2\sigma_\mu^2}\sum_x(\beta_x^2\kappa_t^2 - 2\beta_x(\log\mu_{xt}-\alpha_x)\kappa_t + (\log\mu_{xt}-\alpha_x)^2)\right] \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}\left[\kappa_t^2 - 2\kappa_t(\eta_t+\rho(\kappa t-1-\eta_{t-1})) + (\eta_t+\rho(\kappa_{t-1}-\eta_{t-1}))^2\right]\right\} \\
&\propto \exp\left[-\frac{1}{2\sigma_\mu^2}\left(\kappa_t^2\sum_x\beta_x^2 - 2\sum_x\beta_x(\log\mu_{xt}-\alpha_x)\kappa_t\right)\right] \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}\left[\kappa_t^2 - 2\kappa_t(\eta_t+\rho(\kappa t-1-\eta_{t-1}))\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\left(\frac{\sum_x\beta_x^2}{\sigma_\kappa^2}+\frac{1}{\sigma_\kappa^2}\right)\kappa_t^2 - 2\kappa_t\left(\frac{\sum_x\beta_x(\log\mu_{xt}-\alpha_x)}{\sigma_\mu^2}+\frac{\eta_t+\rho(\kappa_{t-1}-\eta_{t-1})}{\sigma_\kappa^2}\right)\right]\right\} \\
&\propto N(\mu'_\kappa, (\sigma'_\kappa)^2),
\end{aligned}$$

where

$$\begin{aligned}
(\sigma'_\kappa)^2 &= \left[\frac{\sum_x\beta_x^2}{\sigma_\mu^2}+\frac{1}{\sigma_\kappa^2}\right]^{-1}, \\
\mu'_\kappa &= (\sigma'_\kappa)^2 \times \left[\frac{\sum_x\beta_x(\log\mu_{xt}-\alpha_x)}{\sigma_\mu^2}+\frac{\eta_t+\rho(\kappa_{t-1}-\eta_{t-1})}{\sigma_\kappa^2}\right].
\end{aligned}$$

ii. $1 < t < T$,

$$\begin{aligned}
&f(\kappa_t|\boldsymbol{\kappa}_{-1,t},\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{D},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\rho,\sigma_\mu^2) \\
&\propto f(\log\boldsymbol{\mu}_t|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\kappa_t,\sigma_\mu^2)f(\kappa_t|\kappa_{t-1},\boldsymbol{\psi},\rho,\sigma_\kappa^2)f(\kappa_{t+1}|\kappa_t,\boldsymbol{\psi},\rho,\sigma_\kappa^2) \\
&\propto \left\{\prod_x\exp\left[-\frac{1}{2\sigma_\mu^2}(\log\mu_{xt}-\alpha_x-\beta_x\kappa_t)^2\right]\right\} \\
&\quad \times \exp\left[-\frac{1}{2\sigma_\kappa^2}(\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1}))^2\right] \times \exp\left[-\frac{1}{2\sigma_\kappa^2}(\kappa_{t+1}-\eta_{t+1}-\rho(\kappa_t-\eta_t))^2\right]
\end{aligned}$$

$$\begin{aligned}
&\propto \exp\left\{-\frac{1}{2\sigma_\mu^2}\left[\kappa_t^2\sum_x\beta_x^2 - 2\kappa_t\sum_x\beta_x(\log\mu_{xt}-\alpha_x)\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}\left[\kappa_t^2 - 2\kappa_t(\eta_t+\rho(\kappa_{t-1}-\eta_{t-1}))\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}\left[\rho^2\kappa_t^2 - 2\rho\kappa_t(\kappa_{t+1}-\eta_{t+1}+\rho\eta_t)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\kappa_t^2\left(\frac{\sum_x\beta_x^2}{\sigma_\mu^2}+\frac{1+\rho^2}{\sigma_\kappa^2}\right) - 2\kappa_t\left(\frac{\sum_x\beta_x(\log\mu_{xt}-\alpha_x)}{\sigma_\mu^2}\right.\right.\right. \\
&\quad \left.\left.\left. + \frac{\eta_t+\rho(\kappa_{t-1}-\eta_{t-1})+\rho(\kappa_{t+1}-\eta_{t+1}+\rho\eta_t)}{\sigma_\kappa^2}\right)\right]\right\} \\
&\propto N(\mu^*_\kappa, (\sigma^*_\kappa)^2),
\end{aligned}$$

where

$$\begin{aligned}
(\sigma_\kappa^*)^2 &= \left[\frac{\sum_x \beta_x^2}{\sigma_\mu^2} + \frac{1+\rho^2}{\sigma_\kappa^2}\right]^{-1}, \\
\mu_\kappa^* &= (\sigma_\kappa^*)^2 \times \left[\frac{\sum_x \beta_x(\log\mu_{xt} - \alpha_x)}{\sigma_\mu^2} + \frac{\eta_t + \rho(\kappa_{t-1} - \eta_{t-1}) + \rho(\kappa_{t+1} - \eta_{t+1} + \rho\eta_t)}{\sigma_\kappa^2}\right].
\end{aligned}$$

A.3 Conditional Posterior Distribution of β_x

Denoting $\boldsymbol{\mu}_x = (\mu_{x1}, \mu_{x2}, \ldots, \mu_{xT})^\top$ as the mortality rates corresponding to age x, we have for $2 \le x \le A$,

$$\begin{aligned}
& f(\beta_x|\boldsymbol{\beta}_{-1,x}, \sum_x \beta_x = 1, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\boldsymbol{\beta}_{-1}, \sum_x \beta_x = 1, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
=\ & f(\boldsymbol{d}|\log\boldsymbol{\mu}) f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \sum_x \beta_x = 1, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\beta}_{-1}, \sum_x \beta_x = 1|\sigma_\beta^2) \\
& \times f(\boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \sum_x \beta_x = 1, \boldsymbol{\kappa}, \sigma_\mu^2) f(\sum_x \beta_x = 1|\boldsymbol{\beta}_{-1}, \sigma_\beta^2) f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
\propto\ & f(\log\boldsymbol{\mu}_x|\alpha_x, \beta_x, \boldsymbol{\kappa}, \sigma_\mu^2) f(\log\boldsymbol{\mu}_1|\alpha_1, \boldsymbol{\beta}_{-1}, \sum_x \beta_x = 1, \boldsymbol{\kappa}, \sigma_\mu^2) \\
& \times f(\beta_1 = 1 - \beta_2 - \ldots - \beta_A|\boldsymbol{\beta}_{-1}, \sigma_\beta^2) f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
\propto\ & \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\log\mu_{xt} - \alpha_x - \beta_x\kappa_t]^2\right\} \times \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\log\mu_{1t} - \alpha_1 - \kappa_t(1 - \beta_2 - \ldots - \beta_A)]^2\right\} \\
& \times \exp\left[-\frac{1}{2\sigma_\beta^2}(1 - \beta_2 - \ldots - \beta_A)^2\right] \exp\left[-\frac{1}{2\sigma_\beta^2}\beta_x^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[-\beta_x\kappa_t\log\mu_{xt}-\alpha_x]^2\right\} \times \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\log\mu_{1t}-\alpha_1-\kappa_t(1-\sum_{i\neq 1,x}\beta_i)]^2\right\} \\
&\quad \times \exp\left[-\frac{1}{2\sigma_\beta^2}(-\beta_x+1-\sum_{i\neq 1,x}\beta_i)^2\right]\exp\left[-\frac{1}{2\sigma_\beta^2}\beta_x^2\right] \\
&\propto \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\beta_x^2\kappa_t^2-2\beta_x\kappa_t(\log\mu_{xt}-\alpha_x)]\right\} \\
&\quad \times \prod_t \exp\{-\frac{1}{2\sigma_\mu^2}[\beta_x^2\kappa_t^2-2\beta_x\kappa_t(-\log\mu_{1t}+\alpha_1+(1-\sum_{i\neq 1,x}\beta_i)\kappa_t)]\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\beta^2}[\beta_x^2-2\beta_x(1-\sum_{i\neq 1,x}\beta_i)]\right\}\exp\left[-\frac{1}{2\sigma_\beta^2}\beta_x^2\right] \\
&= \exp\left\{-\frac{1}{2\sigma_\mu^2}\left[\beta_x^2\sum_t\kappa_t^2-2\beta_x\sum_t\kappa_t(\log\mu_{xt}-\alpha_x)\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\mu^2}\left[\beta_x^2\sum_t\kappa_t^2-2\beta_x\sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+(1-\sum_{i\neq 1,x}\beta_i)\kappa_t)\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma_\beta^2}[\beta_x^2-2\beta_x(1-\sum_{i\neq 1,x}\beta_i)]\right\}\times\exp\left[-\frac{1}{2\sigma_\beta^2}\beta_x^2\right] \\
&= \exp\left\{-\frac{1}{2}\left[\beta_x^2\left(\frac{2\sum_t\kappa_t^2}{\sigma_\mu^2}+\frac{2}{\sigma_\beta^2}\right)-2\beta_x\left(\frac{\sum_t\kappa_t(\log\mu_{xt}-\alpha_x)}{\sigma_\mu^2}\right.\right.\right. \\
&\quad \left.\left.\left.+\frac{\sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t(1-\sum_{i\neq 1,x}\beta_i))}{\sigma_\mu^2}+\frac{(1-\sum_{i\neq 1,x}\beta_i)}{\sigma_\beta^2}\right)\right]\right\} \\
&\propto N(\mu_\beta^*,(\sigma_\beta^*)^2),
\end{aligned}
$$

where

$$
\begin{aligned}
(\sigma_\beta^*)^2 &= \left[\frac{2\sum_t\kappa_t^2}{\sigma_\mu^2}+\frac{2}{\sigma_\beta^2}\right]^{-1}, \\
\mu_\beta^* &= (\sigma_\beta^*)^2\times\left[\frac{\sum_t\kappa_t(\log\mu_{xt}-\alpha_x)}{\sigma_\mu^2}+\frac{\sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t(1-\sum_{i\neq 1,x}\beta_i))}{\sigma_\mu^2}\right. \\
&\quad \left.+\frac{(1-\sum_{i\neq 1,x}\beta_i)}{\sigma_\beta^2}\right].
\end{aligned}
$$

Therefore,

$$
\beta_x|\boldsymbol{\beta}_{-1,x},\sum_x\beta_x=1,\boldsymbol{\kappa}_{-1},\boldsymbol{\alpha},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\rho,\sigma_\mu^2\sim N(\mu_\beta^*,(\sigma_\beta^*)^2).
$$

A.4 Conditional Posterior Distribution of α_x

Using the prior distribution $\alpha_x \overset{\text{ind}}{\sim} N(\alpha_0, \sigma_\alpha^2)$, then

$$\begin{aligned}
& f(\alpha_x|\boldsymbol{\alpha}_{-x}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
=\ & f(\boldsymbol{d}|\log\boldsymbol{\mu}) f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}, \sigma_\mu^2) f(\alpha_x) f(\boldsymbol{\alpha}_{-x}) f(\boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log\boldsymbol{\mu}_x|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\alpha_x) \\
\propto\ & \prod_t \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\log\mu_{xt} - \alpha_x - \beta_x\kappa_t)^2\right]\right\} \exp\left[-\frac{1}{2\sigma_\alpha^2}(\alpha_x - \alpha_0)^2\right] \\
\propto\ & \prod_t \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\alpha_x^2 - 2\alpha_x\log\mu_{xt} + 2\alpha_x\beta_x\kappa_t)\right]\right\} \exp\left[-\frac{1}{2\sigma_\alpha^2}(\alpha_x^2 - 2\alpha_0\alpha_x)\right] \\
=\ & \exp\left\{-\frac{1}{2}\left[\alpha_x^2\left(\frac{T}{\sigma_\mu^2} + \frac{1}{\sigma_\alpha^2}\right) - 2\alpha_x\left(\frac{\sum_t \log\mu_{xt} - \beta_x\sum_t\kappa_t}{\sigma_\mu^2} + \frac{\alpha_0}{\sigma_\alpha^2}\right)\right]\right\} \\
\propto\ & N(\alpha_x^*, (\sigma_\alpha^*)^2),
\end{aligned}$$

where

$$\begin{aligned}
(\sigma_\alpha^*)^2 &= \left[\frac{T}{\sigma_\mu^2} + \frac{1}{\sigma_\alpha^2}\right]^{-1}, \\
\alpha_x^* &= (\sigma_\alpha^*)^2 \times \left(\frac{\sum_t \log\mu_{xt} - \beta_x\sum_t\kappa_t}{\sigma_\mu^2} + \frac{\alpha_0}{\sigma_\alpha^2}\right) \\
&= \frac{\sigma_\alpha^2(\sum_t \log\mu_{xt} - \beta_x\sum_t\kappa_t) + \sigma_\mu^2\alpha_0}{\sigma_\mu^2 + T\sigma_\alpha^2}.
\end{aligned}$$

A.5 Conditional Posterior Distribution of σ_μ^2

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

Therefore,

$$
\begin{aligned}
& \sigma_\mu^{-2} | \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho \\
\sim \; & \text{Gamma} \left(a_\mu + \frac{AT}{2}, b_\mu + \frac{1}{2} \sum_{x,t} (\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2 \right).
\end{aligned}
$$

A.6 Conditional Posterior Distribution of σ_κ^2

$$
\begin{aligned}
& f(\sigma_\kappa^2 | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto \; & f(\sigma_\kappa^2, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
= \; & f(\boldsymbol{d} | \log \boldsymbol{\mu}) f(\log \boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\kappa}_{-1} | \rho, \sigma_\kappa^2, \boldsymbol{\psi}) f(\sigma_\kappa^2) f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto \; & f(\boldsymbol{\kappa}_{-1} | \rho, \sigma_\kappa^2, \boldsymbol{\psi}) f(\sigma_\kappa^2) \\
= \; & \prod_{t=2}^{T} [f(\kappa_t | \kappa_{t-1}, \rho, \sigma_\kappa^2, \boldsymbol{\psi})] f(\sigma_\kappa^2) \\
= \; & \prod_{t=2}^{T} \left\{ \frac{1}{\sqrt{2\pi\sigma_\kappa^2}} \exp\left[-\frac{1}{2\sigma_\kappa^2} (\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1}))^2 \right] \right\} \times \frac{b_\kappa^{a_\kappa}}{\psi(a_\kappa)} (\sigma_\kappa^2)^{-a_\kappa - 1} \exp\left(-\frac{b_\kappa}{\sigma_\kappa^2} \right) \\
\propto \; & (\sigma_\kappa^2)^{-(a_\kappa + \frac{T-1}{2}) - 1} \exp\left\{ -\frac{1}{\sigma_\kappa^2} \left[b_\kappa + \frac{1}{2} \sum_{2}^{T} (\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1}))^2 \right] \right\} \\
\propto \; & \text{Inverse Gamma} \left(a_\kappa + \frac{T-1}{2}, b_\kappa + \frac{1}{2} \sum_{2}^{T} (\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1}))^2 \right).
\end{aligned}
$$

Therefore,

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

A.7 Conditional Posterior Distribution of σ_β^2

$$\begin{aligned}
& f(\sigma_\beta^2|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\sigma_\beta^2, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
=\ & f(\boldsymbol{d}|\log\boldsymbol{\mu}) f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) f(\sigma_\beta^2) f(\boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \rho, \boldsymbol{\psi}, \sigma_\mu^2) \\
\propto\ & f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) f(\sigma_\beta^2).
\end{aligned}$$

The prior distribution of $\boldsymbol{\beta}_{-1}$ is given by

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

where in terms of density function,

$$\begin{aligned}
& f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
=\ & \frac{1}{\sqrt{(2\pi)^{A-1}\left|\sigma_\beta^2\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)\right|}} \\
& \times \exp\left\{-\frac{1}{2}[\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}]^\top [\sigma_\beta^2(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1})]^{-1}[\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}]\right\} \\
\propto\ & (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top \left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right\}.
\end{aligned}$$

Hence, the conditional posterior distribution of σ_β^2 can be written as

$$\begin{aligned}
& f(\sigma_\beta^2|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \rho, \boldsymbol{\psi}, \sigma_\mu^2) \\
\propto\ & (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top \left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right\} \\
& \times \frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)}(\sigma_\beta^2)^{-a_\beta - 1} \exp\left(-\frac{b_\beta}{\sigma_\beta^2}\right) \\
\propto\ & (\sigma_\beta^2)^{-(a_\beta + \frac{A-1}{2}) - 1} \exp\left\{-\frac{1}{\sigma_\beta^2}\left[b_\beta + \frac{1}{2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top \left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right]\right\} \\
\propto\ & \text{Inverse Gamma}\left(a_\beta + \frac{A-1}{2}, b_\beta + \frac{1}{2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top \left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right).
\end{aligned}$$

Whence,

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

A.8 Conditional Posterior Distribution of ρ

$$\begin{aligned}
& f(\rho|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\sigma_\mu^2) \\
\propto\ & f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2,\rho,\boldsymbol{\psi})f(\rho) \\
=\ & \prod_{t=2}^{T}[f(\kappa_t|\kappa_{t-1},\sigma_\kappa^2,\rho,\boldsymbol{\psi})]f(\rho) \\
\propto\ & \prod_{t=2}^{T}\exp\left\{-\frac{1}{2\sigma_\kappa^2}[\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})]^2\right\}\exp\left(-\frac{\rho^2}{2\sigma_\rho^2}\right) \\
\propto\ & \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\rho^2\sum_{t=2}^{T}(\kappa_{t-1}-\eta_{t-1})^2-2\rho\sum_{t=2}^{T}(\kappa_t-\eta_t)(\kappa_{t-1}-\eta_{t-1})]-\frac{\rho^2}{2\sigma_\rho^2}\right\}.
\end{aligned}$$

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

$$\begin{aligned}
& f(\rho|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\sigma_\mu^2) \\
\propto\ & \exp\left[-\frac{1}{2\sigma_\kappa^2}(a_\rho\rho^2-2b_\rho\rho)-\frac{\rho^2}{2\sigma_\rho^2}\right] \\
=\ & \exp\left\{-\frac{1}{2}\left[\left(\frac{a_\rho}{\sigma_\kappa^2}+\frac{1}{\sigma_\rho^2}\right)\rho^2-\frac{2b_\rho}{\sigma_\kappa^2}\rho\right]\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\frac{a_\rho}{\sigma_\kappa^2}+\frac{1}{\sigma_\rho^2}\right)\left[\rho^2-\frac{2b_\rho}{\sigma_\kappa^2\left(\frac{a_\rho}{\sigma_\kappa^2}+\frac{1}{\sigma_\rho^2}\right)}\rho\right]\right\} \\
\propto\ & \exp\left\{-\frac{1}{2\left(\frac{a_\rho}{\sigma_\kappa^2}+\frac{1}{\sigma_\rho^2}\right)^{-1}}\left[\rho-\frac{b_\rho}{a_\rho+\frac{\sigma_\kappa^2}{\sigma_\rho^2}}\right]^2\right\} \\
\propto\ & N(\mu_\rho^*,(\sigma_\rho^*)^2),
\end{aligned}$$

where

$$\begin{aligned}
(\sigma_\rho^*)^2 &= \left(\frac{a_\rho}{\sigma_\kappa^2}+\frac{1}{\sigma_\rho^2}\right)^{-1}=\frac{\sigma_\kappa^2}{a_\rho+\frac{\sigma_\kappa^2}{\sigma_\rho^2}}, \\
\mu_\rho^P &= \frac{b_\rho}{a_\rho+\frac{\sigma_\kappa^2}{\sigma_\rho^2}}.
\end{aligned}$$

A.9 Conditional Posterior Distribution of $\boldsymbol{\psi}$

The AR(1) prior on $\boldsymbol{\kappa}_{-1}$ is written multivariately as

$$(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2,\rho,\boldsymbol{\psi}) \sim MVN_{T-1}(\boldsymbol{Y}_{-1}\boldsymbol{\psi}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi},\sigma_\kappa^2\boldsymbol{Q}^{-1}),$$

with a density representation

$$\begin{aligned}
& f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2,\rho,\boldsymbol{\psi}) \\
= \;& \frac{1}{\sqrt{2\pi\left|\sigma_\kappa^2\boldsymbol{Q}^{-1}\right|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\kappa}_{-1}-\boldsymbol{Y}_{-1}\boldsymbol{\psi}+\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi})^\top(\sigma_\kappa^2\boldsymbol{Q}^{-1})^{-1}(\boldsymbol{\kappa}_{-1}-\boldsymbol{Y}_{-1}\boldsymbol{\psi}+\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1\boldsymbol{\psi})\right\} \\
\propto \;& (\sigma_\kappa^2)^{-\frac{T-1}{2}}\exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1}-(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\boldsymbol{\psi}]^\top\boldsymbol{Q}[\boldsymbol{\kappa}_{-1}-(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\boldsymbol{\psi}]\right\}.
\end{aligned}$$

Notice how the $\boldsymbol{\psi}$ vector appears conveniently as a multiplicative factor in the mean vector of the above probability density function (in contrast to dealing with the univariate conditional distribution of κ_t). The conditional posterior distribution of $\boldsymbol{\psi}$ can now be derived as

$$\begin{aligned}
& f(\boldsymbol{\psi}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\rho,\sigma_\mu^2) \\
\propto \;& f(\boldsymbol{\psi},\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\rho,\sigma_\mu^2) \\
= \;& f(\boldsymbol{d}|\log\boldsymbol{\mu})f(\log\boldsymbol{\mu}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\sigma_\mu^2)f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2,\rho,\boldsymbol{\psi})f(\boldsymbol{\psi})f(\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\sigma_\kappa^2,\sigma_\beta^2,\rho,\sigma_\mu^2) \\
\propto \;& f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2,\rho,\boldsymbol{\psi})f(\boldsymbol{\psi}) \\
\propto \;& \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1}-(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\psi]^\top\boldsymbol{Q}[\boldsymbol{\kappa}_{-1}-(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\psi]\right\} \\
& \times\exp\left[-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\right] \\
= \;& \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1}^\top\boldsymbol{Q}\boldsymbol{\kappa}_{-1}-\boldsymbol{\kappa}_{-1}^\top\boldsymbol{Q}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\psi-\boldsymbol{\psi}^\top(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)^\top\boldsymbol{Q}\boldsymbol{\kappa}_{-1}\right. \\
& \left.+\boldsymbol{\psi}^\top(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)^\top\boldsymbol{Q}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)\psi]\right\} \\
& \times\exp\left\{-\frac{1}{2}[\boldsymbol{\psi}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\psi}-\boldsymbol{\psi}^\top\boldsymbol{\Sigma}_0^{-1}\psi_0-\psi_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\psi}+\psi_0^\top\boldsymbol{\Sigma}_0^{-1}\psi_0]\right\} \\
\propto \;& \exp\left\{-\frac{1}{2}\left[\boldsymbol{\psi}^\top\left(\frac{1}{\sigma_\kappa^2}(\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)\boldsymbol{\psi}\right.\right. \\
& \left.\left.-\boldsymbol{\psi}^\top\left(\frac{1}{\sigma_\kappa^2}(\boldsymbol{X}-\rho\boldsymbol{R}^{-1}\boldsymbol{X}_1)^\top\boldsymbol{Q}\boldsymbol{\kappa}_{-1}+\boldsymbol{\Sigma}_0^{-1}\psi_0\right)-\left(\frac{1}{\sigma_\kappa^2}\boldsymbol{\kappa}_{-1}^\top\boldsymbol{Q}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_1)+\psi_0^\top\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\psi}\right]\right\} \\
& \text{(proportional to an exponentiated quadratic form)} \\
\propto \;& N_2\left(\boldsymbol{\psi}^*,\boldsymbol{\Sigma}_\psi^*\right),
\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\psi}^{*} &= \left[\frac{1}{\sigma_{\kappa}^{2}}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_{1})^{\top}\boldsymbol{Q}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_{1})+\boldsymbol{\Sigma}_{0}^{-1}\right]^{-1}\\ \boldsymbol{\psi}^{*} &= \boldsymbol{\Sigma}_{\psi}^{*}\times\left[\frac{1}{\sigma_{\kappa}^{2}}(\boldsymbol{Y}_{-1}-\rho\boldsymbol{R}^{-1}\boldsymbol{Y}_{1})^{\top}\boldsymbol{Q}\boldsymbol{\kappa}_{-1}+\boldsymbol{\Sigma}_{0}^{-1}\boldsymbol{\psi}_{0}\right].\end{aligned}$$

Thus,

$$\boldsymbol{\psi}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_{\kappa}^{2},\sigma_{\beta}^{2},\rho,\sigma_{\mu}^{2}\sim N_{2}(\boldsymbol{\psi}^{*},\boldsymbol{\Sigma}_{\psi}^{*}).$$

Note that if we did not implicitly restrict $\kappa_1 = 0$, and suppose we put a prior on it as $\kappa_1 \sim N(\eta_1 = \psi_1 + \psi_2 \cdot 1, \sigma_{\kappa}^2)$, then the conditional posterior distribution reduces considerably to a much simpler expression, i.e.

$$\boldsymbol{\psi}|\boldsymbol{\alpha},\boldsymbol{\beta}_{-1},\boldsymbol{\kappa},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_{\kappa}^{2},\sigma_{\beta}^{2},\rho,\sigma_{\mu}^{2}\sim(\boldsymbol{\psi}',\boldsymbol{\Sigma}_{\psi}'),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\psi}' &= \left(\frac{1}{\sigma_{\kappa}^{2}}\boldsymbol{Y}^{\top}\boldsymbol{Q}\boldsymbol{Y}+\boldsymbol{\Sigma}_{0}^{-1}\right)^{-1},\\ \boldsymbol{\psi}' &= \boldsymbol{\Sigma}_{\psi}'\times\left(\frac{1}{\sigma_{\kappa}^{2}}\boldsymbol{Y}^{\top}\boldsymbol{Q}\boldsymbol{\kappa}+\boldsymbol{\Sigma}_{0}^{-1}\boldsymbol{\psi}_{0}\right),\\ \boldsymbol{Y} &= \begin{pmatrix}1 & 1\\ 1 & 2\\ \vdots & \vdots\\ 1 & T\end{pmatrix}.\end{aligned}$$

Using this distribution, at the same time setting $\kappa_1 = 0$ within the sampling algorithm provides an alternative way to generate conditional posterior of $\boldsymbol{\psi}$.

Appendix B

Conditional Posterior Distributions of the PLNLC Model with Blocking

B.1 Conditional Posterior Distribution of $\boldsymbol{\alpha}$

$$\begin{aligned}
& f(\boldsymbol{\alpha}|\boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log\boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
=\ & f(\boldsymbol{d}|\log\boldsymbol{\mu}) f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\alpha}) \\
=\ & \prod_t f(\log\boldsymbol{\mu}_t|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) f(\boldsymbol{\alpha}).
\end{aligned}$$

Since we have $\log\boldsymbol{\mu}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t + \boldsymbol{\nu}_t \sim N_A\left(\boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t, \sigma_\mu^2 \cdot \boldsymbol{I}_A\right)$, thus

$$\begin{aligned}
f(\log\boldsymbol{\mu}_t|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) &\propto \exp\left\{-\frac{1}{2}[\log\boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t]^\top (\sigma_\mu^2 \cdot \boldsymbol{I}_A)^{-1} [\log\boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t]\right\} \\
&= \exp\left[-\frac{1}{2\sigma_\mu^2}[\log\boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t]^\top [\log\boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t]\right].
\end{aligned}$$

With the prior distribution $\alpha_x \overset{\text{ind}}{\sim} N(\alpha_0, \sigma_\alpha^2)$ for $x = 1, 2, \ldots, A$, then

$$\begin{aligned}
\boldsymbol{\alpha} &\sim N_A\left(\alpha_0 \mathbf{1}_A = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_A \end{pmatrix}, \sigma_\alpha^2 \cdot \boldsymbol{I}_A\right), \\
\Rightarrow f(\boldsymbol{\alpha}) &\propto \exp\left[-\frac{1}{2\sigma_\alpha^2}(\boldsymbol{\alpha} - \alpha_0 \mathbf{1}_A)^\top (\boldsymbol{\alpha} - \alpha_0 \mathbf{1}_A)\right].
\end{aligned}$$

The conditional posterior distribution of $\boldsymbol{\alpha}$ then becomes

$$\begin{aligned}
& f(\boldsymbol{\alpha} | \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log \boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & \prod_t \left[f(\log \boldsymbol{\mu}_t | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2)\right] f(\boldsymbol{\alpha}) \\
\propto\ & \prod_t \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\log \boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t)^\top (\log \boldsymbol{\mu}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\kappa_t)\right]\right\} \exp\left[-\frac{1}{2\sigma_\alpha^2}(\boldsymbol{\alpha} - \alpha_0 \mathbf{1}_A)^\top (\boldsymbol{\alpha} - \alpha_0 \mathbf{1}_A)\right] \\
=\ & \prod_t \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t) - (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)^\top \boldsymbol{\alpha} + (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)^\top (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t))\right]\right\} \\
& \times \exp\left[-\frac{1}{2\sigma_\alpha^2}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \alpha_0 \mathbf{1}_A - \alpha_0 \mathbf{1}_A^\top \boldsymbol{\alpha} + \alpha_0^2 \mathbf{1}_A^\top \mathbf{1}_A)\right] \\
\propto\ & \exp\left\{-\frac{1}{2\sigma_\mu^2}[T\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \sum_t (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t) - \sum_t (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)^\top \boldsymbol{\alpha}]\right\} \\
& \times \exp\left[-\frac{1}{2\sigma_\alpha^2}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \alpha_0 \mathbf{1}_A - \alpha_0 \mathbf{1}_A^\top \boldsymbol{\alpha})\right] \\
=\ & \exp\left\{-\frac{1}{2}\left[\boldsymbol{\alpha}^\top \left(\frac{T \cdot \boldsymbol{I}_A}{\sigma_\mu^2} + \frac{\boldsymbol{I}_A}{\sigma_\alpha^2}\right)\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \left(\frac{\sum_t (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)}{\sigma_\mu^2} + \frac{\alpha_0 \mathbf{1}_A}{\sigma_\alpha^2}\right)\right.\right. \\
& \left.\left. - \left(\frac{\sum_t (\log \boldsymbol{\mu}_t - \boldsymbol{\beta}\kappa_t)^\top}{\sigma_\mu^2} + \frac{\alpha_0 \mathbf{1}_A^\top}{\sigma_\alpha^2}\right)\boldsymbol{\alpha}\right]\right\} \\
\propto\ & N_A(\boldsymbol{\mu}_\alpha^*, \boldsymbol{\Sigma}_\alpha^*),
\end{aligned}$$

where

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

Therefore, $\boldsymbol{\alpha} | \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \log \boldsymbol{\mu}, \boldsymbol{d}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2 \sim N_A(\boldsymbol{\mu}_\alpha^*, \boldsymbol{\Sigma}_\alpha^*)$.

B.2 Conditional Posterior Distribution of $\boldsymbol{\beta}_{-1}$

Suppose that $\boldsymbol{\mu}_{-x,t}(\mu_{1\,t}, \ldots, \mu_{x-1\,t}, \mu_{x+1\,t}, \ldots, \mu_{A\,t})^\top$ denotes a vector of mortality rates corresponding to year t excluding the x^{th} component, the conditional posterior distribution of $\boldsymbol{\beta}_{-1}$ can be derived as

$$\begin{aligned}
& f(\boldsymbol{\beta}_{-1}|\boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \boldsymbol{D}, \log\boldsymbol{\mu}, \sigma^2_\kappa, \sigma^2_\beta, \boldsymbol{\psi}, \rho, \sigma^2_\mu) \\
\propto\ & f(\log\boldsymbol{\mu}|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) f(\boldsymbol{\beta}_{-1}|\sigma^2_\beta) \\
=\ & f(\log\boldsymbol{\mu}_{-1,1}, \ldots, \log\boldsymbol{\mu}_{-1,T}|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}) f(\log\boldsymbol{\mu}_1|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}) f(\boldsymbol{\beta}_{-1}|\sigma^2_\beta) \\
=\ & \prod_t [f(\log\boldsymbol{\mu}_t^{-1}|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \sigma^2_\mu)] \times \prod_t [f(\log\mu_{1t}|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \sigma^2_\mu)] f(\boldsymbol{\beta}_{-1}|\sigma^2_\beta).
\end{aligned}$$

Now we know that

$$\log\boldsymbol{\mu}_{-1,t} = \boldsymbol{\alpha}_{-1} + \boldsymbol{\beta}_{-1}\kappa_t + \boldsymbol{\nu}_{-1,t} \sim N_{A-1}(\boldsymbol{\alpha}_{-1} + \boldsymbol{\beta}_{-1}\kappa_t, \sigma^2_\mu I_{A-1}),$$

where $\boldsymbol{\nu}_{-x,t} = (\nu_{1\,t}, \ldots, \nu_{x-1\,t}, \nu_{x+1\,t}, \ldots, \nu_{A\,t})^\top$. Therefore,

$$\begin{aligned}
& f(\log\boldsymbol{\mu}_{-1,t}|\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \sigma^2_\mu) \\
\propto\ & \exp\left\{-\frac{1}{2}[\log\boldsymbol{\mu}_{-1,t} - \boldsymbol{\alpha}_{-1} - \boldsymbol{\beta}_{-1}\kappa_t]^\top (\sigma^2_\mu \boldsymbol{I}_{A-1})^{-1} [\log\boldsymbol{\mu}_{-1,t} - \boldsymbol{\alpha}_{-1} - \boldsymbol{\beta}_{-1}\kappa_t]\right\} \\
=\ & \exp\left[-\frac{1}{2\sigma^2_\mu}(\log\boldsymbol{\mu}_{-1,t} - \boldsymbol{\alpha}_{-1} - \boldsymbol{\beta}_{-1}\kappa_t)^\top (\log\boldsymbol{\mu}_{-1,t} - \boldsymbol{\alpha}_{-1} - \boldsymbol{\beta}_{-1}\kappa_t)\right].
\end{aligned}$$

Moreover, the log mortality rate for infant at time t, $\log\mu_{1t}$ can be expressed in terms of $\boldsymbol{\beta}_{-1}$ by

$$\begin{aligned}
\log\mu_{1t} &= \alpha_1 + (1 - (\beta_2 + \ldots + \beta_A))\kappa_t + \nu_{1t} \\
&= \alpha_1 + \kappa_t - \kappa_t \mathbf{1}^\top_{A-1}\boldsymbol{\beta}_{-1} + \nu_{1t} \\
\Rightarrow \log\mu_{1t} &\sim N\left(\alpha_1 + \kappa_t - \kappa_t \mathbf{1}^\top_{A-1}\boldsymbol{\beta}_{-1}, \sigma^2_\mu\right),
\end{aligned}$$

The density function of $\log\mu_{1t}$ is then

$$\begin{aligned}
& f(\log\mu_{1t}|\alpha_1, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \sigma^2_\mu) \\
\propto\ & \exp\left\{-\frac{1}{2\sigma^2_\mu}\left[\log\mu_{1t} - \alpha_1 - \kappa_t + \kappa_t \mathbf{1}^\top_{A-1}\boldsymbol{\beta}_{-1}\right]^\top \left[\log\mu_{1t} - \alpha_1 - \kappa_t + \kappa_t \mathbf{1}^\top_{A-1}\boldsymbol{\beta}_{-1}\right]\right\}.
\end{aligned}$$

Again, $\boldsymbol{\beta}_{-1}|\sigma^2_\beta \sim N_{A-1}\left(\frac{1}{A}\mathbf{1}_{A-1}, \sigma^2_\beta(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1})\right)$, with density function

$$f(\boldsymbol{\beta}_{-1}|\sigma^2_\beta) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)^\top \sigma_\beta^{-2}\left(\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\left(\boldsymbol{\beta}_{-1} - \frac{1}{A}\mathbf{1}_{A-1}\right)\right\}.$$

Finally,

$$
\begin{aligned}
& f(\boldsymbol{\beta}_{-1}|\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\boldsymbol{d},\log\boldsymbol{\mu},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\rho,\sigma_\mu^2) \\
\propto\ & \prod_t [f(\log\boldsymbol{\mu}_{-1,t}|\boldsymbol{\beta}_{-1},\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\sigma_\mu^2)] \\
& \times \prod_t [f(\log\boldsymbol{\mu}_{1t}|\boldsymbol{\beta}_{-1},\boldsymbol{\alpha},\boldsymbol{\kappa}_{-1},\sigma_\mu^2)] f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
\propto\ & \prod_t \exp\left[-\frac{1}{2\sigma_\mu^2}(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1}-\boldsymbol{\beta}_{-1}\kappa_t)^\top(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1}-\boldsymbol{\beta}_{-1}\kappa_t)\right] \\
& \times \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\log\mu_{1t}-\alpha_1-\kappa_t+\kappa_t\boldsymbol{\beta}_{-1}^\top\mathbf{1}_{A-1}][\log\mu_{1t}-\alpha_1-\kappa_t+\kappa_t\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1}]\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\beta^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\} \\
=\ & \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[-\boldsymbol{\beta}_{-1}^\top\kappa_t+(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1})^\top][-\boldsymbol{\beta}_{-1}\kappa_t+(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1})]\right\} \\
& \times \prod_t \exp\left\{-\frac{1}{2\sigma_\mu^2}[\kappa_t\boldsymbol{\beta}_{-1}^\top\mathbf{1}_{A-1}+(\log\mu_{1t}-\alpha_1-\kappa_t)][\kappa_t\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1}+(\log\mu_{1t}-\alpha_1-\kappa_t)]\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\beta^2}\left(\boldsymbol{\beta}_{-1}^\top-\frac{1}{A}\mathbf{1}_{A-1}^\top\right)\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\}
\end{aligned}
$$

$$
\begin{aligned}
\propto\ & \prod_t \exp\Bigg\{-\frac{1}{2\sigma_\mu^2}\Big[\boldsymbol{\beta}_{-1}^\top\boldsymbol{\beta}_{-1}\kappa_t^2 - \boldsymbol{\beta}_{-1}^\top\kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1}) - \kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1})^\top\boldsymbol{\beta}_{-1}\Big]\Bigg\} \\
& \times\prod_t \exp\Bigg\{-\frac{1}{2\sigma_\mu^2}\Big[\kappa_t^2\boldsymbol{\beta}_{-1}^\top\mathbf{1}_{A-1}\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1} \\
& \qquad -\boldsymbol{\beta}_{-1}^\top\kappa_t\mathbf{1}_{A-1}(-\log\mu_{1t}+\alpha_1+\kappa_t) - \kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t)\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1}\Big]\Bigg\} \\
& \times\exp\Bigg\{-\frac{1}{2\sigma_\beta^2}\Bigg[\boldsymbol{\beta}_{-1}^\top\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\boldsymbol{\beta}_{-1} - \frac{1}{A}\boldsymbol{\beta}_{-1}^\top\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\mathbf{1}_{A-1} \\
& \qquad -\frac{1}{A}\mathbf{1}_{A-1}^\top\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\boldsymbol{\beta}_{-1}\Bigg]\Bigg\} \\
=\ & \exp\Bigg\{-\frac{1}{2\sigma_\mu^2}\Bigg[\boldsymbol{\beta}_{-1}^\top\boldsymbol{\beta}_{-1}\sum_t(\kappa_t^2) - \boldsymbol{\beta}_{-1}^\top\sum_t\kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1}) - \sum_t\kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1})^\top\boldsymbol{\beta}_{-1}\Bigg]\Bigg\} \\
& \times\exp\Bigg\{-\frac{1}{2\sigma_\mu^2}\Bigg[\sum_t(\kappa_t^2)\boldsymbol{\beta}_{-1}^\top\mathbf{1}_{A-1}\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1} \\
& -\boldsymbol{\beta}_{-1}^\top\sum_t\kappa_t\mathbf{1}_{A-1}(-\log\mu_{1t}+\alpha_1+\kappa_t) - \sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t)\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1}\Bigg]\Bigg\} \\
& \times\exp\Bigg\{-\frac{1}{2\sigma_\beta^2}\Bigg[\boldsymbol{\beta}_{-1}^\top\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{-1}^\top\frac{1}{A}\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\mathbf{1}_{A-1} \\
& -\frac{1}{A}\mathbf{1}_{A-1}^\top\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\boldsymbol{\beta}_{-1}\Bigg]\Bigg\} \\
=\ & \exp\Bigg\{-\frac{1}{2}\Bigg[\boldsymbol{\beta}_{-1}^\top\Bigg(\frac{\sum_t\kappa_t^2}{\sigma_\mu^2}\boldsymbol{I}_{A-1} + \frac{1}{\sigma_\mu^2}\left(\sum_t\kappa_t^2\right)\boldsymbol{J}_{A-1} + \frac{1}{\sigma_\beta^2}\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\Bigg)\boldsymbol{\beta}_{-1} \\
& -\boldsymbol{\beta}_{-1}^\top\Bigg(\frac{1}{\sigma_\mu^2}\sum_t\kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1}) + \frac{1}{\sigma_\mu^2}\sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t)\mathbf{1}_{A-1} \\
& +\frac{1}{A\sigma_\beta^2}\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\mathbf{1}_{A-1}\Bigg) \\
& -\Bigg(\frac{1}{\sigma_\mu^2}\sum_t\kappa_t(\log\boldsymbol{\mu}_{-1,t}-\boldsymbol{\alpha}_{-1})^\top + \frac{1}{\sigma_\mu^2}\sum_t\kappa_t(-\log\mu_{1t}+\alpha_1+\kappa_t)\mathbf{1}_{A-1}^\top \\
& +\frac{1}{A\sigma_\beta^2}\mathbf{1}_{A-1}^\top(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1})^{-1}\Bigg)\boldsymbol{\beta}_{-1}\Bigg]\Bigg\} \\
\propto\ & N_{A-1}(\boldsymbol{\mu}_\beta^*,\boldsymbol{\Sigma}_\beta^*),
\end{aligned}
$$

where

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

Hence,

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

B.3 Conditional Posterior Distribution of $\boldsymbol{\kappa}_{-1}$

First, the AR(1) process on κ_t can be written as in multivariate form as

$$\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} = \boldsymbol{P}(\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi}) + (-\rho\eta_1, 0, \ldots, 0)^\top + \boldsymbol{\epsilon}_{-1},$$

where

$$\boldsymbol{\epsilon}_{-1} = (\epsilon_2, \ldots, \epsilon_T)^\top \sim N_{T-1}(\mathbf{0}, \sigma^2_\kappa \boldsymbol{I}_{T-1}).$$

Thus,

$$\begin{aligned}
&\Rightarrow\ (\boldsymbol{I}_{T-1} - \boldsymbol{P})(\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi}) \sim N_{T-1}\left((-\rho\eta_1, 0, \ldots, 0)^\top, \sigma^2_\kappa \boldsymbol{I}_{T-1}\right) \\
&\Rightarrow\ (\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi}) \sim N_{T-1}\left((\boldsymbol{I}_{T-1} - \boldsymbol{P})^{-1}(-\rho\eta_1, 0, \ldots, 0)^\top, (\boldsymbol{I}_{T-1} - \boldsymbol{P})^{-1}\sigma^2_\kappa \boldsymbol{I}_{T-1}(\boldsymbol{I}_{T-1} - \boldsymbol{P})^{-T}\right) \\
&\Rightarrow\ (\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi}) \sim N_{T-1}\left(\boldsymbol{R}^{-1}(-\rho\eta_1, 0, \ldots, 0)^\top, \sigma^2_\kappa(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right) \\
&\Rightarrow\ \boldsymbol{\kappa}_{-1} \sim N_{T-1}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}(-\rho\eta_1, 0, \ldots, 0)^\top, \sigma^2_\kappa \boldsymbol{Q}^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{Q} &= \boldsymbol{R}^\top \boldsymbol{R} = (\boldsymbol{I}_{T-1} - \boldsymbol{P})^\top (\boldsymbol{I}_{T-1} - \boldsymbol{P}) \\
&= \begin{pmatrix} 1 & & & & \\ -\rho & 1 & & \mathbf{0} & \\ & -\rho & \ddots & & \\ & \mathbf{0} & \ddots & \ddots & \\ & & & -\rho & 1 \end{pmatrix}^\top \begin{pmatrix} 1 & & & & \\ -\rho & 1 & & \mathbf{0} & \\ & -\rho & \ddots & & \\ & \mathbf{0} & \ddots & \ddots & \\ & & & -\rho & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1+\rho^2 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ & \mathbf{0} & \ddots & \ddots & -\rho \\ & & & -\rho & 1+\rho^2 \end{pmatrix}_{(T-1)\times(T-1)}.
\end{aligned}$$

Therefore,

$$\boldsymbol{\kappa}_{-1}|\rho, \boldsymbol{\psi}, \sigma_\kappa^2 \sim N_{T-1}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}(-\rho\eta_1, 0, \ldots, 0)^\top, \sigma_\kappa^2 \boldsymbol{Q}^{-1}).$$

For simplicity, let $\boldsymbol{a} = (\rho\eta_1, 0, \ldots, 0)^\top = \rho \boldsymbol{Y}_1 \boldsymbol{\psi}$, the joint probability density function of $\boldsymbol{\kappa}_{-1}$ can be expressed as

$$\begin{aligned}
f(\boldsymbol{\kappa}_{-1}|\boldsymbol{\psi}, \sigma_\kappa^2, \rho) &\propto \exp\left\{-\frac{1}{2}[\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]^\top (\sigma_\kappa^2 \boldsymbol{Q}^{-1})^{-1} [\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]^\top \boldsymbol{Q} [\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]\right\}.
\end{aligned}$$

As in previously, denote $\boldsymbol{\mu}_{x,-1} = (\mu_{x1}, \ldots, \mu_{x\,t-1}, \mu_{x\,t+1}, \ldots, \mu_{xT})^\top$ as a vector of mortality rates corresponding to age x excluding the t^{th} component. We have

$$\begin{aligned}
& f(\boldsymbol{\kappa}_{-1}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log\boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto\ & f(\log\boldsymbol{\mu}|\boldsymbol{\kappa}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \sigma_\mu^2) f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2, \boldsymbol{\psi}, \rho) \\
\propto\ & f(\log\boldsymbol{\mu}_{1,-1}, \log\boldsymbol{\mu}_{2,-1}, \ldots, \log\boldsymbol{\mu}_{A,-1}|\boldsymbol{\kappa}_{-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \sigma_\mu^2) f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2, \boldsymbol{\psi}, \rho) \\
& \text{(Just delete the 1}^{\text{st}}\text{ column of } \boldsymbol{\mu} \text{ corresponding to } t = 1) \\
=\ & \prod_x \left[f(\log\boldsymbol{\mu}_{x,-1}|\boldsymbol{\kappa}_{-1}, \alpha_x, \beta_x, \sigma_\mu^2)\right] \times f(\boldsymbol{\kappa}_{-1}|\sigma_\kappa^2, \boldsymbol{\psi}, \rho).
\end{aligned}$$

Notice that since

$$\log\boldsymbol{\mu}_{x,-1} = \alpha_x \mathbf{1}_{T-1} + \beta_x \boldsymbol{\kappa}_{-1} + \boldsymbol{\nu}_{x,-1},$$

where $\boldsymbol{\nu}_{x,-1} = (\nu_{x2}, \dots, \nu_{xT})^\top \sim N_{T-1}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I}_{T-1})$, hence

$$\log \boldsymbol{\mu}_{x,-1} | \boldsymbol{\kappa}_{-1}, \alpha_x, \beta_x, \sigma_\mu^2 \sim N_{T-1}(\alpha_x \mathbf{1}_{T-1} + \beta_x \boldsymbol{\kappa}_{-1}, \sigma_\mu^2 \boldsymbol{I}_{T-1}).$$

The probability density function of $\log \boldsymbol{\mu}_{x,-1}$ is the following,

$$\begin{aligned}
& f(\log \boldsymbol{\mu}_{x,-1} | \alpha_x, \beta_x, \boldsymbol{\kappa}_{-1}, \sigma_\mu^2) \\
\propto \quad & \exp\left\{-\frac{1}{2}[\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1}]^\top (\sigma_\mu^2 \boldsymbol{I}_{T-1})^{-1} [\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1}]\right\} \\
= \quad & \exp\left\{-\frac{1}{2\sigma_\mu^2}[\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1}]^\top [\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1}]\right\}.
\end{aligned}$$

The conditional posterior distribution of $\boldsymbol{\kappa}_{-1}$ is then given by,

$$\begin{aligned}
& f(\boldsymbol{\kappa}_{-1} | \boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2) \\
\propto \quad & \prod_x [f(\log \boldsymbol{\mu}_{x,-1} | \boldsymbol{\kappa}_{-1}, \alpha_x, \beta_x, \sigma_\mu^2)] f(\boldsymbol{\kappa}_{-1} | \sigma_\kappa^2, \boldsymbol{\psi}, \rho) \\
\propto \quad & \prod_x \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1})^\top (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1} - \beta_x \boldsymbol{\kappa}_{-1})\right]\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]^\top \boldsymbol{Q} [\boldsymbol{\kappa}_{-1} - \boldsymbol{Y}_{-1}\boldsymbol{\psi} - \boldsymbol{R}^{-1}\boldsymbol{a}]\right\} \\
\propto \quad & \prod_x \left\{\exp\left[-\frac{1}{2\sigma_\mu^2}(\beta_x^2 \boldsymbol{\kappa}_{-1}^\top \boldsymbol{\kappa}_{-1} - \boldsymbol{\kappa}_{-1}^\top \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1}) - \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1})^\top \boldsymbol{\kappa}_{-1})\right]\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}[\boldsymbol{\kappa}_{-1}^\top \boldsymbol{Q} \boldsymbol{\kappa}_{-1} - \boldsymbol{\kappa}_{-1}^\top \boldsymbol{Q}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a}) - (\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a})^\top \boldsymbol{Q} \boldsymbol{\kappa}_{-1}]\right\} \\
= \quad & \exp\left\{-\frac{1}{2\sigma_\mu^2}\left[\left(\sum_x \beta_x^2\right) \boldsymbol{\kappa}_{-1}^\top \boldsymbol{\kappa}_{-1} - \boldsymbol{\kappa}_{-1}^\top \sum_x \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1})\right.\right. \\
& \left.\left. - \sum_x \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1})^\top \boldsymbol{\kappa}_{-1}\right]\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\kappa^2}\left[\boldsymbol{\kappa}_{-1}^\top \boldsymbol{Q} \boldsymbol{\kappa}_{-1} - \boldsymbol{\kappa}_{-1}^\top \boldsymbol{Q}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a}) - (\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a})^\top \boldsymbol{Q} \boldsymbol{\kappa}_{-1}\right]\right\} \\
= \quad & \exp\left\{-\frac{1}{2}\left[\boldsymbol{\kappa}_{-1}^\top \left(\frac{\sum_x \beta_x^2}{\sigma_\mu^2} \boldsymbol{I}_{T-1} + \frac{1}{\sigma_\kappa^2}\boldsymbol{Q}\right) \boldsymbol{\kappa}_{-1}\right.\right. \\
& - \boldsymbol{\kappa}_{-1}^\top \left(\frac{1}{\sigma_\mu^2} \sum_x \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1}) + \frac{1}{\sigma_\kappa^2}\boldsymbol{Q}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a})\right) \\
& \left.\left. - \left(\frac{1}{\sigma_\mu^2} \sum_x \beta_x (\log \boldsymbol{\mu}_{x,-1} - \alpha_x \mathbf{1}_{T-1})^\top + \frac{1}{\sigma_\kappa^2}(\boldsymbol{Y}_{-1}\boldsymbol{\psi} + \boldsymbol{R}^{-1}\boldsymbol{a})^\top \boldsymbol{Q}\right) \boldsymbol{\kappa}_{-1}\right]\right\} \\
\propto \quad & N_{T-1}(\boldsymbol{\mu}_\kappa^*, \boldsymbol{\Sigma}_\kappa^*),
\end{aligned}$$

where

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

Therefore, $\boldsymbol{\kappa}_{-1}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{d}, \log \boldsymbol{\mu}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \sigma_\mu^2 \sim N_{T-1}(\boldsymbol{\mu}_\kappa^*, \boldsymbol{\Sigma}_\kappa^*)$.

Appendix C

Conditional Posterior Distribution of μ_{xt} under the PGLC Model

Let $\boldsymbol{\mu}_{-xt}$ denotes the vector of all the mortality rates excluding the xt^{th} element. Then we have

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

Hence,

$$\mu_{xt}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{-1}, \boldsymbol{\kappa}_{-1}, \boldsymbol{d}, \boldsymbol{\mu}_{-xt}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi \sim \text{Gamma}\left(d_{xt} + \phi, e_{xt} + \frac{\phi}{\exp(\alpha_x + \beta_x\kappa_t)}\right).$$

Appendix D

Hessian Matrix for the NBLC Model

The full joint posterior density is computed by multiplying the model likelihood and prior distributions,

$$
\begin{aligned}
& f(\boldsymbol{\kappa}_{-1}, \boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi|\boldsymbol{d}) \\
= \; & f(\boldsymbol{d}|\boldsymbol{\alpha}, \boldsymbol{\kappa}_{-1}, \phi) f(\boldsymbol{\kappa}_{-1}|\rho, \boldsymbol{\psi}, \sigma_\kappa^2) f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) f(\boldsymbol{\alpha}) f(\sigma_\kappa^2, \sigma_\beta^2, \rho, \boldsymbol{\psi}, \phi) \\
= \; & \prod_{x,t} [f(d_{xt}|\alpha_x, \beta_x, \kappa_t, \phi)] \prod_{t=2}^{T} f(\kappa_t|\kappa_{t-1}, \rho, \boldsymbol{\psi}, \sigma_\kappa^2) f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
& \times \prod_{x=1}^{A} [f(\alpha_x)] f(\sigma_\kappa^2) f(\sigma_\beta^2) f(\rho) f(\boldsymbol{\psi}) f(\phi).
\end{aligned}
$$

First, we alter the form of $f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2)$ to facilitate the subsequent computations,

$$\begin{aligned}
& f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2) \\
= \;& \frac{1}{\sqrt{(2\pi)^{A-1}\left|\sigma_\beta^2(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1})\right|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{-1}-\frac{1}{A}\mathbf{1}_{A-1}\right)^\top \sigma_\beta^{-2}\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)^{-1}\right. \\
& \left.\left(\boldsymbol{\beta}_{-1}-\frac{1}{A}\mathbf{1}_{A-1}\right)\right\} \\
\propto \;& \frac{1}{\sqrt{(\sigma_\beta^2)^{A-1}\left|\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right|}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left(\boldsymbol{\beta}_{-1}-\frac{1}{A}\mathbf{1}_{A-1}\right)^\top (\boldsymbol{I}_{A-1}+\boldsymbol{J}_{A-1})\left(\boldsymbol{\beta}_{-1}-\frac{1}{A}\mathbf{1}_{A-1}\right)\right\} \\
\propto \;& (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left[\boldsymbol{\beta}_{-1}^\top(\boldsymbol{I}_{A-1}+\boldsymbol{J}_{A-1})\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{-1}^\top(\boldsymbol{I}_{A-1}+\boldsymbol{J}_{A-1})\frac{1}{A}\mathbf{1}_{A-1}\right.\right. \\
& \left.\left.-\frac{1}{A}\mathbf{1}_{A-1}^\top(\boldsymbol{I}_{A-1}+\boldsymbol{J}_{A-1})\boldsymbol{\beta}_{-1}+\frac{1}{A^2}\mathbf{1}_{A-1}^\top(\boldsymbol{I}_{A-1}+\boldsymbol{J}_{A-1})\mathbf{1}_{A-1}\right]\right\} \\
= \;& (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left[(\boldsymbol{\beta}_{-1}^\top+(\sum_{x=2}^{A}\beta_x)\mathbf{1}_{A-1}^\top)\boldsymbol{\beta}_{-1}-\frac{1}{A}(\boldsymbol{\beta}_{-1}^\top+(\sum_{x=2}^{A}\beta_x)\mathbf{1}_{A-1}^\top)\mathbf{1}_{A-1}\right.\right. \\
& \left.\left.-\frac{1}{A}(\mathbf{1}_{A-1}^\top+(A-1)\mathbf{1}_{A-1}^\top)\boldsymbol{\beta}_{-1}+\frac{1}{A^2}(\mathbf{1}_{A-1}^\top+(A-1)\mathbf{1}_{A-1}^\top)\mathbf{1}_{A-1}\right]\right\}
\end{aligned}$$

$$\begin{aligned}
= \;& (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left[\boldsymbol{\beta}_{-1}^\top\boldsymbol{\beta}_{-1}+\left(\sum_{x=2}^{A}\beta_x\right)\mathbf{1}_{A-1}^\top\boldsymbol{\beta}_{-1}-\frac{1}{A}\left(\sum_{x=2}^{A}\beta_x+\sum_{x=2}^{A}\beta_x(A-1)\right)\right.\right. \\
& \left.\left.-\frac{1}{A}\left(\sum_{x=2}^{A}\beta_x+(A-1)\sum_{x=2}^{A}\beta_x\right)+\frac{1}{A^2}(A-1+(A-1)(A-1))\right]\right\} \\
= \;& (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left[\sum_{x=2}^{A}\beta_x^2+\left(\sum_{x=2}^{A}\beta_x\right)^2-2\sum_{x=2}^{A}\beta_x+\frac{1}{A^2}A(A-1)\right]\right\} \\
= \;& (\sigma_\beta^2)^{-\frac{A-1}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left[\sum_{x=2}^{A}\beta_x^2+\left(\sum_{x=2}^{A}\beta_x\right)^2-2\sum_{x=2}^{A}\beta_x+\frac{A-1}{A}\right]\right\}.
\end{aligned}$$

Note that the above expression for $f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2)$ can be equivalently derived by realizing that

$$f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2)=\frac{f(\sum_x\beta_x=1|\boldsymbol{\beta}_{-1},\sigma_\beta^2)f(\boldsymbol{\beta}_{-1}|\sigma_\beta^2)}{f(\sum_x\beta_x=1|\sigma_\beta^2)}.$$

Secondly, for $\boldsymbol{\psi}$, we have

$$f(\boldsymbol{\psi}) \;=\; \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_0|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\right\}.$$

Now suppose we let $\boldsymbol{\psi}_0 = \begin{pmatrix} \psi_{10} \\ \psi_{20} \end{pmatrix}$ and $\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{\Sigma_{22}}{|\boldsymbol{\Sigma}_0|} & -\frac{\Sigma_{12}}{|\boldsymbol{\Sigma}_0|} \\ -\frac{\Sigma_{21}}{|\boldsymbol{\Sigma}_0|} & \frac{\Sigma_{11}}{|\boldsymbol{\Sigma}_0|} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}$, then

$$\begin{aligned}
f(\boldsymbol{\psi}) \;&\propto\; \exp\left\{-\frac{1}{2}(\psi_1-\psi_{10} \;\; \psi_2-\psi_{20})\begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}\begin{pmatrix} \psi_1-\psi_{10} \\ \psi_2-\psi_{20} \end{pmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}[f_{11}(\psi_1-\psi_{10})+f_{21}(\psi_2-\psi_{20}) \;\; f_{12}(\psi_1-\psi_{10})+f_{22}(\psi_2-\psi_{20})]\right. \\
&\quad \left.\times\begin{pmatrix} \psi_1-\psi_{10} \\ \psi_2-\psi_{20} \end{pmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}[(f_{11}(\psi_1-\psi_{10})+f_{21}(\psi_2-\psi_{20}))(\psi_1-\psi_{10})\right. \\
&\quad \left.+(f_{12}(\psi_1-\psi_{10})+f_{22}(\psi_2-\psi_{20}))(\psi_2-\psi_{20})]\right\} \\
&= \exp\left\{-\frac{1}{2}[f_{11}(\psi_1-\psi_{10})^2+(f_{12}+f_{21})(\psi_1-\psi_{10})(\psi_2-\psi_{20})+f_{22}(\psi_2-\psi_{20})^2]\right\}.
\end{aligned}$$

The joint posterior density is then

$$\begin{aligned}
& f(\boldsymbol{\kappa}_{-1},\boldsymbol{\beta}_{-1},\boldsymbol{\alpha},\sigma_\kappa^2,\sigma_\beta^2,\boldsymbol{\psi},\rho,\phi|\boldsymbol{d}) \\
\propto\; & \left\{\prod_{x,t}\frac{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)]^{d_{xt}}}{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi]^{d_{xt}+\phi}}\right\}\exp\left\{-\frac{1}{2\sigma_\kappa^2}\sum_{t=2}^{T}[\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})]^2\right\} \\
& \times\exp\left\{-\frac{1}{2\sigma_\beta^2}\left[\sum_{x=2}^{A}\beta_x^2+\left(\sum_{x=2}^{A}\beta_x\right)^2-2\sum_{x=2}^{A}\beta_x+\frac{A-1}{A}\right]\right\} \\
& \times\exp\left[-\frac{1}{2\sigma_\alpha^2}\sum_x(\alpha_x-\alpha_0)^2\right]\exp\left(-\frac{\rho^2}{2\sigma_\rho^2}\right) \\
& \times(\sigma_\kappa^2)^{-\frac{T-1}{2}-a_\kappa-1}\exp\left(-\frac{b_\kappa}{\sigma_\kappa^2}\right)\times(\sigma_\beta^2)^{-\frac{A-1}{2}-a_\beta-1}\exp\left(-\frac{b_\beta}{\sigma_\beta^2}\right)\exp\left(\frac{1}{2A\sigma_\beta^2}\right) \\
& \times\exp\left\{-\frac{1}{2}[f_{11}(\psi_1-\psi_{10})^2+(f_{12}+f_{21})(\psi_1-\psi_{10})(\psi_2-\psi_{20})+f_{22}(\psi_2-\psi_{20})^2]\right\} \\
& \times\phi^{AT\phi+a_\phi-1}\exp(-b_\phi\phi)\times\frac{\prod_{x,t}\Gamma(d_{xt}+\phi)}{\Gamma(\phi)^{AT}}.
\end{aligned}$$

The log joint posterior density is thus given by

$$
\begin{aligned}
& \log f(\boldsymbol{\kappa}_{-1}, \boldsymbol{\beta}_{-1}, \boldsymbol{\alpha}, \sigma_\kappa^2, \sigma_\beta^2, \boldsymbol{\psi}, \rho, \phi | \boldsymbol{d}) \\
= & \sum_{x,t} \{ d_{xt} \log e_{xt} + d_{xt}(\alpha_x + \beta_x \kappa_t) - (d_{xt} + \phi) \log[e_{xt} \exp(\alpha_x + \beta_x \kappa_t) + \phi] \} \\
& - \frac{1}{2\sigma_\kappa^2} \sum_{t=2}^{T} [\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1})]^2 - \frac{1}{2\sigma_\beta^2} \left[\sum_{x=2}^{A} \beta_x^2 + \left(\sum_{x=2}^{A} \beta_x \right)^2 - 2 \sum_{x=2}^{A} \beta_x + \frac{A-1}{A} \right] \\
& - \frac{1}{2\sigma_\alpha^2} \sum_x (\alpha_x - \alpha_0)^2 - \frac{\rho^2}{2\sigma_\rho^2} - \left(\frac{T-1}{2} + a_\kappa + 1 \right) \log(\sigma_\kappa^2) - \frac{b_\kappa}{\sigma_\kappa^2} - \left(\frac{A-1}{2} + a_\beta + 1 \right) \log(\sigma_\beta^2) \\
& - \frac{b_\beta}{\sigma_\beta^2} + \frac{1}{2A\sigma_\beta^2} - \frac{1}{2} [f_{11}(\psi_1 - \psi_{10})^2 + (f_{12} + f_{21})(\psi_1 - \psi_{10})(\psi_2 - \psi_{20}) + f_{22}(\psi_2 - \psi_{20})^2] \\
& + (AT\phi + a_\phi - 1) \log \phi - b_\phi \phi + \sum_{x,t} \log \Gamma(d_{xt} + \phi) - AT \log \Gamma(\phi).
\end{aligned}
$$

The second order partial derivatives of the log joint posterior density, which form the components of the Hessian matrix are as follows:

$$
\begin{aligned}
\frac{\partial^2 \log f}{\partial \kappa_j \partial \kappa_i} &= \begin{cases} -\sum_{x=1}^{A} \frac{(d_{xi}+\phi)\phi\beta_x^2 e_{xi}\exp(\alpha_x+\beta_x\kappa_i)}{[e_{xi}\exp(\alpha_x+\beta_x\kappa_i)+\phi]^2} - \frac{1}{\sigma_\kappa^2} - \frac{\rho^2}{\sigma_\kappa^2} & \text{for } i = j \neq T, \\ -\sum_{x=1}^{A} \frac{(d_{xi}+\phi)\phi\beta_x^2 e_{xi}\exp(\alpha_x+\beta_x\kappa_i)}{[e_{xi}\exp(\alpha_x+\beta_x\kappa_i)+\phi]^2} - \frac{1}{\sigma_\kappa^2} & \text{for } i = j = T, \\ \rho/\sigma_\kappa^2 & \text{for } j = i-1 \text{ or } j = i+1, \\ 0 & \text{otherwise,} \end{cases} \\
\frac{\partial^2 \log f}{\beta_j \beta_i} &= \begin{cases} -\sum_{t=1}^{T} \frac{(d_{it}+\phi)\phi\kappa_t^2 e_{it}\exp(\alpha_i+\beta_i\kappa_t)}{[e_{it}\exp(\alpha_i+\beta_i\kappa_t)+\phi]^2} & \\ \quad -\sum_{t=1}^{T} \frac{(d_{1t}+\phi)\phi\kappa_t^2 e_{1t}\exp(\alpha_1+\beta_1\kappa_t)}{[e_{1t}\exp(\alpha_1+\beta_1\kappa_t)+\phi]^2} - \frac{2}{\sigma_\beta^2} & \text{for } i = j \neq 1, \\ -\sum_{t=1}^{T} \frac{(d_{1t}+\phi)\phi\kappa_t^2 e_{1t}\exp(\alpha_1+\beta_1\kappa_t)}{[e_{1t}\exp(\alpha_1+\beta_1\kappa_t)+\phi]^2} - \frac{1}{\sigma_\beta^2} & \text{for } i \neq j \text{ and } i, j \neq 1, \end{cases} \\
\frac{\partial^2 \log f}{\partial \alpha_j \partial \alpha_i} &= \begin{cases} -\sum_t \left[\frac{(d_{it}+\phi)\phi e_{it}\exp(\alpha_i+\beta_i\kappa_t)}{[e_{it}\exp(\alpha_i+\beta_i\kappa_t)+\phi]^2}\right] - \frac{1}{\sigma_l^2} & \text{for } i = j \text{ and } i = 1, 2, \ldots, A, \\ 0 & \text{for } i \neq j \text{ and } i = 1, 2, \ldots, A, \end{cases} \\
\frac{\partial^2 \log f}{\partial \alpha_j \partial \kappa_i} &= -\frac{(d_{ji}+\phi)\phi\beta_j e_{ji}\exp(\alpha_j+\beta_j\kappa_i)}{[e_{ji}\exp(\alpha_j+\beta_j\kappa_i)+\phi]^2} \text{ for } i = 2, 3, \ldots, T \text{ and } j = 1, 2, \ldots, A, \\
\frac{\partial^2 \log f}{\partial \beta_j \partial \kappa_i} &= d_{ji} - d_{1i} - (d_{ji}+\phi)\left[1 - \frac{\phi}{e_{ji}\exp(\alpha_j+\beta_j\kappa_i)+\phi}\right] \\
&\quad +(d_{1i}+\phi)\left[1 - \frac{\phi}{e_{1i}\exp(\alpha_1+\beta_1\kappa_i)+\phi}\right] - \frac{(d_{ji}+\phi)\phi\beta_j\kappa_i e_{ji}\exp(\alpha_j+\beta_j\kappa_i)}{[e_{ji}\exp(\alpha_j+\beta_j\kappa_i)+\phi]^2} \\
&\quad +\frac{(d_{1i}+\phi)\phi\beta_1\kappa_i e_{1i}\exp(\alpha_1+\beta_1\kappa_i)}{[e_{1i}\exp(\alpha_1+\beta_1\kappa_i)+\phi]^2} \text{ for } i = 2, 3, \ldots, T \text{ and } j = 2, 3, \ldots, A, \\
\frac{\partial^2 \log f}{\partial \beta_j \partial \alpha_i} &= \begin{cases} \sum_t \frac{(d_{1t}+\phi)\phi\kappa_t e_{1t}\exp(\alpha_1+\beta_1\kappa_t)}{[e_{1t}\exp(\alpha_1+\beta_1\kappa_t)+\phi]^2} & \text{for } i = 1 \text{ and } j = 2, 3, \ldots, A, \\ -\sum_t \frac{(d_{it}+\phi)\phi\kappa_t e_{it}\exp(\alpha_i+\beta_i\kappa_t)}{[e_{it}\exp(\alpha_i+\beta_i\kappa_t)+\phi]^2} & \text{for } i = j \neq 1, \\ 0 & \text{for } i \neq j \text{ and } i \neq 1, \end{cases} \\
\frac{\partial^2 \log f}{\partial (\sigma_\kappa^2)^2} &= -\frac{1}{(\sigma_\kappa^2)^3} \sum_{t=2}^{T} [\kappa_t - \eta_t - \rho(\kappa_{t-1} - \eta_{t-1})]^2 + \frac{a_\kappa + 1}{(\sigma_\kappa^2)^2} - \frac{2b_\kappa}{(\sigma_\kappa^2)^3}, \\
\frac{\partial^2 \log f}{\partial (\sigma_\beta^2)^2} &= -\frac{\beta_1^2}{(\sigma_\beta^2)^3} - \frac{1}{(\sigma_\beta^2)^3} \sum_{x=2}^{A} \beta_x^2 + \frac{a_\beta + 1 + \frac{A-1}{2}}{(\sigma_\beta^2)^2} - \frac{2b_\beta}{(\sigma_\beta^2)^3} + \frac{1}{A(\sigma_\beta^2)^3},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \log f}{\partial \rho^2} &= -\frac{1}{\sigma_\kappa^2}\sum_{t=2}^{T}(\kappa_{t-1}-\eta_{t-1})^2 - \frac{1}{\sigma_\rho^2},\\
\frac{\partial^2 \log f}{\partial \phi^2} &= \sum_{x,t}\left\{-\frac{1}{e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi} + \frac{d_{xt}-e_{xt}\exp(\alpha_x+\beta_x\kappa_t)}{[e_{xt}\exp(\alpha_x+\beta_x\kappa_t)+\phi]^2}\right\} - \frac{a_\phi-1}{\phi^2} + \frac{AT}{\phi}\\
&\quad \sum_{x,t}\text{Trigamma}(d_{xt}+\phi) - AT\text{Trigamma}(\phi),
\end{aligned}
$$

(Trigamma function is simply the 2^{nd} derivative of the log of the gamma function.)

$$
\begin{aligned}
\frac{\partial^2 \log f}{\partial \psi_1^2} &= -\frac{(1-\rho)^2}{\sigma_\kappa^2}(T-1) - f_{11},\\
\frac{\partial^2 \log f}{\partial \psi_2^2} &= -\frac{1}{\sigma_\kappa^2}\sum_{t=2}^{T}[t(\rho-1)-\rho]^2 - f_{22},\\
\frac{\partial^2 \log f}{\partial \sigma_\beta^2 \partial \sigma_\kappa^2} &= 0,\\
\frac{\partial^2 \log f}{\partial \rho \partial \sigma_\kappa^2} &= -\frac{1}{(\sigma_\kappa^2)^2}\sum_{t=2}^{T}(\kappa_{t-1}-\eta_{t-1})[\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})],\\
\frac{\partial^2 \log f}{\partial \psi_1 \partial \sigma_\kappa^2} &= \frac{\rho-1}{(\sigma_\kappa^2)^2}\sum_{t=2}^{T}[\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})],\\
\frac{\partial^2 \log f}{\partial \psi_2 \partial \sigma_\kappa^2} &= \frac{1}{\sigma_\kappa^2}\sum_{t=2}^{T}[t(\rho-1)-\rho][\kappa_t-\eta_t-\rho(\kappa_{t-1}-\eta_{t-1})],\\
\frac{\partial^2 \log f}{\partial \phi \partial \sigma_\kappa^2} &= \frac{\partial^2 \log f}{\partial \rho \partial \sigma_\beta^2} = \frac{\partial^2 \log f}{\partial \psi_1 \partial \sigma_\beta^2} = \frac{\partial^2 \log f}{\partial \psi_2 \partial \sigma_\beta^2} = \frac{\partial^2 \log f}{\partial \phi \partial \sigma_\beta^2} = 0,\\
\frac{\partial^2 \log f}{\partial \psi_1 \partial \rho} &= -\frac{1}{\sigma_\kappa^2}\sum_{t=2}^{T}[(\kappa_t-\eta_t)-(2\rho-1)(\kappa_{t-1}-\eta_{t-1})],\\
\frac{\partial^2 \log f}{\partial \psi_2 \partial \rho} &= -\frac{1}{\sigma_\kappa^2}\sum_{t=2}^{T}[(t-1)(\kappa_t-\eta_t)-((2\rho-1)t-2\rho)(\kappa_{t-1}-\eta_{t-1})],\\
\frac{\partial^2 \log f}{\partial \phi \partial \rho} &= 0,\\
\frac{\partial^2 \log f}{\partial \psi_1 \partial \psi_2} &= \frac{1-\rho}{\sigma_\kappa^2}\sum_{t=2}^{T}[(\rho-1)t-\rho] - \frac{1}{2}(f_{12}+f_{21}),\\
\frac{\partial^2 \log f}{\partial \phi \partial \psi_1} &= 0,\\
\frac{\partial^2 \log f}{\partial \phi \partial \psi_2} &= 0,
\end{aligned}
$$

$$\frac{\partial^2 \log f}{\partial \sigma_\kappa^2 \partial \kappa_i} = \begin{cases} \frac{1}{(\sigma_\kappa^2)^2}[\kappa_i - \eta_i - \rho(\kappa_{i-1} - \eta_{i-1})] & \\ \quad -\frac{\rho}{(\sigma_\kappa^2)^2}[\kappa_{i+1} - \eta_{i+1} - \rho(\kappa_i - \eta_i)] & \text{for } i = 2, 3, \ldots, T-1, \\ \frac{1}{(\sigma_\kappa^2)^2}[\kappa_i - \eta_i - \rho(\kappa_{i-1} - \eta_{i-1})] & \text{for } i = T, \end{cases}$$

$$\frac{\partial^2 \log f}{\partial \sigma_\beta^2 \partial \kappa_i} = 0 \text{ for } i = 2, 3, \ldots, T,$$

$$\frac{\partial^2 \log f}{\partial \rho \partial \kappa_i} = \begin{cases} \frac{1}{\sigma_\kappa^2}[(\kappa_{i-1} - \eta_{i-1}) - 2(\rho(\kappa_i - \eta_i) + (\kappa_{i+1} - \eta_{i+1}))] & \text{for } i = 2, 3, \ldots, T-1, \\ \frac{1}{\sigma_\kappa^2}(\kappa_{i-1} - \eta_{i-1}) & \text{for } i = T, \end{cases}$$

$$\frac{\partial^2 \log f}{\partial \psi_1 \partial \kappa_i} = \begin{cases} \frac{(\rho-1)^2}{\sigma_\kappa^2} & \text{for } i = 2, 3, \ldots, T-1, \\ \frac{1-\rho}{\sigma_\kappa^2} & \text{for } i = T, \end{cases}$$

$$\frac{\partial^2 \log f}{\partial \psi_2 \partial \kappa_i} = \begin{cases} \frac{i(\rho-1)^2}{\sigma_\kappa^2} & \text{for } i = 2, 3, \ldots, T, \\ \frac{i-(i-1)\rho}{\sigma_\kappa^2} & \text{for } i = T, \end{cases}$$

$$\frac{\partial^2 \log f}{\partial \phi \partial \kappa_i} = \sum_x \frac{e_{xi}\beta_x \exp(\alpha_x + \beta_x \kappa_i)[d_{xi} - e_{xi}\exp(\alpha_x + \beta_x \kappa_i)]}{[e_{xi}\exp(\alpha_x + \beta_x \kappa_i) + \phi]^2} \text{ for } i = 2, 3, \ldots, T,$$

$$\frac{\partial^2 \log f}{\partial \sigma_\kappa^2 \partial \beta_i} = 0,$$

$$\frac{\partial^2 \log f}{\partial \sigma_\beta^2 \partial \beta_i} = -\frac{\beta_1}{(\sigma_\beta^2)^2} + \frac{\beta_i}{(\sigma_\beta^2)^2},$$

$$\frac{\partial^2 \log f}{\partial \rho \partial \beta_i} = \frac{\partial^2 \log f}{\partial \psi_1 \partial \beta_i} = \frac{\partial^2 \log f}{\partial \psi_2 \partial \beta_i} = 0,$$

$$\begin{aligned} \frac{\partial^2 \log f}{\partial \phi \partial \beta_i} = {} & \sum_t \kappa_t e_{it} \exp(\alpha_i + \beta_i \kappa_t) \frac{d_{it} - e_{it}\exp(\alpha_i + \beta_i \kappa_t)}{[e_{it}\exp(\alpha_i + \beta_i \kappa_t) + \phi]^2} \\ & - \sum_t \kappa_t e_{1t} \exp(\alpha_1 + \beta_1 \kappa_t) \frac{d_{1t} - e_{1t}\exp(\alpha_1 + \beta_1 \kappa_t)}{[e_{1t}\exp(\alpha_1 + \beta_1 \kappa_t) + \phi]^2}, \end{aligned}$$

$$\frac{\partial^2 \log f}{\partial \sigma_\kappa^2 \partial \alpha_i} = \frac{\partial^2 \log f}{\partial \sigma_\beta^2 \partial \alpha_i} = \frac{\partial^2 \log f}{\partial \rho \partial \alpha_i} = \frac{\partial^2 \log f}{\partial \psi_2 \partial \alpha_i} = \frac{\partial^2 \log f}{\partial \psi_2 \partial \alpha_i} = 0,$$

$$\frac{\partial^2 \log f}{\partial \phi \partial \alpha_i} = \sum_t \frac{e_{it}\exp(\alpha_i + \beta_i \kappa_t)[d_{it} - e_{it}\exp(\alpha_i + \beta_i \kappa_t)]}{[e_{it}\exp(\alpha_i + \beta_i \kappa_t) + \phi]^2}.$$

Subsequently, the Hessian matrix under the NBLC model can be computed by filling out the matrix entries using the corresponding second order partial derivatives provided above, that is

$$[H_{\text{NBLC}}(\boldsymbol{\theta})]_{ij} = \frac{\partial^2 \log f(\boldsymbol{\theta}|\boldsymbol{d})}{\partial \theta_j \partial \theta_i},$$

and noting that it should be a 247 × 247 symmetric square matrix $\left(\frac{\partial^2 \log f}{\partial \theta_j \partial \theta_i} = \frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j}\right)$.

Appendix E

Iterative Conditional Modes Search Algorithm for the NBLC Model

1. Set initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}_{-1}^{(0)}, \boldsymbol{\kappa}_{-1}^{(0)}, (\sigma_\kappa^{(0)})^2, (\sigma_\beta^{(0)})^2, \boldsymbol{\psi}^{(0)}, \rho^{(0)}, (\sigma_\mu^{(0)})^2)^\top$.

2. Set $i = 1$.

3. Numerically optimise the joint posterior distribution with respect to $\boldsymbol{\alpha}$, given the rest of the parameters stay fixed at current iteration to yield $\boldsymbol{\alpha}^{(i)}$ using "optim" function in R.

4. Numerically optimise the joint posterior distribution with respect to $\boldsymbol{\beta}_{-1}$, given the rest of the parameters stay fixed at current iteration to yield $\boldsymbol{\beta}_{-1}^{(i)}$ using "optim" function in R.

5. Numerically optimise the joint posterior distribution with respect to $\boldsymbol{\kappa}$, given the rest of the parameters stay fixed at current iteration to yield $\boldsymbol{\kappa}_{-1}^{(i)}$ using "optim" function in R.

6. Numerically optimise the joint posterior distribution with respect to ϕ given the rest of the parameters stay fixed at current iteration to yield $\phi^{(i)}$ using "optimize" function in R.

7. Compute

$$(\sigma_\kappa^{(i)})^2 = \frac{b_\kappa + \frac{1}{2}\sum_{t=2}^{T}[\kappa_t^{(i)} - \psi_1^{(i-1)} - \psi_2^{(i-1)}t - \rho(\kappa_{t-1}^{(i)} - \psi_1^{(i-1)} - \psi_2^{(i-1)}(t-1))]^2}{a_\kappa + \frac{T-1}{2} + 1}.$$

8. Compute

$$(\sigma_\beta^{(i)})^2 = \frac{b_\beta + \frac{1}{2}(\boldsymbol{\beta}_{-1}^{(i)} - \frac{1}{A}\mathbf{1}_{A-1})^\top (\boldsymbol{I}_{A-1} - \frac{1}{A}\boldsymbol{J}_{A-1})^{-1}(\boldsymbol{\beta}_{-1}^{(i)} - \frac{1}{A}\mathbf{1}_{A-1})}{a_\beta + \frac{A-1}{2} + 1}.$$

9. Compute

$$\rho^{(i)} = \frac{\sum_{t=2}^{T}(\kappa_t^{(i)} - \psi_1^{(i-1)} - \psi_2^{(i-1)}t)(\kappa_{t-1}^{(i)} - \psi_1^{(i-1)} - \psi_2^{(i-1)}(t-1))}{\sum_{t=2}^{T}(\kappa_{t-1}^{(i)} - \psi_1^{(i-1)} - \psi_2^{(i-1)}(t-1))^2 + \frac{(\sigma_\kappa^{(i)})^2}{\sigma_\rho^2}}.$$

10. Finally, compute $\boldsymbol{\psi}^{(i)} = (\boldsymbol{\psi}^*)^{(i)}$, where

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

   $\boldsymbol{R}^{(i)}$, $\boldsymbol{Q}^{(i)}$ are $\boldsymbol{R}$ and $\boldsymbol{Q}$ evaluated at $\rho^{(i)}$.

11. These complete a full iteration to produce $\boldsymbol{\theta}^{(1)}$. Set $i = i + 1$, repeat steps 3-10 until convergence, i.e. $\boldsymbol{\theta}^{(n)}$ such that $|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}^{(n-1)}| <$ tol, where $|\cdot|$ is some distance measure and tol is the tolerance level set by the user.

Upon convergence, the last iteration $\boldsymbol{\theta}^{(n)}$ can then serve as the joint mode of our posterior distribution.

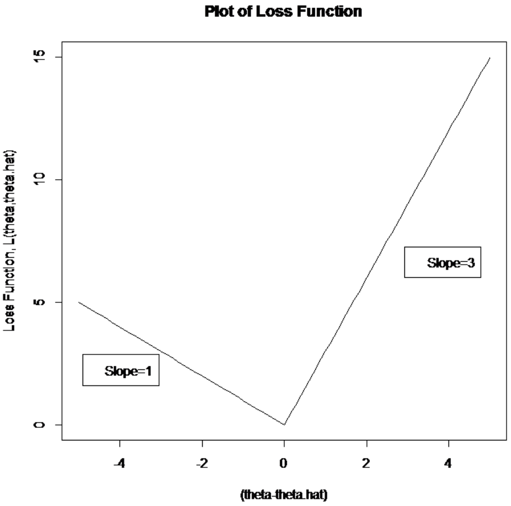
Appendix F

Bayesian Point Estimation

Suppose that the loss function penalises underestimation three times more than overestimation (linearly),

$$L(\theta,\hat{\theta}) = \begin{cases} 3(\theta-\hat{\theta}) & \text{for } \theta > \hat{\theta} \\ -(\theta-\hat{\theta}) & \text{for } \theta \leq \hat{\theta} \end{cases} ,$$

where $\hat{\theta}$ is the point estimator of the parameter θ. Or graphically,



The posterior expected loss can be written down as

$$\begin{aligned}
\mathbb{E}_\theta[L(\theta,\hat{\theta})|\boldsymbol{d}] &= \int_{-\infty}^{\infty} L(\theta,\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta \\
&= \int_{-\infty}^{\hat{\theta}} L(\theta,\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} L(\theta,\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta \\
&= \int_{-\infty}^{\hat{\theta}} [-(\theta-\hat{\theta})] f(\theta|\boldsymbol{d})\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} 3(\theta-\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta \\
&= 3\int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta - \int_{-\infty}^{\hat{\theta}} (\theta-\hat{\theta}) f(\theta|\boldsymbol{d})\mathrm{d}\theta.
\end{aligned}$$

Using Leibniz's rule for differentiation under integral sign that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\int_{a(\theta)}^{b(\theta)} f(\theta,x)\mathrm{d}x\right] = \int_{a(\theta)}^{b(\theta)} \frac{\partial f(\theta,x)}{\partial\theta}\mathrm{d}x + f(\theta,b(\theta))\cdot\frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta} - f(\theta,a(\theta))\cdot\frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta},$$

the first derivative of the posterior expected loss function is then given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\mathbb{E}_\theta[L(\theta,\hat{\theta})|\boldsymbol{d}] &= -3\int_{\hat{\theta}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta - 3(\hat{\theta}-\hat{\theta})f(\hat{\theta}|\boldsymbol{d}) + \int_{-\infty}^{\hat{\theta}} f(\theta|\boldsymbol{d})\mathrm{d}\theta - (\hat{\theta}-\hat{\theta})f(\hat{\theta}|\boldsymbol{d}) \\
&= -3\int_{\hat{\theta}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta + \int_{-\infty}^{\hat{\theta}} f(\theta|\boldsymbol{d})\mathrm{d}\theta.
\end{aligned}$$

Now set $\frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\mathbb{E}[L(\theta,\hat{\theta})|\boldsymbol{d}] = 0$ to obtain the optimal point estimate of $\hat{\theta}$, $\hat{\theta}_{opt}$

$$\int_{-\infty}^{\hat{\theta}_{opt}} f(\theta|\boldsymbol{d})\mathrm{d}\theta = 3\int_{\hat{\theta}_{opt}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta.$$

Note that

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta = \int_{-\infty}^{\hat{\theta}_{opt}} f(\theta|\boldsymbol{d})\mathrm{d}\theta + \int_{\hat{\theta}_{opt}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta = 1 \\
\Rightarrow\ & 3\int_{\hat{\theta}_{opt}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta + \int_{\hat{\theta}_{opt}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta = 1 \\
\Rightarrow\ & \int_{\hat{\theta}_{opt}}^{\infty} f(\theta|\boldsymbol{d})\mathrm{d}\theta = 0.25 \text{ and } \int_{-\infty}^{\hat{\theta}_{opt}} f(\theta|\boldsymbol{d})\mathrm{d}\theta = 1 - 0.25 = 0.75,
\end{aligned}$$

which clearly indicates that $\hat{\theta}_{opt}$ is the 75^{th} percentile of the posterior distribution of θ. To check that the point estimate is indeed a minimum, the second derivative is computed,

$$\frac{\mathrm{d}^2}{\mathrm{d}\hat{\theta}^2}\mathbb{E}_\theta[L(\theta,\hat{\theta})|\boldsymbol{d}] = [3f(\hat{\theta}|\boldsymbol{d}) + f(\hat{\theta}|\boldsymbol{d})] = 4f(\hat{\theta}|\boldsymbol{d}),$$

which is obviously positive when evaluated at the 75^{th} percentile of the posterior distribution since a proper pdf has non-negative values (or zero in some special cases but has to be positive when evaluated at 75th percentile at least for a continuous function).

Appendix G

Bartlett's Paradox

Bartlett's paradox (sometimes referred to as Lindley's paradox) arises whenever we wish to assert arbitrarily diffuse or improper uniform priors in one or more of the models during model comparison. Following is an example to illustrate the paradox. Consider two models for the response variable, Y, with only one available observation, y,

1. $$M_1 : Y \sim \text{Poisson}(\mu = 1).$$

   That is, the model is deterministic (parameter is constant, not random).

2. $$M_2 : Y \sim \text{Poisson}(\mu) \text{ with } \mu \sim \text{Gamma}(\alpha, \beta),$$

   where α and β are constants. (Sometimes, we impose $\alpha = \beta$ so that $\mathbb{E}(\mu) = 1$, making the two models similar in fitted values, but second models with heavier tail as its mean is allowed to be random. However, this is not necessarily needed for the paradox to occur). The posterior distribution of μ is given by

   $$\begin{aligned} f(\mu|y) &\propto f(y|\mu)f(\mu) \\ &= \frac{e^{-\mu}\mu^y}{y!} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu} \\ &\propto \mu^{y+\alpha-1}e^{-(1+\beta)\mu} \\ &\propto \text{Gamma}(y+\alpha, 1+\beta). \end{aligned}$$

The marginal likelihood of each model is computed as follows:

1. $$f(y|M_1) = f(y|\mu = 1) = \frac{e^{-1}1^y}{y!} = \frac{e^{-1}}{y!}.$$

2.

$$\begin{aligned}
f(y|M_2) &= \int_0^\infty f(y|\mu, M_2) f(\mu|M_2)\mathrm{d}\mu \\
&= \int_0^\infty \frac{e^{-\mu}\mu^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu} \\
&= \frac{\beta^\alpha}{y!\Gamma(\alpha)}\int_0^\infty \mu^{\alpha+y-1}e^{-(1+\beta)\mu}\mathrm{d}\mu \\
&= \frac{\beta^\alpha}{y!\Gamma(\alpha)}\frac{\Gamma(\alpha+y)}{(1+\beta)^{\alpha+y}}\underbrace{\int_0^\infty \frac{(1+\beta)^{y+\alpha}}{\Gamma(y+\alpha)}\mu^{y+\alpha-1}e^{-(1+\beta)\mu}\mathrm{d}\mu}_{=1} \\
&= \frac{\beta^\alpha\Gamma(y+\alpha)}{y!\Gamma(\alpha)(1+\beta)^{y+\alpha}}.
\end{aligned}$$

A common practice to obtain a non-informative prior for gamma distribution, is to assume that $\alpha = \beta = 0$ (or more formally, $\alpha \to 0$ and $\beta \to 0$), which then implies that $\mathrm{Var}(\mu) \to \infty$. As a result, for $y \neq 0$,

$$f(y|M_2) \to \frac{0^0\Gamma(0+y)}{y!\Gamma(0)(0+1)^{0+y}} = \frac{1\times\Gamma(y)}{y!\Gamma(0)1^y} = \frac{1}{y\Gamma(0)} = \frac{1}{\infty} = 0.$$

Hence, the Bayes factor in favour of M_1 is

$$\frac{f(y|M_1)}{f(y|M_2)} = \frac{\frac{e^{-1}}{y!}}{0} = \infty,$$

which is infinitely large. This implies that the simpler model, M_1, is always favoured **regardless** of the observation y, as long as we choose a sufficiently diffuse prior. Clearly, this property violates the fundamental principle of Bayesian analysis because ultimately, we would like the data to express their own preference for models when the priors are non-informative. This example demonstrated how the data is ignored during the computation of Bayes factor in model selection, and always heavily penalize the more complicated model. This phenomenon is known as the Bartlett's Paradox in model selection, which occurs whenever the specified priors are too "flat" relative to the likelihood.

The result is not specific to this example, and in fact, is not restricted to the Poisson likelihood. The following is an example of the paradox for normal distributions (see for example Shafer, 1979). Consider again, two models for Y, with only one available sample, y,

1.

$$M_1 : Y \sim N(\theta = 0, \nu^2).$$

2.

$$M_2 : Y \sim N(\theta, \nu^2) \text{ with } \theta \sim N(\mu, \sigma^2),$$

where the variance, ν^2 is assumed to be known.

The marginal likelihoods are:

1.

$$f(y|M_1) = f(y|\theta = 0) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-y^2/2\nu^2}.$$

2.

$$\begin{aligned} f(y|M_2) &= \int_{-\infty}^{\infty} f(y|\theta, M_2) f(\theta|M_2) \mathrm{d}\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\nu^2}} \exp\left[-\frac{(y-\theta)^2}{2\nu^2}\right] \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \mathrm{d}\theta \\ &\vdots \\ &= \frac{1}{\sqrt{2\pi}\nu} \sqrt{\frac{1}{\sigma^2\nu^2} + \frac{1}{(\sigma^2)^2}} \exp\left[-\frac{1}{2}\left(\frac{y^2}{\nu^2} + \frac{\mu^2}{\sigma^2}\right) + \frac{\left(\frac{y}{\nu^2} + \frac{\mu}{\sigma^2}\right)^2}{\left(\frac{1}{\nu^2} + \frac{1}{\sigma^2}\right)^3}\right]. \end{aligned}$$

As $\sigma^2 \to \infty$ relative to a finite ν^2, we have

$$\begin{cases} \sqrt{\frac{1}{\sigma^2\nu^2} + \frac{1}{(\sigma^2)^2}} \to \sqrt{0+0} = 0, \\ \exp\left[-\frac{1}{2}\left(\frac{y^2}{\nu^2} + \frac{\mu^2}{\sigma^2}\right)\right] \to \exp\left[-\frac{1}{2}\left(\frac{y^2}{\nu^2} + 0\right)\right] = \exp\left(-\frac{y^2}{2\nu^2}\right), \\ \exp\left[\frac{\left(\frac{y}{\nu^2}+\frac{\mu}{\sigma^2}\right)^2}{2(\frac{1}{\nu^2}+\frac{1}{\sigma^2})^3}\right] \to \exp\left[\frac{(\frac{y}{\nu^2}+0)^2}{2(1+0)^3}\right] = \exp\left(\frac{y^2}{2\nu^2}\right). \end{cases}$$

Therefore,

$$f(y|M_2) \to \frac{1}{\sqrt{2\pi}\nu} \times 0 \times \exp\left(-\frac{y^2}{2\nu^2}\right) \times \exp\left(\frac{y^2}{2\nu^2}\right) = 0,$$

as $\sigma^2 \to \infty$. The Bayes factor in favour of M_1 is again

$$\frac{f(y|M_1)}{f(y|M_2)} = \frac{\frac{1}{\sqrt{2\pi\nu^2}} e^{-y^2/2\nu^2}}{0} = \infty,$$

irrespective of the data observed.

Fundamentally, Bartlett's paradox occurs whenever a prior that is "flat" relative to the data likelihood is specified. Effectively, the prior is giving negligible probability to a region where there is non-negligible likelihood, irrespective of the data collected. This is explained in a more general set up by Dellaportas et al. (2012), where the occurrence of the Bartlett's paradox is illustrated by making the prior parameter dispersion (the scale multiplicative factor of the prior variance) tending to infinity.

Appendix H

The Test Quantities

An expression of the test quantity under each model is given by

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

Appendix I

Sub-Hessian Matrix of the Joint Posterior Distribution of the NBLL Model

The log joint posterior density of the NBLL model that is dependent of κ_t is

$$\begin{aligned}\log f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d}) &= \text{constant} + \sum_{x,t} d_{xt}(\alpha_x + \beta_x t + \kappa_t) \\ &\quad - \sum_{x,t}(d_{xt}+\phi)\log[e_{xt}\exp(\alpha_x+\beta_x t+\kappa_t)+\phi] - \frac{1}{4\sigma_\kappa^2}\boldsymbol{\kappa}_{-1,2}^\top \boldsymbol{D}^{-1}\boldsymbol{\kappa}_{-1,2},\end{aligned}$$

where the constant involves terms that are independent of κ_t, κ_1 and κ_2 are computed from $\boldsymbol{\kappa}_{-1,2}$ through Equation (5.3). Suppose that the matrix $\boldsymbol{ID}_{(T-2)\times(T-2)}$ denotes the inverse of matrix $\boldsymbol{D}$, $\boldsymbol{D}^{-1}$, and it has general element

$$[\boldsymbol{ID}]_{ij}.$$

Then, the ij^{th} element of $\boldsymbol{H}^{\kappa}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})$ for $i,j=1,\ldots,T-2$ is given by

$$[\boldsymbol{H}^{\kappa}_{\text{NBLL}}(\boldsymbol{\theta}_{\text{NBLL}})]_{ij} = \frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{NBLL}}|\boldsymbol{d})}{\partial\kappa_{j+2}\partial\kappa_{i+2}},$$

where for $i \neq j$, we have

$$\begin{aligned}\frac{\partial^2 \log f}{\partial\kappa_{j+2}\kappa_{i+2}} &= -i\times j\sum_{x=1}^{A}\frac{(d_{x1}+\phi)\phi e_{x1}\exp(\alpha_x+\beta_x t_1+\kappa_1)}{[e_{x1}\exp(\alpha_x+\beta_x t_1+\kappa_1)+\phi]^2} \\ &\quad -(i+1)(j+1)\sum_{x=1}^{A}\frac{(d_{x2}+\phi)\phi e_{x2}\exp(\alpha_x+\beta_x t_2+\kappa_2)}{[e_{x2}\exp(\alpha_x+\beta_x t_2+\kappa_2)+\phi]^2} - \frac{1}{2\sigma_\kappa^2}[\boldsymbol{ID}]_{ij},\end{aligned}$$

and for $i = j$, we have

$$\begin{aligned}
\frac{\partial^2 \log f}{\partial \kappa_{j+2} \kappa_{i+2}} &= -\sum_{x=1}^{A} \frac{(d_{x\,i+2} + \phi)\phi e_{x\,i+2} \exp(\alpha_x + \beta_x t_{i+2} + \kappa_{i+2})}{[e_{x\,i+2} \exp(\alpha_x + \beta_x t_{i+2} + \kappa_{i+2}) + \phi]^2} \\
&\quad -i^2 \sum_{x=1}^{A} \frac{(d_{x1} + \phi)\phi e_{x1} \exp(\alpha_x + \beta_x t_1 + \kappa_1)}{[e_{x1} \exp(\alpha_x + \beta_x t_1 + \kappa_1) + \phi]^2} \\
&\quad -(i+1)^2 \sum_{x=1}^{A} \frac{(d_{x2} + \phi)\phi e_{x2} \exp(\alpha_x + \beta_x t_2 + \kappa_2)}{[e_{x2} \exp(\alpha_x + \beta_x t_2 + \kappa_2) + \phi]^2} - \frac{1}{2\sigma_\kappa^2} [\boldsymbol{ID}]_{ii}.
\end{aligned}$$

Appendix J

Kurtosis of the Bessel Distribution

Here, we derive the kurtosis of the Bessel distribution. In particular, for $W = UV$ with $U \sim N(0, \sigma_u^2)$ and $V \sim N(0, \sigma_v^2)$, we want to compute

$$\mathrm{kurt}[W] = \frac{\mathbb{E}[(W - \mathbb{E}(W))^4]}{(\mathbb{E}[(W - \mathbb{E}(W))^2])^2}, \tag{J.1}$$

where kurt[ ] denotes the kurtosis operator. First, notice that due to independence,

$$\mathbb{E}(W) \quad = \quad \mathbb{E}(UV) = \mathbb{E}(U)\mathbb{E}(V) = 0 \times 0 = 0.$$

Equation (J.1) then reduces to

$$\begin{aligned} \mathrm{kurt}[W] &= \frac{\mathbb{E}[W^4]}{(\mathbb{E}[W^2])^2} \\ &= \frac{\mathbb{E}[U^4 V^4]}{(\mathbb{E}[U^2 V^2])^2} \\ &= \frac{\mathbb{E}[U^4]\mathbb{E}[V^4]}{(\mathbb{E}[U^2]\mathbb{E}[V^2])^2} \qquad \text{(since } U \text{ and } V \text{ are independent).} \end{aligned}$$

Knowing that the r^{th} cumulant of a normal distribution is 0 for $r > 2$ and using the relationship between the cumulants and central moments (see for example Garthwaite et al., 2002), we have

$$\mathbb{E}[U^4] = \mathbb{E}[(U - \mathbb{E}[U])^4] = 0 + 3(\sigma_u^2)^2 = 3\sigma_u^4.$$

Similarly,

$$\mathbb{E}[V^4] = \mathbb{E}[(V - \mathbb{E}[V])^4] = 0 + 3(\sigma_v^2)^2 = 3\sigma_v^4.$$

Of course, we also have

$$\mathbb{E}[U^2] = \mathrm{Var}[U] + (\mathbb{E}[U])^2 = \sigma_u^2 + 0^2 = \sigma_u^2,$$

and

$$\mathbb{E}[V^2] = \sigma_v^2 + 0^2 = \sigma_v^2.$$

Therefore,

$$\mathrm{kurt}[W] \quad = \quad \frac{3\sigma_u^4 \times 3\sigma_v^4}{(\sigma_u^2)^2 \times (\sigma_v^2)^2} = 9,$$

which is a constant, irrespective of the values of σ_u^2 and σ_v^2.

Appendix K

Deriving Time Series Priors for κ_t and γ_c with Constraints

The projection model for κ_t for the Negative-Binomial Log-linear (NBLL) model is an AR(1) model without drift:

$$\begin{cases} \kappa_t = \rho\kappa_{t-1} + \epsilon_t & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \epsilon_1 \end{cases},$$

where $\epsilon_t \sim N(0, \sigma_\kappa^2)$ are independent Gaussian errors. This model can be expressed multivariately as

$$\boldsymbol{\kappa} = \boldsymbol{P}\boldsymbol{\kappa} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_T)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)^\top$ and

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \rho & 0 & & & \vdots \\ 0 & \rho & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho & 0 \end{pmatrix}_{T\times T}.$$

Rearranging the above equation and using the result of linear transformation on a normal distribution, we see that

$$\begin{aligned} & (\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{\kappa} = \boldsymbol{\epsilon} \sim N_T(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{I}_T) \\ \Rightarrow\quad & \boldsymbol{\kappa} \sim N_T((\boldsymbol{I}_T - \boldsymbol{P})^{-1} \times \boldsymbol{0}, (\boldsymbol{I}_T - \boldsymbol{P})^{-1}(\sigma_\kappa^2 \boldsymbol{I}_T)(\boldsymbol{I}_T - \boldsymbol{P})^{-\top}) \\ \Rightarrow\quad & \boldsymbol{\kappa} \sim N_T(\boldsymbol{0}, \sigma_\kappa^2[(\boldsymbol{I}_T - \boldsymbol{P})^\top(\boldsymbol{I}_T - \boldsymbol{P})]^{-1}) \\ \Rightarrow\quad & \boldsymbol{\kappa} \sim N_T(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{Q}^{-1}), \end{aligned}$$

where $\boldsymbol{Q} = (\boldsymbol{I}_T - \boldsymbol{P})^\top(\boldsymbol{I}_T - \boldsymbol{P})$. Defining

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & 3 & \cdots & T-1 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{T\times T},$$

we have

$$\boldsymbol{A\kappa} = \left(\sum_t \kappa_t, \sum_t t\kappa_t, \kappa_3, \ldots, \kappa_T\right)^\top \sim N_T(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{B}),$$

where $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^\top$. Finally, using the conditional property of a normal distribution (see for example Kotz et al., 2000) and suppose $\boldsymbol{B}$ is partitioned such that

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11_{2\times 2}} & \boldsymbol{B}_{12_{2\times(T-2)}} \\ \boldsymbol{B}_{21_{(T-2)\times 2}} & \boldsymbol{B}_{22_{(T-2)\times(T-2)}} \end{pmatrix},$$

it can be shown that the constrained prior on the κ parameters is

$$\boldsymbol{\kappa}_{-1,2} \left| \left(\sum \kappa_t, \sum t\kappa_t\right)^\top = (0,0)^\top \sim N_{T-2}(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{D}), \right.$$

where $\boldsymbol{\kappa}_{-1,2} = (\kappa_3, \ldots, \kappa_T)^\top$ and $\boldsymbol{D} = [\boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}]$, as required.

The time series prior of γ_c for the NBLL-C model can be derived in a similar fashion as above. Recall that the projection model for NBLL-C model is an ARIMA(1,1,0) model:

$$\begin{cases} (\gamma_c - \gamma_{c-1}) = \rho_\gamma(\gamma_{c-1} - \gamma_{c-2}) + \epsilon_c^\gamma, & \text{for } c = 3, \ldots, C, \\ \gamma_2 - \gamma_1 = \frac{1}{\sqrt{1-\rho_\gamma^2}}\epsilon_2^\gamma, \\ \gamma_1 = 100\epsilon_1^\gamma, \end{cases}$$

where ρ_γ is the auto-regressive coefficient and $\epsilon_c^\gamma \overset{\text{ind}}{\sim} N(0, \sigma_\gamma^2)$ are random errors. By defining the matrix

$$\boldsymbol{R}^\gamma = \begin{pmatrix} 1/100 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -\sqrt{1-\rho_\gamma^2} & \sqrt{1-\rho_\gamma^2} & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \rho_\gamma & -(1+\rho_\gamma) & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & \rho_\gamma & -(1+\rho_\gamma) & 1 & 0 & \cdots & 0 \end{pmatrix}_{C\times C},$$

the ARIMA(1,1,0) prior for γ_c can be expressed multivariately as

$$\boldsymbol{R}^\gamma \boldsymbol{\gamma} = \boldsymbol{\epsilon}^\gamma,$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_C)^\top$ and $\boldsymbol{\epsilon}^\gamma = (\epsilon_1^\gamma, \ldots, \epsilon_C^\gamma)^\top$. Again, rearranging the above equation using linear properties of a normal distribution, we obtain

$$\begin{aligned}
& \boldsymbol{R}^\gamma \boldsymbol{\gamma} = \boldsymbol{\epsilon}^\gamma \sim N_C(\boldsymbol{0}, \sigma_\gamma^2 \boldsymbol{I}_C) \\
\Rightarrow \quad & \boldsymbol{\gamma} \sim N_C((\boldsymbol{R}^\gamma)^{-1} \times \boldsymbol{0}, (\boldsymbol{R}^\gamma)^{-1} \times (\sigma_\gamma^2 \boldsymbol{I}_C) \times (\boldsymbol{R}^\gamma)^{-\top}) \\
\Rightarrow \quad & \boldsymbol{\gamma} \sim N_C(\boldsymbol{0}, \sigma_\gamma^2 (\boldsymbol{Q}^\gamma)^{-1}),
\end{aligned}$$

where $\boldsymbol{Q}^\gamma = (\boldsymbol{R}^\gamma)^\top \boldsymbol{R}^\gamma$. By defining

$$\boldsymbol{A}^\gamma = \begin{array}{l} \text{r}ow1 \\ \text{r}ow2 \\ \text{r}ow3 \\ \text{r}ow4 \\ \text{r}ow5 \\ \vdots \\ \text{r}ow73 \\ \text{r}ow74 \\ \vdots \\ \text{r}owC \end{array} \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \\
1 & 2 & 3 & 4 & 5 & \cdots & \cdots & \cdots & \cdots & C \\
1^2 & 2^2 & 3^2 & 4^2 & 5^2 & \cdots & \cdots & \cdots & \cdots & C^2 \\
0 & 1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 1 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\
0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & & & & & \ddots & \ddots & \ddots & & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 1 & 0
\end{pmatrix}_{C \times C},$$

we have

$$\boldsymbol{A}^\gamma \boldsymbol{\gamma} = \left(\sum \gamma_c, \sum c\gamma_c, \sum c^2 \gamma_c, (\boldsymbol{\gamma}')^\top \right)^\top \sim N_C(\boldsymbol{0}, \sigma_\gamma^2 \boldsymbol{B}^\gamma),$$

where $\boldsymbol{\gamma}' = (\gamma_2, \ldots, \gamma_{71}, \gamma_{73}, \ldots, \gamma_{C-1})^\top$ and $\boldsymbol{B}^\gamma = \boldsymbol{A}^\gamma (\boldsymbol{Q}^\gamma)^{-1} (\boldsymbol{A}^\gamma)^\top$. Now suppose that $\boldsymbol{B}^\gamma$ is partitioned such that

$$\boldsymbol{B}^\gamma = \begin{pmatrix} \boldsymbol{B}^\gamma_{11_{3\times 3}} & \boldsymbol{B}^\gamma_{12_{3\times(C-3)}} \\ \boldsymbol{B}^\gamma_{21_{(C-3)\times 3}} & \boldsymbol{B}^\gamma_{22_{(C-3)\times(C-3)}} \end{pmatrix},$$

it can then be shown (using conditional property of a normal distribution) that the constrained prior for the cohort parameters is

$$\boldsymbol{\gamma}' \bigg| \left(\sum \gamma_c, \sum c\gamma_c, \sum c^2 \gamma_c \right)^\top = (0,0,0)^\top \sim N_{C-3}(\boldsymbol{0}, \sigma_\gamma^2 \boldsymbol{D}^\gamma),$$

where $\boldsymbol{D}^\gamma = [\boldsymbol{B}^\gamma_{22} - \boldsymbol{B}^\gamma_{21} (\boldsymbol{B}^\gamma_{11})^{-1} \boldsymbol{B}^\gamma_{12}]$ as required. The three cohort components removed from the parameter space can be deterministically computed from the rest by solving the equations for the constraints, $\sum \gamma_c = \sum c\gamma_c = \sum c^2\gamma$, leading to

$$\begin{aligned}
\gamma_1 &= \frac{1}{71 \times (C-1)} \sum_{c \neq 1,72,C} (c-72)(C-c)\gamma_c \ , \\
\gamma_{72} &= -\frac{1}{69 \times 71} \sum_{c \neq 1,72,C} (C-c)(c-1)\gamma_c \ , \\
\gamma_C &= \frac{1}{69 \times (C-1)} \sum_{c \neq 1,72,C} (c-1)(72-c)\gamma_c \ .
\end{aligned}$$

For computational stability, $\{\gamma_1, \gamma_{72}, \gamma_C\}$ are chosen to be removed from the parameter space instead of other combinations such as $\{\gamma_1, \gamma_2, \gamma_3\}$. This is to ensure the positive definiteness of the matrix $\boldsymbol{D}^\gamma$ for any value of ρ^γ. The exact reason behind this phenomenon appears to be unclear at present, but can be informally explained as follows. Intuitively, the points to be removed (due to the constraints) can be thought of as the "hinges" upon where the constraints are applied. The more spread out the "hinges" are across the cohort parameters, the more stable the computation involved because the correlations induced by each of the constraints are less conflicting as the constraints act at points that are far apart. In other words, the method by which the constraints are implemented does not influence the fitted values ultimately, but it affects the way by which exploration of posterior distributions is carried out by inducing different correlation structures, and hence, the different computational stability and efficiency. On a related matter, each of rows 1-3 of matrix $\boldsymbol{A}^\gamma$ is standardized by subtracting their corresponding row mean and their row standard deviation for similar purposes (not shown in the matrix representation above).

References

Alders, M. and J. de Beer (2005). An Expert Knowledge Approach to Stochastic Mortality Forecasting in the Netherlands. *Stockholm: Swedish Social Insurance Agency, In: N. Keilman (ed.), Perspectives on mortality forecasting. II. Probabilistic models. Social Insurance Studies 2*, 39–64.

Alho, J. M. (1991). Effect of aggregation on the estimation of trend in mortality. *Mathematical Population Studies,* **3**(1), 53–67.

Alho, J. M. (1992a). Estimating the strength of expert judgement: The case of US mortality forecasts. *Journal of Forecasting, 11*(2), 157–167.

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

Alho, J. M. and B. D. Spencer (1990). Error models for official mortality forecasts. *Journal of the American Statistical Association, 85*(411), 609–616.

Altomare, D. F., G. Serio, O. C. Pannarale, L. Lupo, N. Palasciano, V. Memeo, and M. Rubino (1990). Prediction of mortality by logistic regression analysis in patients with postoperative enterocutaneous fistulae. *Br J Surg. 77*(4), 450–453.

Andreozzi, L., B. M. T., and N. Arnesi (2008). The Lee Carter Method for Estimating and Forecasting Mortality: An Application for Argentina.

Azose, J. J., H. Ševčíková, and A. E. Raftery (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences, 113*(23), 6460–6465.

Balanda, K. P. (1987). Kurtosis Comparisons of the Cauchy and Double Exponential Distributions. *Communications in Statistics-Theory and Methods, 16*(2), 579–592.

Bartlett, M. S. (1957). A comment on d. v. lindley's statistical paradox. *Biometrika 44*(3/4), 533–534.

Bayarri, M. J. and J. O. Berger (2000a). P values for composite null models. *Journal of the American Statistical Association, 95*(452), 1127–1142.

Bayarri, M. J. and J. O. Berger (2000b). Rejoinder. *Journal of the American Statistical Association, 95*(452), 1168–1170.

Bell, W. and B. Monsell (1991). Using principal components in time series modelling and forecasting of age-specific mortality rates. *American Statistical Association, Proc. Social Statistic Section*, 154–159.

Berger, J. O. and L. R. Pericchi (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association, 91*(433), 109–122.

Berger, J. O. and L. R. Pericchi (2001). Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series 38*, 135–207.

Booth, H., J. Maindonald, and L. Smith (2002). Applying lee-carter under conditions of variable mortality decline. *Population Studies: A Journal of Demography, 56*(3), 325–336.

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

Börger, M. and C. Aleksic, M (2014). Coherent Projections of Age, Period, and Cohort Dependent Mortality Improvements. A paper presented at Living to 100 Symposium, Orlando, Fla..

Brilinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics, 42*(4), 693–734.

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.

Butler, W. J. and R. M. Park (1987). Use of Logistic Regression model for The Analysis of Proportionate Mortality Data. *American Journal of Epidemiology,* **125**(3), 515–523.

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

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

Cairns, A., D. Blake, K. Dowd, G. Coughlan, D. Epstein, A. Ong, and I. Balevich (2007). A Quantitative Comparison of Stochastic Mortality Models Using Data from England & Wales and the United States. *North American Actuarial Journal,* **13**(1), 1–35.

Cairns, A., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah (2010). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics,* **48**(3), 355–367.

Cairns, A. J. G., D. Blake, K. Dowd, and A. R. Kessler (2016). Phantoms never die: living with unreliable population data. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **179**(4), 975–1005.

Carlin, B. P. and S. Chib (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Ser. B,* **57**, 473–484.

Carlin, B. P. and T. A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.). Chapman and Hall CRC Press.

Carriere, J. F. (1992). Parametric Models for Life Tables. *Transactions of the Society of Actuaries,* **44**, 77–99.

Caselli, G. and V. Egidi (1992). New frontiers in survival. Presented at the conference: Human resources in Europe at the dawn of the 21-st century. 27-29 November, 1991. Luxembourg: Eurostat, 91–120.

Chatfield, C. (1984). *The Analysis of Time Series: An Introduction* (3rd ed.). Chapman and Hall Ltd.

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

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association,* **90**(432), 1313–1321.

Chib, S. and I. Jeliazkov (2001). Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association,* **96**(453), 270–281.

Coale, A. and E. E. Kisker (1990). Defects in data on old age mortality in the united states: New procedures for calculating approximately accurate mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum,* **4**(1), 1–31.

Continuous Mortality Investigation Bureau (2004). Projecting Future Mortality: A Discussion Paper. CMI Working Paper no. 3, London: Institute and Faculty of Actuaries.

Continuous Mortality Investigation Bureau (2006). Stochastic Projection Methodologies: Further Progress and P-Spline Model Features, Example Results and Implications. CMI Working Paper no. 20, London: Institute and Faculty of Actuaries.

Continuous Mortality Investigation Bureau (2007). Stochastic Projection Methodologies: Lee-Carter Model Features, Example Results and Implications. CMI Working Paper no. 25, London: Institute and Faculty of Actuaries.

Continuous Mortality Investigation Bureau (2009). A Prototype Mortality Projections Model: Part One - An Outline of the Proposed Approach. CMI Working Paper no. 38, London: Institute and Faculty of Actuaries.

Continuous Mortality Investigation Bureau (2016). CMI Mortality Projections Model consultation. CMI Working Paper no. 90, London: Institute and Faculty of Actuaries.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* **34**(2), 187–220.

Craig, C. C. (1936). On the Frequency Function of xy. *Annals of Mathematical Statistics,* **7**, 1–15.

Cramer, H. and H. Wold (1935). Mortality variations in sweeden. *Scandinavian Actuarial Journal,* (18), 161–241.

Crimmins, E. M. (1981). The Changing Pattern of American Mortality Decline, 1940-77, and Its Implication for the Future. *Population and Development Review,* **7**(2), 229–254.

Currie, I. D. (2012). Forecasting with the age-period-cohort model? *Proceedings of* 27^{th} *International Workshop on Statistical Modelling, Prague*, 87–92.

Currie, I. D., M. Durban, and P. H. C. Eilers (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling,* **4**(4), 279–298.

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

de Beer, J. (2000). Dealing with Uncertainty in Population Forecasting. *Voorburg: Department of Population, Statistics Netherlands*.

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2002). On bayesian model and variable selection using MCMC. *Applied Stochastic Models in Business and Industry, 12*(1), 27–36.

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2012). Joint Specification of Model Space and Parameter Space Prior Distributions. *Statistical Science, 27*(2), 232–246.

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.

Diaconis, P. and D. Ylvisaker (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics, 7*(2), 269–281.

Frome, E. L. (1983). The Analysis of Rates Using Poisson Regression Models. *Biometrics,* **39**, 665–674.

Fu, W. J. (2008). A smoothing cohort model in age-period-cohort analysis with applications to homicide arrest rates and lung cancer mortality rates. *Sociological Methods and Research, 36*(3), 327–361.

Garthwaite, P. H., I. T. Jolliffe, and B. Jones (2002). *Statistical Inference* (2$^{\text{nd}}$ ed.). Oxford: Oxford University Press.

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. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics, 7*, 2595–2602.

Gelman, A. and X. L. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13*(2), 163–185.

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.

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6**(6), 721–741.

Giacometti, R., M. Bertocchi, S. T. Rachev, and F. J. Fabozzi (2012). A Comparison of the Lee-Carter Model and AR-ARCH Model for Forecasting Mortality Rates. *Insurance: Mathematics and Economics,* **50**(1), 85–93.

Girosi, F. and G. King (2007). Understanding the lee-carter mortality forecasting method.

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

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality etc. *Phil. Trans. Roy. Soc.* , **115**, 513–585.

Government Actuary's Department (2001). National population projections: Review of methodology for projecting mortality. Technical report, Government Actuary's Department: London.

Government Actuary's Department (2006). National population projections 2004-based. Technical report, Government Actuary's Department: London.

Granger, C. W. J. and P. Newbold (1986). *Forecasting Economic Time Series* (2nd ed.). Academic Press, Inc.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* **82**, 711–732.

Gunning-Schepers, L. J. (1988). *The health benefits of prevention, a simulation approach.* Ph. D. thesis, Rotterdamn: Erasmus University.

Haberman, S. and M. Russolillo (2005). Lee Carter Mortality Forecasting: Application to the Italian Population. *Actuarial Research Paper,* No. 167.

Hagnell, M. (1991). A multivariate time series analysis of fertility, adult mortality, nupitality, and real wages in sweden 1751-1850: a comparison of two different approaches. *Journal of Official Statistics, 7*(4), 437–455.

Hartmann, M. (1987). Past and recent attempts to model mortaltiy at all ages. *Journal of Official Statistics,* **3**(1), 19–36.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika,* **57**(1), 97–109.

Heathcote, C. R. and I. M. McDermid (1994). Projections of cohort life expectancy based on weighted least squares methods. In: C. Mathers, J. McCallum and J-M Robine (eds.) (1994). *Advances in Health Expectancies*, 96–114.

Heligman, L. and J. H. Pollard (1980). The age pattern of mortality. *Journal of the Institute of Actuaries,* **107**, 49–80.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science,* **14**(4), 382–417.

Hollmann, F. W., T. J. Mulder, and J. E. Kallan (2000). Methodology and assumptions for the population projections of the united states: 1999 to 2100. Working Paper 38, Population Division, U.S. Bureau of the Census.

Hunt, C. E. (2001). Sudden Infant Death Syndrome and Other Causes of Infant Mortality. *American Journal of Respiratory and Critical Care Medicine,* **164**(3), 346–357.

Hyndman, R. J. and H. Booth (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting,* **24**(3), 323–342.

Hyndman, R. J., H. Booth, and F. Yasmeen (2012). Coherent Mortality Forecasting: The Product-Ratio Method with Functional Time Series Models. *Demography,* **50**(1), 261–283.

Hyndman, R. J. and M. S. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis,* **51**(10), 4942–4956.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (2nd ed.). Wiley Series in Probability and Mathematical Statistics.

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

Kass, R. E. and S. Vaidyanathan (1992). Approximate bayes factor and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B,* **54**, 129–144.

Kass, R. E. and L. Wasserman (1994). Formal rules for selecting prior distributions: A review and annotated bibliography. Technical report, *Journal of the American Statistical Association.*

Keilman, N. (1990). *Uncertainty in national population forecasting: Issues, backgrounds, analyses, recommendations.* Swets & Zeitlinger, Amsterdam.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions* (2nd ed.), Volume 1. Wiley Series in Probability and Mathematical Statistics.

Lee, R. and S. Tuljapurkar (2001). Population forecasting for fiscal planning: Issues and innovations. In A. J. Auerbach and R. D. Lee (Eds.), *Demographic Change and Fiscal Policy:*, pp. 7–57. Cambridge University Press.

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, N. and R. D. Lee (2005). Coherent Mortality Forecasts for A Group of Populations: An Extension of the Lee-Carter Method. *Demography,* **42**(3), 575–594.

Li, S. H. and W. S. Chan (2005). Outlier analysis and mortality forecasting: The United Kingdom and Scandinavian countries. *Scandinavian Actuarial Journal,* (3), 187–211.

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.

Lopez, A. and H. Crujisen (1991). Mortality in european community: Trends and perspectives. International Conference on Long-term population scenarios for the European Community, Luxembourg, 27-29 November.

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.

Lutz, W. (1996). *The Future Population of the World: What Can We Assume Today?* Earthscan, London, UK.

Lutz, W., W. C. Sanderson, and S. Scherbov (2004). *The End of World Population Growth in the 21st Century: New Challenges for Human Capital Formation and Sustainable Development.* Earthscan, London, UK.

Mackenbach, J. P. (1988). *Mortality and medical care. Studies of mortality by cause of death in The Netherlands and other European countries.* Ph. D. thesis, Rotterdamn: Erasmus University.

MacMinn, R. (2003). International mortality comparisons. Presentation to the Society of Actuaries, Vancouver. *www.journalofriskandinsurance.org*.

Madigan, D. and A. E. Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association 89*(428), 1535–1546.

Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries,* **8**, 301–310.

Manton, K. G. and E. Stallard (1992). Projecting the future size and health status of the us elderly population. *International Journal of Forecasting,* **8**(3), 433–458.

Manton, K. G., E. Stallard, and H. D. Tolley (1991). Limits to Human Life Expectancy: Evidence, Prospects and Implications. *Population Council,* **17**(4), 603–637.

McCullagh, P. and J. A. Nelder (1989). *Generalised linear models.* London etc.: Chapman and Hall.

McDonald, A. S. (1996). An actuarial survey of statistical models for decrement and transition data. i. multiple state, poisson and binomial models. *British Actuarial Journal, 2*(1), 129–155.

McNown, R. and A. Rogers (1989). Forecasting Mortality: A Parameterized Times Series Approach. *Demography,* **26**(4), 645–660.

McNown, R. and A. Rogers (1992). Forecasting cause-specific mortality using time series methods. *International Journal of Forecasting,* **8**(3), 413–432.

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.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics,* **21**, 1087–1091.

Murphy, M. (1990). Methods for forecasting mortality for population projections. In: OPCS Population Projections: Trends, Methods and Uses. Occupational paper 38, London.

Murphy, M. and D. Wang (2001). Do previous birth interval and mother's education influence infant survival? a Bayesian model averaging analysis of chinese data. *Population Studies,* **55**(1), 37–47.

Murray, C. J. L. and A. D. Lopez (1997a). Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. *The Lancet,* **349**(9064), 1498–1504.

Murray, C. J. L. and A. D. Lopez (1997b). Global Mortality, Disability and the Contribution of Risk Factors: Global Burden of Disease Study. *The Lancet,* **349**(9063), 1436–1442.

Murray, C. J. L. and A. D. Lopez (1997c). Mortality by cause for eight regions of the world: Global Burden of Disease Study. *The Lancet,* **349**(9061), 1269–1276.

Murray, C. J. L. and A. D. Lopez (1997d). Regional patterns of disability free life expectancy and disability adjusted life expectancy: Global Burden of Disease Study. *The Lancet,* **349**(9062), 1347–1352.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Mehodological),* **56**(1), 3–48.

Oeppen, J. and J. W. Vaupel (2002). Broken limits to life expectancy. *Science,* **296**(5570), 1029–1031.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Mehodological),* **57**(1), 99–138.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). *Uncertain Judgements: Eliciting Experts' Probabilities.* John Wiley & Sons, Ltd.

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

Orbanz, P. and Y. W. Teh (2010). *Bayesian Nonparametric Models*, pp. 81–89. Boston, MA: Springer US.

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.

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

Perks, W. (1932). On some experiments on the graduation of mortality statistics. *Journal of the Institute of Actuaries,* **63**, 12–40.

Philip, H. (1999). Fundamentals of Survival Data. *Biometrics,* **55**(1), 13–22.

Raftery, A. E. (1999). Bayes Factor and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods and Research,* **27**(3), 411–427.

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

Raftery, A. E., G. H. Givens, and J. E. Zeh (1995). Inference from a deterministic population dynamics model about bowhead whale, *Balaena mysticetus*, replacement yield. *Journal of the American Statistical Association,* **90**, 402–416.

Reichmuth, W. H. and S. Samad (2008). Modeling and Forecasting Age-Specific Mortality: A Bayesian Approach. SFB 649 Discussion Paper 2008-052a, Sonderforschungsbereich 649, Humboldt Universität zu Berlin, Germany. available at http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2008-052a.pdf.

Renshaw, A. E. and H. Haberman (2003). Lee-Carter Mortality Forecasting: A Parallel Generalized Linear Modelling Approach for England and Wales Mortality Projections. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **52**(1), 119–137.

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.

Renshaw, A. E., H. Haberman, and P. Hatzopolous (1996). The Modelling of Recent Mortality Trends in United Kingdom Male Assured Lives. *British Actuarial Journal,* **2**(2), 449–477.

Ripley, B. D. (1987). *Stochastic Simulation.* New York: Wiley.

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.

Robins, J. M., A. Van der Vaart, and V. Ventura (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association,* **95**, 1143–1156.

Rosenthal, J. S. (2014). *Optimising and Adapting the Metropolis Algorithm*, Volume Chapter 6 of the SSC volume Statistics in Action: A Canadian Outlook. Chapman & Hall/CRC, Boca Raton, Florida.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics,* **6**(2), 461–464.

Shafer, G. (1979). Lindley's Paradox. *Technical Report No. 125, Department of Statistics, Stanford University.*

Shang, H. L., P. W. F. Smith, J. Bijak, and A. Wiśniowski (2016). A multilevel functional data method for forecasting population, with an application to the United Kingdom. *International Journal of Forecasting,* **32**(3), 629–649.

Shaw, C. (2007). Fifty years of united kingdom national projections: How accurate have they been? *Population Trends,* **128**, 8–23.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Mehodological),* **55**(1), 3–23.

Soneji, S. and G. King (2012). Statistical Security for Social Social Security. *Population Association of America,* **49**, 1037–1060.

Spiegelman, M. (1968). *Introduction to Demography.* Cambridge, Mass.: Harvard University Press.

Stoto, M. A. and J. S. Durch (1993). Forecasting Survival, Health and Disability: Report on a Workshop. *Population Development and Review,* **19**(3), 557–581.

Streftaris, G. and B. J. Worton (2008). Efficient and accurate approximate Bayesian inference with an application to insurance data. *Computational Statistics and Data Analysis,* **52**(5), 2604–2622.

Tabeau, E., P. Ekamper, C. Huisman, and A. Bosch (2001). *Predicting Mortality from Period, Cohort or Cause-specific Trends: A Study of Four European Countries*, pp. 159–187. Dordrecht: Springer Netherlands.

Tabeau, E., A. van den Berg Jeths, and C. Heathcote (2001). *Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective*, Volume 9. Kluwer Academic Publishers, London.

Thatcher, A. (1999). The Long-Term Pattern of Adult Mortality and the Highest Attained Age. *Journal of the Royal Statistical Society, Series A,* **162**(1), 5–43.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*(393), 82–86.

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

Van Oortmarssen, G. J., J. D. F. Habbema, J. T. N. Lubbe, G. A. d. J. Lubbe, and P. J. van der Maas (1981). Predicting the effects of mass screening for disease - a simulation approach. *European Journal of Operational Research,* **6**(4), 399–409.

Wadsworth, M. E. J. (1991). *The Imprint of Time: Childhood, History and Adult Life.* Clarendon Press, Oxford.

Waldron, H. (2005). Literature review of long-term mortality projections. *Social Security Bulletin,* **66**(1), 16–30.

Weakliem, D. L. (1999). A Critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods and Research,* **27**(3), 359–397.

Wetterstrand, W. H. (1978). Recent Mortality Experience Described by Gompertz's and Makeham's Law-Including a Generalization. A paper presented at Actuarial Research Conference at Ball State University,.

Willets, R. C. (1999). *Mortality Trends: An Analysis of Mortality Improvement in the UK.* London: General & Cologne Re.

Willets, R. C. (2004). The Cohort Effects: Insights and Explanations. *British Actuarial Journal,* **10**(4), 833–877.

Wilmoth, J. R. (1990). Variation in vital rates by age, period and cohort. *Sociological Methodology, 20*, 295–335.

Wilmoth, J. R. (1993). Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change. *Technical Report, Department of Demography, University of California at Berkeley*.

Wilmoth, J. R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies, 5*(4), 293–319.

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.

Wolf, S. S. and J. L. Gastwirth (1967). The Effect of Autoregressive Dependence on A Non-Parametric Test. *IEEE Transactions on Information Theory, 13*(2), 311–313.

Wolfson, M. (1994). A framework for modelling and understanding the health of human populations. *World Health Statistics Quarterly, 47*(3-4), 157–176.

Wong-Fupuy, C. and S. Haberman (2004). Projecting mortality trends: recent developments in the United Kingdom and the United States. *North American Actuarial Journal, 8*(2), 56–83.

Yang, S. S., J. C. Yue, and Y. Y. Yeh (2011). Coherent Mortality Modelling for a Group of Populations. A paper presented at Living to 100 Symposium at Orlando, 1–23.

Zhu, Z. and Z. Li (2013). Logistic regression for insured mortality experience studies. SCOR Inform.