Network for Studies on Pensions, Aging and Retirement

Netspar THESES

Hong Li

Finding an Optimal Sample Size for the Lee-Carter Model

A Bayesian Approach

RM Thesis 2012-057

Finding an optimal sample size for the Lee-Carter model: A Bayesian approach

by
Hong Li (U1238660)

(Research Master. Tilburg University 2012)

CentER
Tilburg University

Supervisor
Prof. Bertrand Melenberg

Contents

Acknowledgments

I'd love to express my sincere gratitude to my supervisor, Prof. Bertrand Melenberg, for his original idea that initiates this thesis, his patient and earnest guidance throughout my research, as well as his expertise and professionalism that inspire me all along. Furthermore, I am indebted to my many friends who encouraged and motivated me all along and gave their hands whenever and wherever. In particular, I am heartily thankful to my girlfriend, Zhenzhen Fan, who kindly shared with me her experience in programming and empirical analysis, and provided insightful comments that have shed light on my research. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Abstract

In this thesis, we aim to find an age-specific sample size for the Lee-Carter model that maximizes the likelihood in the out-of-sample forecast. We set up a dynamic sequential Bayesian updating model which explicitly models the sample size as one of its parameters. Markov Chain Monte Carlo methods are used to fit the model and to sample from the posterior predictive distributions. This model is applied to the sex neutral mortality rates of the Dutch population from 1928 to 2009 for ages from 0 to 99, and results for age 0, 25, 50, 75, and 90 are reported. We present a dynamic algorithm as well as a reduced static application. Results from the static model show that the means of the conditional posterior densities of sample size for the 5 ages range from 31 to 34, while in their dynamic counterpart, they turn out to be 33.7 and 32.6 for age 0 and 25, respectively, and from 38 to 43 for the other three ages. In addition, the standard deviations of the conditional posterior density of the sample size obtained from the static model are approximately 11 for all ages. In the dynamic case, nevertheless, the standard deviations range from 3 to 5. Furthermore, the dynamic model delivers narrower confidence intervals than the ones obtained from the static model, but, except for age 0, still mildly wider than the original Lee-Carter model.

Chapter 1

Introduction

Government agents and private insurance companies rely on mortality forecasts when making financial policy and design insurance products. For example, due to the increasing of life expectancy in the Netherlands, the Dutch government is planning on increasing the retirement age from 65 to 67. In the past, the forecasts of mortality rates were based mostly on life table methods and expert opinion. However, due to the change in the trend of the life expectancy, these methods became less and less satisfying. As an illustration, the expected remaining life time for a Dutch male increased from 11.2 years in 1900 to 15.4 years in 2000, while this number increased from 11.8 years to 19.4 years for a Dutch female at the same age (Hári, De Waegenaere, Melenberg, and Nijman 2008). Since evolutions of the life expectancies in most countries have presented a nonlinear pattern, the simple linear projection method or expert opinions tend to underestimate the future life expectancy. Therefore, methods for more precise mortality forecasts are in need.

Over the past two decades, various methods for forecasting mortality have been developed. Among these methods, the Lee-Carter model (Lee and Carter 1992) is perhaps the most widely used one nowadays. The Lee-Carter model describes a time series of age-specific log mortality rates as the sum of a time-invariant and age-specific component and another age-and-time-specific component which is the product of an age-invariant and time varying parameter, which reflects the common trend of development of the log mortality, and an age-specific, time-invariant component, which reflects the sensitivity for each age to the common trend. The Lee-Carter model was first used to fit the historical U.S mortality data. After the original work of Lee and Carter, various extensions of the model are proposed (Lee and Tuljapurkar (1994); Lee (2000); Lee and Miller (2001), to mention a few[1]). Despite its popularity, the Lee-Carter model has some drawbacks. First, the Lee-Carter model does not explicitly consider the optimal sample size. In their original paper, Lee and Carter estimate the model based on the U.S log mortality data from 1900 to 2000. They argue that the length of the sample size is not critical as long as it is longer than 10 to 20 years. However, Lee and Miller (2001) restrict the sample size to start from 1950, in order to avoid structural breaks, and to obtain a better fit. Therefore, we see that the choice of sample size is not a trivial problem in the estimation

[1]Interested readers may refer to Pitacco, Denuit, Haberman, and Olivieri (2009), which includes various discussions and extensions on the Lee-Carter model

of Lee-Carter model. More specifically, if the choice of sample size is so large that the evolvement of mortality is non-linear even within this sample size, then the estimation of the trend of mortality might be inconsistent in the context of the Lee-Carter model. Also, the Lee-Carter model does not account for the variances of the parameters in the model while computing the prediction errors. The omission of such variances may lead to serious underestimation of the uncertainties in the evolvement of the future mortality, especially in the short run (Appendix B in Lee and Carter (1992)).

There are already some attempts to solve either of the aforementioned problems alone. Booth, Maindonald, and Smith (2002) design a procedure to search for a maximal sample size, based on which the trend of mortality development is linear. In particular, they first fix the ending year of the sample size, then look for a starting year that is as early as possible, while trying to assure the linearity of the development of mortality within the whole sample size. To address the second problem, Brouhns, Denuit, and Vermunt (2002) and Brouhns, Denuit, and Van Keilegom (2005) implement a bootstrap method for a log bilinear formulation of the Lee-Carter model bootstrap method. Similarly, Koissi, Shapiro, and Högnäs (2006) use a bootstrap method to compute the confidence intervals of the Lee-Carter model. A more natural way to incorporate all sources of parameter variances into the estimation of the Lee-Carter model is using a Bayesian approach. For example, Pedroza (2002) and Pedroza (2006) present a state space representation of the Lee-Carter model. Pedroza formalizes the Lee-Carter model as a statistical model in those papers, and estimates the model by Bayesian approach. This method also incorporates all sources of parameter variance. Pedroza (2006) applies the Bayesian method to the U.S male mortality, and obtains appropriately wider confidence intervals than those obtained from the original Lee-Carter estimation.

However, so far little research has been carried out to combine the Bayesian approach, which accounts for all sources of parameter variance in a natural way, with the search for a reasonable sample size, which can assure the consistency of the estimator, in the context of the Lee-Carter model. This thesis attempts to combine these two aspects into a single statistical model. To elaborate, our interest is to find a sample size for the Lee-Carter model that yields the largest likelihood in the out-of-sample forecast. Under a Bayesian framework, our objective function is the posterior distribution function, which is proportional to the product of the likelihood function and the prior distribution. Therefore, we are looking for a conditional posterior distribution function (possibly age-specific and conditional on the estimation of other parameters) of the sample size that gives the highest probability during the out-of-sample forecast. In this thesis, we practice two setups for the Bayesian model, namely, a static model and a dynamic one. In the static case, we update the conditional posterior distribution for the sample size only once, while in the dynamic case, we choose an update period and update the conditional posterior distribution at every stage during this period. After obtaining the final conditional posterior distribution, we can then compute the posterior forecasts and do the inference of the Lee-Carter model. The way we proceed is that, for each age, we first estimate the Lee-Carter model and compute the forecasted log mortality rates and confidence intervals based on different sample sizes, then weight these forecasts and confidence intervals by the corresponding conditional posterior density function.

We apply our Bayesian model to the log of sex neutral mortality data of

the Dutch population from year 1928 to 2009 for all ages from 0 to 99, but we present the results of 5 representative ages: 0, 25, 50, 75, and 90. According to our results, the conditional posterior densities of sample size obtained from the static model are flatter and display no clear peak, while the ones obtained from the dynamic model are more condensed and peaked. Also, the posterior means obtained in the static case are in general smaller than those obtained in the dynamic case. The posterior means of the sample size for the 5 ages obtained from the static model range from 31 to 34, while in the dynamic model, the posterior means of the sample size are 33.7 and 32.6 for age 0 and 25, respectively, and between 38 to 43 for the other three ages. In addition, except for age 0, in both cases we obtain wider confidence intervals of the forecasted mortality rates than those obtained from the original Lee-Carter model, while the dynamic model delivers narrower confidence intervals than the static model for all ages.

The thesis also has some future research potentials. A possible extension is to apply the Bayesian model to other sources of data, for instance, male or female log mortality data of the Dutch population, or mortality data from other countries. We expect different patterns of conditional posterior distribution for the sample size with different data. In addition, due to the compatibility of the Bayesian model and the Markov Chain Monte Carlo method, it is possible to apply the Bayesian model to model specifications other than the Lee-Carter model.

This thesis is organized as follows. Chapter 2 presents the static Bayesian model and results obtained with the Dutch mortality data. Section 2.1 introduces briefly the Lee-Carter model and the historical Dutch mortality data. Section 2.2 and 2.3 presents the setup of the static Bayesian model, and section 2.4 presents the empirical results obtained from the model applied to Dutch mortality data. Section 2.5 discusses the Bayesian sample size selection problem. Chapter 3 presents the dynamic Bayesian model. Section 3.1 describes the idea of sequential Bayesian updating. Section 3.2 and 3.3 presents two cases of Bayesian updating and their empirical results applied to Dutch mortality data, respectively. Conclusion is presented in chapter 4.

Chapter 2

Static analysis

In this chapter, we start our investigation by employing the static Bayesian analysis, which could be treated as a reduced form of the sequential Bayesian updating algorithm that we will introduce in chapter 3. This approach is called "static" because we only update the conditional posterior distribution for the sample size once.

First of all, we give a brief introduction to the Lee-Carter model (Lee and Carter 1992), nowadays perhaps the most widely used model to forecast the log of mortality rates, as well as some relevant concepts and methodologies from Bayesian statistics.

2.1 Lee-Carter model

Denote by N and T the set of ages and time periods considered, respectively. A common choice of age group is from age 0 to 110+, thus $N = \{1, 2, ..., 109, 110+\}$.[1] We can estimate the Lee-Carter model using an arbitrary subset n of N. Suppose the set of ages $n \in N$ has k components. Let $m_t = (m_{x_1,t}, m_{x_2,t}, ..., m_{x_k,t})$ be the mortality rates of ages $x_1, x_2, ..., x_k$ at year t with $n \in N$ and $t \in T$. The Lee-Carter model representation for the age group $x_1, x_2, ..., x_k$ at time t can be formulated as

$$\ln m_t = \alpha + \beta\kappa_t + \varepsilon_t, \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{2.1}$$

where $\ln m_t = (\ln m_{x_1,t}, \ln m_{x_2,t}, ..., \ln m_{x_k,t})'$ is a $k \times 1$ vector of the log of mortality rates from age x_1 to x_k at time t. $\alpha = (\alpha_{x_1}, \alpha_{x_2}, ..., \alpha_{x_k})'$ and $\beta = (\beta_{x_1}, \beta_{x_2}, ..., \beta_{x_k})'$ are vectors of age-specific and time invariant parameters. $\kappa = \{\kappa_t : t \in T\}$ is a collection of time-varying parameters. Note that the components of $\beta\kappa_t$ in Equation (2.1) are products of two parameters, and thus cannot be fully identified in general. Therefore, we have to impose some normalization on β and κ, respectively. The common normalization is

$$\sum_x \beta_x = 1, \text{and} \sum_t \kappa_t = 0, \tag{2.2}$$

[1] In most online data sources, such as Human Mortality Database (http://www.mortality.org/), the mortality rates of the population that is older than 109 years old are integrated in the entry "110+". Therefore, in this way, we have 111 age groups.

or

$$\sum_x \beta_x^2 = 1, \text{and} \sum_t \kappa_t = 0. \tag{2.3}$$

Given the above normalization, α_{x_i} can be interpreted as the average log-mortality of age x_i over time, and β_{x_i} measures the sensitivity to the change of the κ process for age x_i. Furthermore, the state vector κ captures the change of the log-morality for all age groups. In literature, κ is often modeled as an $ARIMA(0,1,0)$ process,

$$\kappa_t = d + \kappa_{t-1} + \omega_t. \tag{2.4}$$

In addition, the components of the error vector $\varepsilon_t = (\varepsilon_{x_1,t}, \varepsilon_{x_2,t}, ..., \varepsilon_{x_k,t})'$ are invariant with respect to time and independent among ages, which means that Σ is a $N \times N$ diagonal and time constant covariance matrix.

In Equation (2.4), d is the drift term, i.e., the mean evolvement of κ. ω_t is an i.i.d error process, i.e., $\omega_t \sim \mathcal{N}(0, \sigma_\omega^2)$. Equations (2.1) and (2.4) are called a state space model representation in Pedroza (2006).

However, one should note that the Lee Carter model is a linear approximation of the log mortality. If the development of the log mortality was linear, (2.4) could perfectly capture this change, which means that we could have more and more efficient estimation of the drift term d by extending the sample size. However, we do not have this ideal situation in the real world. In Figure 2.1, where we plot the log of mortality rates of the Dutch population in the past 160 years, we can see that they changed apparently not in a linear way, especially for young age groups. The same case applies to other developed countries such as Japan, France, and the US, etc. We can see that the speed of decline of log-mortality is in general increasing. In fact, the state vector κ obtained by Lee and Carter was not linear either (figure (5) in Lee and Carter 1992). As a consequence, it is probably not the case that we can obtain a consistent estimation of the drift term d by using a linear approximation. Therefore, under the context of Lee-Carter model, it is important to find a (probably age-specific and weighted) sample size that can yield the forecasts of future mortality rates that are closest to the real ones.

There exist several estimation methods of the Lee-Carter model, for example the singular value decomposition (SVD) method (Lee and Carter 1992) and the weighted least square estimation method (Wilmoth 1993). In this thesis, we use the weighted least square method, and the normalization in Equation (2.2).

2.2 Bayesian model

2.2.1 A Bayesian setup

Before getting to the setup of the Lee-Carter model in the Bayesian context, we first introduce some relevant concepts. One distinct feature of Bayesian statistics from the viewpoint of frequentist statistics is that, in a Bayesian framework, we consider a model space which includes all possible model specifications. More specifically, in a parametric frequentist framework, we consider a family of models $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, which contains the "true" model, P_0. Θ is the associated parameter space. We say that the model is identifiable if there exists a one-to-one mapping $\Theta \to \mathcal{P}$ for each $\theta \in \Theta$. In other words, the mapping $\Theta \to \mathcal{P}$ should be invertible. Furthermore, if the model is identifiable and the parameter space is chosen appropriately, there exists a "true" parameter, $\theta_0 \in \Theta$, such

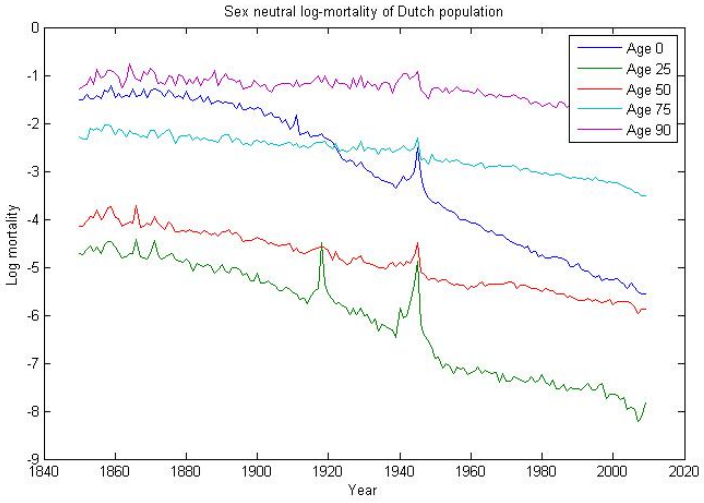


Figure 2.1: Sex neutral log-mortality of Dutch population

that θ_0 is associated with the "true" model, i.e., $P_\theta = P_0$. On the other hand, in a Bayesian framework, the existence of a "true" parameter is not assumed. Instead, we study the parameter space Θ as a whole by imposing probability distributions on it. In particular, we assume a prior distribution on Θ before having access to the sample, and use the sample to "correct" the prior distribution after it is obtained. In the subsequent introduction to Bayesian statistics, we rely ourselves heavily on the lecture notes from Kleijn (2012).

In a Bayesian framework, we may combine sample and model into a product space. Denote by $\mathcal{Y}$ and Θ the sample and model space, respectively, and $\mathcal{B}$ and $\mathcal{G}$ the associated σ-algebra. At this stage, $\mathcal{B}$ and $\mathcal{G}$ can be any σ-algebras that assure our functions of interest to be measurable. Then we can denote the product space by $(\mathcal{Y} \times \Theta, \sigma(\mathcal{B} \times \mathcal{G}))$ and impose a probability distribution (usually according to expert opinion) on it,

$$\Pi : \sigma(\mathcal{B} \times \mathcal{G}) \to [0, 1]. \tag{2.5}$$

Π assigns a joint probability distribution of observations $Y \in \mathcal{Y}$ and parameters $\theta \in \Theta$. In this way, we are able to define a conditional distribution induced by Π: $\Pi_{Y|\theta}(\cdot|\theta) : \mathcal{B} \times \Theta \to [0, 1]$, which gives the distribution of observations Y conditional on any particular $\theta \in \Theta$.[2] A nice property about this conditional distribution is that it can be linked to the frequentist setting: for each $\Pi_{Y|\theta}$ we can find a corresponding element $P_\theta \in \mathcal{P}$. Actually, for any particular $\theta^* \in \Theta$, $\Pi_{Y|\theta^*}$ is an almost sure version of the corresponding P_{θ^*}. This relation can be stated as

$$P_{\theta^*} = \Pi_{Y|\theta^*}(\cdot|\theta^*) : \mathcal{B} \to [0, 1], a.s. \tag{2.6}$$

An immediate consequence from Equation (2.6) is that the notion of "model" in the frequentist setting is represented up to the null set of the marginal distri-

[2]The notation Θ in $\mathcal{B} \times \Theta$ has the interpretation that, we are able to condition on any $\theta \in \Theta$, and then measure any subset of the $\mathcal{B}$. All domains denoted by the product of a σ-algebra and a space in the subsequent context can be interpreted in the same way.

bution of Π with respect to Θ. Denote this marginal distribution by Π_θ, then we have

$$\Pi_\theta(\cdot) : \mathcal{G} \to [0, 1]. \tag{2.7}$$

Equation (2.7) is also called the prior distribution of θ in a Bayesian context. For simplicity of notation, we hereafter drop the subscript θ and denote the prior distribution by $\Pi(\theta)$. The prior distribution can be interpreted as as the degree of belief we attach to any subset of the model *a priori*, that is, before any observation is realized. After observations Y are obtained, we can then correct this prior belief by computing the posterior distribution. Denote the posterior distribution induced by Π by $\Pi(\cdot|Y)$, then we have

$$\Pi(\theta|Y) : \mathcal{G} \times \mathcal{Y} \to [0, 1]. \tag{2.8}$$

The interpretation of equation (2.8) is that, for any $\theta^* \in \Theta$, $\Pi(\theta^*|Y)$ reflects the amended belief we attach to θ^* after observing Y.

Apart from the posterior distribution, the posterior density is another important concept. In fact, a posterior density is often more convenient to deal with in applications. Assuming that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by some σ-finite measure, say μ, on $\mathcal{Y}$, we could define the model in terms of μ-densities, $p_\theta = \frac{dP_\theta}{d\mu} : \mathcal{Y} \to R$. In this way, the posterior distribution can be expressed as

$$\Pi(\theta \in G|Y) = \frac{\int_G p_\theta(Y) d\Pi(\theta)}{\int_\Theta p_\theta(Y) d\Pi(\theta)}, \tag{2.9}$$

and the corresponding posterior density can be written as

$$d\Pi(\theta|Y) = \frac{p_\theta(Y) d\Pi(\theta)}{\int_\Theta p_\theta(Y) d\Pi(\theta)}. \tag{2.10}$$

To sum up, the Bayesian statistics procedure is as follows. First, based on our prior understanding of data Y, we choose a statistical model $\mathcal{P}$ equipped with a parameter space Θ. Second, based on the expertise, we impose a (possibly uninformative) prior distribution $\Pi(\theta)$. After observations are obtained, we can compute the posterior density function according to Equation (2.10).

After the above introduction, we can now turn to the Bayesian Lee-Carter model. In this thesis, we have $\mathcal{Y} = R$ as the collection of the Dutch log-mortality rates and $\mathcal{B} = \mathcal{B}(R)$. In addition, we define $\Theta = \{\theta : \theta = (\alpha', \beta', d, \sigma'_\varepsilon, \sigma_\omega)' \in R^{2N+1} \times R_+^{N+1}\}$ and $\mathcal{G} = \sigma(R^{2N+1} \times R_+^{N+1})$. As the dominating measure, we simply choose the Lebesgue measure λ. We denote the prior and posterior density as $\pi(\cdot)$ and $\pi(\cdot|Y)$, respectively. Then, according to Equation (2.10), for any $\theta^* \in \Theta$, we can write the posterior density as

$$\begin{aligned} \pi(\theta^*|Y) &= \frac{p_{\theta^*}(Y)\pi(\theta^*)}{\int_\Theta p_\theta(Y)\pi(\theta) d\theta} \\ &\propto p_{\theta^*}(Y)\pi(\theta^*). \end{aligned} \tag{2.11}$$

Note that in equation (2.11), $p_{\theta^*}(Y)$ can be treated as the likelihood of observing Y given θ^*.

2.2.2 An introduction to Markov Chain Monte Carlo

Before we can search for an optimal sample size, we should be able to fit the Lee-Carter model given any sample size. For this purpose, we need to obtain a joint

distribution of the parameters $(\alpha', \beta', d, \sigma_\varepsilon^2, \sigma_\omega^2)'$. However, to obtain a posterior predictive distribution of such high dimensionality is difficult. For example, if we include 100 ages, we have 302 parameters, and thus have to integrate out a 302 dimensional function. Integration of this kind is very computationally complicated and inefficient. Therefore, we have to consider an alternative method, preferably short of direct integration, to calculate the desired posterior distributions. A common approach applicable in this case is the Monte Carlo Markov Chain (MCMC) method. We now give a brief introduction to some relevant concepts based on the lecture notes by Walsh (2004).

Monte Carlo integration

The Monte Carlo method was developed originally by physicists to use random number generation to calculate complex integrals. Suppose we want to calculate an integral

$$\int_a^b f(x)dx. \tag{2.12}$$

We can decompose $f(x)$ into the product of a nowhere-zero density function $p(x)$ and another integrable function $g(x)$, given by $g(x) = \frac{f(x)}{p(x)}$, which enables us to write (2.12) as

$$\int_a^b g(x)p(x)dx = E_{p(x)}[g(x)]. \tag{2.13}$$

Therefore, equation (2.12) could be treated as the expectation of $g(x)$ under the density $p(x)$. We could generate, say N draws, from $p(x)$, and approximate (2.12) as

$$\int_a^b f(x)dx = E_{p(x)}[g(x)] \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i). \tag{2.14}$$

Equation (2.14) is called Monte Carlo integration.

Markov chain

First, let $X = (X_1, X_2, ...X_T)$ be a discrete stochastic process. Therefore, for each $t \in \{1, 2, ...T\}$, X_t is an E valued function, where E is a finite dimensional space endowed with the σ-algebra $\mathcal{E}$. X is called a Markov chain with state space $(E^T, \mathcal{E}^T)$ if the transitional probability between any two states in E depends only on the current state, i.e., for any $s \in E$,

$$P(X_{t+1} = s|X_1, X_2, ..., X_t) = P(X_{t+1} = s|X_t). \tag{2.15}$$

The interpretation of Equation (2.15) is that the current state of X includes all past knowledge of X that one could use to predict the future values of X. Therefore, for a discrete time Markov chain, we can define a transitional probability matrix P, with its (i, j)-th component $P_{ij} = P(X_{t+1} = s_j|X_t = s_i)$. Besides the transitional matrix, we also need to specify the starting position of the process. We do it by defining an initial distribution, i.e., the distribution of X_0. In fact the family of initial distribution is very large: it can be any probability that satisfies the mapping $E \to [0, 1]$. We denote this family by $\mathcal{V}$.

A discrete time Markov chain can be defined by the transitional matrix together with the initial distribution.

Furthermore, define $\pi^{\nu}_{s_k}(t) = P^{\nu}(X_t = s_k)$ as the probability that $X_t = s_k$ given a $\nu \in \mathcal{V}$, then by a simple application of the Chapman-Kolmogorov equation (Athanasios 1991), we have

$$\begin{aligned}\pi^{\nu}_{s_j}(t+1) =& P^{\nu}(X_{t+1} = s_j) \\ =& \sum_{s_i \in E} P(X_{t+1} = s_j | X_t = s_i) P^{\nu}(X_t = s_i) \\ =& \sum_{s_i \in E} P_{ij} \pi^{\nu}_{s_i}(t), \end{aligned} \tag{2.16}$$

which means that the probability that X reached stage s_j at time $t+1$ is the sum of the transitional probability from s_i to s_j times the probability that X is in state s_i for all $s_i \in E$. This relation holds for every initial distribution, thus, for simplicity of notation, we drop the subscript ν hereafter.

In fact, we can write the Chapman-Kolmogorov equation more compactly. Assume that E contains n components. Denote by $\pi(t)$ a $1 \times n$ vector, with $\pi_{s_i}(t)$ the i-th component, we then have the matrix form of Chapman-Kolmogorov equation

$$\pi(t+1) = \pi(t)P. \tag{2.17}$$

Using the matrix form, we can immediately obtain the induction

$$\pi(t+2) = \pi(t+1)P = \pi(t)P^2 = ... = \pi(0)P^{t+2}. \tag{2.18}$$

In fact, a Markov chain may reach a stationary distribution π^*, such that

$$\pi^* = \pi^* P, \tag{2.19}$$

and π^* is independent of the initial value X_0. The sufficient condition for the existence of a stationary distribution is that the "detailed balance equation" holds, i.e.,

$$p_{ij}\pi^*_i = p_{ji}\pi^*_j. \tag{2.20}$$

Equation (2.20) is called the reversibility condition, and this condition immediately implies equation (2.19).

The notion of discrete time Markov chain can be generalized to a continuous time framework. The transitional probability matrix would then be an infinite dimensional matrix and thus be infeasible. Instead, we can specify a transitional kernel $P(x, y)$. This kernel is of an infinitesimal sense such that $P(x, y)$ represents the probability of changing to y in the next infinitesimally small time interval while the current state is x. $P(x, y)$ should satisfy

$$\int P(x, y)dy = 1. \tag{2.21}$$

Furthermore, if there exists a continuous time stationary distribution π^*, then it holds that

$$\pi^*(y) = \int \pi^*(x)P(x, y)dx. \tag{2.22}$$

Markov Chain Monte Carlo

We have already mentioned that we can use the Monte Carlo integration to approximate a complex integral. However, one problem of Monte Carlo integration is that sampling from a density function $p(x)$ might still be very difficult. To get over this problem, Markov Chain Monte Carlo is an attractive method. In particular, the Metropolis-Hastings algorithm and its special case, the Gibbs sampler, are the most widely used sampling methods. We hereby give a brief introduction.

The first attempt of Metropolis-Hastings algorithm was to integrate complex functions in the field of mathematical physics, for example, Meteopolis and Ulam (1949), Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), and Hastings (1970). Suppose we want to draw samples from a distribution, $p(x)$, where $p(x)$ is known up to a constant K, where K might be very difficult to compute, say $p(x)=\frac{f(x)}{K}$. We can then proceed as follows;

1. Start with any x_0 which is within the support of $f(x)$.

2. Given the current value $x^{(0)}$, draw a candidate value, $x^{(1)}$, from a proposal distribution $q(x)$. There are two general forms of the proposal distribution: random walks and independent chain. In the former case, the draw of $x^{(1)}$ depends on (and only on) the current value $x^{(0)}$. In this case we can write $q(x)$ as $q(x^{(0)}, x^{(1)})$. In the latter case the draw of $x^{(1)}$ can also be independent of the current value. Note that the support of the proposal distribution should cover the support, or at least the most part of the support of $f(x)$.

3. Given the candidate value x_1, calculate the acceptance probability α. In the random walks sampling case, we have

$$\alpha = \alpha(x^{(0)}, x^{(1)}) = \min(\frac{f(x^{(1)})q(x^{(1)}, x^{(0)})}{f(x^{(0)})q(x^{(0)}, x^{(1)})}, 1), \tag{2.23}$$

   and in the independent chain case, (2.23) becomes

$$\alpha = \alpha(x^{(0)}, x^{(1)}) = \min(\frac{f(x^{(1)})q(x^{(0)})}{f(x^{(0)})q(x^{(1)})}, 1). \tag{2.24}$$

   Note that here we have $\frac{f(x^{(0)})}{f(x^{(1)})} = \frac{p(x^{(0)})}{p(x^{(1)})}$, thus the unknown constant K cancels out.

4. We accept the proposed value $x^{(1)}$ as the new current value with probability α. If $x^{(1)}$ is accepted, we replace $x^{(0)}$ by $x^{(1)}$, otherwise we keep $x^{(0)}$ unchanged.

We then repeat step 2 to step 4 for n times for n large enough. After a sufficient burn-in period, say $\frac{n}{2}$, the recorded vector $(x_{\frac{n}{2}+1}, ...x_n)$ can be treated as a sample from the density $p(x)$. Now we give some intuitions to this conclusion.

In the Metropolis-Hastings algorithm, the acceptance probability from x_0 to x_1 could be interpreted as the ratio of densities penalized by the corresponding proposal density of two draws. Take the independent chain sampling for example, If $\frac{f(x_1)}{q(x_1)}$ is larger than $\frac{f(x_0)}{q(x_0)}$, then we accept x_1 as the new current value

with probability one. Otherwise we still accept x_1 with probability $\frac{f(x_1)q(x_0)}{f(x_0)q(x_1)}$ rather than just simply reject it. The proposal densities, $q(x_0)$ and $q(x_1)$, could be interpreted as a sort of penalty, which could help the chain to better explore the support of $f(x)$, and leads to a quicker convergence.

It can be shown that the Metropolis-Hastings sampling generates a Markov chain whose equilibrium density is our interested density $p(x)$. A sufficient condition is that the Metropolis-Hastings transitional kernel satisfies the detailed equation (2.20) with $p(x)$.

The transition probability from x to y in the Metropolis-Hastings algorithm is

$$p(x,y) = q(x,y)\alpha(x,y) = q(x,y)\min(\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1). \tag{2.25}$$

Therefore, for the detailed equation to hold, we show that

$$q(x,y)\alpha(x,y)p(x) = q(y,x)\alpha(y,x)p(y), \text{ for all } x, y \in \text{Support}(p). \tag{2.26}$$

1. For $q(x,y)p(x) = q(y,x)p(y)$, we have $\alpha(x,y) = \alpha(y,x) = 1$. Thus (2.26) becomes
$$q(x,y)p(x) = q(y,x)p(y),$$
which holds apparently.

2. For $q(x,y)p(x) > q(y,x)p(y)$, we have
$$\alpha(x,y) = \frac{q(y,x)p(y)}{q(x,y)p(x)}, \text{ and } \alpha(y,x) = 1.$$
Thus (2.26) becomes
$$\begin{aligned} q(y,x)p(y) =& q(x,y)\frac{q(y,x)p(y)}{q(x,y)p(x)}p(x) \\ =& q(y,x)p(y) \end{aligned} \tag{2.27}$$

3. The case where $q(x,y)p(x) < q(y,x)p(y)$ is similar to case 2.

To summarize, the Metropolis-Hastings algorithm works as follows. For an m dimensional probability density function $p(x)$, we first specify an initial $x^{(0)} = (x_1^{(0)}, ..., x_m^{(0)})$, and sample a candidate values $x^{(1)} = (x_1^{(1)}, ..., x_m^{(1)})$ from some proposed density $q(x)$. We then decide whether to accept $x^{(1)}$ according to Equation (2.23) or (2.24). One advantage of the Metropolis-Hastings sampling algorithm is that there is barely any restriction: the posterior distribution could contain arbitrary many variables, the posterior distribution and the conditional posterior distributions do not need to be any familiar forms (Gaussian or χ^2, etc), etc. Therefore, it could be treated as a backup method, i.e., when all other sampling methods fail, we can still use the Metropolis-Hastings sampling.

However, there are also drawbacks in applying the Metropolis-Hastings algorithm. The biggest disadvantage of this method might be the lack of efficiency. Sometimes it requires an enormous time of iterations to reach convergence. In addition, it is often difficult to control the acceptance probability α. If α is too high, then the chain could possibly be poorly mixed, which means that the chain would stay in a small region of the parameter space for a long time, instead of exploring the whole space quickly. On the other hand, if α is too low,

then we might reject too many candidate draws, and the sampling could be very inefficient.

Therefore, although the Metropolis-Hastings algorithm can be applied in almost all situations, it is better to treat it as the last choice. In situations where we have sufficient knowledge on the family of the posterior and conditional posterior distribution, there is a more efficient alternative applicable, the Gibbs sampler.

The Gibbs sampler, first introduced by Geman and Geman (1984), is a special case of the Metropolis-Hastings algorithm, where the acceptance rate α is always 1. One distinct feature of the Gibbs sampler is that we only need to consider univariate conditional distributions. Assume again that we have an n dimensional distribution $p(x)$. For each round of sampling, we need to draw n times from n conditional distributions. For the i-th draw, we sample x_i from $p(x_i|x_{-i})$, which means that we fix all other components of x except for x_i. In general, the sampling procedure is as follows.

1. Specify initial value $x^{(0)} = (x_1^{(0)}, ..., x_n^{(0)})$, where $p(x_0) > 0$.

2. For each variable, fix the values of all other variables, and draw from the univariate conditional distribution.

$$
\begin{aligned}
x_1^{(1)} &\sim p(x_1|x_2^{(0)}, ..., x_n^{(0)}); \\
x_2^{(1)} &\sim p(x_2|x_1^{(1)}, x_3^{(0)} ..., x_n^{(0)}); \\
&... \\
x_n^{(1)} &\sim p(x_n|x_1^{(1)}, x_2^{(1)} ..., x_{n-1}^{(0)}).
\end{aligned}
$$

3. Save $x^{(1)} = (x_1^{(1)}, ..., x_n^{(1)})$, and continue step 2 for desired number of times.

A nice property of the Gibbs sampler is that, the Gibbs sequence will eventually converge to an equilibrium distribution that is the target distribution that we want to draw from. What's more, the Gibbs sequence is independent of the starting values (Tierney 1994).

Ergodic Theorem

For both the Metropolis-Hastings algorithm and the Gibbs sampler, after recording x-s that can be treated as draws from the target distribution, it seems like we could now compute the Monte Carlo integration and all quantities of interest. However, one thing we need to notice is that the parameters that we draw during the sampling procedure are rarely independent. Therefore, we are not able to employ the strong law of large numbers, and conclude to equation (2.14).

Fortunately, we have a useful tool, the Ergodic Theorem.[3] To illustrate the Ergodic Theorem, we first introduce some properties about Markov chains. First of all, a Markov chain is called *aperiodic* if, for any state x and y, the minimal number of steps that the chain require to transfer from x to y is equal to one. In other words, the chain is not forced to run with some cycle of fixed length between any two states. Second, a Markov chain is called *irreducible* if the chain is able to commute between any two state. At last, a Markov chain is

[3]There are many textbooks that give extensive discussions on the Ergodic Theorem. Interested readers may refer to, for example, Walters (2000) and Anosov (2001).

said to be *recurrent* if, when a chain is in any state x, it will eventually return to state x with probability one. A Markov chain has the *positive recurrence* property if the expected time of return is finite.

These three properties can be satisfied in many Bayesian statistics applications. With these properties at hand, we could then apply the Ergodic Theorem: If $x^{(0)}$, $x^{(1)}$, ... $x^{(N)}$ are N values drawn from a Markov chain which is aperiodic, irreducible, and positive recurrent, then for any function $f(x)$ that satisfies $E[f(x)] < \infty$, we have

$$\frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \to \int f(x)\pi(x)dx, \tag{2.28}$$

with probability one, as $N \to \infty$, π is density of the stationary distribution. The Ergodic Theorem can be considered as the Markov chain analog of the strong law of large numbers, and thanks to it we are able to compute the Monte Carlo integration by Markov chain Monte Carlo method.

2.2.3 Markov Chain Monte Carlo applied to Lee-Carter model

As stated in Equation (2.1) and (2.4), the Lee-Carter model can be represented as a state space model. Therefore, we can find some nice conditional distribution, such as the Gaussian distribution and the Inverse-Gamma distribution, for the parameters. In particular, for our sampling procedure we rely on Pedroza (2006), but with some changes that will be discussed later.

In order to obtain comparable results to those obtained from the original estimation of the Lee-Carter model, we assume uninformative priors for our parameters: $p(\alpha, \beta, d) \propto 1$, $p(\sigma_\varepsilon^2) \propto \frac{1}{\sigma_\varepsilon^2}$ and $p(\sigma_\omega^2) \propto \frac{1}{\sigma_\omega^2}$. In the state space model, we proceed our sampling in two steps. We first sample the parameters $(\alpha', \beta', d, \sigma_\varepsilon^{2'}, \sigma_\omega^2)'$, and then we sample the state vector κ.

In order to run the Gibbs sampler, we first need to obtain the form of the conditional posterior distributions of each parameter. As mentioned earlier, we first sample the parameters $(\alpha', \beta', d, \sigma_\varepsilon^{2'}, \sigma_\omega^2)'$. Since we have assumed noninformative conditional priors for each parameter, we can update their posterior conditional distributions respectively by the standard routine as we stated in section 2.2.1. The conditional posterior distributions for each parameter is stated here:

1. $$(\alpha_i, \beta_i)|y, \kappa, \sigma_{\varepsilon_i}^2 \sim \mathcal{N}((X'X)^{-1}X'y, \sigma_{\varepsilon_i}^2(X'X)^{-1}).$$

2. $$d|\kappa, \kappa_0, \sigma_\omega^2 \sim \mathcal{N}(\frac{\kappa_T - \kappa_0}{T}, \frac{\sigma_\omega^2}{T}).$$

3. $$\sigma_{\varepsilon_i}^2|\alpha, \beta, \kappa, y \sim \text{Inv-Gamma}(\frac{T}{2}, \frac{\sum_t (y_{i,t} - \alpha_i - \beta_i\kappa_t)^2}{2}).$$

4. $$\sigma_\omega^2|\kappa, \kappa_0, d \sim \text{Inv-Gamma}(\frac{T-2}{2}, \frac{\sum_{t=1}^{t=T}(\kappa_t - \kappa_{t-1} - d)^2}{2}).$$

Next we need to sample the state vector κ. One way to proceed is to use the single-state Gibbs sampler, i.e., for each entry κ_t, we draw κ_t from $p(\kappa_t|y, \kappa_{-\mathbf{t}}, \phi)$, where ϕ denote all other parameters. However, this algorithm could be extremely inefficient. We hereby rely on the method proposed by West and Harrison (1997): we first run the Kalman filter, and then sample κ by the equation

$$p(\kappa_1, ...\kappa_T|y) = p(\kappa_{T-1}|\kappa_T, y)p(\kappa_{T-2}|\kappa_{T-1}, \kappa_T, y)...p(\kappa_1|\kappa_2, ...\kappa_T|y). \quad (2.29)$$

The Kalman filter, first introduced by Kalman et al. (1960), is a natural estimation method for state space model. In the context of the Lee-Carter model, we could imagine the state vector κ as the real trend of the change of log-mortality. This trend evolves as an $ARIMA(0, 1, 0)$ process as stated in Equation (2.4). The error vector ω can be treated as the system error, which represents the deviation of the state variables to a linear trend. However, the state variables are unobservable, so we can only observe the realized log-mortalities, the so called measurements. Therefore, we are actually facing two sources of uncertainties: the system error ω and the measurement error ε. The Kalman filter can be used to obtain minimal standard error estimators, compared with estimation methods that consider only one source of error.

The Kalman filter algorithm works in a recursive manner. In each time period it proceeds in two steps. Before the new observation arrives, we first estimate the state variables *a priori* based on the current information. This step is called a prediction step. Next, when the new observation arrives, we update our estimates by incorporating the estimation error. The updated estimates are called *a posteriori* estimates and could be used to predict the state variables in next time period. We next give a formulation of Kalman filter in the context of Lee-Carter model.

Assume that there are $n \in N$ ages, denote by $y_t = \ln m_t$ a $n \times 1$ vector containing log-mortality rates of all ages at time t, i.e., the measurements. Denote by $a_t = E(\kappa_t|y_1, ...y_{t-1})$ the prediction of κ_t using all information up to time $t-1$, and $R_t = Var(a_t|y_1, ...y_{t-1})$ the prediction variance. The a_t and R_t are called the best predictions, in the sense that a_t is the minimal-quadratic-loss estimator that we could obtain before the measurement y_t arrives. After we obtain y_t, we can compute the prediction error of the measurement vector y_t according to Equation (2.1) and (2.4),

$$\nu_t = y_t - \alpha - \beta a_t, \quad (2.30)$$

and the variance of ν_t

$$Q_t = \beta R_t \beta' + \Sigma. \quad (2.31)$$

Also, with the new observation y_t at hand, we can then update our predictions a and R,

$$a_{t+1} = a_t + d + K_t \nu_t, \quad (2.32)$$

and

$$R_{t+1} = R_t(1 - K_t\beta) + \sigma_\omega^2, \quad (2.33)$$

where

$$K_t = R_t \beta' Q_t^{-1} \quad (2.34)$$

is called the Kalman gain, which we can use to correct our *a priori* estimates. Note that we have to specify the initial value, a_0 and R_0.

Now the procedure of the Gibbs sampler becomes clear. The posterior conditional distribution of each parameter depends on other parameters, thus we can run the Gibbs sampler recursively. After a sufficient iterations, we can reach the stationary conditional posterior distribution for each parameter. We let the initial value of a_0 and R_0 be $\bar{\kappa}$ and $\hat{\sigma}^2_\omega$ obtained from the original weight least square estimation, respectively. Now we can start the Gibbs sampler, the procedure is as follows. [4]

1. Draw κ_T from $\mathcal{N}(a_T, R_T)$. Then for each $t \in \{1, ..., T-1\}$, draw κ_t from

$$\kappa_t | \kappa_{t+1}, y \sim \mathcal{N}(h_t, H_t),$$

where $h_t = a_t + \frac{R_t}{R_{t+1}}(\kappa_{t+1} - a_{t+1})$, and $H_t = R_t - \frac{R_t^2}{R_{t+1}}$.

2. For every age i, draw $\sigma^2_{\varepsilon_i}$ from

$$\sigma^2_{\varepsilon_i} | \alpha, \beta, \kappa, y \sim \text{Inv-Gamma}(\frac{T}{2}, \frac{\sum_t (y_{i,t} - \alpha_i - \beta_i \kappa_t)^2}{2}).$$

3. Let $X = (\iota, \kappa)$, where ι is a $T \times 1$ vector of ones. Then for each age i, draw α_i and β_i from

$$(\alpha_i, \beta_i) | y, \kappa, \sigma^2_{\varepsilon_i} \sim \mathcal{N}((X'X)^{-1}X'y, \sigma^2_{\varepsilon_i}(X'X)^{-1}).$$

4. Draw d from

$$d | \kappa, \kappa_0, \sigma^2_\omega \sim \mathcal{N}(\frac{\kappa_T - \kappa_0}{T}, \frac{\sigma^2_\omega}{T}).$$

5. Draw σ^2_ω from

$$\sigma^2_\omega | \kappa, \kappa_0, d \sim \text{Inv-Gamma}(\frac{T-2}{2}, \frac{\sum_{t=1}^{t=T} (\kappa_t - \kappa_{t-1} - d)^2}{2}).$$

2.3 Forecasting and inference

In a frequentist framework, forecasting future mortality is simply done by computing

$$\hat{\kappa}_{T+\ell} = \kappa_T + \ell d, \tag{2.35}$$

[4] Our procedure of the Gibbs sampler is different from the one in Pedroza (2006). First, on the 10-th page of the paper, the author mentioned that she sample κ_T from $\mathcal{N}(a_T, Q_t)$, and during the sampling of κ_t for $t \in \{1, ..., T-1\}$, she computes

$$h_t = a_t + B_t(\kappa_{t+1} - a_{t+1}),\ H_t = Q_t - B_t R_{t+1} B_t',\ B_t = Q_t R_{t+1}^{-1}.$$

However, based on our computation, Q_t is a $N \times N$ matrix. Moreover, both of h_t and H_t are also $N \times N$ matrix. It's impossible to draw a scalar, κ_t, from a multivariate normal distribution. Hence, we think there is at least a typo in the formulation of Pedroza. Instead, we replace all Q_t with R_t in the above derivation. The reason is that, R_t is the variance of the prediction error of state variables, while Q_t is the covariance matrix of the prediction errors of the measurement equation. Therefore R_t seems to be more appropriate as the variance of the proposal distribution of κ_t. Another different specification in our sampling is that, we assign different variance to the measurement equation of each age, rather than assigning the same variance for a cluster of more than 10 ages (the number of ages in each cluster was not specified by the author).

and

$$\hat{y}_{T+\ell} = \hat{\alpha} + \hat{\beta}\hat{\kappa}_{T+\ell}, \tag{2.36}$$

where y_T denotes the log-mortality rates at time T.

Let $\Delta\kappa_t$=$\kappa_t - \kappa_{t-1}$, we can compute the variance of d as

$$\sigma_d^2 = \frac{\sum_{t=1}^{T}(\Delta\kappa_t - \hat{d})^2}{T}. \tag{2.37}$$

The computation of 95% confidence interval proposed by Lee and Carter (1992) is

$$\hat{y}_{T+\ell} \pm 1.96\hat{\beta}\hat{\sigma}_{\kappa_{T+\ell}}, \tag{2.38}$$

where

$$\hat{\sigma}^2_{\kappa_{T+\ell}} = \ell^2\sigma_d^2 + \ell\sigma_\omega^2. \tag{2.39}$$

One problem about the computation in Equation (2.38), as stated in their original paper, is that it considers only the variance from the $\{\kappa\}$ process, and thus ignores the variance other parameters. Although they stated that $\hat{\sigma}^2_{\kappa_{T+\ell}}$ in Equation (2.39) could capture at least 95% of the standard error based on all sources, the confidence interval estimated from $\hat{\sigma}^2_{\kappa_{T+\ell}}$ would seriously underestimate the errors in age-specific death rates in a short horizon, say 15 years.

In contrast, under the MCMC framework, errors from all sources are naturally incorporated during the estimation of confidence interval. The forecast of log-mortality rates and inference following MCMC is straightforward. We calculate the log-mortality rates for ℓ year-ahead from the posterior predictive distribution

$$p(y_{T+\ell}|Y^T) = \int_\Theta p(y_{T+\ell}|\theta, Y^T)p(\theta|Y^T)d\theta, \tag{2.40}$$

where Y^T is the observed sample up to time T.

The way we proceed is as follows.

1. Run one chain for m iterations, and save the parameters of the latter $\frac{m}{2}$ draws.

2. Denote $\theta^{(k)}$ as the parameters simulated from the k-th draw. For forecasting log-mortality rates at ℓ year-ahead, we first draw

$$\kappa^{(k)}_{T+\ell} \sim \mathcal{N}(\kappa^{(k)}_T + \ell d^{(k)}, \ell(\sigma^{(k)}_\omega)^2),$$

   then draw

$$y^{(k)}_{T+\ell} \sim \mathcal{N}(\alpha^{(k)} + \beta^{(k)}\kappa^{(k)}_{T+\ell}, \Sigma^{(k)}).$$

3. For each ℓ, draw $y^{(k)}_{T+\ell}$ for every $k \in [\frac{m}{2}+1, m]$. These draws then form the empirical posterior predictive distribution for the log-mortality rates at year $T+\ell$.

Given that the chains attain convergence[5], we could closely approximate this distribution, and calculate the posterior mean consistently. For the confidence interval, we could just take the 0.025 and 0.975 quantiles from the empirical distributions. Next we fit our model with real data.

[5]Some methods for testing convergence of Markov Chains will be introduced in section 2.5

2.4 The Bayesian model applied to the Lee-Carter model

In this section, we use the sex neutral central death rates in Netherlands from 1949 to 2009, and ages from 0 to 99, collected from the Human Mortality Database[6]. We use the data from 1949 to 1999 to estimate the models, and use the last 10 years to run the out-of-sample forecast.

A common approach in estimating the Lee-Carter model is to include all ages available, typically from age 0 to 110+. However, as shown in Figure 2.1, the developments of log mortality among different ages in the Dutch population is different. Therefore, as we include more ages, we are taking the risk of incorporating more irrelevant information in the estimation.[7] Therefore, from this section on, we make a specification at the choice of age groups. In particular, we run our estimation for 5 ages at a time. In other words, we first estimate our model using log mortality rates from age 0 to 4, then update the conditional posterior density for the sample size conditional on these 5 ages, and we estimate the model based on log mortality rates from age 5 to 9, and so on. We run the estimation for 20 age groups, i.e., until age groups 94 to 99. We report hereafter the results for 5 ages:0, 25, 50, 75, and 90.

One problem that should be considered carefully is the number of iterations needed for each sample size. If the iteration time is too small, we might end up with a posterior distribution that is away from the stationary one. On the other hand, if we have too many iterations, our program becomes inefficient. Furthermore, it's possible that the number of iterations needed for different age groups are different. Therefore it's necessary to assure convergence for each age group before we could draw any statistical conclusion.

We start with 2000 iterations. For a preliminary check of convergence, we can look at the plots of the parameters. In figure 2.2 and 2.3 we report the plots of part of the parameters for the model based on the sample size of 12.[8] We see that the draws of all parameters from age 25 are well mixed, while α, β, and σ_ε from age 90 are not well mixed. However, even though there exist very large peaks in the plots of parameters obtained from age 90, we can see that the means of these plots are still stable.

One possible explanation is that, since we normalize the sum of β to be 1, it could happen that when the ratios of two or more of the β-s are large, some β would become very large after normalization. To test whether this explanation is valid, we change the normalization of β to the one stated in Equation (2.3), and re-run the program. We can see from figure 2.4 that, after normalization, the plots of parameters simulated from age 90 are well mixed, which indicates of convergence of Markov chain with 2000 iterations for age 90.

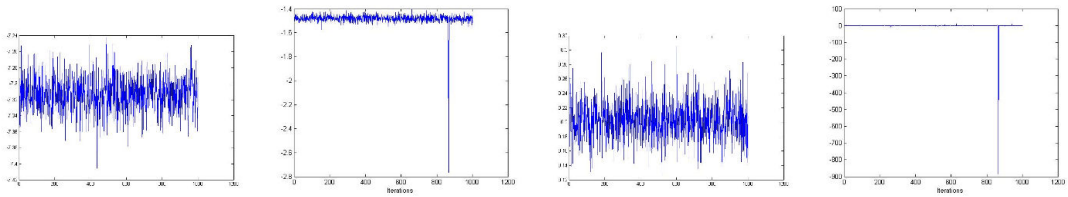
Apart from the intuitive check of convergence, there are more sophisticated statistics, of which the most famous ones are the Gelman-Rubin statistics (Gelman and Rubin 1992) and Geweke diagnostics statistics (Geweke and of Minneapolis 1991). We calculate the Gelman-Rubin statistics for the above two ages. We now give a simple example.

1. Suppose the parameter of interest is α. We run $m \geq 2$ chains, and record

[6] http://www.mortality.org/

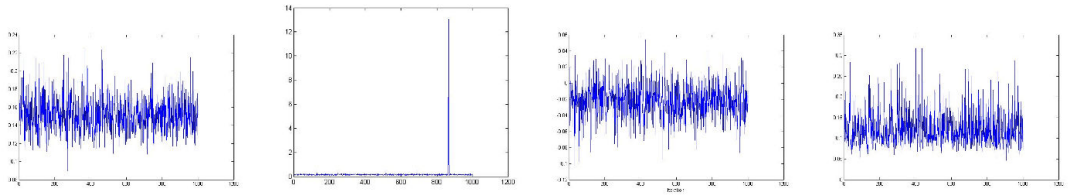
[7] We provide a more detailed discussion of this problem in the Extension.

[8] In fact we also plot the parameters based on other sample sizes, and the results are very similar.



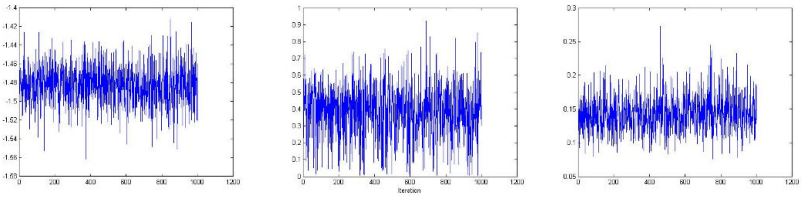
(a) α_{25} (b) α_{90} (c) β_{25} (d) β_{90}

Figure 2.2: Parameters old normalization



(a) $\sigma_{\varepsilon_{25}}$ (b) $\sigma_{\varepsilon_{90}}$ (c) d (d) σ_{ω}

Figure 2.3: Parameters old normalization (cont.)



(a) α_{90} (b) β_{90} (c) $\sigma_{\varepsilon_{90}}$

Figure 2.4: Parameters new normalization

Parameter	Gelman-Rubin Statistics	Gelman-Rubin Statistics under alternative normalization
d	0.9995	0.9999
σ_ω	0.9995	1.0043
σ_{ε_0}	0.9995	0.9996
$\sigma_{\varepsilon_{25}}$	0.9995	0.9995
$\sigma_{\varepsilon_{90}}$	0.9995	0.9995
α_0	0.9996	0.9995
α_{25}	0.9995	0.9995
α_{90}	0.9995	0.9995
β_0	3.7628	0.9995
β_{25}	0.9995	0.9995
β_{90}	1.2554	0.9995

Table 2.1: Genman-Rubin statistics

the last n components of each chain. Denote by α_j^i the i-th parameter from the j-th chain.

2. Calculate the within chain variance

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \tag{2.41}$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\alpha_j^i - \bar{\alpha}_j)^2$, and $\bar{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\alpha_j^i$.

3. Calculate the between chain variance

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\alpha}_j - \bar{\bar{\alpha}})^2, \tag{2.42}$$

where $\bar{\bar{\alpha}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\alpha}_j$.

4. Estimate the variance of α from the stationary distribution as a weighted average of W and B

$$\hat{V}(\alpha) = (1 - \frac{1}{n})W + \frac{1}{n}B. \tag{2.43}$$

5. Calculate the potential scale reduction factor

$$\hat{R} = \sqrt{\frac{\hat{V}(\alpha)}{W}}. \tag{2.44}$$

In general, if convergence is attained, the Gelman-Rubin statistics should be around 1. From table 2.1, we can see that statistics obtained from age 25 are very close to 1 under both normalizations. In addition, Gelman-Rubin statistics for β_0 and β_{90} are different from 1 under the old normalization, but close to 1 under the alternative normalization. This observation confirms our earlier conclusion.

After checking the convergence of the Markov chains, we can look at the results of forecast and inference generated by the Markov chains. Here we

report the forecasts obtained from the Bayesian model as well as the original Lee-Carter model. We see that, for all ages, the confidence intervals generated by the Bayesian model are larger than the one generated by the original Lee-Carter model. This result is consistent with our expectation, since in the Bayesian model we incorporated all parameter variances in the computation process, while in the original Lee-Carter model we only consider the variance of the κ process. Furthermore, we can see that, for age 25, the confidence interval generated by the original Lee-Carter model fails to fully capture the actual log mortality rates, which indicates that the neglect of parameter variance other than κ is a nontrivial problem.

2.5 Bayesian sample size selection

As stated earlier, due to the linear structure of the Lee-Carter model and the nonlinear development of the log mortality rates, using as large a sample size as possible might not be the best idea. Our goal is to find an optimal sample size for the Lee-Carter model. To do this, we consider the sample size as an extra parameter, M, and study its conditional posterior distribution.

For the search of optimal sample size, we extend our parameter space to be $\theta = (M, \alpha', \beta', d, \sigma_\varepsilon, \sigma_\omega)' \in \Theta =\in \mathbf{N} \times \mathbf{R}^{2N+1} \times \mathbf{R}_+^{N+1}$. In particular, we specify $M = \{12, ...51\}$. The reason is that Lee-Carter model behaves poorly, in the sense that the likelihood ratio is much lower, in most cases where the sample size is too small. Denote by M_i the Lee-Carter model estimated based on sample size of i, and Y^T the log mortality rates observed up to time T. For the purpose of comparison, all samples used to estimate end at the same period. For example, at time t, the model M_2 is estimated with sample $\{Y_{t-1}, Y_t\}$, M_3 with $\{Y_{t-2}, Y_{t-1}, Y_t\}$, and so on.

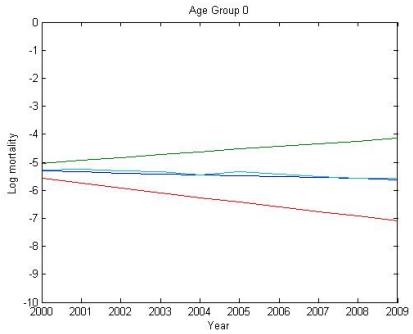
Since we have no prior knowledge that which sample size is better, we impose a uniform conditional prior distribution on M, i.e., $\Pi(M|\theta_{-M}) = \frac{1}{40}$, where θ_{-M} denotes all parameters other than M.

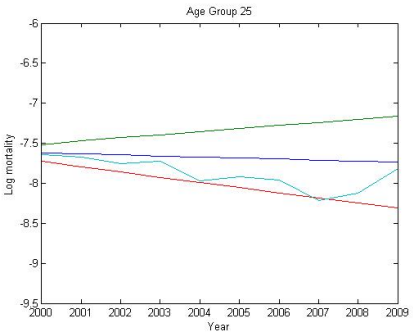
In many demographical and actuarial applications, people are interested in finding the optimal predictions of mortality rates at several years ahead. To be consistent with this purpose, we make one specification during the computation of the conditional posterior distribution of M, that is, at time t, we estimate all the parameters except for M using information only up to time $t - 10$. By doing this, we hope to approximate the situation where we are looking for the optimal "prediction" of the 10-year-ahead log mortality rates at every t.

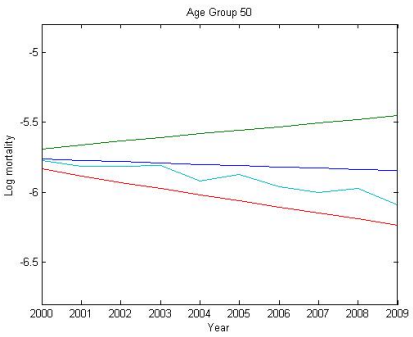
In this section, we use the Dutch log mortality rate from 1949 to 2009. In other words, we estimate the 40 models with data from 1949 to 1999, and update the conditional posterior distribution of M based on the fit of the data at year 2009. The results are shown in Figure 2.6. We can see that there is a peak at around sample size 30 for age 75, but except for that, there is no clear peak for the posterior density of M. Therefore, it seems that the conditional posterior distribution of sample size is still vague with only one update. In addition, the sample size shorter than 15 has significantly lower density, which might indicate that the Bayesian models based on such short sample size do not perform well.

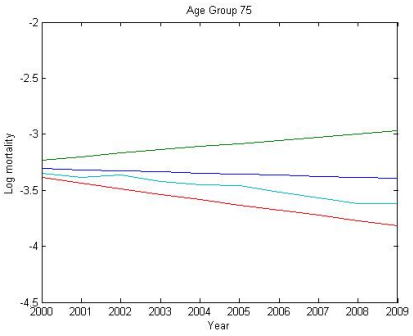
Furthermore, we show some summary statistics of the conditional posterior densities for the sample size in Table 2.2. We see that the posterior means concentrate within 31 to 35. Also, the standard deviations are large, which verifies the conclusion that our current knowledge of the behavior of sample size

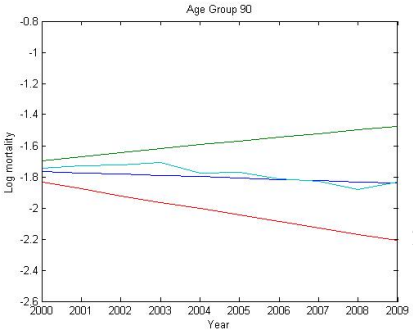
Original Lee-Carter model



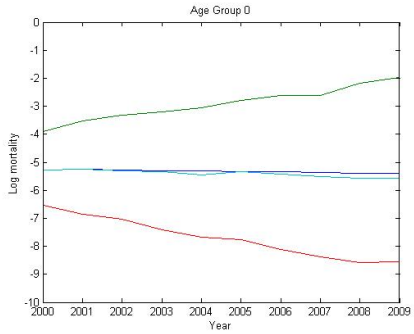


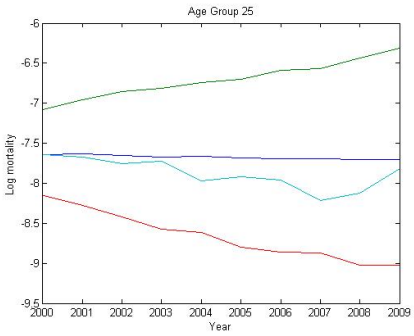


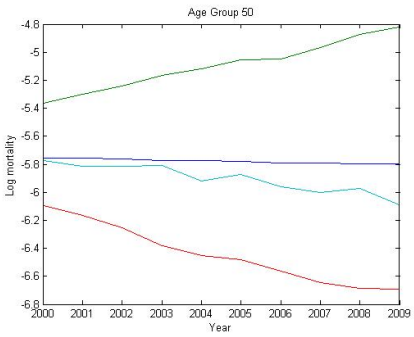


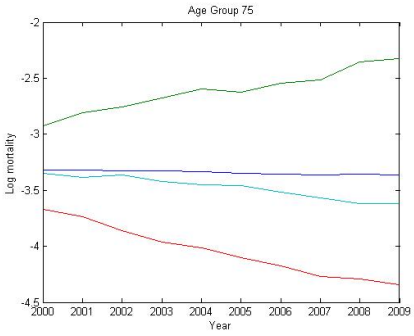


the Bayesian model









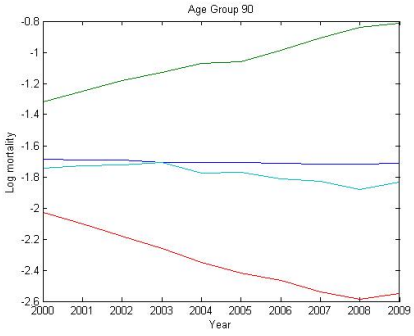


Figure 2.5: Static Bayesian model forecasts

Age Group	Mean	Median	Standard deviation
0	32.6227	33.5	11.0891
25	33.8986	35.5	11.2517
50	34.0199	35.5	10.9156
75	31.5030	30.5	11.1176
90	33.9310	35.5	11.0904

Table 2.2: Summary of Statistics static model

is limited.

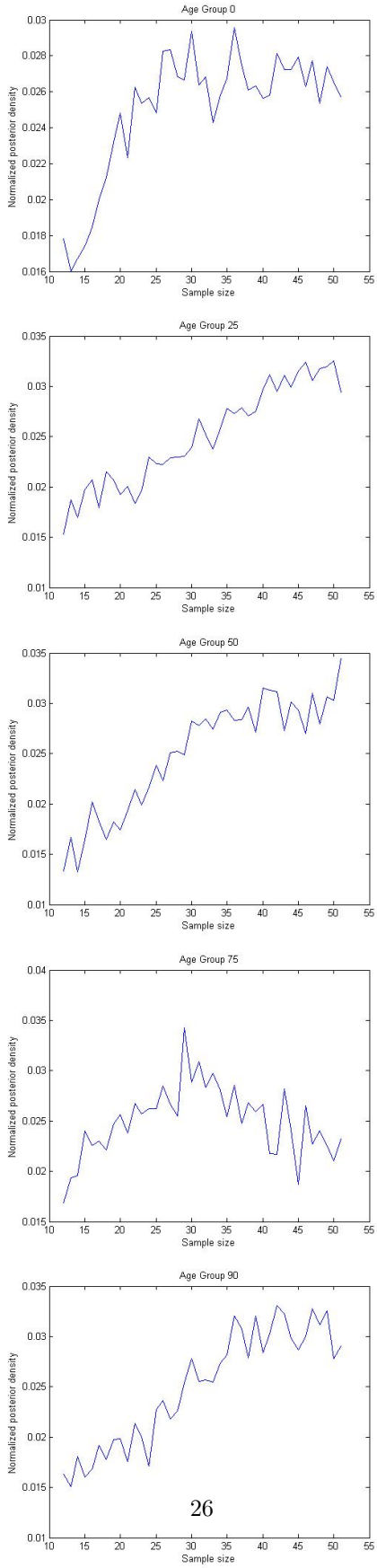


Figure 2.6: Conditional posterior density

Chapter 3

Dynamic analysis

In the previous chapter, we have conducted a static analysis, which means that we only update the posterior distribution once. As we see from the empirical results, by doing only one update, we are not able to find an optimal sample size in general. Therefore, one might wonder what would happen if we update the conditional posterior distribution of M for more than one time, by treating the previous conditional posterior distribution as the conditional prior distribution at every stage. This idea brings us to the dynamic Bayesian model.

As in section 2.5, we consider the sample size, $M = \{12, 13, ..., n\}$, as one component of the extended parameter space $\theta = (M, \alpha', \beta', d, \sigma_\varepsilon, \sigma_\omega)' \in \Theta =\in \mathbf{N}\times\mathbf{R}^{2N+1}\times\mathbf{R}_+^{N+1}$, with the corresponding appropriate σ-algebra. As a starting point, we may first consider a lower dimensional subset of Θ. For example, we may first study the conditional distribution of M, and treat the rest of parameters as known fixed values.

3.1 Sequential Bayesian updating

In the context of Bayesian updating, we consider the situation where we obtain data $Y_1, Y_2, ...$ sequentially at every stage. At time t_0, we impose a prior distribution $\Pi(\theta)$. At time $t,\ t \geq 1$, we can treat the posterior distribution obtained in time $t-1$, $\Pi(\theta|Y_1, ...Y_{t-1})$, as the prior distribution at t. For example, in a discrete case, at time t, the posterior probability of any set $G \in \mathcal{G}$ is

$$\Pi(\theta \in G|Y_1, ..., Y_t) \propto \int_G p_\theta(Y_t) d\Pi(\theta|Y_1, ..., Y_{t-1}). \tag{3.1}$$

Again, denote by $\pi(\theta|Y_1, ..., Y_{t-1})$ the posterior density at time $t-1$. Then

$$\pi(\theta|Y_1, ..., T_t)d\theta \propto p_\theta(Y_t) d\Pi(\theta|Y_1, ..., Y_{t-1}). \tag{3.2}$$

3.2 Case I: conditional optimal sample size, a basic approach

3.2.1 Model setup

To illustrate the idea of sequential Bayesian updating, we first consider the simplest case, where we have only a one-dimensional, discrete parameter space.

Denote by M_n the model estimated with sample size n, and fix all other parameters, our parameter space becomes $\theta = M \in \Theta = \mathbf{N}$. Suppose that we have observed the information up to time T, let $n \in \{m, ..., N\}$ with some $N < T$ and $m > 1$ be the collection of possible lengths of sample size that we consider. Again, to reflect our prior ignorance, we impose a uniform prior on $(\Theta, \mathcal{G})$. Suppose n have k components, then

$$\pi_\theta(M_i) = \frac{1}{k}, \text{ for every } i \in n. \tag{3.3}$$

Suppose we want to find a fit of the log mortality rates at time t that yields minimal quadratic variation using information up to time t. We could then write (3.2) as

$$\pi(M_n|Y_1, ..., Y_t) \propto p_{M_n}(Y_t)\pi(M_n|Y_1, ..., Y_{t-1}), \tag{3.4}$$

for any $n \in \{m, ..., N\}$, with

$$\pi(M_n|Y_1, ...Y_{t-1}) = \frac{p_{M_n}(Y_{t-1})}{\sum_{i=m}^{N} p_{M_i}(Y_{t-1})}. \tag{3.5}$$

We then repeat this computation from time t to T. Finally, we end up with a posterior density function $\pi_{\theta|Y_1,...,Y_T}(\cdot)$, which we can use to predict the log mortality rates in the future. Denote by $\hat{Y}_{x,\tau}(M_n)$ the forecasted log mortality rate of age x at time $\tau > T$ using model M_n, the weighted forecasted log mortality rate can be computed up to a constant,

$$\bar{Y}_{x,\tau} \propto \sum_{i=1}^{T} \pi_{\theta|Y_T}(M_i)\hat{Y}_{x,\tau}(M_i) \tag{3.6}$$

In order to yield comparable results to the ones shown in chapter 2, we also make the specification that, at any time t, we estimate all the parameters except for M using information only up to time $t - 10$.

3.2.2 Application to the Lee-Carter model

In this section, we base our analysis on the log mortality rates from Dutch population from 1928 to 2009, and the ages from 0 to 99. In particular, we use the log mortality rates from 1928 to 1979 to estimate and save θ_{-M} using the weighted least square estimation. Then we use samples from 1980 to 1999 to update the posterior densities. At every year, we first re-estimate and save θ_{-M}, then use θ_{-M} to compute the likelihood and update the conditional posterior density of M.

In this chapter we also choose $M = \{12, 13, ..., 51\}$, denote by M_i the Lee-Carter model estimated based on sample size of i, and Y^T the log mortality rates observed up to time T. Also, all samples used to estimate end at the same period. For example, at time t, the model M_2 is estimated with sample $\{Y_{t-1}, Y_t\}$, M_3 with $\{Y_{t-2}, Y_{t-1}, Y_t\}$, and so on.

The normalized conditional posterior densities for M obtained in the end of our sample are presented in Figure 3.1. We can see that, different from the static Bayesian model, the conditional posterior densities now have clear peaks, which means that we now have more information on the behavior of sample size.

Age Group	Mean	Median	Standard deviation
0	24.1468	24	4.0298
25	21.1002	21.5	1.8310
50	34.7548	35.5	2.9057
75	25.8562	26.5	2.9474
90	29.0189	29.5	4.5116

Table 3.1: Summary of Statistics dynamic model (basic)

Furthermore, the peaks are obviously different among different ages. However, the sample sizes of length 20 to 40 seems to cover most probability mass for all ages. This result verifies the preliminary observation in Figure 2.1: the development of life expectancy of the Dutch population is non-linear, thus the largest sample size does not have to be the optimal one.

From Table 3.1, we see that the posterior means of the conditional distributions of sample size are more dispersed compared with the ones obtained from the static model. Also, the standard deviations are much smaller in this case. Therefore, it seems like we have gained more knowledge on the posterior behavior of sample size after sequential updating.

Note that the results shown in Figure 3.1 should be interpreted with caution. In particular, the whole computation process is implicitly conditioned on the point estimations of all parameters except for M, i.e., the true value of $\theta_{-M} = (\alpha', \beta', d, \sigma_\varepsilon, \sigma_\omega)'$ is assumed known. Therefore, it ignores the parameter variance and may probably induce biases.

3.3 Case II: conditional optimal sample size, an extended approach

3.3.1 Model setup

In this section, we extend the search for an optimal sample size. In particular, we combine the method from the static analysis with the simple dynamic method introduced in the last section. Again, define $\Theta = \{\theta : \theta = (M, \alpha', \beta', d, \sigma_\varepsilon, \sigma_\omega)'\}$ and $\Theta_{-M} = \{\theta : \theta = (\alpha', \beta', d, \sigma_\varepsilon, \sigma_\omega)'\}$.

First of all , we define a conditional distribution induced by Π in Equation (2.5),

$$\Pi_{\theta_{-M}|M}(\cdot|M) : \sigma(\Theta_{-M}) \times \mathcal{N} \to [0, 1], \tag{3.7}$$

with $\Theta_{-M} = \{\mathcal{R}^{2N+1} \times \mathcal{R}_+^{N+1}\}$, and $\sigma(\Theta_{-M})$ the corresponding Borrel σ-algebra. Equation (3.7) can be interpreted as the "conditional prior" distribution of the parameter $\theta_{-M} \in \Theta_{-M}$. Therefore, we can construct the corresponding (conditional) posterior distribution in the similar way as we did in Chapter 2. Denote this posterior distribution by

$$\Pi_{\theta_{-M}|M,Y}(\cdot|M, Y) : \sigma(\Theta_{-M}) \times \mathcal{N} \times \mathcal{Y} \to [0, 1]. \tag{3.8}$$

Note that, Equation (3.7) and (3.8) can be interpreted as the extensions of the prior and posterior distributions that we used in static analysis. Conditional upon M and Y, the study of θ_{-M} is no different from that in the static case by

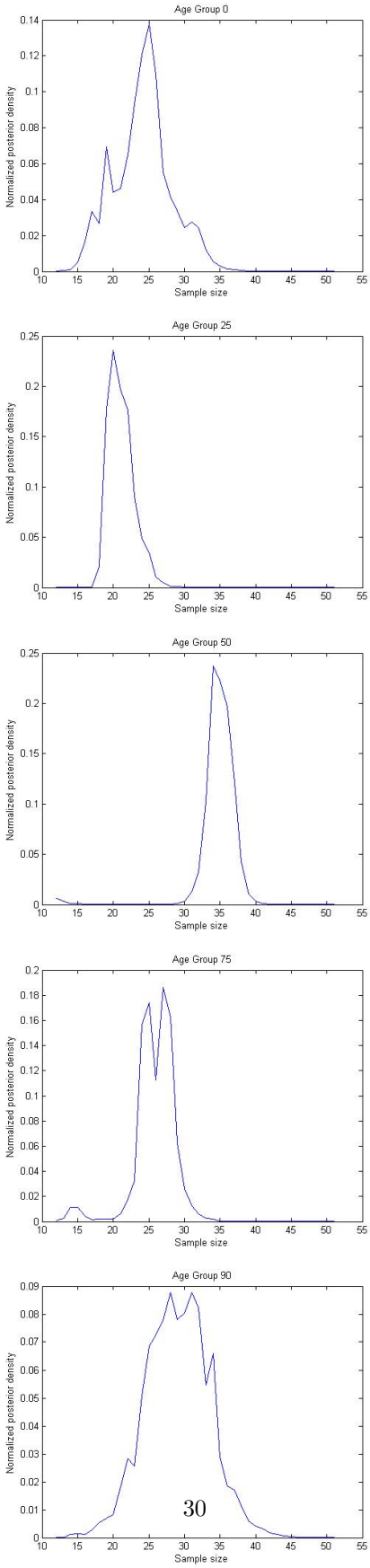


Figure 3.1: Normalized posterior densities (basic)

using these two equations. The additional property of Equation (3.7) and (3.8) is that different values of M-s and Y-s can be explicitly considered.

Now we are able to proceed. The procedure is as follows.

1. Define M, and adopt the choice of sample as in the previous section. Start from any $t \geq 45$, run the Gibbs sampler introduced in section 2.2.3 for each M_i. We set the number of iterations as 2000 for every run, and save the parameters drawn in the last 1000 iterations.

2. Impose a uniform conditional prior of M, as in the previous section. Denote by $\theta^{(j)}_{(M_i,T)}$ the j-th draw of parameter for model M_i at time T, the posterior conditional distribution of M at time T can be written as

$$\Pi_{M_n|Y_1,...Y_T} \propto \frac{1}{1000} \sum_{j=1}^{1000} p_{\theta^{(j)}_{(M_i,T)}}(Y_T)\Pi(M_n|Y_1,...,Y_{T-1}). \tag{3.9}$$

3. Continue step 2 up to the end of observed sample, which is year 2009 in our case. We then obtain the evolution of posterior densities of each M_i. Furthermore, we compute the weighted average of the normalized posterior densities of M_i-s as the optimal conditional sample size.

After obtaining the optimal conditional sample size, we can use it to do the prediction and inference in a similar way as in section 2.3.

3.3.2 Application to the Lee-Carter model

We use here the same data as in last section, i.e., we use log mortality rates from 1928 to 1979 to estimate θ_{-M}, and use samples from 1980 to 1999 to update the posterior densities. The results are presented in Figure 3.2. We can see that the normalized posterior densities are different among ages, and are different than the ones obtained from the static Bayesian model. We can see that the conditional posterior density have also clear peaks. In addition, from Table 3.2, we see that the posterior means are on average larger than the one obtained in section 3.1.2. In particular, except for age 0, the optimal interval of sample size increases as the age increases. This fact is consistent with the observation in Figure /reffig:dutchmortality. The development of the log mortality rates of ages 75 and 90 are almost linear, thus the optimal sample sizes for these ages tend to be large. Meanwhile, the log mortality of age 25 and 50 develop in a less linear way, thus the largest sample size is not the best one. In addition, the log mortality rates of infants decreased the most during the last one hundred years, and displayed a somehow linear trend in the last 60 years. Therefore, the optimal sample size for age 0 is in middle of all ages.

In addition, the forecasts and inference generated by the dynamic Bayesian model are presented in Figure 3.3. We can see that the confidence intervals generated by the dynamic Bayesian model is much smaller than the ones generated by the static model, which indicates that the dynamic model is indeed more efficient. Furthermore, we obtain a smaller confidence interval for age 0 compared with the one obtained by the original Lee-Carter model. Also, the realized log mortality of age 25 does not exceed the confidence interval, as in the original Lee-Carter case.

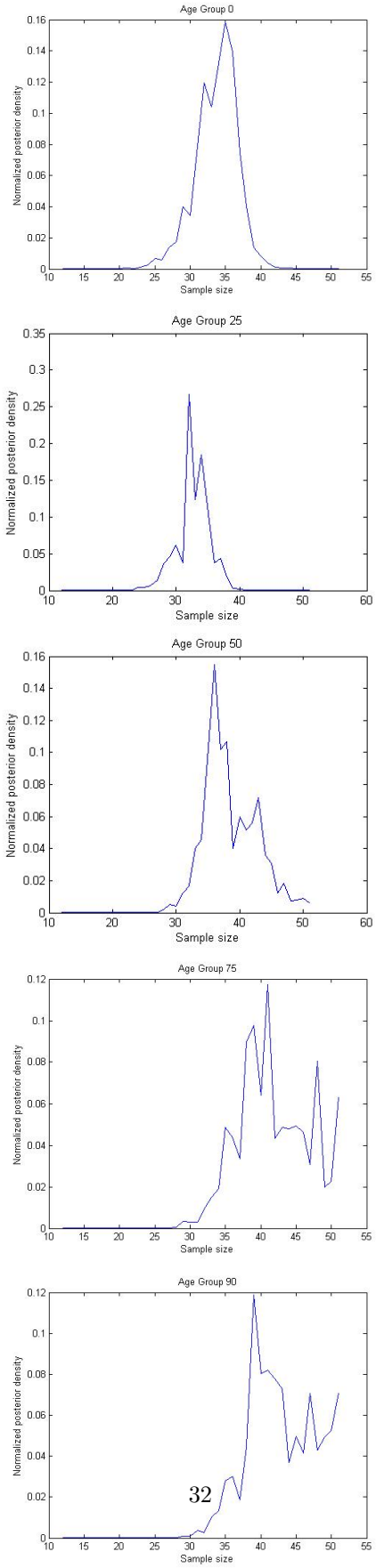


Figure 3.2: Normalized posterior densities (extended)

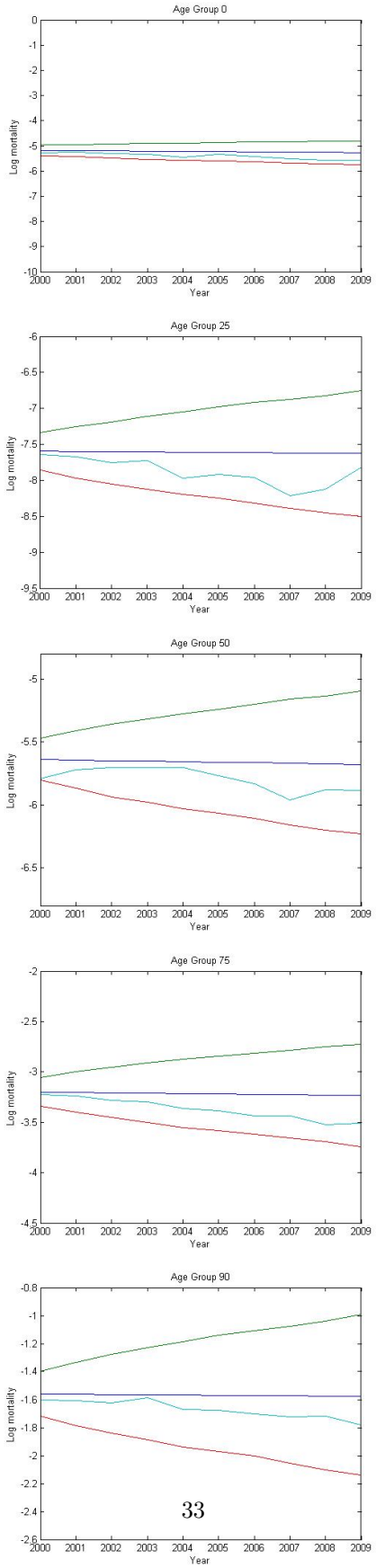


Figure 3.3: Dynamic Bayesian model forecasts

Age Group	Mean	Median	Standard deviation
0	33.7209	34.5	2.9696
25	32.6860	33.5	2.5962
50	38.5340	38.5	4.2238
75	41.8286	41.5	5.0002
90	42.9300	42.5	4.8163

Table 3.2: Summary of Statistics dynamic model

Chapter 4

Conclusion

In this thesis we investigate the choice of the sample size of the Lee-Carter model by using a Bayesian approach. In particular, we formalize the Lee-Carter model to be a statistical model, and explicitly consider the sample size as a parameter of the model, and finally compute the conditional posterior density of the sample size. We estimate both the static and the dynamic Bayesian model; in the former case we update the conditional posterior density once, and in the latter case we update the conditional posterior density at every stage. Furthermore, we do the out-of-sample forecast and carry out inference in both cases. Specifically , we estimate the Lee-Carter model based on the different sample sizes, then forecast the future log mortality rate and confidence intervals for each sample size. We then weighted these forecasts and confidence intervals by the conditional posterior density for sample size. We report the results for 5 ages: 0, 25, 50, 75, and 90.

In the static case we observe no clear peak in the conditional posterior density of sample size for all ages except for age 75, while in the dynamic case we obtain clear peaks in the conditional posterior density for all ages. Moreover, the standard deviations are around 11 for the conditional posterior density for the sample size for all ages from the static model, while in the dynamic case they range from 3 to 5. Therefore, it is likely that we gain more information on the conditional posterior density of the sample size after sequential updating. In particular, in the dynamic case, the posterior means of sample size range from 32 to 43, and display an increasing trend except for age 0. This finding is consistent with the preliminary observation of the development of log mortality rate of the Dutch population in the past 160 years.

In addition, in both cases, we obtain wider confidence intervals in all ages except for age 0 compared to the original Lee-Carter estimation. However, in the dynamic case, we observe significantly narrower confidence interval than in the static case for all ages. Furthermore, for age 25, the realized log mortality exceeds the 95% confidence interval for the original Lee-Carter estimation, but not for either Bayesian estimation. Therefore, it seems that the dynamic Bayesian model delivers appropriately wider confidence intervals than the original Lee-Carter model, and we obtain more accurate and reasonable forecasts and inference results by using the dynamic Bayesian model.

The thesis also has some future research potentials. A possible extension is to apply the Bayesian model to other sources of data, for instance, male or

female log mortality data of the Dutch population, or mortality data from other countries. We expect different patterns of conditional posterior distribution for the sample size with different data. In addition, due to the compatibility of the Bayesian model and the Markov Chain Monte Carlo method, it is possible to apply the Bayesian model to model specifications other than the Lee-Carter model.

Chapter 5

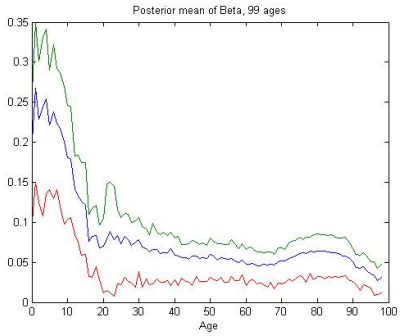
Extension: An examination of the choice of age groups

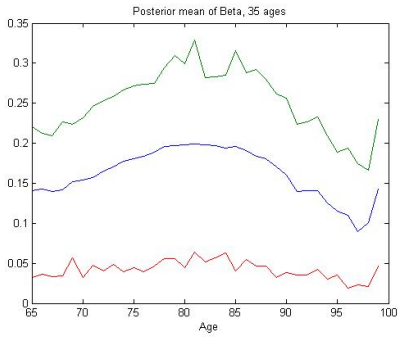
In the previous chapters, we estimate the Bayesian model for 5 ages at a time, and thus run the estimation for 20 age groups. However, one may wonder whether the Bayesian model is, especially computational, capable of the estimation of all ages at a time. In this chapter, for the purpose of illustration, we estimate the dynamic Bayesian model for 5 times, including ages 0-98, 65-99, 65-84, 75-79, and 75-76, respectively, and study the posterior distributions of the relevant parameters.

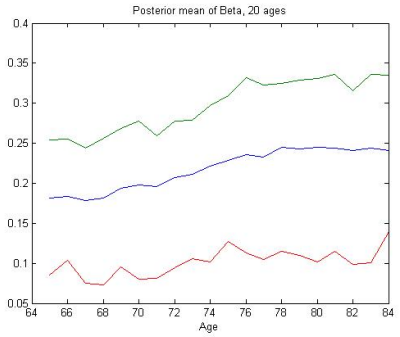
Figure 5.1 reports the posterior mean, as well as the 95% confidence intervals of β for the 5 estimations. We can see that the posterior means of β obtained from the 5 estimations are very consistent, in the sense that they are very similar to each other within the same age interval. Due to the normalization we make in Equation (2.3), the posterior means of β are of different magnitude, nevertheless they display almost the same shape in the same age interval. Similarly, the posterior means of σ_ε obtained from the 5 estimations are also very consistent, as shown in Figure 5.2. Also, from the first sub-figure in Figure 5.2, we can see that the posterior means of σ_ε are not significantly different among different ages. Therefore, it does not harm in assuming that the ε_t-s in Equation (2.1) are homoscedastic.

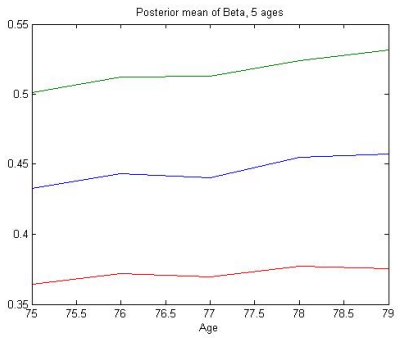
Figure 5.3 plots the posterior density of σ_ω obtained from the 5 estimations. We can see that as the number of ages included decrease from 99 to 5, the posterior density of σ_ω becomes narrower and more concentrated at the region near 0, reflecting more precise knowledge about the behavior of the κ process. One plausible explanation is that, as shown in Figure 2.1, the developments of log mortalities are different among different ages, therefore the more ages we include, the more irrelevant information is taken into account, and thus the variance of ω_t becomes larger. However, we notice that the posterior densities obtained using 5 ages and 2 ages are very similar.

Similar results can be observed in Figure 5.4. As the number of ages included decreases from 99 to 5, the 95% confidence intervals of the forecasted log mortality become narrower and narrower. Meanwhile, the forecasted confidence intervals obtained using 2 ages are significantly wider than those obtained using 5 ages. The reason might be that, when we use only 2 ages for estimation, we are using too few data so that the results are more noisy.









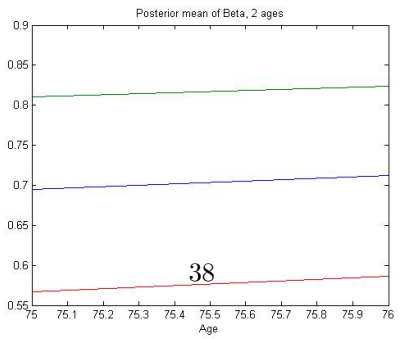
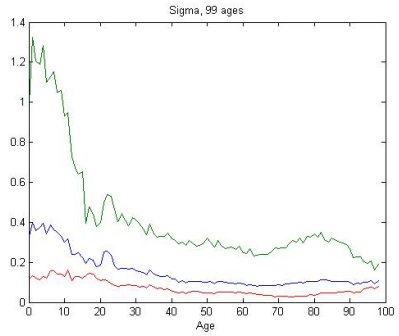
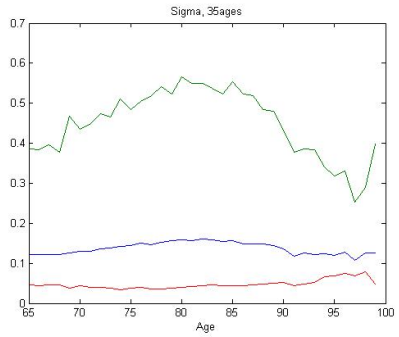
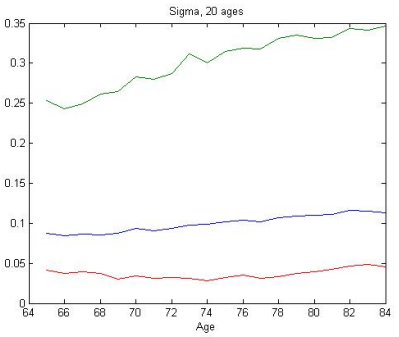
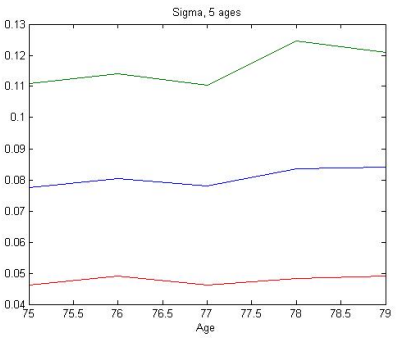


Figure 5.1: Posterior mean of β









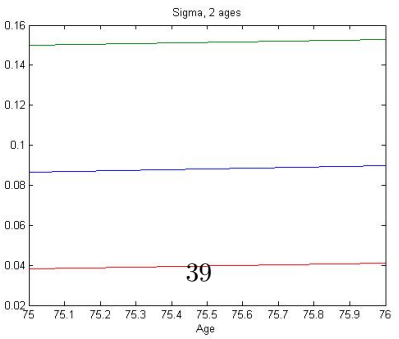
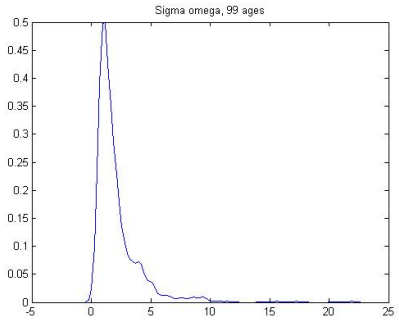
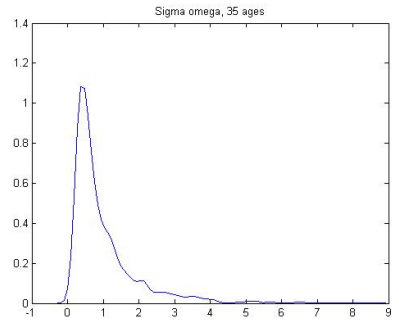
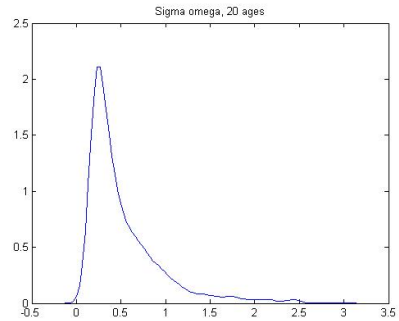
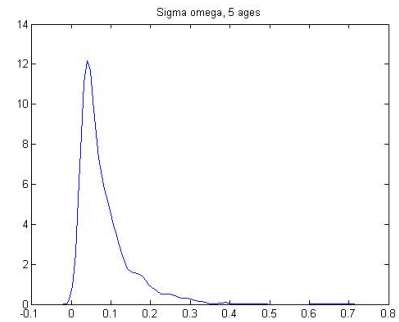


Figure 5.2: Posterior means of σ_ε









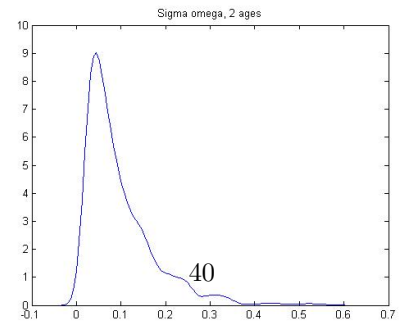


Figure 5.3: Posterior densities of σ_ω

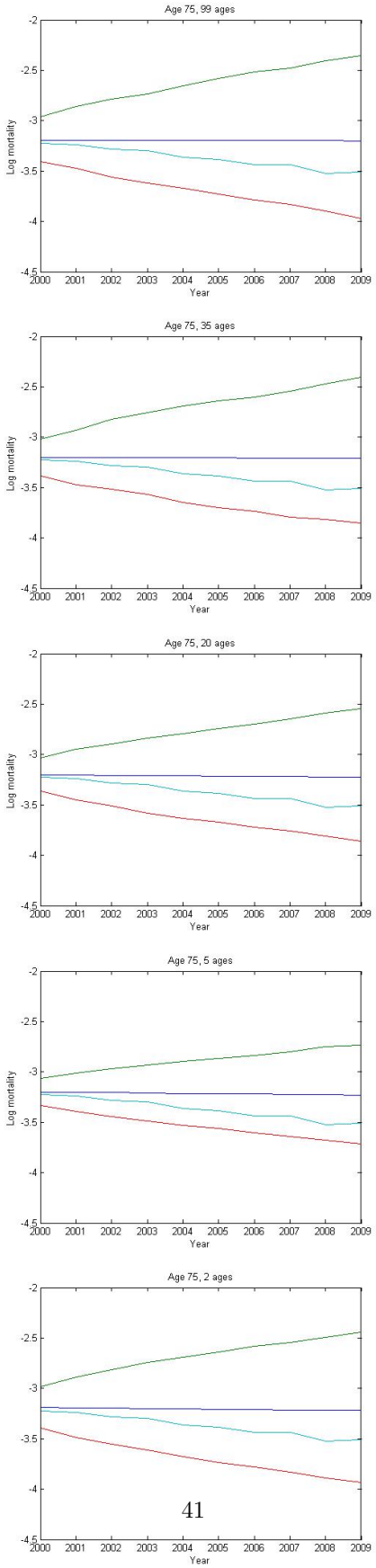


Figure 5.4: Out-of-Sample forecast

References

Anosov, D. (2001). Ergodic theory. *Hazewinkel, Michiel, Encyclopaedia of Mathematics, Kluwer Academic Publishers, ISBN*, 978–1556080104.

Athanasios, P. (1991). *Probability, random variables and stochastic processes*. McGraw-Hill.

Booth, H., J. Maindonald, and L. Smith (2002). Applying lee-carter under conditions of variable mortality decline. *Population Studies 56*(3), 325–336.

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

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

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

Geweke, J. and F. R. B. of Minneapolis (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department.

Hári, N., A. De Waegenaere, B. Melenberg, and T. Nijman (2008). Longevity risk in portfolios of pension annuities. *Insurance: Mathematics and Economics 42*(2), 505–519.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Kalman, R. et al. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Kleijn, B. (2012). Lecture notes in Universiteit van Amsterdam, downloadable from http://home.medewerker.uva.nl/b.j.k.kleijn/page1.html.

Koissi, M., A. Shapiro, and G. Högnäs (2006). Evaluating and extending the lee–carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics 38*(1), 1–20.

Lee, R. (2000). The lee-carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal 4*(1), 80–93.

Lee, R. and L. Carter (1992). Modeling and forecasting us mortality. *Journal of the American Statistical Association*, 659–671.

Lee, R. and T. Miller (2001). Evaluating the performance of the lee-carter method for forecasting mortality. *Demography 38*(4), 537–549.

Lee, R. and S. Tuljapurkar (1994). Stochastic population forecasts for the united states: Beyond high, medium, and low. *Journal of the American Statistical Association*, 1175–1189.

Meteopolis, N. and S. Ulam (1949). The monte carlo method. *Journal of the American Statistical Association 44*(247), 335–341.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 1087.

Pedroza, C. (2002). *Bayesian Hierarchical Time Series Modeling of Mortality Rates*. Ph. D. thesis, Havard University.

Pedroza, C. (2006). A bayesian forecasting model: predicting us male mortality. *Biostatistics 7*(4), 530–550.

Pitacco, E., M. Denuit, S. Haberman, and A. Olivieri (2009). *Modelling longevity dynamics for pensions and annuity business*. Oxford University Press, USA.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 1701–1728.

Walsh, B. (2004). Markov chain monte carlo and gibbs sampling. downloadable from http://web.mit.edu/ wingated/www/introductions/mcmc-gibbs-intro.pdf.

Walters, P. (2000). *An introduction to ergodic theory*, Volume 79. Springer Verlag.

Wilmoth, J. (1993). Computational methods for fitting and extrapolating the lee-carter model of mortality change. Technical report, Technical report, Department of Demography, University of California, Berkeley.