A Particle Markov Chain Monte Carlo Approach for the Estimation of CBD-Type Models

by

Xueyi Xu

B.Sc., Simon Fraser University, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Declaration of Committee

Name: Xueyi Xu

Degree: Master of Science

Thesis title: A Particle Markov Chain Monte Carlo Approach for the Estimation of CBD-Type Models

Committee:

Chair: Joan Hu
Professor, Statistics and Actuarial Science

Barbara Sanders
Senior Supervisor
Associate Professor, Statistics and Actuarial Science

Jean-François Bégin
Co-Supervisor
Assistant Professor, Statistics and Actuarial Science

Himchan Jeong
External Examiner
Assistant Professor, Statistics and Actuarial Science

Abstract

The current literature on mortality has mainly focused on model specification, giving less regard to parameter estimation. Indeed, over the last three decades, multiple mortality models have been introduced, most being extensions of the well-known Lee-Carter model or the Cairns-Black-Dowd (CBD) model. However, the estimation of these models has been somewhat overlooked; most papers focus on frequentist methods, such as the (two-stage) maximum likelihood estimation method that estimates the mortality parameters first and then the parameters of the mortality improvement dynamics second. In this report, we present a new Bayesian-based estimation procedure for CBD-type models that relies on the particle Markov chain Monte Carlo (pMCMC) method of Andrieu et al. (2010). This methodology captures the dynamic nature of the mortality improvement factors (and their

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors, Prof. Barbara Sanders and Dr. Jean-François Bégin, without whom this project would not have been successful. Their invaluable knowledge and continuous guidance are incredibly useful and are greatly appreciated.

I would like to extend my thanks to Dr. X. Joan Hu for chairing my defence, and Dr. Himchan Jeong for investing his precious time in serving the examining committee, making insightful suggestions and helping refine the project. I also would like to thank all sta and faculty members of the Department of Statistics and Actuarial Science for their dedication, especially Dr. Cary Tsai, Dr. Yi Lu, Dr. Rachel Altman, Dr. Tim Swartz, and Dr. Haolun Shi, whose courses greatly broadened my knowledge in actuarial science and statistics. Furthermore, I would like to take this opportunity to thank the Graduate Student Society and the Department of Statistics and Actuarial Science at Simon Fraser University for o ering financial supports throughout my graduate study.

Finally, I am always grateful to my academic fellows and loving friends for sharing their experiences and being with me during my studies at Simon Fraser University. Special thanks to my cat, Dummy, for companying with me when I stay up for studying, coding, and writing up the report, although he always falls asleep after few minutes.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Longevity risk is one of the most signi cant risks that insurance companies, pension plan sponsors, and government are exposed to. Longevity risk arises when people live longer than expected, which means pensions or annuities need to be paid for a much longer time than expected, resulting in higher pension plan and insurance liabilities.

To capture and understand longevity risk, a wide variety of stochastic mortality models has been developed during the last three decades. Most of these are extensions of the well-known Lee-Carter (LC) model or of the Cairns-Black-Dowd (CBD) model. Lee and Carter (1992) introduce the rst stochastic mortality model, and since then, multiple LC-type mortality models were proposed; for instance, Renshaw and Haberman (2006) extend the LC model to incorporate cohort e ect. Currie et al. (2006) introduce a simpler age-period-cohort model, which can be seen as a special case of the Renshaw and Haberman (2006) model.

Cairns et al. (2006) contribute the main mortality modelling alternative to the LC model. Their framework is also extended; for instance, Cairns et al. (2009) propose new CBD-type models that include cohort-e ect factors and other features. Plat (2009) combines the CBD model with cohort e ect with some features of the LC model. Inspired by Plat (2009), Dowd et al. (2020) extend the CBD-type models of Cairns et al. (2009) by including a static age function $_x$; these are called CBD-X models.

All the mortality models above are constructed from a mixture of independent deter-

following relationship:

$$q_{x;t} = 1 \quad \exp(\quad _{x;t}) = 1 \quad \exp(\quad m_{x;t}):$$

Age e□ects represent mortality variation by age, regardless of birth cohort. Period e□ects capture mortality changes over time that equally a□ect all ages during a particular calendar year. Cohort e□ects represent mortality variations resulting from di□erent generations represented by the year of birth. Throughout this report, we use $c = t \quad x$ to represent the year of birth or cohort year. According to Hunt and Blake (2021), most of the existing stochastic mortality models in the literature can be written as the following age-period-cohort e□ect model,

$$_{x;t} = \quad _x + \sum_{i=1}^{N} \quad _x^{(i)} \; _t^{(i)} \; _c^{(i)}; \tag{2.1}$$

where

- $\hat{}$ $_{x;t}$ can be the death rate on logarithmic scale (i.e., $_{x;t} = \log(m_{x;t})$) or the mortality rate in logit scale (i.e., $_{x;t} = \log \frac{q_{x;t}}{1 \quad q_{x;t}} = \text{logit}(q_{x;t})$) for people aged x in year t;
- $\hat{}$ $_x$ is a static age function, which is the basic age e□ect and can be treated as the mortality table without mortality improvement;
- $\hat{}$ N is the number of period and/or cohort terms within the model;
- $\hat{}$ $_x^{(i)}$ is the i^{th} age-e□ect factor, which variations are associated with physiological and social ageing processes;
- $\hat{}$ $_t^{(i)}$ is the i^{th} period-e□ect factor that a□ects mortality trend for people of all ages that are alive in period t; for example, this can be related to environmental conditions, medical developments, and living conditions;
- $\hat{}$ $_c^{(i)}$ represents the i^{th} cohort e□ect; that is, the impact of the past conditions on current mortality rate. For instance, generation-specific habits, wars, and catastrophes, for individuals born in the same year.

The two major mortality model families used in actuarial science are nested within this general model. They are the LC-type models and the CBD-type models.

2.1 Generalized Lee and Carter-Type Models

Lee and Carter (1992) present an original model that fits and predicts mortality rates for the United States. It is a two-factor model that has one period-e□ect factor and two age-e□ect

widely used for various countries' demographic and actuarial applications (e.g., Lundström and Qvist, 2004; Haqqi Anna Zili et al., 2018; Kamaruddin and Noriszura, 2018).

2.1.1 The Original LC Model

Lee and Carter (1992) propose the following two-factor model for death rates:

$$_{x;t} = \ _{x} + \ ^{(1)}_{x} \ ^{(1)}_{t} \ ;$$

where $_{x;t} = \log(m_{x;t})$ and parameters $_{x}$ describe the basic pattern of $\log(m_{x;t})$ averaged over time for each age. Parameters $^{(1)}_{t}$ express the overall evolution of mortality over time;

where $\bar{x} = \frac{\sum_{x=x_{min}}^{x=x_{max}} x}{x_{max} - x_{min} + 1}$ is the average of fitted ages. This model can be written as a nested version of Equation (2.1) by setting $N = 2$, $_x = 0$ for all x, $^{(1)}_x = 1$ for all x, $^{(2)}_x = (x - \bar{x})$, and $^{(1)}_c = {}^{(2)}_c = 1$ for all c. The first period-effect factor $^{(1)}_t$ represents the overall mortality improvement over time, and the second $^{(2)}_t$ captures the different mortality improvement by ages; specifically, people aged below the mean age $\bar{x}$ have larger mortality improvements than people aged above the mean age $\bar{x}$. They use two period-effect factors because the quantification of mortality improvement through only one period-effect factor is not adequate. The same authors point out that having only one period-effect factor $^{(1)}_t$ in the LC model implies perfect correlation among the changes of central death rates for

where $_{x;t} = \text{logit}(q_{x;t})$ and $^2_x = \frac{\sum_{x=x_{max}}^{x=x_{min}} (x \quad \bar{x})^2}{x_{max} \quad x_{min} + 1}$. This model can be written as a nested case of Equation (2.1) by setting $N = 4$, $_x = 0$ for all x, $^{(1)}_x = \quad ^{(4)}_x = 1$ for all x, $^{(2)}_x = (x \quad \bar{x})$, $^{(3)}_x = ((x \quad \bar{x})^2 \quad ^2_x)$, $^{(4)}_t = 1$ for all t, and $^{(1)}_c = \quad ^{(2)}_c = \quad ^{(3)}_c = 1$ for all c. Parameters $^{(1)}_t$ and $^{(2)}_t$ have the same meaning as their counterparts in the original CBD model, and $^{(3)}_t$ captures non-linearity with respect to age in year t.

Figure 2.1: US mortality rates for the year 2003. The dots are $\text{logit}(q_{x;t})$ calculated from the observed data and the solid line is the model-based estimated mortality rate in logit scale, $\text{logit}(\hat{q}_{x;t})$.

2.3 The CBD-X Models

Several extensions of the CBD-type model are introduced by Dowd et al. (2020), which originate from Plat (2009). These models combine elements of LC-type and CBD-type models. The resulting models can be used for full age ranges while capturing the cohort e ect.

2.3.1 The Plat Model

Plat (2009) assumes a static age function $_x$, three period-e ect terms, and cohort e ect. The Plat model is thus given by:

$$_{x;t} = \quad _x + \quad ^{(1)}_t + \quad ^{(2)}_t (x \quad \bar{x}) + \quad ^{(3)}_t (x \quad \bar{x})^+ + \quad ^{(4)}_c ;$$

where $\kappa_{x;t} = \log(m_{x;t})$ and $(x - \bar{x})^+ = \max(0; x - \bar{x})$. This model can be written as a nested case of the Equation (2.1) by setting $N = 4$, $\beta_x^{(1)} = \beta_x^{(4)} = \kappa_t^{(4)} = \gamma_c^{(1)} = \gamma_c^{(2)} = \gamma_c^{(3)} = 1$ for all x, t, or c, $\beta_x^{(2)} = (x - \bar{x})$, and $\beta_x^{(3)} = (x - \bar{x})^+$.

Plat (2009) suggests removing the third period-effect term if the user is only interested in the older ages. The two-factor reduced Plat model then becomes:

$$\kappa_{x;t} = \alpha_x + \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_c^{(3)}.$$

This model is virtually the same as the CBD-X(2) model that we will discuss in the following section.

2.3.2 The CBD-X Models

Inspired by Plat (2009), Dowd et al. (2020) investigate a new type of mortality model which combines the static age function α_x with the CBD-type models. Models of this type are called CBD-X[1] models. The CBD-X models are used to model mortality of groups of adults over a wider range of ages than usually advisable for the CBD-type models. The CBD-X models create a hybrid between the LC and the CBD models, just like the Plat model, as the LC model can be written in the CBD-X form. The CBD-X(1) model is the CBD-X model with one period effect, and the same naming logic applies to the other CBD-X models in this report. The CBDX model in Hunt and Blake (2020) is the CBD-X(2) model without cohort effect. Also, the CBD-X model in Cairns et al. (2019) can be treated as a multi-population version of the CBD-X(2) model without cohort effect.

The CBD-X models are given by the following:

$$\text{CBD-X(1)}: \quad \kappa_{x;t} = \alpha_x + \kappa_t^{(1)} + \gamma_c^{(2)}; \tag{2.2}$$

$$\text{CBD-X(2)}: \quad \kappa_{x;t} = \alpha_x + \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_c^{(3)}; \tag{2.3}$$

$$\text{CBD-X(3)}: \quad \kappa_{x;t} = \alpha_x + \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \beta_x^{(2)}) + \gamma_c^{(4)}; \tag{2.4}$$

where $\kappa_{x;t} = \log(m_{x;t})$. The CBD-X(1) model can be written as a nested case of Equation (2.1) by setting $N = 2$, and $\beta_x^{(1)} = \beta_x^{(2)} = \kappa_t^{(2)} = \gamma_c^{(1)} = 1$ for all x, t, or c. The CBD-X(1) model without cohort effect is essentially the special case of the LC model with $\beta_x^{(1)} = 1$. γ_c 483(373([(c)473

$\mu_x^{(3)} = (x - \bar{x})^2$ σ_x^2. Moreover, σ_x, $\mu_x^{(i)}$, $\mu_t^{(i)}$, and

Chapter 3

Particle Filters for State-Space Models

3.1 State-Space Model

The estimation of the mortality models mainly adopts two-stage frequentist methods, which use a maximum likelihood-based approach first, and then estimate a model for the period e□ects for forecasting purposes. However, by using this two-stage method, the dynamics of the period e□ects are not directly incorporated in the first step. Fung et al. (2017) argue that recasting di□erent classes of mortality models in a state-space formulation allows for state-

The measurement equation can be written as:

$$y_t = g(\ _t; "_t); \quad \text{for } t \ 2 \ f \ 1; : : : ; T g; \tag{3.2}$$

where $y_t \ 2 \ R^n$ is the vector of observed data at time t, $\ _t$ are the latent variables, and $"_t \ 2 \ R^n$ is the vector of error terms used in the measurement equation. The function $g : R^N \ R^n \ ! \ R^n$ handles the relationship between the current state of the latent variables and the observations. If g and f are both linear functions, then the SSM is linear; otherwise, it is non-linear. If the error terms $"_t$ and Z_t are modelled with Gaussian distributions, then we have a Gaussian SSM; otherwise, we have a non-Gaussian SSM.

The transition and measurement equations usually involve some unknown parameters . In the model, both the latent variables $\ _t$ and these unknown parameters need to be estimated. The di□erence between them is that latent variables vary from time to time while parameters are fixed. They should, therefore, be estimated di□erently.

3.1.1 State-Space Formulation of the CBD-X Models

The mortality models introduced in Chapter 2 are cast into a state-space representation. The latent variables are the period e□ects. The observed data are the logged estimated central death rates, $\hat{m}_{x;t}$.

Recall that the CBD-X(3) model can be written as in Equation (2.4):

$$\log(m_{x;t}) = \ _x + \ ^{(1)}_t + \ ^{(2)}_t (x \ \ x) + \ ^{(3)}_t \ (x \ \ x)^2 \ \ ^2_x :$$

Let the logged observed central death rate, $\log(\hat{m}_{x;t})$, be the true logged central death rate plus some noise:

$$\log(\hat{m}_{x;t}) = \log(\ m_{x;t}) + \ _{x;t} :$$

Let $y_t = [\log(\hat{m}_{x_{min};t}) \ : : : \ \log(\hat{m}_{x_{max};t})]^>$ be the set of $\log(\hat{m}_{x;t})$ for all the age x in year t. The latent variables are the period e□ects $\ _t = \ ^{(1)}_t \ \ ^{(2)}_t \ \ ^{(3)}_t \ ^>$: The error terms within the measurement equation are $"_t = ["_{x_{min};t} \ ::: \ "_{x_{max};t}]^>$. Thus, by using matrix notation, we obta Td [(t)r835 10.9701 Tf86.533 -29.729 Td [(y)]TJ/F36 7.9701 Tf 6.44 -1.636 Td [(t)]TJ

The transition equation is

$$\kappa_t = \quad + \quad \kappa_{t\ 1} + Z_t; \quad Z_t \quad N\ (0; \quad);$$

where

$$\kappa_t = \begin{bmatrix} \kappa_t^{(1)} & \kappa_t^{(2)} & \kappa_t^{(3)} \end{bmatrix}^{>}; \qquad = [\ _1 \quad _2 \quad _3]^{>}; \qquad = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{12} & v_{22} & v_{23} \\ v_{13} & v_{23} & v_{33} \end{bmatrix}:$$

A key component of the state-space representation is the distribution of the measurement errors. We assume that $\log(\hat{m}_{x;t})$ N $(\log(m_{x;t}); \frac{1}{\hat{D}_{x;t}})$ yields $_{x;t}$ N $(0; \frac{1}{\hat{D}_{x;t}})$, where $\hat{D}_{x;t}$ is the observed number of deaths at age x in year t. See Appendix A for the justification of this assumption.

By writing the CBD-X models in the state-space form, the dynamics of the observations y_t and the dynamics of the period effect κ_t are combined into one system. Furthermore, since logged observed central death rates follow normal distribution, the measurement density for single population mortality models listed in Chapter 2 can be written as:

$$p(y_{1:}$$

$D_{x;t}$

and the transition density:

$$\underbrace{p(\quad_t j y_{1:t\ 1};\quad)}_{\text{Predictive Distribution}} = \int \underbrace{p(\quad_t j \quad_{t\ 1};\quad)}_{\text{Transition Density}} \underbrace{p(\quad_{t\ 1} j y_{1:t\ 1};\quad)}_{\text{Filtering Distribution}} d\quad_{t\ 1};$$

where the transition density $p(\quad_t j \quad_{t\ 1};\quad)$ is implied by the transition equation as $\quad_t = f(\quad_{t\ 1}; Z_t)$.

Second, in the update step, we apply Bayes' rule to update the filtering distribution by combining the predicted $\quad_t$ with the additional time-t observations y_t:

$$p(\quad_t j y_{1:t};\quad) = \frac{p(y_t;\quad_t j y_{1:t\ 1};\quad)}{p(y_t j y_{1:t\ 1};\quad)}$$
$$= p(y_{1:t};$$

If we sample J independent random variables, $_{0:t}$ $p(\ _{0:t} j y_{1:t};\ )$, then standard Monte Carlo methods approximate $p(\ _{0:t} j y_{1:t};\ )$ by generating an empirical distribution made up of J samples of $^{j}_{0:t}$:

$$\hat{p}(\ _{0:t} j y_{1:t};\ ) = \frac{1}{J} \sum_{j=1}^{J} \ _{^{j}_{0:t}}(\ _{0:t});$$

where $_{x}(x)$ is a Dirac mass function centred at x. However, if $p(\ _{0:t} j y_{1:t};\ )$ is a complex high-dimensional probability distribution, then we cannot sample $_{0:t}$ directly from it. The SMC algorithm addresses this problem by sampling $_{0:t}$ from the proposal distribution $q(\ _{0:t} j y_{1:t};\ )$. The proposal distribution can be any distribution from which it is easy to generate a sample. The approximation of the conditional probability becomes:

$$\hat{p}(\ _{0:t} j y_{1:t};\ ) / \sum_{j=1}^{J} W_t^j \ _{^{j}_{0:t}}(\ _{0:t});$$

where W_t^j is the normalized importance weight associated with particle $^{j}_{0:t}$ to correct for the fact that $_{0:t}$ are not sampled from the right distribution. We define the importance weight w_t^j as

$$w_t^j = \frac{p\ ^{j}_{0:t}\ y_{1:t};}{q\ ^{j}_{0:t}\ y_{1:t};};$$

and the normalized importance weight is given by:

$$W_t^j = \frac{w_t^j}{\sum_{k=1}^{J} w_t^k}: \tag{3.4}$$

To avoid having to recompute the entire expression for the importance weights at each iteration, and to increase computational efficiency, instead of sampling all the particles $^{j}_{0:t}$ from a joint proposal distribution at once, we sample particles from a sequence of conditional distributions. We rewrite the proposal distribution in recursive form:

$$q\ ^{j}_{0:t}\ y_{1:t};\qquad q\ ^{j}_{t}\ ^{j}_{0:t\ 1}; y_{1:t};\qquad q\ ^{j}_{0:t\ 1}\ y_{1:t\ 1}$$

$$= \frac{p(y_t \mid \theta^{(j)}_{0:t}, y_{1:t-1}, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{0:t-1}, y_{1:t-1}, \lambda) \, p(\theta^{(j)}_{0:t-1} \mid y_{1:t-1}, \lambda)}{p(y_t \mid y_{1:t-1}, \lambda)}.$$

In the SSM, latent variables satisfy the Markov property and, given θ_t, the observations y_t are independent of $y_{1:t-1}$, leading to

$$p(\theta^{(j)}_{0:t} \mid y_{1:t}, \lambda) = \frac{p(y_t \mid \theta^{(j)}_t, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, \lambda) \, p(\theta^{(j)}_{0:t-1} \mid y_{1:t-1}, \lambda)}{p(y_t \mid y_{1:t-1}, \lambda)}.$$

Then, the importance weight becomes:

$$w^{(j)}_t = \frac{p(y_t \mid \theta^{(j)}_t, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, \lambda) \, p(\theta^{(j)}_{0:t-1} \mid y_{1:t-1}, \lambda)}{p(y_t \mid y_{1:t-1}, \lambda) \, q(\theta^{(j)}_t \mid \theta^{(j)}_{0:t-1}, y_{1:t}, \lambda) \, q(\theta^{(j)}_{0:t-1} \mid y_{1:t-1}, \lambda)}$$
$$\propto \frac{p(y_t \mid \theta^{(j)}_t, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, \lambda)}{q(\theta^{(j)}_t \mid \theta^{(j)}_{0:t-1}, y_{1:t}, \lambda)} w^{(j)}_{t-1}.$$

We thus define

$$w^{(j)}_t \propto \frac{p(y_t \mid \theta^{(j)}_t, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, \lambda)}{q(\theta^{(j)}_t \mid \theta^{(j)}_{0:t-1}, y_{1:t}, \lambda)},$$

which is known as the incremental importance weight. The denominator of the incremental importance weight is typically reduced to $q(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, y_t, \lambda)$ for computational convenience, so

$$w^{(j)}_t = \frac{p(y_t \mid \theta^{(j)}_t, \lambda) \, p(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, \lambda)}{q(\theta^{(j)}_t \mid \theta^{(j)}_{t-1}, y_t, \lambda)}.$$

As a result, at each iteration, only the incremental importance weight needs to be computed, and a new sequence of particles is obtained by keeping the trajectories of the particles sampled up to time $t-1$.

The SMC produces its approximation by an iterative process. The algorithm presented below is based on Andrieu et al. (2010). At first, we need to generate initial values for our particles θ_0. Then, for each time t, we start by sampling, J random samples of θ_t, denoted by $\theta^{(j)}_t$, from a proposal distribution $q(\theta_t \mid y_t, \theta_{t-1}, \lambda)$ to approximate $p(\theta_t \mid y_t, \theta_{t-1}, \lambda)$: Then, the particles are weighted and the corresponding normalized importance weights are calculated; the latter is needed as we sample particles from the proposal distribution and not the target distribution.

The importance weight of one particle might grow exponentially over time and, as the number of iterations increases, all the probability mass will eventually be allocated to that particle; that is, one particle could end up with normalized importance weight close

to one, while the other particles' normalized importance weights would be close to zero. This is known as weight degeneracy. Adding a resampling step can prevent the weight degeneracy problem. Thus, the last step of the SMC algorithm is resampling. Resampling means a new sequence of particles is replicated from the existing particles based on their normalized importance weights. The simplest resampling method (and the one used in this study) is called multinomial resampling. Specifically, in this study, we draw J random variables with replacement from a multinomial distribution with probabilities W_t, where $W_t = [W_t^1 \ ::: \ W_t^J]$, and W_t^J is obtained by Equation (3.4). As a consequence of this resampling, the particles with small weights will be eliminated while those with large weights will be duplicated. The resampled particle's weights are set equal to $w_t^j = \frac{1}{N}$ for $t \ 2 \ f1; : : : ; Tg$, which forces the weights not to permanently degenerate. If resampling is done at the end of every step, then the importance weight becomes

$$w_t^j \ / \ \frac{p \ y_t \ {}^j_t; \quad p \ {}^j_t \ {}^j_{t \ 1};}{q \ {}^j_t \ {}^j_{0:t \ 1}; y_{1:t};}$$

$$/ \ w_t^j:$$

Andrieu et al. (2010) introduce an ancestor variable, A_t^j, to keep track of the particles. Instead of resampling W_{qq}

Algorithm 1 Sequential Monte Carlo

1: sample $\theta_0^j \sim q(\theta_0 | \lambda)$ and set $A_0^j = j$
2: for $t = 1, \ldots, T$ do
3: sample $\theta_t^j \sim q(\theta_t | y_t, \theta_{t-1}^{A_{t-1}^j}, \lambda)$ and set $\theta_{0:t}^j = (\theta_{0:t-1}^{A_{t-1}^j}, \theta_t^j)$
4: compute the weight:

$$w_t^j = \frac{p(\theta_{0:t}^j, y_{1:t} | \lambda)}{p(\theta_{0:t-1}^j, y_{1:t-1} | \lambda) q(\theta_t^j | y_t, \theta_{t-1}^{A_{t-1}^j}, \lambda)}$$

$$= \frac{p(y_t | \theta_t^j, \lambda) p(\theta_t^j | \theta_{t-1}^{A_{t-1}^j}, \lambda)}{q(\theta_t^j | y_t, \theta_{t-1}^{A_{t-1}^j}, \lambda)}$$

5: normalize the weight: $W_t^j = \frac{w_t^j}{\sum_{k=1}^J w_t^k}$
6: sample A_t^j from a multinomial distribution with support 1 to J and weights W_t
7: end for

where

$$p(y_t | y_{1:t-1}, \lambda) = \int p(\theta_{t-1} | y_{1:t-1}, \lambda) p(\theta_t | \theta_{t-1}, \lambda) p(y_t | \theta_t, \lambda) d\theta_{1:t}$$
$$= \int w_t p(\theta_{t-1} | y_{1:t-1}, \lambda) q(\theta_t | y_t, \theta_{t-1}, \lambda) d$$

Chapter 4

Bayesian Inference

The values of the unknown parameters within the transition and measurement equations need to be estimated. The estimation can be done by frequentist or Bayesian methods. Because Bayesian and frequentist inference differ in their basic philosophies, the core features of both paradigms are reviewed in the next section.

4.1 Comparison of Frequentist and Bayesian Paradigms

In frequentist estimation, any unknown model parameter is generally assumed constant. The rationale is that even if a parameter cannot be observed, there exists one true value, and randomness stems from natural deviations of anything unknown when experiments are repeated. The fundamental measure of such uncertainty is captured by probability, which is the limit of the relative frequency of an event in a very long, theoretically infinite, sequence of the same experiment conducted independently of each other. The results of a frequentist approach can be represented by a confidence interval or a hypothesis test. Confidence intervals use data from a sample to estimate a population parameter. For example, a $100p\%$ confidence interval includes the true but unknown value with confidence $p \in (0;1)$. However, it is wrong to state that the unknown parameter lies with this interval with probability p. Similarly, given a statistic to test a null hypothesis H_0 related to the problem, the corresponding frequentist p-value is not the probability that H_0 is true, but rather the probability of observing a result at least as extreme for the outcome under the null distribution in a sequence of similar inferences.

Bayesian inference is different from frequentist methods in multiple ways. First, the probability actually expresses the chance of an event happening in this case. Second, unknown parameters are treated as random variables that can be described with probability distributions. Under the Bayesian paradigm, a probability expresses a degree of belief in an event. Essentially, this methodology starts with a set of prior beliefs based on scientific knowledge of the underlying problem. Then, one needs to update the prior belief in light of the observed data to come up with posterior beliefs. In the end, one analyzes the model fit

and sensitivity with respect to model assumptions. Because the model is set up via a full probabilistic approach, any probabilistic statements can be immediately interpreted as such without relating it to a sequence of independent repetitions. A 100p% probability interval then expresses a range for the quantity of interest with coverage probability p, and a p-value is interpreted as the probability of replicated data being more extreme than observed data evaluated under a specified test statistic.

We perform Bayesian inference conditional on observations to estimate the unknown parameters as we want to assess the parameter uncertainty. Unlike the frequentist methods that yield point estimates of unknown parameters, Bayesian methods yield a posterior distribution of the unknown parameters, which allows us to understand their uncertainty.

4.2 Bayesian Inference

The fundamental usage of Bayesian inference is based on Bayes' theorem; that is, given the probability distribution for the parameters of interest and the data $y_{1:T}$, the posterior distribution for , on which all inference is based, depends on the observed values. In particular, the likelihood distribution represents the data generating mechanism.

In the case of state-space models, the likelihood has two forms: the marginal likelihood and the complete data likelihood. The marginal likelihood is shown in Equation (3.5). In constructing the marginal likelihood, we consider all possible values of latent variables that can have been generated by observed data. Typically, the marginal likelihood is hard to evaluate in closed form as it involves multidimensional integrals.

The complete data likelihood is constructed assuming that the values of the latent variables are known. Indeed, it is not true, but the value of each latent variables can be imputed as part of the estimation procedure when using Bayesian methods. In our case, the complete data likelihood can be written as

$$\begin{aligned} L(y_{1:T}; \ _{0:T} j \ ) &= p(\ _0 j \ ) \prod_{t=1}^{T} p(\ _t j \ _{t\ 1}; \ ) \prod_{t=1}^{T} p(y_t j \ _t; \ ); \\ &= p(\ _0 j \ ) \underbrace{p(\ _{1:T} j \ )}_{\text{Transition}} \underbrace{p(y_{1:T} j \ _{0:T}; \ )}_{\text{Measurement}}; \end{aligned} \tag{4.1}$$

where $p(\ _t j \ _{t\ 1}; \ )$ is the transition density obtained from Equation (3.1), the measurement density $p(y_t j \ _t; \ )$ is implied by the measurement equation of Equation (3.2), and $p(\ _0 j \ )$ is the initial distribution for the latent factors. The prior belief on are converted into the probability distribution $p(\ )$. Then, Bayes' theorem states that

$$p(\ ; \ _{0:T} j y_{1:T}) = \frac{p(\ ; \ _{0:T}; y_{1:T})}{p(y_{1:T})} = \frac{L(y_{1:T}; \ _{0:T} j \ ) p(\ )}{\int L(y_{1:T}; \ _{0:T} j \ ) p(\ ) \, d \ };$$

where the denominator

$$\int L(y_{1:T}; \ _{0:T} \mid \ ) \, p(\ ) \, d$$

is a constant, leading to the following expression:

$$\underbrace{p(\ ; \ _{0:T} \mid y_{1:T})}_{\text{Posterior}} \propto \underbrace{L \ (y_{1:T}; \ _{0:T} \mid \ )}_{\text{Likelihood}} \ \underbrace{p(\ )}_{\text{Prior}}; \tag{4.2}$$

which summarizes the key elements of Bayesian inference. Each component of Equation (4.2) is discussed in the rest of this section.

4.2.1 The Likelihood

A likelihood function takes the data set as given and gives all of the relevant information to the evaluation of statistical evidence. The likelihood function can be obtained by using Equation (4.1). The transition density can be obtained by using Equation (2.5), and the general form of the measurement part of Equation (4.1) is given in Equation (3.3). Finally, for the CBD-X(3) model, the measurement part of Equation (4.1) is given by:

$$p(y_{1:T} \mid \ _{0:T}; \ ) \propto \exp\left(\sum_{x;t} \frac{D_{x;t}}{2} \log(m_{x;t}) \quad _x \quad {}^{(1)}_t \quad {}^{(2)}_t (x \quad \bar{x}) \quad {}^{(3)}_t ((x \quad \bar{x})^2 \quad {}^2_x) \right)^{\frac{1}{2}} :$$

4.2.2 The Prior

The prior distribution plays a vital role in determining the posterior distribution. In practice, prior distributions are speci ed using available information, such as experts' opinions or the results of previous studies. In this latter case, the prior distribution is called an informative prior distribution. Similarly, if the prior does not contain any information based on prior beliefs, it is called a non-informative prior distribution.

To obtain the posterior distribution within the CBD-X models, uniform prior distributions are assumed throughout this report, except for the variance (i.e., v_{ii}) and covariance (i.e, v_{ij} for $i \neq j$) parameters of the covariance matrix in Equation (2.5).

The variance parameters v_{ii} are assumed to follow the half-normal distribution with a mean parameter equal to 0 and a variance equal to 10 such that the variance parameters v_{ii} have positive values. The variance value is chosen because a large value makes sure that the distribution is proper and yet non-informative. The prior probability density function (pdf) of the variance parameter v_{ii} is

$$f(v_{ii}) \propto e^{-\frac{v_{ii}^2}{20}}; \quad v_{ii} \in (0; 1):$$

The prior distribution of the covariance parameters v_{ij} follows a truncated-normal distribution with the same mean and variance as the variance parameters, but are assumed to be within the interval ($\sqrt{v_{ii}\,v_{jj}}$; $\sqrt{v_{ii}\,v_{jj}}$) ensuring the correlation coe□cients between the series are within (1; 1). The prior pdf of the covariance parameter v_{ij} is

$$f\ v_{ij}\ \ f v_{ss} g_{s=1}^{3}\ \ / \ \frac{\frac{v_{ij}}{\sqrt{10}}}{2\ \ \sqrt{v_{ii}\,v_{jj}}\ \ \sqrt{10}};\quad v_{ij}\ 2\ (\ \sqrt{v_{ii}\,v_{jj}};\quad \sqrt{v_{ii}\,v_{jj}});$$

where (x) is the standard normal distribution evaluated at x, and (x) is its cumulative distribution function evaluated at x as well.

4.2.3 The Posterior

The posterior distribution is the probabi4(the)-333(coeF36 7.nng)-348(the)(the)-3o0

main sampling algorithms used in this study: the Metropolis-Hastings algorithm and the Gibbs sampler.

4.3.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is one of the most useful methods to construct MCMC samples. The Markov chain constructed from this method asymptotically reaches a unique stationary distribution $(\ ; \ _{0:T} j y_{1:T})$, such that $(\ ; \ _{0:T} j y_{1:T})$ approaches to the target distribution $p(\ ; \ _{0:T} j y_{1:T})$.

This algorithm has two main ingredients: a proposal distribution and an acceptance probability. The algorithm starts by setting initial parameters $^{(0)}$. Then, at iteration i and depending on the previous parameter value $^{(i \ 1)}$, the algorithm generates a candidate for the new parameter value $^{(New)}$ from a proposal distribution $q \ ^{(New)} \ ^{(i \ 1)}$. Same as the proposal distribution in the SMC algorithm, it can be any distribution. We define the transition density $p \ ^{(New)} \ ^{(i \ 1)}$ as the conditional probability of moving to $^{(New)}$ from $^{(i \ 1)}$.

A su□cient but not necessary condition for the existence of stationary distribution is that the Markov chain be reversible. A Markov chain is reversible if each transition is reversible; that is, for every pair of parameters $^{(New)}$ and $^{(i \ 1)}$, the probability of being in state $^{(i \ 1)}$ and transitioning to state $^{(New)}$ must be equal to the probability of being in state $^{(New)}$ and transitioning to state $^{(i \ 1)}$:

$$p \ ^{(New)} \ ^{(i \ 1)} \ p \ ^{(i \ 1)} = p \ ^{(i \ 1)} \ ^{(New)} \ p \ ^{(New)} \ ^{(i \ 1)}$$

4.4.1 Conditional Sequential Monte Carlo

The PG sampler, introduced by Andrieu et al. (2010), is an extension of the Gibbs sampler. Instead of generating random variables sequentially at each iteration, the PG sampler generates all random variables at the same time. The main idea of the PG sampler is to run a conditional sequential Monte Carlo (cSMC) algorithm iteratively. The cSMC algorithm is similar to the standard SMC algorithm introduced in Chapter 3, except that a pre-specified path $_{0:T}$ is retained in all the resampling steps, whereas the remaining $J \quad 1$ particles are generated as in the standard SMC algorithm. For simplicity, we set the last particle $^{J}_{t} = \ _{t}$, and its ancestor variable $A^{J}_{t} = J$ for all t. The cSMC algorithm is summarized in Algorithm 4.

Algorithm 4 Conditional Sequential Monte Carlo

1: sample $^{j}_{0} \quad q(\ _{0}j\ )$ for $j = 1;:::;J \quad 1$ and set $^{J}_{0} = \ _{0}$
2: set $A^{j}_{0} = j$ for $j = 1;:::;J$
3: for $t = 1;:::;T$ do
4: sample $^{j}_{t} \quad q\ \ _{t}\ y_{t};\ ^{A^{j}_{t}\ 1}_{t\ 1};$
5: set $^{j}_{0:t} = \ ^{A^{j}_{t}\ 1}_{0:t\ 1};\ ^{j}_{t}$ for $j = 1;:::;J \quad 1$, and set $^{J}_{t} = \ _{t}$
6: compute the weight for $j = 1;:::;J$:

$$w^{j}_{t} = \frac{p\ \ ^{j}_{0:t};y_{1:t};}{p\ \ ^{j}_{0:t\ 1}y^{\ \ 1:t\ 1};\ \ ;\ \ 1=\ \ t;}$$

of $_{0:T}$, we sample an index

Algorithm 5 Particle Gibbs

1: set initial values $\theta^{(0)}$ and $x_{0:T}^{(0)}$
2: for $i = 1, \ldots, M$ do
3: for $g = 1, \ldots, d$ do
4: $\theta_g^{(i)} = [x_1^{(i)}, \ldots, x_{g-1}^{(i)}, \theta_{g+1}^{(i-1)}, \ldots, \theta_d^{(i-1)}]$
5: sample

$$/ \exp \left(\sum_{t=1}^{T} \frac{\hat{D}_{x;t}}{2} \log(\hat{m}_{x;t}) \right) \quad {}_{x}^{(1)}{}_{t} \quad {}_{t}^{(2)}(x \quad x) \quad {}_{t}^{(3)}(x \quad \bar{x})^2 \quad {}_{x}^{2} \quad {}^{!}{}_{2}$$

$$/ \exp \left(\sum_{t=1}^{T} \frac{\hat{D}_{x;t}}{2} \right) \quad {}_{x}^{2} \quad 2 \; {}_{x} \quad \log(\hat{m}_{x;t}) \quad {}_{t}^{(1)} \quad {}_{t}^{(2)}(x \quad x) \quad {}_{t}^{(3)}(x \quad \bar{x})^2 \quad {}_{x}^{2} \quad !!$$

$$/ \exp \left(\sum_{t=1}^{T} \frac{\hat{D}_{x;t}}{2} \right) {}_{x}^{2} + 2 \; {}_{x} \sum_{t=1}^{T} \frac{\hat{D}_{x;t}}{2} \log(\hat{m}_{x;t}) \quad {}_{t}^{(1)} \quad {}_{t}^{(2)}(x \quad x) \quad {}_{t}^{(3)}((x \quad \bar{x})^2 \quad {}_{x}^{2})$$

$$+ (a_{12}b_t^{(1)} + a_{22}b_t^{(2)} + a_{23}b_t^{(3)})b_t^{(2)}$$
$$+ (a_{13}b_t^{(1)} + a_{23}b_t^{(2)} + a_{33}b_t^{(3)})b_t^{(3)};$$

and the full conditional posterior distribution of $_1$ with uniform prior can be written as

$$\begin{aligned}
& f(\ _1 j\ \ _1;\ _{0:T}) \\
& / \prod_{t=1}^{T} \frac{1}{(2\ )^2 j\ j} \exp\ \ \frac{1}{2}(\ _t\ \ \ _t)^{>}\ \ \ ^1(\ _t\ \ \ _t) \\
& / \exp\ \ \frac{1}{2}\sum_{t=1}^{T}(a_{11}b_t^{(1)} + a_{12}b_t^{(2)} + a_{13}b_t^{(3)})b_t^{(1)} + (a_{12}b_t^{(1)} + a_{22}b_t^{(2)} + a_{23}b_t^{(3)})b_t^{(2)} \\
& \qquad + (a_{13}b_t^{(1)} + a_{23}b_t^{(2)} + a_{33}b_t^{(3)})b_t^{(3)} \Big) \\
& / \exp\ \ \frac{1}{2}\sum_{t=1}^{T} a_{11}\ \ b_t^{(1)}\ ^2 + 2a_{12}b_t^{(1)}b_t^{(2)} + 2a_{13}b_t^{(1)}b_t^{(3)} \Big) \\
& / \exp\ \ \frac{1}{2}\sum_{t=1}^{T} a_{11}\ \ _1^2\ \ 2\ _1\ \ _t^{(1)}\ \ \ _{t\ 1}^{(1)} + 2a_{12}\ \ \ _1\ \ _t^{(2)}\ \ \ _{t\ 1}^{(2)}\ \ \ _2 \\
& \qquad + 2a_{13}\ \ \ _1\ \ _t^{(3)}\ \ \ _{t\ 1}^{(3)}\ \ \ _3 \Big) \\
& / \exp\ \ \frac{1}{2}\sum_{t=1}^{T} a_{11}\ \ _1^2\ \ 2\ _1\ a_{11}(\ _t^{(1)}\ \ \ _{t\ 1}^{(1)}) + a_{12}(\ _t^{(2)}\ \ \ _{t\ 1}^{(2)}\ \ \ _2) \\
& \qquad + a_{13}(\ _t^{(3)}\ \ \ _{t\ 1}^{(3)}\ \ \ _3) \Big) \\
& / \exp\ \ \frac{1}{2}\ a_{11}T\ _1^2\ \ 2\ _1\sum_{t=1}^{T} a_{11}(\ _t^{(1)}\ \ \ _{t\ 1}^{(1)}) + a_{12}(\ _t^{(2)}\ \ \ _{t\ 1}^{(2)}\ \ \ _2) \\
& \qquad + a_{13}(\ _t^{(3)}\ \ \ _{t\ 1}^{(3)}\ \ \ _3) \Big) \\
& / \exp\ \ \frac{1}{2\frac{1}{a_{11}T}}\ \ _1^2 \\
& \qquad 2\ _1\sum_{t=1}^{T} \frac{a_{11}(\ _t^{(1)}\ \ \ _{t\ 1}^{(1)}) + a_{12}(\ _t^{(2)}\ \ \ _{t\ 1}^{(2)}\ \ \ _2) + a_{13}(\ _t^{(3)}\ \ \ _{t\ 1}^{(3)}\ \ \ _3)}{a_{11}T} \Big);
\end{aligned}$$

so

$$\begin{aligned}
& _1 j\ \ _1;\ _{0:T} \\
& N\ \ @\sum_{t=1}^{T} \frac{a_{11}(\ _t^{(1)}\ \ \ _{t\ 1}^{(1)}) + a_{12}(\ _t^{(2)}\ \ \ _{t\ 1}^{(2)}\ \ \ _2) + a_{13}(\ _t^{(3)}\ \ \ _{t\ 1}^{(3)}\ \ \ _3)}{a_{11}T};\ \frac{1}{a_{11}T}A :
\end{aligned}$$

Similarly,

$$\alpha_2 \mid \alpha_{-2}, \sigma^2, y_{0:T} \sim N\left(\sum_{t=1}^{T} \frac{a_{22}(\alpha_t^{(2)} - \alpha_{t-1}^{(2)}) + a_{12}(\alpha_t^{(1)} - \alpha_{t-1}^{(1)} - \alpha_1) + a_{23}(\alpha_t^{(3)} - \alpha_{t-1}^{(3)} - \alpha_3)}{a_{22}T}, \frac{1}{a_{22}T}\right);$$

and

$$\alpha_3 \mid \alpha_{-3}, \sigma^2, y_{0:T} \sim N\left(\sum_{t=1}^{T} \frac{a_{33}(\alpha_t^{(3)} - \alpha_{t-1}^{(3)}) + a_{13}(\alpha_t^{(1)} - \alpha_{t-1}^{(1)} - \alpha_1) + a_{23}(\alpha_t^{(2)} - \alpha_{t-1}^{(2)} - \alpha_2)}{a_{33}T}, \frac{1}{a_{33}T}\right);$$

For the variance (i.e., $v_{11}; v_{22};$ and v_{33}) and covariance parameters (i.e, v_{12}; $v_{13};$ and v_{23}) of the covariance matrix , there is no closed-form solution for the full conditional posterior distribution. We thus apply the MH algorithm to update them with the prior distributions discussed in the previous section. The proposal distributions for the variance parameters are a truncated normal distribution centred at the values from the previous iteration and with lower bound at zero to make sure the value of variance is positive. The proposal distribution for the covariance parameters is a normal distribution centred at the last observation in the chain.

5.1, the early part of the chain for $\theta_1^{(1)}$ behaves differently from the remaining part and is discarded.

Figure 5.1: Trace plot of $\theta_1^{(1)}$ with 3,000 particles and 120,000 iterations for the CBD-X(3) model.

The remaining samples in the chain produced by the pMCMC method yield the posterior distribution of each model parameter. For example, the posterior distribution of θ_1 shown in Figure 5.2.

When constructing a pMCMC algorithm, the speed of convergence should be within the practical constraints of time and computational power. The speed of convergence is mainly influenced by the number of iterations and the number of particles. The number of iterations within a pMCMC algorithm represents the total sample size simulated for each parameter, and the number of particles determines how many particles are used in the cSMC algorithm at each iteration. The number of iterations required for convergence varies from application to application. However, several tests can assess convergence. We choose the number of iterations and the number of particles by first looking at the trace plots. We then apply the Gelman-Rubin test to perform further convergence diagnostics. These tests will be discussed in Section 5.2.1.

The algorithm set-up for each model is summarized in Table 5.1. There are more particles in the CBD-X(3) model than for the other two models because it has three period-effect factors. Thus, it needs more particles in the cSMC algorithm to generate consistent approximations. Usually, we will set more iterations for models with more parameters as convergence tends to be slower in these cases. However, we use fewer iterations for the CBD-X(3)

model than the CBD-X(2) model to save on the run-time as we have more particles in the CBD-X(3) model.

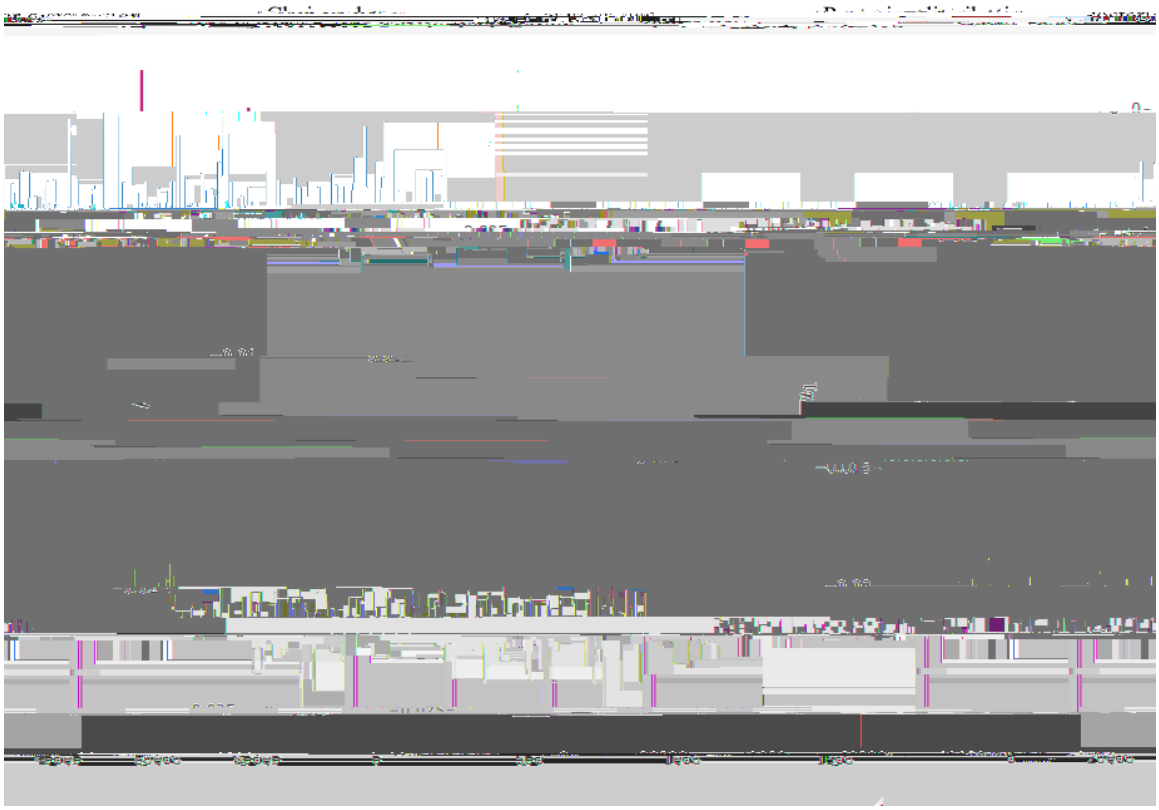

Figure 5.2: Trace plot of $_1$ with 3,000 particles after burn-in period within the CBD-X(3) model (left-hand side) and the corresponding posterior distribution (right-hand side).

Model	Numbers of Iterations (M)	Burn-In Period	Numbers of Particles (J)
CBD-X(1)	30;000	6;000	500
CBD-X(2)	300;000	60;000	500
CBD-X(3)	120;000	24;000	3;000

Table 5.1: Algorithm set-up for each model.

5.2.1 Convergence Diagnostics

As noted above, the outputs of the CBD-X models under the pMCMC algorithm must be diagnosed for convergence before performing any type of statistical inference. To assess convergence, we look at the trace plot first. Trace plots give insight into the behaviour of the Markov chain and point out possible flaws in the algorithm. If the pMCMC chain is stuck in some part of the state space, the trace plots show flat bits indicating slow convergence.

shown in Figure 5.3. Therefore, using 1,000 particles is not enough for the CBD-X(3) model to converge, and we increased the number of particles to 3,000 for this model. If too many proposed values are accepted consecutively, trace plots may move slowly and not explore the rest of the state space. If a trace plot exhibits rapid up-and-down variation with no long-term trends or drifts, then the parameter appears to have converged, as in Figure 5.4.

which measures how the means in each chain vary around the overall mean. The within-chain variance is given by:

$$W = \frac{1}{H} \sum_{h=1}^{H} \hat{\sigma}_h^2,$$

which is the averaged variances of the chains. Under certain stationary conditions, the combined variance

$$\hat{V} = \frac{M-1}{M} W + \frac{B}{MN}$$

5.2.2 Estimated Parameters

A few 95% posterior credible intervals for each parameter are shown in Figues 5.5 to 5.8. Credible intervals are an important concept in Bayesian statistics. Their core purpose is to describe and summarize the uncertainty related to the unknown estimated parameters. A credible interval is an interval that contains a value with a certain probability. For instance, a 95% posterior credible intervals for $x_{x95\%}$

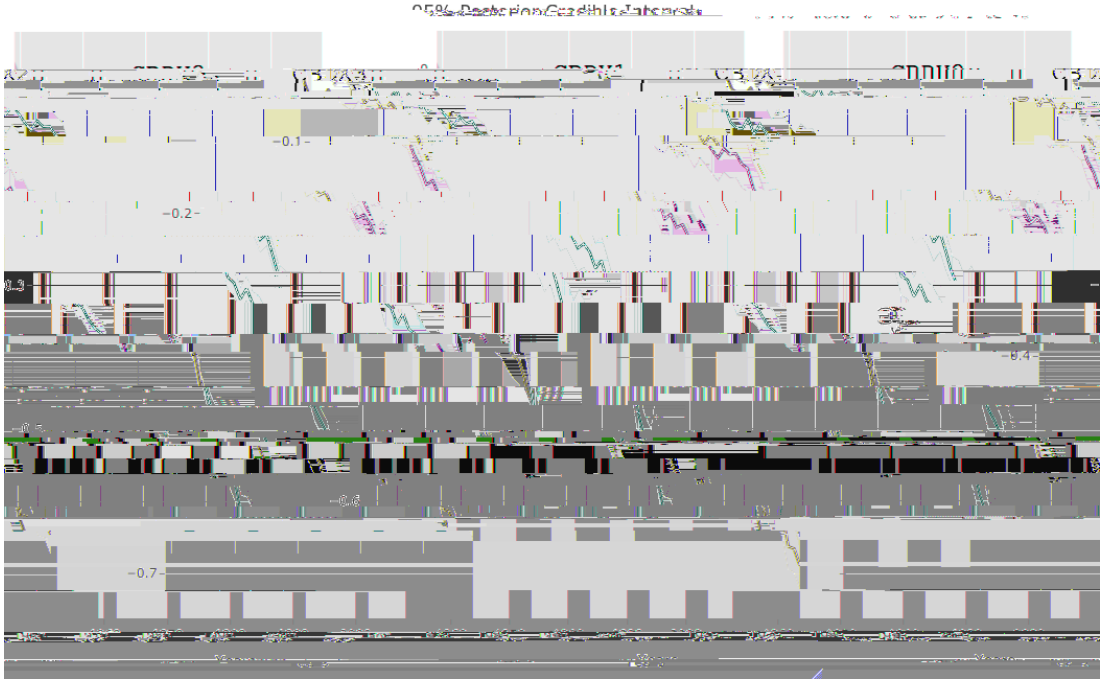

Figure 5.6: 95% posterior credible intervals for $\theta_t^{(1)}$ from 1959 to 2008 The solid line is the mean value of $\theta_t^{(1)}$.

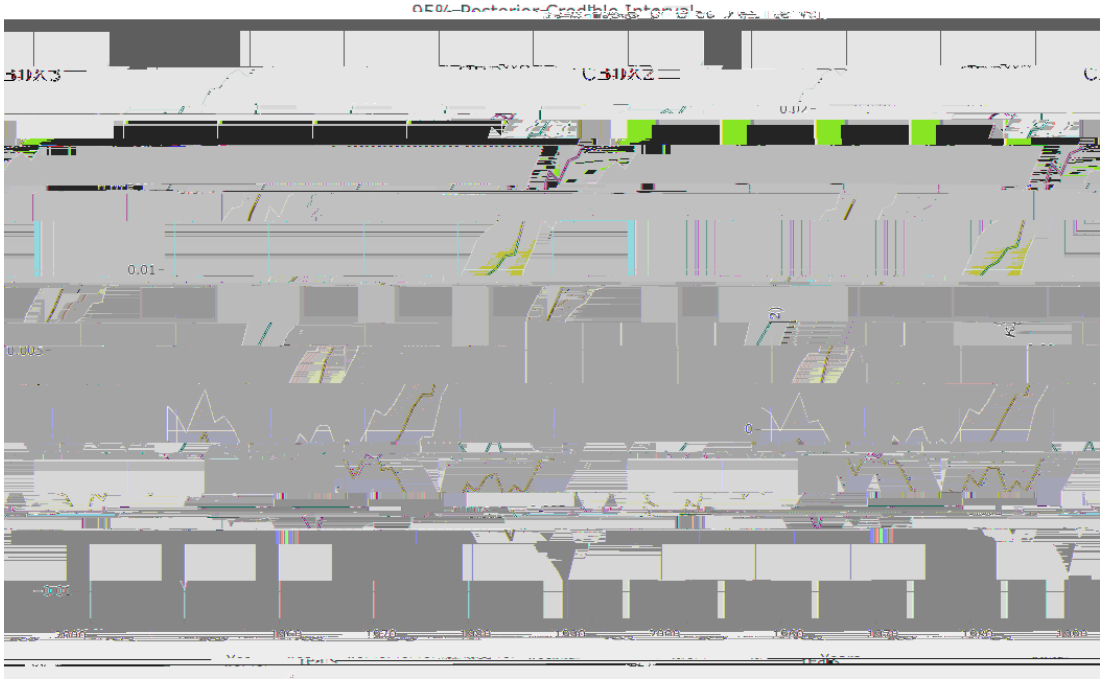

Figure 5.7: 95% posterior credible intervals for $\theta_t^{(2)}$ from 1959 to 2008 The solid line is the mean value of $\theta_t^{(2)}$.

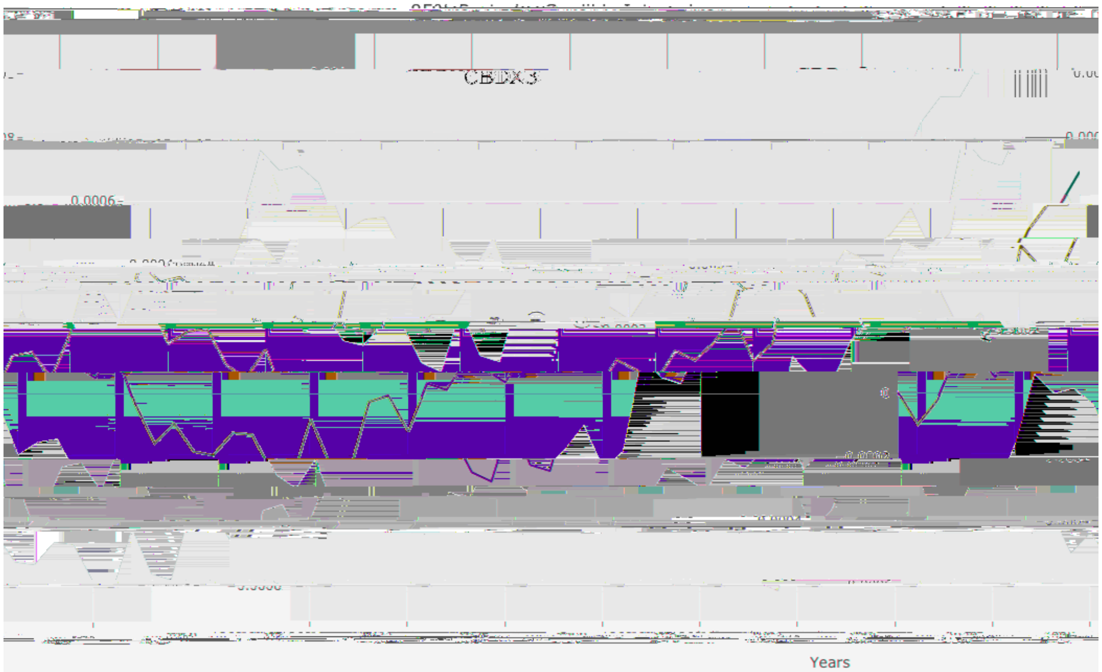

Figure 5.8: 95% posterior credible intervals for $_t^{(3)}$ from 1959 to 2008. The solid line is the mean value of $_t^{(3)}$.

5.2.3 Forecasting Death Rates

Once a su ciently large sample of the posterior distribution of the latent variables $_t$ and parameters are simulated, they are used to forecast death rates. We incorporate di erent sources of uncertainty in forecast intervals by using the following steps. We draw simulated values of the parameters and the latent variables at the final year (i.e., $_T$) at random from the pMCMC outputs after discarding the burn-in period. Then, we predict the future period-e ect factor $_{T+h}$, where h are years after T, by using the random walk structure we imposed for the period e ect in Equation (2.5). We compute the forecasted death rates m_{T+h} by using the predicted latent variables $_{T+h}$ and the parameters sampled from the pMCMC output. We repeat this process 10,000 times to get a distribution of the forecasted death rates.

The forecasting performance of the CBD-ufpi0 p isasscesled trought the 1-(y)2821185-9112(ut-he)]TJ to16.206

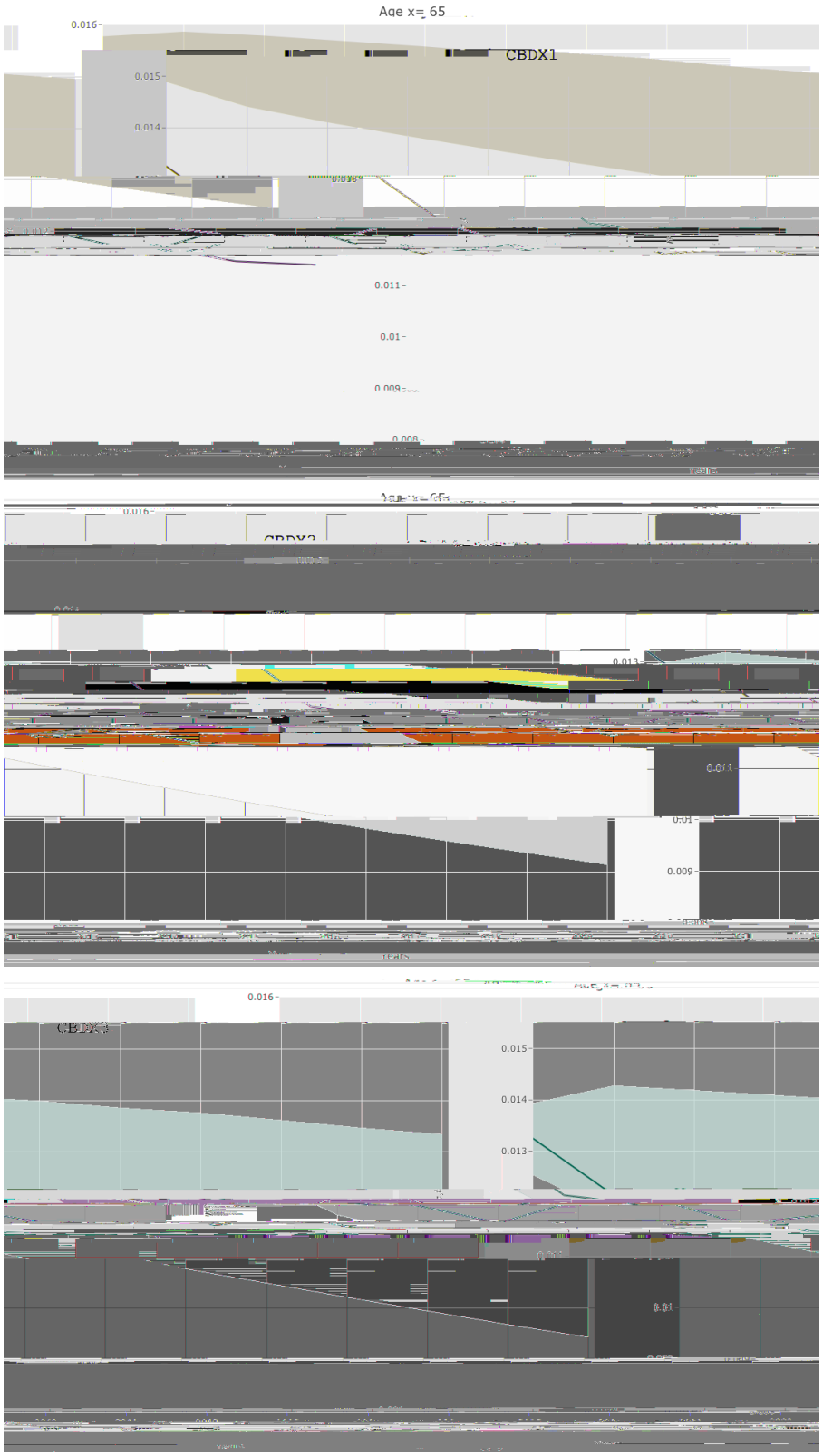

Figure 5.9: 10-year out-of-sample forecasted death rates for the Canadian male population under CBD-X models. The solid line shows the observed $\hat{m}_{x;t}$ for age 65 from 2009 to 2018. Shaded areas are 95% credible intervals.

Figure 5.10: 10-year out-of-sample forecasted death rates for the Canadian male population under CBD-X models. The solid line shows the observed $\hat{m}_{x;t}$ for age 75 from 2009 to 2018. Shaded areas are 95% credible intervals.

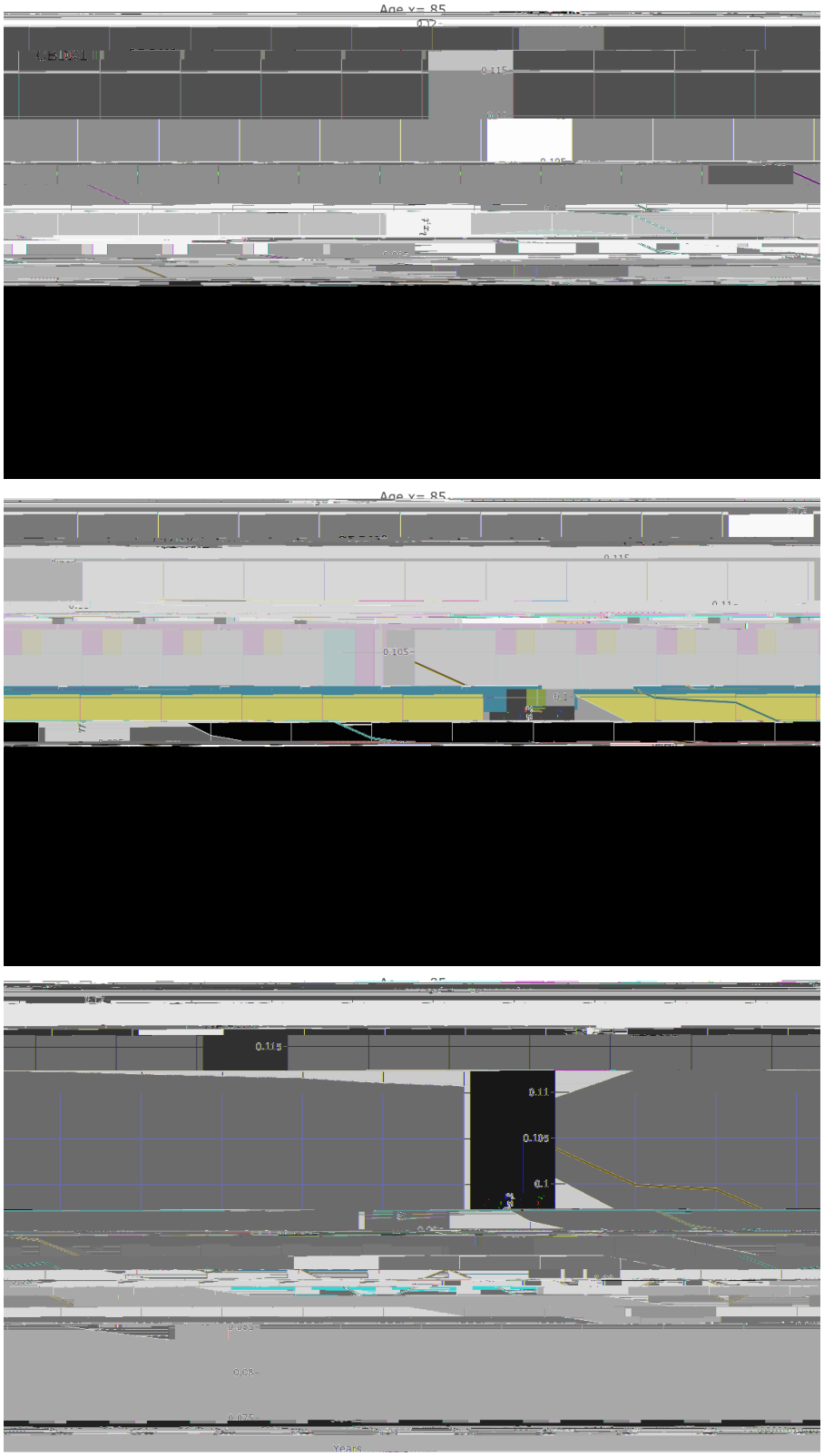

Figure 5.11: 10-year out-of-sample forecasted death rates for the Canadian male population under CBD-X models. The solid line shows the observed $\hat{m}_{x;t}$ for age 85 from 2009 to 2018. Shaded areas are 95% credible intervals.

5.2.4 Model Selection

The problem of model selection requires us to consider two competing notions: one is a measure of model fit that promotes selecting more accurate models, and the other is a measure of model complexity. The deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) is a useful method in Bayesian model selection. Under the DIC, the goodness of fit is measured by the posterior mean deviance, denoted by $\overline{\text{Deviance}(\quad)}$. The deviance is defined as

$$\text{Deviance}(\quad) = \quad 2\log(L(y_{1:T};\quad_{0:T} \mid \quad)) + C;$$

where C is a constant that can be canceled out when comparing different models as it is

Model	DIC	p_D	Posterior Mean Deviance
CBD-X(1)	12;547:210	87:718	12;459:490
CBD-X(2)	1;563:776	184:093	1;392:704
CBD-X(3)	766:372	387:507	378:864

Table 5.2: Deviance information criterion for CBD-X models.

Chapter 6

Forecasting Performance In Different Estimation Methods

One of the advantages of using the Bayesian method is that we can incorporate all the uncertainty within our model estimation and forecasting. This chapter investigates the impact of including the uncertainty by comparing the forecasting performance of the CBD-X(3) model estimated under the maximum likelihood method to that estimated under the pMCMC method. We only compare the performance in the CBD-X(3) model because it has the best performance in both model selection and forecasting, as shown in the previous chapter.

The maximum likelihood method in this chapter is different from the two-stage maximum likelihood estimation method we introduced at the beginning of Chapter 3. Instead of estimating all the parameters (i.e., and $_{0:T}$) by using a frequentist method as in the two-stage maximum likelihood estimation method, the maximum likelihood method in this chapter is only used to estimate the unknown parameters by means of the PF. That is to say, the mortality model is still in a state-space representation as in Chapter 3 but instead of applying the Bayesian method to estimate the unknown parameters , we use a frequentist method and specifically, the MLE. First, we use the mean of the posterior distribution of from pMCMC as the starting value to maximize the likelihood function obtained in the SMC algorithm (i.e., Equation (3.5)) of the CBD-X(3) model by using the Nelder and Mead (1965) method. Then, the latent variables $_{0:T}$ are estimated from the SMC algorithm with the optimal value of obtained from MLE.

Similar to the forecasting algorithm in the pMCMC estimation introduced in Chapter 5, the latent variables $_{T+h}$ are predicted using the random walk model for the period effect; that is, Equation (2.5). Then, we compute the forecasted death rate m_{T+h} by using the predicted latent variables $_{T+h}$ and the constant parameters obtained by MLE. We repeat this process 10,000 times to get a distribution of the forecasted death rate.

The 95% confidence intervals under MLE and the 95% credible intervals under pMCMC of the 10

ages 65, 70, 75, 80, 85, and 89

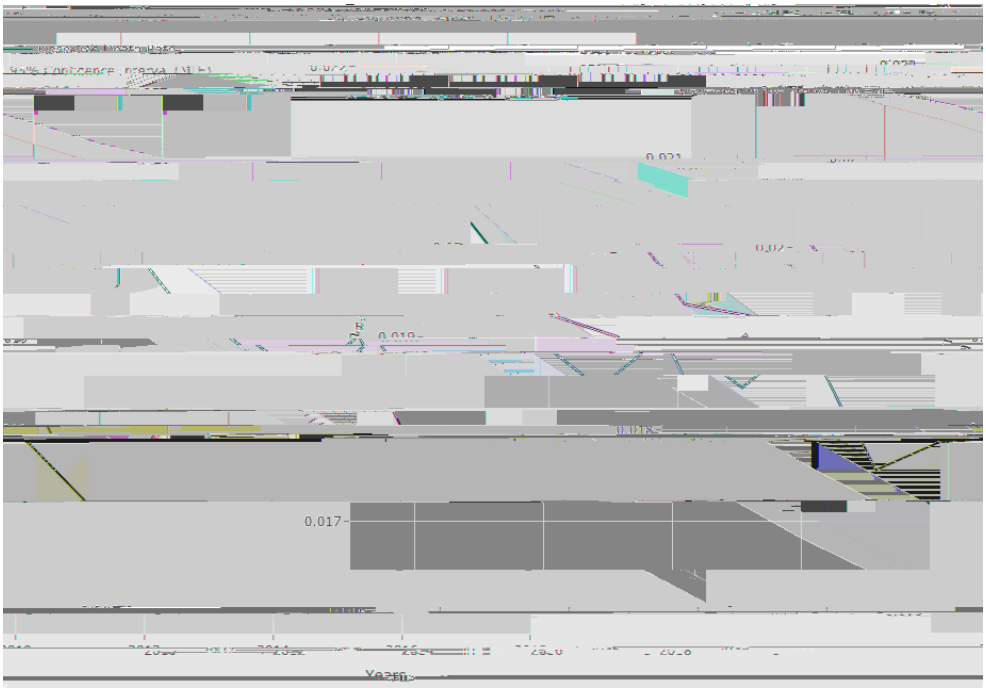

Figure 6.2: 10-year out-of-sample forecasted death rates for the Canadian male population under MLE method (shaded purple area) and pMCMC method (shaded green area). The solid line shows the observed $\hat{m}_{x;t}$ for age 70 from 2009 to 2018.

Figure 6.3: 10-year out-of-sample forecasted death rates for the Canadian male population under MLE method (shaded purple area) and pMCMC method (shaded green area). The solid line shows the observed $\hat{m}_{x;t}$ for age 75 from 2009 to 2018.

Figure 6.4: 10-year out-of-sample forecasted death rates for the Canadian male population under MLE method (shaded purple area) and pMCMC method (shaded green area). The solid line shows the observed $\hat{m}_{x;t}$ for age 80 from 2009 to 2018.

Figure 6.5: 10-year out-of-sample forecasted death rates for the Canadian male population under MLE method (shaded purple area) and pMCMC method (shaded green area). The solid line shows the observed $\hat{m}_{x;t}$ for age 85 from 2009 to 2018.

Figure 6.6: 10-year out-of-sample forecasted death rates for the Canadian male population under MLE method (shaded purple area) and pMCMC method (shaded green area). The solid line shows the observed $\hat{m}_{x;t}$ for age 89 from 2009 to 2018.

Chapter 7

Conclusion

This report presented a Bayesian method—specifically the pMCMC method—to estimate CBD-X models combining features of the LC and CBD models. By using this method, the dynamics of the period effect $_{0:T}$ can be incorporated within the model estimation. Also, parameter uncertainty can be obtained readily and used in calculating forecasting intervals.

First, the CBD-X model was recast into state-space formulation. By doing so, we were able to incorporate the dynamics of period effects into the model estimation. Although the dynamics of the period effects in our models are linear and Gaussian, we still applied the SMC method to allow for more flexibility in choosing the structure of period effects in the future.

Second, a Bayesian approach was used to estimate the unknown parameters and their uncertainty, as Bayesian approaches yield posterior distributions of model parameters as well as mortality rates. A sampling-based approach called MCMC was used because we cannot derive the closed-from solution of joint posterior distributions for unknown parameters. The models were estimated based on Canadian male mortality data.

After fitting the models, we were able to perform model comparison and predictions. The CBD-X(3) model is the best model as it leads to the smallest DIC values and the best forecasting performance. Its 95% credibility intervals of the 10-year out-of-sample forecasted death rate capture the observed death rates in most cases.

To assess how including the parameter uncertainty influenced the forecasting performance, we compared the 95% confidence intervals obtained with pMCMC to those obtained with MLE for 10-year out-of-sample forecasted death rates. We observed that the 95% confidence intervals are generally smaller than the 95% credibility intervals. Thus, in contrast to frequentist estimation methods, the Bayesian approach captures more uncertainty in forecasting—consistently with the fact that mortality models' parameter uncertainty is large.

The Bayesian-based estimation approach proposed in this report is flexible and easy to implement as it can be applied to other mortality models. For future extensions, one possible direction would be applying this estimation approach for multi-population mortality

models. Due to globalization, different populations' mortality are closely linked together. Therefore, developing multi-population mortality models to analyze and forecast the mortality of more than one population in a joint way, such as modelling population between different countries, is of paramount importance. Another possible extension of our work would include the addition of cohort effects within model estimation as it might improve the model performance. Changing the structure of period effects might also be another interesting avenue for future research.

Bibliography

Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72:269–342.

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 31:373–393.

Cairns, A. J. 2011. Modelling and management of longevity risk: Approximations to survivor functions and dynamic hedging. Insurance: Mathematics and Economics49:438–453.

Cairns, A. J. G., D. Blake, and K. Dowd. 2006. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. Journal of Risk and Insurance 73:687–718.

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, and M. Khalaf-Allah. 2011. Bayesian stochastic mortality modelling for two populations. ASTIN Bulletin 41:29–59.

Cairns, A. J. G., D. Blake, K. Dowd, G. D.Coughlan, D. Epstein, A. Ong, and I. Balevich. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal 13:1–35.

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) 179:975–1005.

Cairns, A. J. G., M. Kallestrup-Lamb, C. Rosenskjold, D. Blake, and K. Dowd. 2019. Modelling socio-economic di□erences in mortality using a new a□uence index. ASTIN Bulletin 49:555–590.

Chan, W., J. S. Li, and J. Li. 2014. The CBD mortality indexes: Modeling and applications. North American Actuarial Journal 18:38–58.

Currie, S. J. Richards, J. G. Kirkby, and I. D. 2006. The importance of year of birth in two-dimensional mortality data. British Actuarial Journal 12:5–38.

Czado, C., A. Delwarde, and M. Denuit. 2005. Bayesian Poisson log-bilinear mortality projections. Insurance: Mathematics and Economics36:260–284.

Dowd, K., Cairns, A. J. G., and D. Blake. 2020. CBDX: A workhorse mortality model from the Cairns-Blake-Dowd family. Annals of Actuarial Science 14:445–460.

Fung, M. C., G. W. Peters, and P. V. Shevchenko. 2017. A unified approach to mortality modelling using state-space framework: Characterisation, identification, estimation and forecasting. Annals of Actuarial Science 11:343–389.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis, Second Edition. Chapman and Hall/CRC.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7:457 – 472.

Georgios, M., Cairns, A. J. G., S. George, and K. Torsten. 2017. Stochastic mortality modelling: Key drivers and dependent residuals. North American Actuarial Journal 21:1–26.

Gompertz, B. 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London 115:513–583.

Haqqi Anna Zili, A., S. Mardiyati, and D. Lestari. 2018. Forecasting Indonesian mortality rates using the Lee-Carter model and ARIMA method. AIP Conference Proceedings 2023:020212.

Hunt, A., and D. Blake. 2020. Identifiability in age/period mortality models. Annals of Actuarial Science 14:461–499.

Hunt, A., and D. Blake. 2021. On the structure and classification of mortality models. North American Actuarial Journal 25:215–234.

Kamaruddin, H., and I. Noriszura. 2018. Forecasting selected specific age mortality rate of Malaysia by using Lee-Carter model. Journal of Physics: Conference Series 974:012003.

Koissi, M.-C., A. F. Shapiro, and G. Högnäs. 2006. Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. Insurance: Mathematics and Economics 38:1–20.

Lee, R. D., and L. R. Carter. 1992. Modeling and forecasting US mortality. Journal of the American Statistical Association 87:659–671.

Lundström, H., and J. Qvist. 2004. Mortality forecasting and trend shifts: An application of the Lee-Carter Model to Swedish mortality data. International Statistical Review 72:37–50.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. The Computer Journal 7:308–313.

Pedroza, C. 2006. A Bayesian forecasting model: Predicting U.S. male mortality. Biostatistics 7:530–550.

Renshaw, A. E., and S. Haberman. 2006. A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insurance: Mathematics and Economics38:556–570.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64:583–639.

Appendix A

Distribution of the Error Term

The standard actuarial approach in mortality modelling assumes that, conditional on the model central death rate $m_{x;t}$ and the observed exposures $\hat{E}_{x;t}$, the observed number of deaths $\hat{D}_{x;t}$ has a Poisson distribution with mean and variance both equal to $m_{x;t}\hat{E}_{x;t}$.[1] Cairns et al. (2016) assumes that for each t and x, the observed number of deaths $\hat{D}_{x;t}$ is conditionally independent and has a lognormal distribution; that is, $\log(\hat{D}_{x;t})$ has a normal distribution with mean $_d$ and variance ^2_d. The lognormal distribution is chosen for computational convenience. As in Cairns et al. (2016), we equate the mean and variance of the lognormal distribution to the mean and variance of a matching Poisson distribution. By matching the mean, we obtain:

$$m_{x;t}\hat{E}_{x;t} = \exp\left(_d + \frac{^2_d}{2}\right);$$

which is equivalent to

$$_d = \log(m_{x;t}\hat{E}_{x;t}) \quad \frac{^2_d}{2}:$$

By matching the variance, on the other hand, we have

$$m_{x;t}\hat{E}_{x;t} = \exp(^2_d) \quad 1 \exp 2 _d + ^2_d ;$$

so

$$\exp(^2_d) \quad 1 = \frac{m_{x;t}\hat{E}_{x;t}}{\exp 2\log(m_{x;t}\hat{E}_{x;t}) \quad ^2_d + ^2_d}$$

$$= \frac{1}{m_{x;t}\hat{E}_{x;t}} = \frac{1}{m_{x;t}E_{x;t}}$$

[1] In this report, the observed and theoretical exposures are assumed to be the same, i.e. $E_{x;t} = \hat{E}_{x;t}$

$$= \frac{1}{D_{x;t}}:$$

By using the first-order Taylor series expansion about zero, we have that:

$$\exp(x) \quad 1 \quad x;$$

thus, $\overset{2}{d} \quad \frac{1}{D_{x;t}}$. Since $D_{x;t}$ is unobserved, we use the observed number of deaths $\hat{D}_{x;t}$ to

Appendix B

Summary Tables

The Gelman-Rubin test for each CBD-X models are summarized in the following tables.

CBD-X(1) Model	Mean	Standard Deviation	Gelman-Rubin Test
1	3:529		

CBD-X(1) Model	Mean	Standard Deviation	Gelman-Rubin Test
21	1:712	0:012	1:

CBD-X(1) Model	Mean	Standard Deviation	Gelman-Rubin Test
$\underset{29}{(1)}$	0		

CBD-X(2) Model	Mean	Standard Deviation	Gelman-Rubin Test
1	3:438	0:025	1:000
2	3:372	0:024	1:000
3	3:276	0:023	1:000
4	3:198	0:022	1:000
5	3:111	0:021	1:000
6	3:021	0:021	1:000
7	2:956	0:020	1:000
8	2:852	0:019	1:000
9	2:772	0:019	1:000
10	2:683	0:018	1:000
11	2:600	0:018	1:000
12	2:518	0:018	1:000
13	2:434	0:018	1:000

CBD-X(2) Model	Mean	Standard Deviation	Gelman-Rubin Test
(1)			

CBD-X(2) Model	Mean	Standard Deviation	Gelman-Rubin Test
$\binom{(1)}{42}$	0:479	0:018	1:000
$\binom{(1)}{43}$	0:513	0:018	1:000
$\binom{(1)}{44}$	0:532	0:018	1:000
$\binom{(1)}{45}$	0:553	0:018	1:000
$\binom{(1)}{46}$	0:592	0:018	1:000
$\binom{(1)}{47}$	0:616	0:018	1:000
$\binom{(1)}{48}$	0:663	0:018	1:000
$\binom{(1)}{49}$	0:666	0:018	1:000
$\binom{(1)}{50}$	0:690	0:018	1:000
$\binom{(2)}{1}$	0:001	0:001	1:000
$\binom{(2)}{2}$	0:002	0:001	1:000
$\binom{(2)}{3}$	0:003	0:001	1:000
$\binom{(2)}{4}$	0:002	0:001	1:000
$\binom{(2)}{5}$	0:001	0:001	1:000
$\binom{(2)}{6}$	0:002	0:001	1:000
$\binom{(2)}{7}$	0:002	0:001	1:000
$\binom{(2)}{8}$	0:003	0:001	1:000
$\binom{(2)}{9}$	0:005	0:001	1:000
$\binom{(2)}{10}$	0:005	0:001	1:000
$\binom{(2)}{11}$	0:006	0:001	1:000
$\binom{(2)}{12}$	0:007	0:001	1:000
$\binom{(2)}{13}$	0:004	0:001	1:000
$\binom{(2)}{14}$	0:005	0:001	1:000
$\binom{(2)}{15}$	0:003	0:001	1:000
$\binom{(2)}{16}$	0:004	0:001	1:000
$\binom{(2)}{17}$	0:003	0:001	1:000
$\binom{(2)}{18}$	0:003	0:001	1:000
$\binom{(2)}{19}$	0:004	0:001	1:000
$\binom{(2)}{20}$	0:003	0:001	1:000
$\binom{(2)}{21}$	0:004	0:001	1:000
$\binom{(2)}{22}$	0:003	0:001	1:000
$\binom{(2)}{23}$	0:002	0:001	1:000
$\binom{(2)}{24}$	0:002	0:001	1:000

CBD-X(2) Model	Mean	Standard Deviation	Gelman-Rubin Test
$\beta^{(2)}_{25}$	0:001	0:001	1:000
$\beta^{(2)}_{26}$	0:000	0:001	1:000
$\beta^{(2)}_{27}$	0:001	0:001	1:000
$\beta^{(2)}_{28}$	0:002	0:001	1:000
$\beta^{(2)}_{29}$	0:002	0:001	1:000
$\beta^{(2)}_{30}$	0:004	0:001	1:000
$\beta^{(2)}_{31}$	0:005	0:001	1:000
$\beta^{(2)}_{32}$	0:006	0:001	1:000
$\beta^{(2)}_{33}$	0:008	0:001	1:000
$\beta^{(2)}_{34}$	0:009	0:001	1:000
$\beta^{(2)}_{35}$	0:010	0:001	1:000
$\beta^{(2)}_{36}$	0:010	0:001	1:000
$\beta^{(2)}_{37}$	0:011	0:001	1:000
$\beta^{(2)}_{38}$	0:011	0:001	1:000
$\beta^{(2)}_{39}$	0:013	0:001	1:000
$\beta^{(2)}_{40}$	0:014	0:001	1:000
$\beta^{(2)}_{41}$	0:016	0:001	1:000
$\beta^{(2)}_{42}$	0:016	0:001	1:000
$\beta^{(2)}_{43}$	0:017	0:001	1:000
$\beta^{(2)}_{44}$	0:017	0:001	1:000
$\beta^{(2)}_{45}$	0:018	0:001	1:000
$\beta^{(2)}_{46}$	0:018	0:001	1:000
$\beta^{(2)}_{47}$	0:019	0:001	1:000
$\beta^{(2)}_{48}$	0:020	0:001	1:000
$\beta^{(2)}_{49}$	0:019	0:001	1:000
$\beta^{(2)}_{50}$	0:020	0:001	1:000
v_{11}	0:000	0:000	1:000
v_{22}	0:000	0:000	1:000
v_{12}	0:000	0:000	1:000
β_1	0:013	0:003	1:000
β_2	0:000	0:000	1:000

CBD-X(3) Model	Mean	Standard Deviation	

CBD-X(3) Model	Mean	Standard Deviation	Gelman-Rubin Test
$^{(1)}_{42}$	0:438	0:016	1:006
$^{(1)}_{43}$	0:471	0:016	1:006
$^{(1)}_{44}$	0:487	0:016	1:006
$^{(1)}_{45}$	0:508	0:016	1:006
$^{(1)}_{46}$	0:547	0:016	1:006
$^{(1)}_{47}$	0:570	0:016	1:006
$^{(1)}_{48}$	0:617	0:016	1:006
$^{(1)}_{49}$	0:620	0:016	1:006
$^{(1)}_{50}$	0:643	0:016	1:006
$^{(2)}_{1}$	0:001	0:001	1:002
$^{(2)}_{2}$	0:002	0:002	1:002
$^{(2)}_{3}$	0:003	0:002	1:002
$^{(2)}_{4}$	0:001	0:002	1:002
$^{(2)}_{5}$	0:000	0:002	1:002
$^{(2)}_{6}$	0:002	0:002	1:002
$^{(2)}_{7}$	0:002	0:002	1:002
$^{(2)}_{8}$	0:002	0:002	1:002
$^{(2)}_{9}$	0:004	0:002	1:002
$^{(2)}_{10}$	0:004	0:002	1:002
$^{(2)}_{11}$	0:005	0:002	1:002
$^{(2)}_{12}$	0:006	0:002	1:002
0:			

CBD-X(3) Model	Mean	Standard Deviation	Gelman-Rubin Test
$^{(2)}_{25}$	0:001	0:002	1:002
$^{(2)}_{26}$	0:000	0:002	1:002
$^{(2)}_{27}$	0:001	0:002	1:002
$^{(2)}_{28}$	0:002	0:002	1:002
$^{(2)}_{29}$	0:002	0:002	1:002
$^{(2)}_{30}$	0:005	0:002	1:002
$^{(2)}_{31}$	0:006	0:002	1:002
$^{(2)}_{32}$	0:007	0:002	1:002
$^{(2)}_{33}$	0:008	0:	

CBD-X(3) Model	Mean	Standard Deviation	Gelman-Rubin Test
$\binom{3}{8}$	0:000	0:000	1:001
$\binom{3}{9}$	0:000	0:000	1:001
$\binom{3}{10}$	0:000	0:000	1:001
$\binom{3}{11}$	0:000	0:000	1:001
$\binom{3}{12}$	0:000	0:000	1:001
$\binom{3}{13}$	0:000	0:000	1:001
$\binom{3}{14}$	0:000	0:000	1:001
$\binom{3}{15}$	0:000	0:000	1:001
$\binom{3}{16}$	0:000	0:000	1:001
$\binom{3}{17}$	0:000	0:000	1:001
$\binom{3}{18}$	0:000	0:000	1:001
$\binom{3}{19}$	0:000	0:000	1:001
$\binom{3}{20}$	0:000	0:000	1:001
$\binom{3}{21}$	0:000	0:000	1:001
$\binom{3}{22}$	0:000	0:000	1:001
$\binom{3}{23}$	0:000	0:000	1:001
$\binom{3}{24}$	0:000	0:000	1:001
$\binom{3}{25}$	0:000	0:000	1:001
$\binom{3}{26}$	0:000	0:000	1:001
$\binom{3}{27}$	0:000	0:000	1:001
$\binom{3}{28}$	0:000	0:000	1:001
$\binom{3}{29}$	0:000	0:000	1:001
$\binom{3}{30}$	0:000	0:000	1:001
$\binom{3}{31}$	0:000	0:000	1:001
$\binom{3}{32}$	0:000	0:000	1:001
$\binom{3}{33}$	0:000	0:000	1:002
$\binom{3}{34}$	0:000	0:000	1:001
$\binom{3}{35}$	0:000	0:000	1:001
$\binom{3}{36}$	0:000	0:000	1:001
$\binom{3}{37}$	0:000	0:000	1:001
$\binom{3}{38}$	0:000	0:000	1:001
$\binom{3}{39}$	0:000	0:000	1:001
$\binom{3}{40}$	0:000	0:000	1:001

CBD-X(3) Model	Mean	Standard Deviation	Gelman-Rubin Test
$^{(3)}_{41}$	0:000	0:000	1:001
$v_{11}^{(3)}$	0:000	0:000	1:002
$^{(3)}_{43}$	0:000	0:000	1:001
$^{(3)}_{44}$	0:000	0:000	1:001
$^{(3)}_{45}$	0:000	0:000	1:001
$^{(3)}_{46}$	0:000	0:000	1:001
$^{(3)}_{47}$	0:001	0:000	1:001
$^{(3)}_{48}$	0:001	0:000	1:001
$^{(3)}_{49}$	0:001	0:000	1:001
$^{(3)}_{50}$	0:001	0:000	1:001
v_{11}	0:000		