7-22-2022 2:00 PM

New Developments on the Estimability and the Estimation of Phase-Type Actuarial Models

Cong Nie, *The University of Western Ontario*

Supervisor: Provost, Serge B, *The University of Western Ontario*
Co-Supervisor: Ren, Jiandong, *The University of Western Ontario*
Co-Supervisor: Liu, Xiaoming, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

This thesis studies the estimability and the estimation methods for two models based on Markov processes: the phase-type aging model (PTAM), which models the human aging process, and the discrete multivariate phase-type model (DMPTM), which can be used to model multivariate insurance claim processes.

The principal contributions of this thesis can be categorized into two areas. First, an objective measure of estimability is proposed to quantify estimability in the context of statistical models. Existing methods for assessing estimability require the subjective specification of thresholds, which potentially limits their usefulness. Unlike these methods, the proposed measure of estimability is objective. In particular, this objectivity is achieved via a carefully designed distribution function sensitivity measure, under which the threshold will become an experiment-based quantity. The proposed measure which is validated to be innately sound, is then applied to assess and improve the estimability of several statistical models, the focus being placed on the PTAM.

Secondly, Markov chain Monte Carlo (MCMC) algorithms are proposed for inference on the PTAM and the DMPTM. Up to now, the MCMC algorithms for continuous phase-type distributions have been applied via the Gibbs sampler which consists of two iterative steps: a data augmentation step and a posterior sampling step. However, owing to unique structures of the PTAM and the DMPTM, this Gibbs sampler turns out to be inadequate, giving rise to problems occurring in either the data augmentation step or the posterior sampling step. To circumvent these difficulties, we methodologically extend the existing Gibbs sampling methodology in terms of rejection sampling and data cloning. The proposed algorithms are then applied to calibrate the PTAM and the DMPTM based on simulated and real-life data. Experimental results show that the proposed MCMC algorithms, as a stochastic approximation technique, achieve estimation results that are comparable to those obtained by deterministic approximation techniques, which can also be seen as a contribution made to the field of approximate inference.

Keywords: phase-type distributions, identifiability, estimability, Markov chain Monte Carlo, data augmentation Gibbs sampler, rejection sampling, data cloning.

Summary for Lay Audience

This thesis principally contributes to two areas of studies.

In statistics, it is well-known that a statistical model is identifiable if parameters can be uniquely inferred from data. However, identifiability does not imply estimability. For example, if the number of observations is low or the numerical algorithm is not sufficiently accurate, then the parameters can only be roughly estimated, even if the model is identifiable. Identifiability has a rigorous mathematical definition. However, estimability is usually measured subjectively and an objective measure appears to be lacking in the context of statistical models. Accordingly, the first contribution of this thesis is to propose an objective measure to quantify estimability in the context of statistical models.

Secondly, Markov chain Monte Carlo (MCMC) algorithms are proposed for inference on two actuarial models: the phase-type aging model (PTAM) and the discrete multivariate phase-type model (DMPTM), where the former models aging processes and the latter models multivariate insurance claim processes. MCMC is a methodology for sampling complicated distributions, where one constructs a carefully designed Markov chain whose stationary distribution agrees with the target distribution. Then, sampling from the target distribution is replaced with sampling from the designed Markov chain. In this thesis, we develop MCMC algorithms for the PTAM and DMPTM, where the models' special structures are utilized.

Co-Authorship Statement

Chapter 4 titled "Markov chain Monte Carlo for Bayesian inference on the phase-type aging model" was co-authored with Dr. Xiaoming Liu, Dr. Serge Provost and Dr. Jiandong Ren. It has been submitted to *North American Actuarial Journal*. Professor Liu motivated the research topic. Professor Provost and Professor Ren provided corrections and suggestions regarding the format and the wording of the paper. The paper is otherwise almost entirely the product of my own work, including the determination of contributions, the establishment of theories, the implementation of algorithms and the completion of numerical experiments and the manuscript.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my former supervisor Dr. Xiaoming Liu and my supervisors Dr. Serge B. Provost and Dr. Jiandong Ren. As the first supervisor in my Ph.D. studies, Professor Liu's pragmatic and rigorous attitude towards scientific research deeply impressed me and led me to become a qualified scholar. She was always being patient, caring and tolerant, which I have truly appreciated. Without her long-standing support, I would not have been able to overcome all the difficulties. Professor Provost provided insightful guidance on various aspects of the material introduced in connection with the phase-type aging model and the concept of estimability. Professor Ren suggested to work on the discrete multivariate phase-type model after examining my previous research, so that an additional chapter was initiated to make the thesis more complete. It has truly been my pleasure to have them as supervisors. Without their guidance, I would not have been able to complete all the required milestones and finish the thesis within the expected timeline.

Secondly, I would like to thank my thesis defence committee members: Dr. Qiming He, Dr. Pei Yu, Dr. Kristina Sendova and Dr. Shu Li for reviewing my thesis and providing valuable feedback.

My special thanks go to the following scholars who inspired me and strengthened my confidence during my research: Dr. Louis J. M. Aslett, Dr. Wil Michiels, Dr. Andreas Raue and Dr. Jean Rizk.

I also thank my Master's program supervisor Dr. Gordon E. Willmot for providing a solid recommendation, which, according to Professor Liu, immediately finalized her decision to take me as her student.

My thanks also go to my friends Yawo Kobara and Zelin Zhang for supporting me during my struggles. In particular, I am grateful to Marilyn Nixon, my friendly neighbour in Waterloo, I will never forget her prayers for me at a crossing of my life. I also thank career coach Jade Shi for giving me a professional evaluation which strengthened my determination to complete a Ph.D. degree after working.

Last but not least, I am thankful for my family's endless love, support and confidence in me.

To my family

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Research motivation and objectives

This thesis investigates two phase-type actuarial models: the phase-type aging model (PTAM) and the discrete multivariate phase-type model (DMPTM). The motivations behind these two models are slightly different, which are discussed separately in Sections 1.1.1 and 1.1.2.

1.1.1 Motivation in connection with the PTAM

The PTAM belongs to a class of Coxian Markovian models that was proposed in Cheng et al. (2021). The purpose of the PTAM is to provide a quantitative description of well-known aging characteristics that are part of a genetically determined, progressive and irreversible process. It has already been utilized to fit simulated and real-life data in Cheng et al. (2021) and Cheng (2021).

However, previous model fitting results reveal a major issue: the profile likelihood functions are flat for certain parameters. This phenomenon hinders the reliability of the parameter estimates as a wide range of estimates can produce nearly the same profile likelihood value. Consequently, we aim to seek possible ways to address this issue. Two research avenues are then pursued: Bayesian inference and estimability.

Bayesian inference

An intuitive approach for fixing the issue of flat profile likelihood functions is to utilize Bayesian inference. As the (profile) likelihood functions are flat, the posterior distributions will highly depend on the prior distributions. In that case, the reliability of the parameter estimates can be improved via sound prior information, which produces narrow credible intervals for the parameter estimates. In the context of continuous phase-type distributions, the literature suggests that Bayesian inference be applied via the Gibbs sampler [Bladt et al. (2003), Aslett and Wilson (2011), Okamura et al. (2014)], which belongs to the Markov chain Monte Carlo method. Thus, the research objective consists of applying MCMC-based

Bayesian inference to the PTAM with a view to improve parameter estimability via sound prior information. This gives rise to the contributions presented in Chapter 4.

Estimability

Interestingly, based on another school of thoughts, it turns out that the problem resulting from flat profile likelihood functions is related to the concept of estimability, which is also interchangeably referred to as "practical identifiability". The concept of estimability originates from system biology where ODE models are utilized to model dynamic biological systems. However, an objective definition of estimability appears to be lacking in the context of statistical models. Thus, the research objective consists of initiating a novel definition of estimability to objectively quantify estimability for statistical models, particularly that of the PTAM. After having objectively quantified estimability by redefining it, we may then make improvements on the basis of the proposed definition. This gives rise to the contributions presented in Chapter 3.

However, this research objective cannot be readily implemented. According to the literature, identifiability must be assessed before estimability [Raue et al. (2009); Hengl et al. (2007); Miao et al. (2011); Petersen et al. (2001); McLean and McAuley (2012); Brun et al. (2001); Holmberg (1982); Gontier and Pfister (2020)]. Thus, the identifiability of the PTAM must first be determined, which gives rise to the contributions presented in Chapter 2.

1.1.2 Motivation in connection with the DMPTM

The discrete multivariate phase-type model (DMPTM) is a class of discrete phase-type distributions based on discrete time Markov chains with marked transitions. He and Ren (2016b) established an EM algorithm with respect to its parameter estimation.

Similarly to the Bayesian inference on the PTAM, the aim is to develop an MCMC algorithm for inference on the DMPTM. However, the motivation of this research project is slightly different. The goal is not to utilize Bayesian inference as a means of improving estimability, but to provide an alternative way of determining MLEs, which does not involve the EM algorithm. This can be achieved by combining the MCMC algorithm with the data cloning method [Lele et al. (2007, 2010)], which gives rise to the contributions in Chapter 5.

As well, there are convincing reasons for applying MCMC algorithms to the DMPTM from the perspective of approximate inference. In the approximate inference, the EM algorithm is classified as deterministic approximation, whereas the MCMC algorithm is classified as stochastic approximation. These two categories are parallel and partitions the field of approximate inference [Bishop and Nasrabadi (2006)]. From the perspective of approximate inference, while the application of a deterministic approximation to the DMPTM has been researched via the EM algorithm in He and Ren (2016b), the stochastic approximation counterpart remains unexplored to this day. The development of the MCMC algorithm will then fill this gap and contribute to the field of approximate inference, focusing on the DMPTM.

1.2 Structure of the thesis

The chapters of this thesis are organized as follows.

- In Chapter 1 (current chapter), the research motivation and objectives are discussed, and the structure of the thesis is described.
- In Chapter 2, the identifiability of the PTAM is investigated, which is required for investigation on its estimability in Chapter 3.
- In Chapter 3, a novel definition of estimability is introduced in the context of statistical models. The proposed definition which is validated to be innately sound, will then be applied to assess the estimability of several statistical models, and in particular that of the PTAM.
- In Chapter 4, an MCMC-based Bayesian estimation method is proposed and applied to the PTAM, with a view to improving its estimability. While numerical experimental results indicate that the proposed methodology improves estimability for the PTAM as opposed to the MLE method, this approach may also be utilized as a standalone model fitting technique.
- In Chapter 5, an MCMC algorithm is developed in conjunction with data cloning technique for inference on the DMPTM. While the existing EM algorithm determines the MLEs of the DMPTM based on a deterministic approximation approach, the proposed algorithm provides an alternative way of obtaining the MLEs from a stochastic approximation approach, which directly contributes to the field of approximate inference.
- In Chapter 6, we make general comments and concluding remarks related to our contributions and state certain possible avenues for further research.

The appendices of this thesis are organized as follows.

- Appendix A rigorously investigates several mathematical properties of the dying rates of the PTAM, which are utilized in establishing the identifiability of the PTAM in Chapter 2.
- Appendix B presents rigorous proofs regarding sub-models of the PTAM, which are utilized in assessing the estimability of the PTAM in Chapter 3.
- Appendix C provides the derivation of likelihood function of the PTAM for left-truncated data, which supports the MCMC-based Bayesian inference on the PTAM developed in Chapter 4.
- Appendix D provides algorithms for rejection sampling on a logarithmic scale, which supports the MCMC-based Bayesian inference on the PTAM in Chapter 4.
- Appendix E presents the technical details of the proposed MCMC algorithm for the DMPTM presented in Chapter 5.

Chapter 2

Identifiability of the Phase-Type Aging Model

In this chapter, identifiability of the phase-type aging model (PTAM) is thoroughly investigated. This provides supplementary support to Chapter 3 regarding estimability of the PTAM, because identifiability must be assessed before estimability [Raue et al. (2009); Hengl et al. (2007); Miao et al. (2011); Petersen et al. (2001); McLean and McAuley (2012); Brun et al. (2001); Holmberg (1982); Gontier and Pfister (2020)]. An ad-hoc mathematical proof is introduced, which establishes that the PTAM is identifiable when the number of states is greater or equal to six.

2.1 Preliminaries on the phase-type aging model

The phase-type aging model (PTAM) stems from the phase-type mortality model proposed in Lin and Liu (2007). The motivation for proposing the phase-type mortality model is that it lends itself to linking its parameters to biological and physiological mechanisms of aging, so that the longevity risk facing annuity products can be measured more accurately. Experimental results showed that the phase-type mortality model with a four-state developmental period and a subsequent aging period achieved very satisfactory fitting results with respect to the Swedish and USA cohort mortality data [Lin and Liu (2007)]. Later on, Su and Sherris (2012) applied the phase-type mortality model to an Australian cohort mortality data.

Subsequently, the PTAM proposed in Cheng et al. (2021) developed the aging period of the phase-type mortality model, the difference being that a parsimonious yet flexible representation was adopted for modeling various aging patterns. Similarly, the main objective of the PTAM is to describe the human aging process in terms of the evolution of the distribution of physiological ages, utilizing mortality rates as aging-related variables. Therefore, although the PTAM can reproduce mortality patterns, it ought not to be treated as a mortality model. In this context, the PTAM is most applicable at human ages beyond the attainment of adulthood, where relatively speaking the aging process is the most significant factor that contributes to the variability in lifetimes [Cheng et al. (2021)].

Definition 2.1.1. *Let $\{X(t)\}_{t\geq 0}$ be a continuous time Markov chain (CTMC) defined on a finite state space $\mathcal{S} = \mathcal{E} \cup \Delta = \{1, 2, \ldots, m\} \cup \Delta$, where $\Delta = \{m+1\}$ is the absorbing state and $\mathcal{E}$ is the set of transient states. Let $\{X(t)\}_{t\geq 0}$ have initial distribution $\boldsymbol{\pi}' = (\pi_1, \pi_2, \ldots, \pi_m)$ over the transient states such that $\boldsymbol{\pi}'\boldsymbol{e} = 1$, and let the transition intensity matrix be*

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{S} & \boldsymbol{h} \\ \boldsymbol{0} & 0 \end{bmatrix}, \tag{2.1}$$

where $\boldsymbol{h} = -\boldsymbol{S}\boldsymbol{e}$ and $\boldsymbol{e}$ is a column vector of ones. Define $T := inf\{t \geq 0 | X(t) = m + 1\}$ as the time until absorption. Then, T is said to follow a continuous phase-type (CPH) distribution denoted by $CPH(\boldsymbol{\pi}, \boldsymbol{S})$ of order m, and $\boldsymbol{h}$ is defined as the exit vector.

There is a long history of using phase-type distributions for survival modelling in the category of "absorbing time" distributions; see Aalen (1995); Asmussen et al. (1996); Lin and Liu (2007); Su and Sherris (2012).

Result 1. *Given that $T \sim CPH(\boldsymbol{\pi}, \boldsymbol{S})$ of order m,*

- *The p.d.f. of T is $f_T(t) = \boldsymbol{\pi}' e^{\boldsymbol{S}t}\boldsymbol{h}$.*
- *The c.d.f. of T is $F_T(t) = 1 - \boldsymbol{\pi}' e^{\boldsymbol{S}t}\boldsymbol{e}$.*

It is well-known that, if $\boldsymbol{S}$ of a CPH distribution of order m follows structure specified in (2.2) and $\boldsymbol{\pi}' = (1, 0, \ldots, 0)$, then that distribution belongs to a Coxian distribution with no probability mass at zero [Cox (1955b,a)].

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda_1 + h_1) & \lambda_1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & -(\lambda_2 + h_2) & \lambda_2 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -(\lambda_{m-1} + h_{m-1}) & \lambda_{m-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & -h_m \end{bmatrix}, \tag{2.2}$$

where $\lambda_i > 0, h_j > 0$, $i = 1, 2, \ldots, m-1$, and $j = 1, 2, \ldots, m$.

Then, a phase diagram such as that displayed in Figure 2.1 can often be visualized:

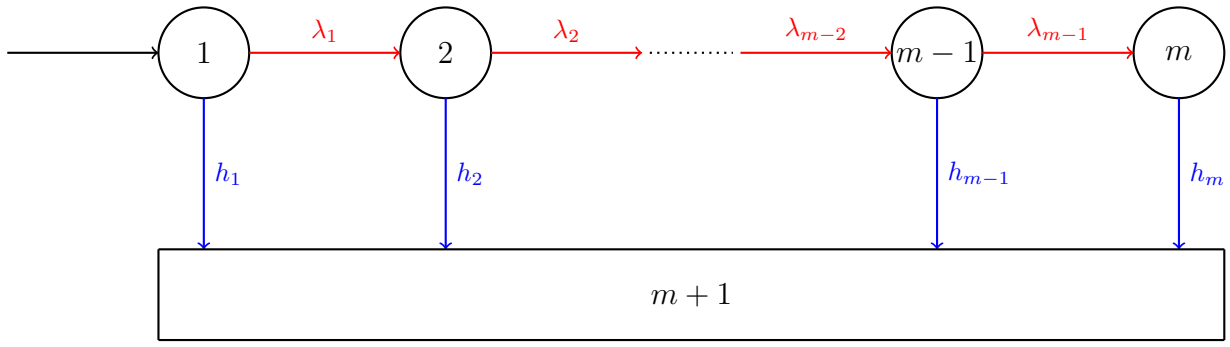

Figure 2.1: Phase diagram for a Coxian distribution with no probability mass at zero.

Definition 2.1.2. *Given that $T > 0$, the PTAM of order m is defined as a Coxian distribution of order m with transition intensity matrix $\boldsymbol{S}$ and exit rate vector $\boldsymbol{h}$ such that*

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda + h_1) & \lambda & 0 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda + h_2) & \lambda & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -(\lambda + h_{m-1}) & \lambda \\ 0 & 0 & 0 & 0 & \dots & 0 & -h_m \end{bmatrix}, \quad \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{m-1} \\ h_m \end{bmatrix}, \tag{2.3}$$

where $\lambda > 0$, $h_m > h_1 > 0$ and

$$h_i = \begin{cases} \left(\frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s\right)^{\frac{1}{s}}, & s \neq 0, \\ h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}}, & s = 0, \end{cases} \tag{2.4}$$

$i = 1, 2, \dots, m$. This is denoted by $PTAM(h_1, h_m, s, \lambda, m)$.

As can be seen from Figure 2.2, the PTAM has a phase diagram similar to that of the Coxian distribution shown in Figure 2.1, the difference being the constant transition rate and the functionally related exit rates defined in (2.4).

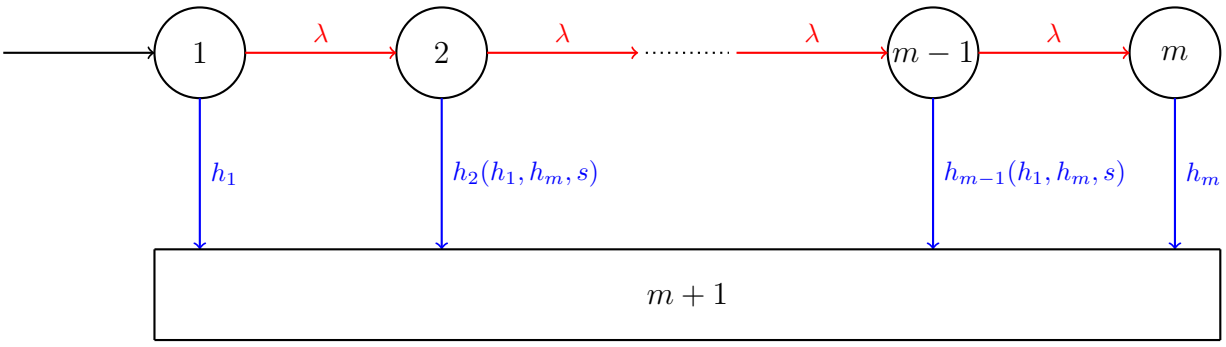

Figure 2.2: Phase diagram for the PTAM.

(i) In Figure 2.2, each state in the Markov process represents the physiological age - a variable that reflects an individual's health condition or frailty level. As the aging process proceeds, the frailty level will increase, until the last state where the individual's health conditions have deteriorated to the point of causing death.

(ii) The transition rate λ is assumed to be constant. The exiting rates h_i's are the dying rates or force of dying. With this setup, an individual will be randomly located in a certain state at a given calendar age. This mathematically describes the fact that the individuals involved will have different physiological ages given the same calendar age.

(iii) The assumption that dying rates have the structure given in (2.4) is somewhat reminiscent of the well-known Box-Cox transformation introduced by Box and Cox (1964). The first and last dying rates h_1 and h_m are included in the model parameters, whereas the remaining in-between rates are interpolated in terms of the parameter s which is a model parameter related to the curvature of the exit rate pattern. To verify this, Figure 2.3 presents the effect of s on the pattern of the exit rates. When $s = 1$, the dying rates have a linear relationship. When $s > 1$, rates are concave, and when $s < 1$, rates are convex. In particular, when $s = 0$, rates behave exponentially. In practice, we believe that it is likely that $s < 1$ when calibrating to mortality data [Cheng et al. (2021)]. That is, the dying rates increase faster than linearly as an individual ages. Throughout this chapter, it will be assumed that h_i will follow the structure given in (2.4), for $i = 1, 2, \ldots, m$.

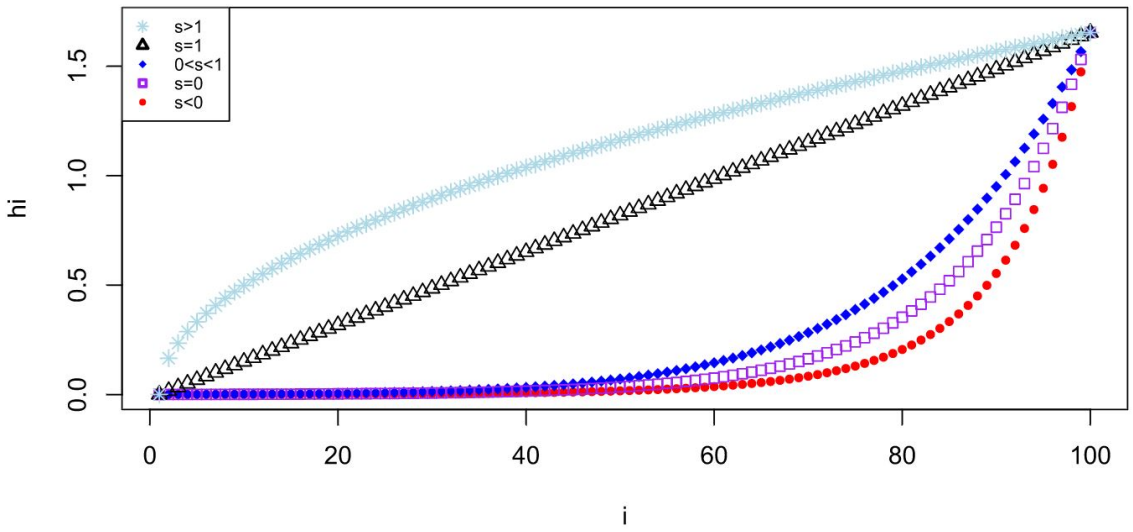

Figure 2.3: Behaviour of the exit rate vector for various values of s with $m = 100$.

(iv) The value of λ needs to be appropriate to the value of m, otherwise there is no need to have many m states if λ is small. We therefore let their ratio be a constant, that is,

$$\frac{m}{\lambda} = \psi. \tag{2.5}$$

(2.5) can be seen as a reparameterization of the PTAM as it still needs five parameters as (h_1, h_m, s, ψ, m). However, such a reparameterization will establish a positive covariance between λ and m, which is more in line with the biological interpretation. Throughout this chapter, we will utilize these two parameterizations of the PTAM interchangeably as needed.

The parameter structure of the PTAM proves to be parsimonious and flexible, which allows one to model the internal aging process explicitly. Further information is available in Cheng et al. (2021).

2.2 Literature review on identifiability of the CPH distributions

Identifiability is also referred to as *a priori* identifiability or uniqueness of model representation in other sources. It relates to the property of the model structure itself before model fitting. The definition of identifiability of a statistical model given in Lehmann and Casella (1998) is presented in Definition 2.2.1.

Definition 2.2.1. *Let $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a statistical model with either finite- or infinite-dimensional parameter space $\boldsymbol{\Theta}$. We say that $\mathcal{M}$ is identifiable if*

$$\left(f(x;\boldsymbol{\theta_1}) = f(x;\boldsymbol{\theta_2})\right) \Rightarrow \left(\boldsymbol{\theta_1} = \boldsymbol{\theta_2}\right), \forall \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \boldsymbol{\Theta}. \tag{2.6}$$

Based on Definition 2.2.1, if a model is non-identifiable, then there exists two p.d.f.-equivalent models with different parameters. On the other hand, if a model is identifiable, then such models do not exist. In other sources, the term "c.d.f.-equivalent" is used, but it is equivalent to "p.d.f.-equivalent". Both terms will be used interchangeably in this thesis.

2.2.1 Identifiability of acyclic phase-type distributions

The PTAM belongs to the class of so-called acyclic phase-type (APH) distributions, which is defined as the class of CPH distributions whose associated Markov process is irreversible. Cumani (1982) derived three c.d.f.-equivalent canonical forms for the APH distributions and proved that any APH distribution with m transient states could be represented by one of the equivalent forms with $2m-1$ as the minimum number of parameters, assuming no probability mass at zero. Each one of the canonical forms is identifiable. The same conclusion was also obtained in Telek and Horváth (2007). We briefly recall their results in Theorem 2.2.2.

Theorem 2.2.2. *Suppose that X follows an APH$(\boldsymbol{\pi}, \boldsymbol{S})$ of order m, and let the eigenvalues, or diagonal terms[1], of $-\boldsymbol{S}$ be denoted by $(D_1, D_2, \ldots, D_{m-1}, D_m)$. Let π_i be the initial probability that the Markov process starts at state i, assuming no probability mass at zero. Then, the minimal number of parameters required to represent the distribution of X is $2m-1$, and there are three c.d.f.-equivalent canonical forms, with different underlying Markov processes $\{X(t)\}_{t\geq 0}$, as shown in Figures 2.4, 2.5 and 2.6, where $D_1 \geq D_2 \geq \cdots \geq D_m$ are the ordered eigenvalues of $-\boldsymbol{S}$ and $\boldsymbol{\pi}$ with $\boldsymbol{\pi'e} = 1$ are the initial probabilities. Moreover, $x_i := \pi_{m+1-i} D_1$ and $y_i := D_i \frac{\pi_{m+1-i}}{\sum_{j=1}^{m+1-i} \pi_j}$ are transition rates in canonical forms 2 and 3, respectively. With these relationships, they will be c.d.f.-equivalent to canonical form 1.*

The proof is available from Cumani (1982).

[1]Since $-\boldsymbol{S}$ is upper triangular, its eigenvalues are its diagonal elements.

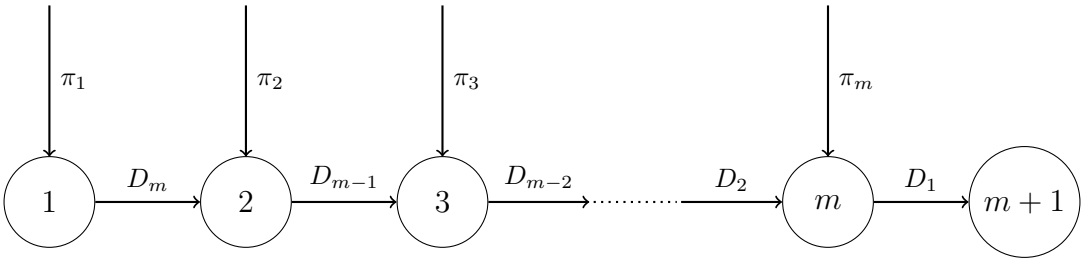

Figure 2.4: Canonical Form 1 of the APH.

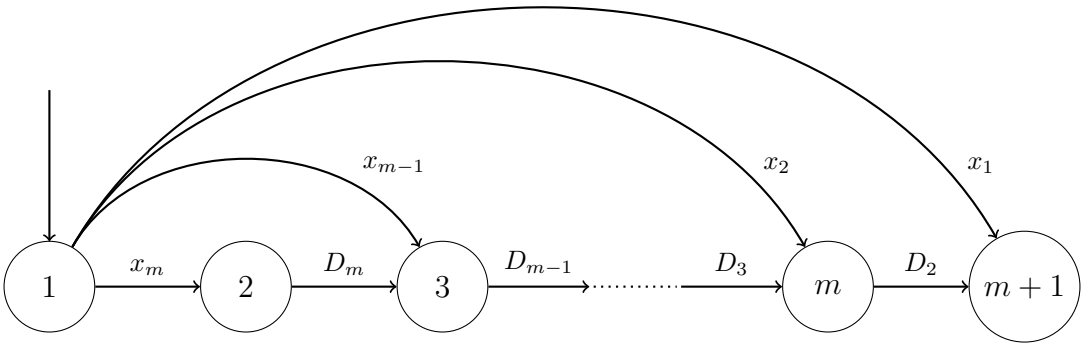

Figure 2.5: Canonical Form 2 of the APH.

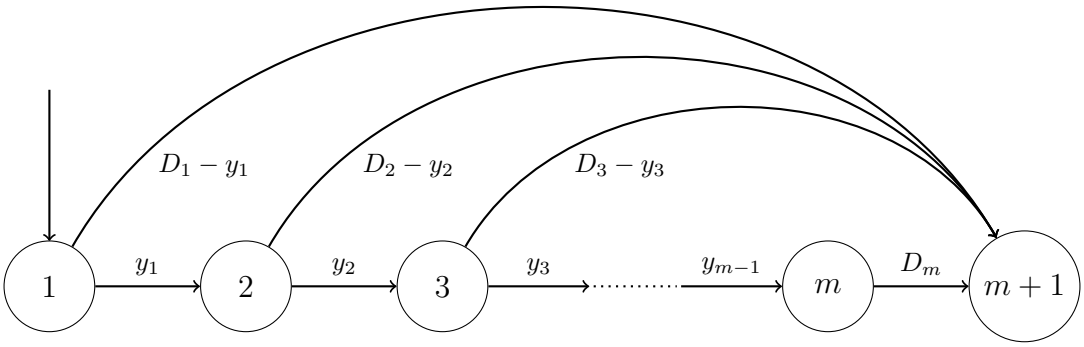

Figure 2.6: Canonical Form 3 of the APH.

Accordingly, their representations can be obtained. For the canonical form 1,

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{m-1} \\ \pi_m \end{bmatrix}, \ \boldsymbol{S} = \begin{bmatrix} -D_m & D_m & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -D_{m-1} & D_{m-1} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -D_3 & D_3 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -D_2 & D_2 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -D_1 \end{bmatrix}. \tag{2.7}$$

For the canonical form 2,

$$\boldsymbol{\pi} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{S} = \begin{bmatrix} -\sum_{i=1}^{m} x_i & x_m & x_{m-1} & x_{m-2} & \dots & x_4 & x_3 & x_2 \\ 0 & -D_m & D_m & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -D_4 & D_4 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -D_3 & D_3 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -D_2 \end{bmatrix}. \tag{2.8}$$

For the canonical form 3,

$$\boldsymbol{\pi} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{S} = \begin{bmatrix} -D_1 & y_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -D_2 & y_2 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -D_{m-2} & y_{m-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -D_{m-1} & y_{m-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -D_m \end{bmatrix}. \tag{2.9}$$

In light of Theorem 2.2.2, it is possible for different APH distributions with different underlying Markov processes to be c.d.f.-equivalent, even with the minimal number of parameters of $2m-1$. This implies that the APH distribution will be non-identifiable as it can be represented by different canonical forms. Fortunately, each one of the canonical forms is identifiable. Therefore, the APH distribution is regarded as an appropriate subclass of phase-type distributions for modelling purposes, provided that one specifies which of the canonical forms is to be utilized [Okamura and Dohi (2016); Bobbio and Cumani (1992)].

2.2.2 Identifiability of the Coxian distribution

In view of Figure 2.6, the canonical form 3 belongs to a Coxian distribution with the constraint that the states are ordered. In fact, the ordering of states is an indispensable condition for canonical form 3 to be identifiable. Without such ordering, it will be possible to create c.d.f.-equivalent representations even within the class of Coxian distribution. Rizk et al. (2019) investigated this issue in depth. We briefly recall their results in Theorem 2.2.3.

Theorem 2.2.3. *Consider two Coxian distributions of the same order m with different representations $(\boldsymbol{\pi}_a, \boldsymbol{S}_a)$ and $(\boldsymbol{\pi}_b, \boldsymbol{S}_b)$, where $\boldsymbol{\pi}'_a = \boldsymbol{\pi}'_b = (1, 0, \ldots, 0)$. Then, they are c.d.f.-equivalent only if following two conditions hold true.*

(i) The exit rates from the first state are equal in both models, that is, $h_{1a} = h_{1b}$.

(ii) The diagonal entries in $\boldsymbol{S}_a$ are a permutation of those in $\boldsymbol{S}_b$.

Proof. See Rizk et al. (2019). □

Following Theorem 2.2.3, an approach for constructing a non-identifiable candidate of a Coxian distribution was developed in Rizk et al. (2019). It is presented as Algorithm 1:

Algorithm 1 Construction of a non-identifiable candidate for a Coxian distribution [Rizk et al. (2019)]

Require: Same order m.
Input: $\boldsymbol{\pi}_a, \boldsymbol{S}_a$, where $\boldsymbol{\pi}'_a = (1, 0, \ldots, 0)$.
Output: $(\boldsymbol{\pi}_b, \boldsymbol{S}_b)$ such that $\boldsymbol{\pi}_b = \boldsymbol{\pi}_a$ and $\boldsymbol{S}_b \neq \boldsymbol{S}_a$.

1: Fix the exit rate from the first state, that is, $h_{1a} = h_{1b}$.
2: Consider a possible permutation of the diagonal entries in $\boldsymbol{S}_a$. This will result in the diagonal terms of $\boldsymbol{S}_b$.
3: Solve for off-diagonal terms in $\boldsymbol{S}_b$ based on moment-matching method[2].

Algorithm 1 will be the theoretical foundation for the proposed mathematical proof for the identifiability of the PTAM, which will be discussed in the next section.

2.3 Identifiability of the PTAM

The principal contribution of this chapter is contained in this section. Since the PTAM belongs to a special class of APH distributions, it also needs to be compared with the three canonical forms when investigating its identifiability. According to Definition 2.1.2, the following two characteristics of the PTAM can be observed when comparing it with the canonical forms:

(i) The number of parameters is $2m - 1$.

(ii) The first $2m - 1$ states are ordered except the last state.

Clearly, the PTAM does not belong to any one of the three canonical forms. In fact, it almost falls into canonical form 3, the only violation being the freedom of the last state. Due to this violation, the identifiability of the PTAM cannot be immediately guaranteed by the canonical forms. On the other hand, the two parameter constraints on the PTAM - constant

[2]The off-diagonal terms are solved based on equating the k^{th} moments of the two models, where $k = 1, 2, 3, \ldots$. We start with $k = 1$ until all the terms are solved. The details of this moment-matching method are available in Rizk et al. (2019).

λ and the constraint (2.4) on $h_i's$ - will increase the model identifiability, counteracting the freedom associated with the last state. So, we could ask ourselves: which aspect takes over? The non-identifiability concern brought about by the the freedom of last state or the identifiability improvement resulting from imposing parameter constraints on λ and $h_i's$? If it is the former, then the PTAM will be non-identifiable; if it is the latter, then the PTAM will be identifiable.

We approach this problem by considering an ad-hoc mathematical proof based on Rizk et al. (2019). The main methodology being utilized is 'proof by contradiction'. Given the PTAM, we will examine all the possibilities of constructing non-identifiable candidates. If there is no possible way to construct any non-identifiable candidates, then we will have established that the PTAM is identifiable.

2.3.1 Identifiability of the PTAM for fixed large m

In this section, we initially fix the number of states m assuming it is large, and then construct non-identifiable candidate first by fixing the exit rate and then permuting the diagonal terms of the intensity matrix.

Before beginning the proof, a modified algorithm needs to be developed for constructing a non-identifiable candidate for the PTAM. We present it as Algorithm 2.

Algorithm 2 Construction of a non-identifiable candidate for the PTAM

Require: Same order m.
Input: $\boldsymbol{\pi}_a, h_{1a}, h_{m_a}, s_a, \lambda_a$, where $\boldsymbol{\pi}_a' = (1, 0, \ldots, 0)$.
Output: $(\boldsymbol{\pi}_b, h_{1b}, h_{m_b}, s_b, \lambda_b)$ such that $\boldsymbol{\pi}_b = \boldsymbol{\pi}_a$ and $(h_{1a}, h_{m_a}, s_a, \lambda_a) \neq (h_{1b}, h_{m_b}, s_b, \lambda_b)$.

1: Fix the exit rate from the first state, that is, $h_{1a} = h_{1b}$.
2: Consider a possible permutation of the diagonal entries in $\boldsymbol{S}_a$. This will result in the diagonal terms for $\boldsymbol{S}_b$.
3: Verify that the parameter constraints - constant transition rate and (2.4) - are satisfied in $\boldsymbol{S}_b$, so that the candidate is eligible to be a PTAM.
4: Verify that the moments are matched between two models.

Unlike Algorithm 1, Step 3 in Algorithm 2 aims at verifying whether the model is still eligible as a PTAM after fixing the first exit rate and permuting the diagonal terms in the transition matrix. This step is needed because the parameters are independent in the case of a Coxian distribution as defined in Rizk et al. (2019), whereas the parameters for the PTAM have constraints. Thus, one must verify if the model is still eligible as a PTAM. In fact, as will be pointed out further, it is generally difficult to construct a non-identifiable candidate whose dying rates satisfy (2.4), which prevents the algorithm from going through in Step 3. In that case, one does not even need to consider verifying the moments in Step 4, because the construction has already failed at Step 3. We now begin the proof.

Proof of the identifiability of the PTAM for large m

Proof. In this proof, we will discuss three cases separately as Cases 1, 2 and 3. They correspond to different treatments on the permutations of the diagonal terms of the transition matrix.

Case 1

In Case 1, we will consider whether $\boldsymbol{S}_b$, as a non-identifiable candidate, can be possibly produced after ordering the diagonal terms of $\boldsymbol{S}_a$ in descending order. According to Definition 2.1.2, the $h_i's$ are increasing. Thus, the first $m-1$ terms on the diagonal entries in $\boldsymbol{S}_a$ are decreasing: $-h_1-\lambda > -h_2-\lambda > \cdots > -h_{m-1}-\lambda$, which is satisfactory as they are already ordered. However, the last term $-h_m$ can fall anywhere, its location depending on the value of λ. Therefore, we will split Case 1 into sub-cases as explained below.

Case 1.1

Case 1.1 considers the instance where $-h_m$ is neither the smallest nor the largest diagonal term. Mathematically, the diagonal terms of $\boldsymbol{S}_b$ after ordering become

$$-h_1-\lambda > -h_2-\lambda > \cdots > -h_n-\lambda \geq -h_m \geq -h_{n+1}-\lambda \cdots > -h_{m-1}-\lambda, \tag{2.10}$$

where $m \geq 6$, $1 \leq n \leq m-2$. Then $\boldsymbol{h}_b$, if it exists, should be

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \\ h_m-\lambda \\ h_{n+1} \\ \vdots \\ h_{m-1}+\lambda \end{bmatrix}. \tag{2.11}$$

Notice that extra points $h_m-\lambda$ and $h_{m-1}+\lambda$ break the smooth structure of $h_1, h_2, \ldots, h_{m-1}$, as shown in Figure 2.3. Therefore, by Proposition A.2.1 in Appendix A, $\boldsymbol{h}_b$ cannot be described by any alternative parameter s_b. Thus, it is impossible to satisfy (2.4). In this case, Step 3 in Algorithm 2 failed.

Case 1.2

Case 1.2 considers the instance where $-h_m$ is the largest diagonal term. Mathematically, the diagonal terms of $\boldsymbol{S}_b$ after ordering become

$$-h_m \geq -h_1-\lambda > -h_2-\lambda > \cdots > -h_{m-1}-\lambda, \tag{2.12}$$

where $m \geq 6$. Then $\boldsymbol{h}_b$, if it exists, should be

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_1 + \lambda + h_1 - h_m \\ \vdots \\ h_{m-2} + \lambda + h_1 - h_m \\ h_{m-1} + \lambda \end{bmatrix}. \tag{2.13}$$

Note that the second to $(m-1)^{th}$ terms in $\boldsymbol{h}_b$ can be seen as the corresponding components of $\boldsymbol{h}_a$ being shifted upwards by $\lambda + h_1 - h_m$. Thus, the last term breaks the smooth pattern as shown in Figure 2.3. Therefore, by Proposition A.3.1 in Appendix A, $\boldsymbol{h}_b$ cannot be described by any alternative parameter s_b. Thus, it is impossible to satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 1.3

Case 1.3 considers the instance where $-h_m$ is the smallest diagonal element. Mathematically, the diagonal terms of $\boldsymbol{S}_b$ after ordering become

$$-h_1 - \lambda > -h_2 - \lambda > \cdots > -h_{m-1} - \lambda \geq -h_m, \tag{2.14}$$

where $m \geq 6$. Notice that the status quo holds, so that Case 1.3 is unnecessary. However, it is included for the sake of completeness.

Case 1 Conclusion

Therefore, in Case 1, when we intend to construct a non-identifiable candidate, $\boldsymbol{S}_b$, after ordering the diagonal terms in descending order, there is no possible way to construct a valid candidate as Step 3 in Algorithm 2 will fail.

Case 2

We will consider whether $\boldsymbol{S}_b$, as a non-identifiable candidate, can be possibly produced after ordering the diagonal terms of $\boldsymbol{S}_a$ in descending order and putting an arbitrary term last. It is allowed to put a term last because the PTAM requires freedom of the last state. In Case 2, we will consider moving a term preceding $-h_m$ last.

Case 2.1

In Case 2.1, we require that the term being moved satisfies following conditions.

- The term is not right before $-h_m$.
- The term is not the first term, $-h_1 - \lambda$.

Mathematically: $-h_1-\lambda > -h_2-\lambda > \cdots > -h_k-\lambda > \cdots > -h_n-\lambda \geq -h_m \geq -h_{n+1}-\lambda > \cdots > -h_{m-1}-\lambda$, where $3 \leq n \leq m-1$, $1 < k < n$, $m \geq 6$. After moving $-h_k - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ \vdots \\ h_{k-1} \\ h_{k+1} \\ \vdots \\ h_n \\ h_m - \lambda \\ h_{n+1} \\ \vdots \\ h_{m-1} \\ h_k + \lambda \end{bmatrix}. \tag{2.15}$$

Notice that the terms $h_1, \ldots, h_{k-1}, h_{n+1}, \ldots, h_{m-1}$, lie on the original curve. However, the terms $h_{k+1}, \ldots, h_n$, are shifted leftwards by one position from the original curve. Therefore, by Proposition A.2.1, Case 2.1 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed[3].

Case 2.2

In Case 2.2, we require that the term being moved satisfies following conditions.

- The term is not right before $-h_m$.
- The term is the first term, $-h_1 - \lambda$.
- $-h_m$ is not the smallest term among the diagonal terms.

Mathematically: $-h_1-\lambda > -h_2-\lambda > \cdots > -h_n-\lambda \geq -h_m \geq -h_{n+1}-\lambda > \cdots > -h_{m-1}-\lambda$, where $1 \leq n \leq m-2$, $m \geq 6$. After moving $-h_1 - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_3 + h_1 - h_2 \\ \vdots \\ h_n + h_1 - h_2 \\ h_m - \lambda + h_1 - h_2 \\ h_{n+1} + h_1 - h_2 \\ \vdots \\ h_{m-1} + h_1 - h_2 \\ h_1 + \lambda \end{bmatrix} = \begin{bmatrix} h_2 + (h_1 - h_2) \\ h_3 + (h_1 - h_2) \\ \vdots \\ h_n + (h_1 - h_2) \\ h_m + (h_1 - h_2) - \lambda \\ h_{n+1} + (h_1 - h_2) \\ \vdots \\ h_{m-1} + (h_1 - h_2) \\ h_1 + \lambda \end{bmatrix}. \tag{2.16}$$

[3]If $n = m-1$, then the points $h_{n+1}, \ldots, h_{m-1}$ will disappear. However, the argument remains unchanged.

We have the terms $h_2 + (h_1 - h_2), h_3 + (h_1 - h_2), \ldots, h_{n-1} + (h_1 - h_2)$ as the original curve shifted upwards by $(h_1 - h_2)$. In addition, an extra entry $h_m + (h_1 - h_2) - \lambda$ breaks the pattern shown in Figure 2.3. Therefore, either by Proposition A.2.1 or A.3.1, Case 2.2 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 2.3

In Case 2.3, we require that the term being moved satisfies following conditions.

- The term is not right before $-h_m$.
- The term is the first term, $-h_1 - \lambda$.
- $-h_m$ is the smallest term among the diagonal terms.

Mathematically: $-h_1 - \lambda > -h_2 - \lambda > \cdots > -h_{m-1} - \lambda \geq -h_m$, $m \geq 6$. After moving $-h_1 - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_2 + (h_1 - h_2) \\ h_3 + (h_1 - h_2) \\ \vdots \\ h_{m-1} + (h_1 - h_2) \\ h_m + (h_1 - h_2) - \lambda \\ h_1 + \lambda \end{bmatrix}. \tag{2.17}$$

We have the terms $h_2 + (h_1 - h_2), h_3 + (h_1 - h_2), \ldots, h_{n-1} + (h_1 - h_2)$ as the original curve shifted upwards by $(h_1 - h_2)$. In addition, an extra entry $h_m + (h_1 - h_2) - \lambda$ breaks the pattern shown in Figure 2.3. Therefore, either by Proposition A.2.1 or A.3.1, Case 2.2 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 2.4

In Case 2.4, we require that the term being moved satisfies following conditions.

- The term is right before $-h_m$.
- The term is not the first term, $-h_1 - \lambda$.

Mathematically: $-h_1-\lambda > -h_2-\lambda > \cdots > -h_n-\lambda \geq -h_m \geq -h_{n+1}-\lambda > \cdots > -h_{m-1}-\lambda$, where $2 \leq n \leq m-1$, $m \geq 6$. After moving $-h_n - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ \vdots \\ h_{n-1} \\ h_m - \lambda \\ h_{n+1} \\ \vdots \\ h_{m-1} \\ h_n + \lambda \end{bmatrix}. \tag{2.18}$$

Since at least three points determine the structure of $\boldsymbol{h}_b$, the terms $h_1, h_2, \ldots, h_{n-1}, h_{n+1}, \ldots, h_{m-1}$ require that $\boldsymbol{h}_b$ has to have same structure as $\boldsymbol{h}_a$. Therefore, we will have $\lambda = h_m - h_n$. Substituting that into $\boldsymbol{S}_b$ and $\boldsymbol{h_n}$ will yield an identical distribution with identical parameters. Therefore, Case 2.4 does not work out.

Case 2.5

In Case 2.5, we require that the term being moved satisfies following conditions.

- The term is right before $-h_m$.
- The term is the first term, $-h_1 - \lambda$.

Mathematically: $-h_1 - \lambda \geq -h_m \geq -h_2 - \lambda > \cdots > -h_{m-1} - \lambda$, where $m \geq 6$. After moving $-h_1 - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_2 + \lambda + h_1 - h_m \\ \vdots \\ h_{m-1} + \lambda + h_1 - h_m \\ h_1 + \lambda \end{bmatrix}. \tag{2.19}$$

We have the terms $h_2 + \lambda + h_1 - h_m, \ldots, h_{m-1} + \lambda + h_1 - h_m$ as the original curve shifted upwards by $(\lambda + h_1 - h_m)$, which breaks the pattern shown in Figure 2.3. Therefore, either by Proposition A.3.1, Case 2.2 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 2 Conclusion

In Case 2, we considered moving an arbitrary term before $-h_m$ to the last position after the diagonal terms of $\boldsymbol{S}_b$ are ordered. There is no possible way to construct a valid non-identifiable candidate as Step 3 in Algorithm 2 will fail.

Case 3

In Case 3, similar to Case 2, we will consider moving an arbitrary term after $-h_m$ last.

Case 3.1

In Case 3.1, we require that the term being moved satisfies following conditions.

- $-h_m$ is not the largest diagonal terms.
- The term is not right after $-h_m$.

Mathematically: $-h_1 - \lambda > -h_2 - \lambda > \cdots > -h_n - \lambda \geq -h_m \geq -h_{n+1} - \lambda > \cdots > -h_j - \lambda > \cdots > -h_{m-1} - \lambda$, where $1 \leq n \leq m-4$, $n+2 \leq j \leq m-2$, $m \geq 6$. After moving $-h_j - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \\ h_m - \lambda \\ h_{n+1} \\ \vdots \\ h_{j-1} \\ h_{j+1} \\ \vdots \\ h_{m-1} \\ h_j + \lambda \end{bmatrix}. \tag{2.20}$$

Notice that $h_1, h_2, \ldots, h_n, h_{j+1}, \ldots, h_{m-1}$ lie on the original curve. However, $h_{n+1}, \ldots, h_{j-1}$ is shifted rightwards by one position from the original curve, because of an extra point $h_m - \lambda$. Therefore, by Proposition A.2.1, Case 3.1 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 3.2

In Case 3.2, we require that the term being moved satisfies following conditions.

- $-h_m$ is not the largest diagonal terms.
- The term is right after $-h_m$.

Mathematically: $-h_1-\lambda > -h_2-\lambda > \cdots > -h_n-\lambda \geq -h_m \geq -h_{n+1}-\lambda > \cdots > -h_{m-1}-\lambda$, where $1 \leq n \leq m-3$, $m \geq 6$. After moving $-h_{n+1}-\lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \\ h_m - \lambda \\ h_{n+2} \\ \vdots \\ h_{m-1} \\ h_{n+1} + \lambda \end{bmatrix}. \tag{2.21}$$

Since at least three points determine the structure of $\boldsymbol{h}_b$, the terms $h_1, h_2, \ldots, h_n, h_{n+2}$, $\ldots, h_{m-1}$, require that $\boldsymbol{h}_b$ has same structure as $\boldsymbol{h}_a$. We will then have $\lambda = h_m - h_{n+1}$. Substituting that into $\boldsymbol{S}_b$ will yield an identical distribution with identical parameters. Therefore, Case 3.2 does not work out.

Case 3.3

In Case 3.3, we require that the term being moved satisfies following conditions.

- $-h_m$ is the largest diagonal terms.

Mathematically: $-h_m \geq -h_1 - \lambda > -h_2 - \lambda > \cdots > -h_n - \lambda > \cdots > -h_{m-1} - \lambda$, where $1 \leq n \leq m-2$, $m \geq 6$. After moving $-h_n - \lambda$ last, we have

$$\boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_1 + (\lambda + h_1 - h_m) \\ \vdots \\ h_{n-1} + (\lambda + h_1 - h_m) \\ h_{n+1} + (\lambda + h_1 - h_m) \\ \cdots \\ h_{m-1} + (\lambda + h_1 - h_m) \\ h_n + \lambda \end{bmatrix}. \tag{2.22}$$

We have the terms $h_1+(\lambda+h_1-h_m), \ldots, h_{j-1}+(\lambda+h_1-h_m), h_{j+1}+(\lambda+h_1-h_m), \ldots, h_{m-1}+(\lambda+h_1-h_m)$ as the original curve shifted upwards by $(\lambda+h_1-h_m)$. Moreover, the missing point $h_j + (\lambda + h_1 - h_m)$ breaks the pattern shown in Figure 2.3. Therefore, either by Proposition A.2.1 or A.3.1, Case 3.3 cannot satisfy (2.4) and Step 3 in Algorithm 2 failed.

Case 3 Conclusion

In Case 3, we considered moving an arbitrary term after $-h_m$ to the last position from the ordered diagonal terms, and found that there is no possible way to construct a valid non-identifiable candidate as Step 3 in Algorithm 2 will fail.

Overall conclusion

By combining Cases 1, 2 and 3, we have shown that it is impossible to construct a non-identifiable candidate as $\boldsymbol{S}_b$ because the corresponding exit rates $\boldsymbol{h}_b$ fail to satisfy (2.4) which is a requirement in Step 3 in Algorithm 2. The proof is now complete. □

Because Step 3 in Algorithm 2 failed, we do not have to attempt to go to Step 4 to determine the parameters by moment matching. However, it is crucial to stress that the proof presented in this section works only when $m \geq 6$. In fact, when $m < 6$, it will be relatively easier to find a non-identifiable candidate whose dying rates satisfy constraint (2.4) because there are fewer elements in $\boldsymbol{h}$. In that case, several sub-cases in Cases 1, 2 or 3 will go through, which will give rise to non-identifiable candidates after verifying the moments in Step 4. This will be discussed later in Section 2.3.3.

After having provided a thorough proof in this section, we may now present Theorem 2.3.1 regarding the identifiability of the PTAM for fixed m and $m \geq 6$:

Theorem 2.3.1. *Consider the PTAM of order m with parameters $\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m)$. Given the value of m and $m \geq 6$, then the PTAM is identifiable.*

2.3.2 Identifiability of the PTAM for different m

In Section 2.3.1, we have determined the identifiability of the PTAM when m is fixed and $m \geq 6$. In this section, we will continue investigating the identifiability of the PTAM, treating the number of states m as one of the model parameters.

According to Rizk et al. (2021), the survival function of a Coxian distribution has the following form:

$$S_X(t) = \sum_{k=1}^{m} p_k S_{Y_k}(t), \tag{2.23}$$

where

$$p_k = \begin{cases} \left(\frac{h_1}{\lambda+h_1}\right), & k=1, \\ \left(\frac{\lambda}{\lambda+h_1}\right)\cdots\left(\frac{\lambda}{\lambda+h_{k-1}}\right)\left(\frac{h_k}{\lambda+h_k}\right), & k=2,\ldots,m-1, \\ \left(\frac{\lambda}{\lambda+h_1}\right)\ldots\left(\frac{\lambda}{\lambda+h_{m-1}}\right), & k=m, \end{cases} \tag{2.24}$$

$$Y_k = Z_1 + \cdots + Z_k, \tag{2.25}$$

$$Z_j \sim exp(\lambda + h_j),\ j = 1, 2, \ldots, m-1, \tag{2.26}$$

$$Z_m \sim exp(h_m). \tag{2.27}$$

Each Y_k follows a general Erlang distribution. Moreover, it is a classical result that Y_k can further be expressed as a mixture of exponential distributions. That is,

$$S_{Y_k}(t) = \sum_{i=1}^{k} q_i(k) S_{Z_i}(t), \tag{2.28}$$

where $q_1(1) = 1$, with

$$q_i(k) = \prod_{\substack{j=1 \\ j\neq i}}^{k} \frac{\lambda + h_j}{h_j - h_i},\ i = 1, 2, \ldots, k, \tag{2.29}$$

$$\tag{2.30}$$

for $1 < k < m$, and

$$q_i(m) = \begin{cases} \left(\frac{h_m}{h_m - h_i - \lambda}\right) \prod_{\substack{j=1 \\ j\neq i}}^{m-1} \frac{\lambda+h_j}{h_j - h_i}, & i = 1, 2, \ldots, m-1, \\ \prod_{j=1}^{m-1} \frac{\lambda+h_j}{\lambda+h_j-h_m}, & i = m, \end{cases} \tag{2.31}$$

for $k = m$.

In other words, there are two layers of mixtures. We can combine these two layers and express the survival function as a final mixture of exponential distributions:

$$S_X(t) = \sum_{k=1}^{m} p_k S_{Y_k}(t) = \sum_{k=1}^{m} p_k \sum_{i=1}^{k} q_i(k) S_{Z_i}(t) = \sum_{i=1}^{m} \tilde{p}_i S_{Z_i}, \tag{2.32}$$

where

$$\tilde{p}_i = \sum_{k=i}^{m} p_k q_i(k),\ i = 1, 2, \ldots, m. \tag{2.33}$$

In the context of the PTAM, since it is possible to have $\lambda + h_i = h_m$ for certain values of λ and i, (2.32) can have $m-1$ summands as the two summands with the same exponential terms can be grouped together.

Lemma 2.3.2. *Consider two Coxian distributions with no probability mass at zero. Let their orders be m_a and m_b and let their transition intensity matrices be $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$, respectively. Suppose that the diagonal terms of $-\boldsymbol{S}_a$ have n_a distinct elements and are denoted by $\lambda_1^{(a)} < \lambda_2^{(a)} < \cdots < \lambda_{n_a}^{(a)}$. The diagonal terms of the matrix for $-\boldsymbol{S}_b$ have n_b distinct elements and are denoted by $\lambda_1^{(b)} < \lambda_2^{(b)} < \cdots < \lambda_{n_b}^{(b)}$. We have that if $n_a \neq n_b$, then $f_{S_a}(t) \neq f_{S_b}(t)$.*

Proof. In light of (2.32) in conjunction with grouping the terms having the same exponent, we can write

$$S_{\boldsymbol{S}_a}(t) = \sum_{i=1}^{n_a} \tilde{p}_{i,n_a} S_{Z_i^{(a)}}(t) = \sum_{i=1}^{n_a} \tilde{p}_{i,n_a} e^{-\lambda_i^{(a)} t},$$

where $Z_i^{(a)}$ is exponential with rate $\lambda_i^{(a)}$ and $\tilde{p}_{i,n_a}$ is the mixing probability corresponding to $S_{Z_i^{(a)}}(t)$; and

$$S_{\boldsymbol{S}_b}(t) = \sum_{i=1}^{n_b} \tilde{p}_{i,n_b} S_{Z_i^{(b)}}(t) = \sum_{i=1}^{n_b} \tilde{p}_{i,n_b} e^{-\lambda_i^{(b)} t},$$

where $Z_i^{(b)}$ is exponential with rate $\lambda_i^{(b)}$ and $\tilde{p}_{i,n_b}$ is the mixing probability corresponding to $S_{Z_i^{(b)}}(t)$.

- If $\lambda_1^{(a)} > \lambda_1^{(b)}$, then

$$\begin{aligned}
&\lim_{t\to\infty} \frac{S_{\boldsymbol{S}_a}(t)}{S_{\boldsymbol{S}_b}(t)} \\
&= \lim_{t\to\infty} \frac{S_{\boldsymbol{S}_a}(t) e^{\lambda_1^{(b)} t}}{S_{\boldsymbol{S}_b}(t) e^{\lambda_1^{(b)} t}} \\
&= \lim_{t\to\infty} \frac{\sum_{i=1}^{n_a} \tilde{p}_{i,n_a} e^{-(\lambda_i^{(a)} - \lambda_1^{(b)})t}}{\tilde{p}_{1,n_b} + \sum_{i=2}^{n_b} \tilde{p}_{i,n_b} e^{-(\lambda_i^{(b)} - \lambda_1^{(b)})t}} \\
&= 0 \\
&\neq 1.
\end{aligned}$$

  Therefore, $S_{\boldsymbol{S}_a}(t) \neq S_{\boldsymbol{S}_b}(t)$.

- If $\lambda_1^{(a)} < \lambda_1^{(b)}$, then

$$\begin{aligned}
&\lim_{t\to\infty} \frac{S_{\boldsymbol{S}_a}(t)}{S_{\boldsymbol{S}_b}(t)} \\
&= \lim_{t\to\infty} \frac{S_{\boldsymbol{S}_a}(t)e^{\lambda_1^{(a)}t}}{S_{\boldsymbol{S}_b}(t)e^{\lambda_1^{(a)}t}} \\
&= \lim_{t\to\infty} \frac{\tilde{p}_{1,n_a} + \sum_{i=2}^{n_a} \tilde{p}_{i,n_a} e^{-(\lambda_i^{(a)} - \lambda_1^{(a)})t}}{\sum_{i=1}^{m_b} \tilde{p}_{i,n_b} e^{-(\lambda_i^{(b)} - \lambda_1^{(a)})t}} \\
&= \pm\infty \\
&\neq 1.
\end{aligned}$$

  Therefore, $S_{\boldsymbol{S}_a}(t) \neq S_{\boldsymbol{S}_b}(t)$.

- If $\lambda_1^{(a)} = \lambda_1^{(b)}$ and $\tilde{p}_{1,n_a} \neq \tilde{p}_{1,n_b}$, then

$$\lim_{t\to\infty} \frac{S_{\boldsymbol{S}_a}(t)}{S_{\boldsymbol{S}_b}(t)} = \frac{\tilde{p}_{1,n_a} + \sum_{i=2}^{n_a} \tilde{p}_{i,n_a} e^{-(\lambda_i^{(a)} - \lambda_1^{(a)})t}}{\tilde{p}_{1,n_b} + \sum_{i=2}^{n_b} \tilde{p}_{i,n_b} e^{-(\lambda_i^{(b)} - \lambda_1^{(b)})t}} = \frac{\tilde{p}_{1,n_a}}{\tilde{p}_{1,n_b}} \neq 1.$$

  Therefore, $S_{\boldsymbol{S}_a}(t) \neq S_{\boldsymbol{S}_b}(t)$.

- If $\lambda_1^{(a)} = \lambda_1^{(b)}$ and $\tilde{p}_{1,n_a} = \tilde{p}_{1,n_b}$, then we continue to compare the remaining terms. That is, $S_{\boldsymbol{S}_a}(t) - \tilde{p}_{1,n_a} S_{Z_1^{(a)}}(t)$ and $S_{\boldsymbol{S}_b}(t) - \tilde{p}_{1,n_b} S_{Z_1^{(b)}}(t)$. To achieve this, we compare $\lambda_2^{(a)}$ and $\lambda_2^{(b)}$ in the similar fashion as above.

Continuing this process, we see that in order for $S_{\boldsymbol{S}_b}(t) = S_{\boldsymbol{S}_b}(t)$, all the exponential rates and the corresponding coefficients must be the same. However, this is not possible because $n_a \neq n_b$. □

Theorem 2.3.3. *Given two PTAMs with p.d.f.'s $f_{\boldsymbol{S}_a}(t)$ and $f_{\boldsymbol{S}_b}(t)$, if $m_a \neq m_b$, then $f_{\boldsymbol{S}_a}(t) \neq f_{\boldsymbol{S}_b}(t)$.*

Proof.

- If the diagonal terms of $-\boldsymbol{S}_a$ and $-\boldsymbol{S}_b$ are all distinct, then $n_a = m_a$ and $n_b = m_b$. Since $n_a \neq n_b$, then the result immediately follows by applying Lemma 2.3.2.

- If the diagonal terms of $-\boldsymbol{S}_a$ and $-\boldsymbol{S}_b$ are not all distinct for both $-\boldsymbol{S}_a$ and $-\boldsymbol{S}_b$, then we must have $\lambda_a + h_{ia} = h_{ma}$ and $\lambda_b + h_{ib} = h_{mb}$ for certain values of λ_a, λ_b and i. In that case, $n_a = m_a - 1$ and $n_b = m_b - 1$. Since $n_a \neq n_b$, then the result still follows by applying Lemma 2.3.2.

- If for instance the diagonal terms of $-\boldsymbol{S}_a$ and $-\boldsymbol{S}_b$ are not all distinct, then we must have $n_a = m_a - 1$ and $n_b = m_b$.

– If $m_b \neq m_a - 1$, then $n_a \neq n_b$, and the result still follows by applying Lemma 2.3.2.

– If $m_b = m_a - 1$, then $n_a = n_b$. In that case, the representations of $f_{\boldsymbol{S}_a}(t)$ and $f_{\boldsymbol{S}_b}(t)$ will consist of same number of exponential terms with corresponding rates as presented in (2.34) and (2.35):

$$h_{1a} + \lambda_a < h_{2a} + \lambda_a < \cdots < h_{ia} + \lambda_a = h_m < \cdots < h_{m_a-1} + \lambda_a \tag{2.34}$$

and

$$h_{1b} + \lambda_b < h_{2b} + \lambda_b < \cdots < h_{m_b} < \cdots < h_{m_b-1} + \lambda_b. \tag{2.35}$$

Observe that (2.34) exhibits a smooth pattern as the dying rates $h_{1a}, h_{2a}, \ldots, h_{m_a-1}$ are simply shifted up by λ_a. However, this observation does not hold true for (2.35) as the term h_{m_b} breaks the smoothness of the pattern. Since having the same pattern is a necessary condition for $f_{\boldsymbol{S}_a}(t)$ and $f_{\boldsymbol{S}_b}(t)$ to have the same exponential rates, it is therefore not possible to have the same exponential rates. Then, we can still conclude that $f_{\boldsymbol{S}_a}(t) \neq f_{\boldsymbol{S}_b}(t)$.

Therefore, based on all the possible cases discussed above, it is not possible to have $f_{\boldsymbol{S}_a}(t) = f_{\boldsymbol{S}_b}(t)$ if $m_a \neq m_b$.

□

Corollary 2.3.4 can be obtained by taking the contrapositive statement of Theorem 2.3.3.

Corollary 2.3.4. *Given two PTAMs with p.d.f.'s $f_{\boldsymbol{S}_a}(t)$ and $f_{\boldsymbol{S}_b}(t)$, if $f_{\boldsymbol{S}_a}(t) \equiv f_{\boldsymbol{S}_b}(t)$, then $m_a = m_b$.*

We may now present Theorem 2.3.5 regarding the identifiability of the PTAM.

Theorem 2.3.5. *Consider the PTAM of order m with parameters $\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m)$, then the PTAM is identifiable when $m \geq 6$.*

Proof. By Theorem 2.3.1, the PTAM is identifiable for fixed m and $m \geq 6$. Therefore, if we still want to construct any non-identifiable candidate, we must relax the condition that m is fixed. If we do it, we need to determine if there exists an $\boldsymbol{S}_b$ that qualifies for different m. However, by Corollary 2.3.4, it is not possible for two p.d.f.-equivalent PTAMs to have different values of m. Therefore, we may conclude that it is impossible to find an $\boldsymbol{S}_b$, even the condition that m is fixed is relaxed. □

2.3.3 Non-identifiability of the PTAM when $m < 6$

In Sections 2.3.1 and 2.3.2, we have proved that the PTAM is identifiable when $m \geq 6$. In this section, we will explain why the PTAM is non-identifiable for $m < 6$. Moreover, we will provide detailed illustrative examples regarding non-identifiable candidates constructed by applying Algorithm 2.

The case $m = 2$

When $m = 2$, the parameter s does not even need to be specified. We have

$$\boldsymbol{S}_a = \begin{bmatrix} -(h_1 + \lambda) & \lambda \\ 0 & -h_2 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}. \tag{2.36}$$

According to Algorithm 2, we can construct

$$\boldsymbol{S}_b = \begin{bmatrix} -h_2 & (h_2 - h_1) \\ 0 & -(h_1 + \lambda) \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_1 + \lambda \end{bmatrix}. \tag{2.37}$$

Since the parameter s is not specified, (2.4) is automatically satisfied. Since there are no other parameters to be determined based on the moment matching condition, we are done. Therefore, for $m = 2$, the PTAM is not identifiable, with a non-identifiability construction as specified above.

It is straightforward to see that in view of this construction, one can create countless non-identifiable candidates as long as $0 \leq h_1 \leq h_2$ and $\lambda > 0$.

The case $m = 3$

We have

$$\boldsymbol{S}_a = \begin{bmatrix} -(h_1 + \lambda) & \lambda & 0 \\ 0 & -(h_2 + \lambda) & \lambda \\ 0 & 0 & -h_3 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}. \tag{2.38}$$

According to Algorithm 2, we can construct $3! - 1 = 5$ possible candidates for $\boldsymbol{S}_b$ by permuting the diagonal elements of $\boldsymbol{S}_a$. One of the permutations yields

$$\boldsymbol{S}_b = \begin{bmatrix} -h_3 & (h_3 - h_1) & 0 \\ 0 & -(h_1 + \lambda) & (h_3 - h_1) \\ 0 & 0 & -(h_2 + \lambda) \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} h_1 \\ 2h_1 + \lambda - h_3 \\ h_2 + \lambda \end{bmatrix}. \tag{2.39}$$

Since $\boldsymbol{h}$ has three elements, the parameter s will be specified as one of the parameters of the PTAM. However, it still does not play its role in terms of parameter constraint, because given three elements, we can always find a corresponding s. Thus, (2.4) is automatically satisfied. In addition, we need $\lambda > h_3 - h_1$ to maintain the increasing pattern of $\boldsymbol{h}_b$.

We now move to Step 4 of Algorithm 2. We must verify whether the moments are matched, given the parameter constraints of the PTAM. That is, we would like to check

whether

$$\boldsymbol{\pi}_a' \boldsymbol{S}_a{}^{-1} \boldsymbol{e} = \boldsymbol{\pi}_b' \boldsymbol{S}_b{}^{-1} \boldsymbol{e}, \tag{2.40}$$

where $\boldsymbol{\pi}_a' = \boldsymbol{\pi}_b' = (1, 0, \ldots, 0)$. It turns out that only when $\lambda = \frac{(h_3 - h_1)^2}{(h_3 - h_2)}$ can Step 4 be satisfied. Thus, as long as $\lambda = \frac{(h_3 - h_1)^2}{(h_3 - h_2)}$, countless non-identifiable candidates can also be obtained via (2.39).

To verify this, we now present a detailed example following this construction. Consider the PTAM with $\boldsymbol{\theta}_a = (\lambda_a, h_{1a}, h_{ma}, s_a, m_a) = (6.243918, 0.6, 2.8, 2, 3)$, which gives

$$\boldsymbol{S_a} = \begin{bmatrix} -6.843918 & 6.243918 & 0 \\ 0 & -8.268764 & 6.243918 \\ 0 & 0 & -2.8 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} 0.6 \\ 2.024846 \\ 2.8 \end{bmatrix}$$

and another model with $\boldsymbol{\theta}_b = (\lambda_b, h_{1b}, h_{mb}, s_b, m_b) = (2.2, 0.6, 8.268764, 1.109574, 3)$, which yields

$$\boldsymbol{S_b} = \begin{bmatrix} -2.8 & 2.2 & 0 \\ 0 & -6.843918 & 2.2 \\ 0 & 0 & -8.268764 \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} 0.6 \\ 4.643918 \\ 8.268764 \end{bmatrix}.$$

Notice that $h_{1a} = h_{1b} = 0.6$ and $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ have permuted diagonal terms, which is in line with Rizk et al. (2019). What is more, $\lambda_a = \frac{(h_{3a} - h_{1a})^2}{(h_{3a} - h_{2a})} = 6.243918$. It can be verified graphically that the above two PTAMs have equivalent p.d.f.'s.

The case $m = 4$

We have

$$\boldsymbol{S}_a = \begin{bmatrix} -(h_1 + \lambda) & \lambda & 0 & 0 \\ 0 & -(h_2 + \lambda) & \lambda & 0 \\ 0 & 0 & -(h_3 + \lambda) & \lambda \\ 0 & 0 & 0 & -h_4 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}. \tag{2.41}$$

According to Algorithm 2, we can construct $4! - 1 = 23$ possible candidates for $\boldsymbol{S}_b$ by permuting diagonal terms of $\boldsymbol{S}_a$. One of the permutations yields

$$\boldsymbol{S}_b = \begin{bmatrix} -(h_2+\lambda) & h_2+\lambda-h_1 & 0 & 0 \\ 0 & -h_4 & h_2+\lambda-h_1 & 0 \\ 0 & 0 & -(h_3+\lambda) & h_2+\lambda-h_1 \\ 0 & 0 & 0 & -(h_1+\lambda) \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} h_1 \\ h_1+h_4-h_2-\lambda \\ h_1+h_3-h_2 \\ h_1+\lambda \end{bmatrix}. \tag{2.42}$$

Since $\boldsymbol{h}$ has more than three elements, the parameter s starts to play its role as (2.4) in terms of parameter constraint. Therefore, (h_1, h_4, λ, s) has to be carefully chosen to ensure that $\boldsymbol{h}_b$ can be described by a suitable s^*. That is,

$$h_1 + h_4 - h_2 - \lambda = \left[\frac{2}{3}h_1^{s*} + \frac{1}{3}(h_1+\lambda)^{s*}\right]^{\frac{1}{s*}}, \tag{2.43}$$

$$h_1 + h_3 - h_2 = \left[\frac{1}{3}h_1^{s*} + \frac{2}{3}(h_1+\lambda)^{s*}\right]^{\frac{1}{s*}}. \tag{2.44}$$

Moreover, we need to have

$$\max(h_3 - h_2, \ h_4 - h_3) < \lambda < h_4 - h_2 \tag{2.45}$$

in order to maintain the increasing pattern of $\boldsymbol{h}_b$. Finally, as before we would like to test whether the moments are matched, that is,

$$\boldsymbol{\pi}_a' \boldsymbol{S}_a{}^{-k} \boldsymbol{e} = \boldsymbol{\pi}_b' \boldsymbol{S}_b{}^{-k} \boldsymbol{e}, \text{ where } k = 1, 2, 3, \ldots. \tag{2.46}$$

Solving for (h_1, h_4, λ, s) that satisfy the constraints (2.43)–(2.46) is quite involved. A numerical method needs to be considered as there are transcendental equations involved. Accordingly, a tolerance level ϵ for the method has to be specified.

We now present a detailed example following this construction. Consider the PTAM with

$$\boldsymbol{\theta}_a = (\lambda_a, h_{1a}, h_{ma}, s_a, m_a) = (0.05873815, 1.883377, 2.020888, 3.885586, 4),$$

which yields

$$\boldsymbol{S_a} = \begin{bmatrix} -1.942115 & 0.05873815 & 0 & 0 \\ 0 & -1.991135 & 0.05873815 & 0 \\ 0 & 0 & -2.036808 & 0.05873815 \\ 0 & 0 & 0 & -2.020888 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} 1.883377 \\ 1.932397 \\ 1.978070 \\ 2.020888 \end{bmatrix}$$

and another model with

$$\boldsymbol{\theta}_b = (\lambda_b, h_{1b}, h_{mb}, s_b, m_b) = (0.1077578, 1.883377, 1.942115, 37.60599, 4),$$

which yields

$$\boldsymbol{S_b} = \begin{bmatrix} -1.991135 & 0.1077578 & 0 & 0 \\ 0 & -2.020888 & 0.1077578 & 0 \\ 0 & 0 & -2.036808 & 0.1077578 \\ 0 & 0 & 0 & -1.942115 \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} 1.883377 \\ 1.913130 \\ 1.929050 \\ 1.942115 \end{bmatrix}.$$

Similarly, notice that $h_{1a} = h_{1b} = 1.883377$ and $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ have permuted diagonal terms, which is in line with Rizk et al. (2019). What is more, $\boldsymbol{h}_a$ and $\boldsymbol{h}_b$ follow our PTAM structure and can be described by s_a and s_b, respectively. It can be verified graphically that the above two PTAMs have equivalent p.d.f.'s with a tolerance level $\epsilon < 0.00001$ in the numerical method. In addition, it can be verified that this example corresponds to Case 2.2, the difference being that Algorithm 2 went through.

The case $m = 5$

The non-identifiability construction for $m = 5$ is completely analogous to the case where $m = 4$. Thus, we will directly provide a detailed example. Consider the PTAM with

$$\boldsymbol{\theta}_a = (\lambda_a, h_{1a}, h_{ma}, s_a, m_a) = (0.523579, 0.0004428392, 0.5240306, -21.15625, 5),$$

which yields

$$\boldsymbol{S_a} = \begin{bmatrix} -0.5240218 & 0.523579 & 0 & 0 & 0 \\ 0 & -0.5240279 & 0.523579 & 0 & 0 \\ 0 & 0 & -0.5240366 & 0.523579 & 0 \\ 0 & 0 & 0 & -0.5240518 & 0.523579 \\ 0 & 0 & 0 & 0 & -0.5240306 \end{bmatrix}, \ \boldsymbol{h}_a = \begin{bmatrix} 0.0004428392 \\ 0.000448902 \\ 0.0004575883 \\ 0.0004728287 \\ 0.5240306 \end{bmatrix}$$

and another model with

$$\boldsymbol{\theta}_b = (\lambda_b, h_{1b}, h_{mb}, s_b, m_b) = (0.5235851, 0.0004428392, 0.5240218392, -26.34487, 5),$$

which yields

$$\boldsymbol{S_b} = \begin{bmatrix} -0.5240279 & 0.5235851 & 0 & 0 & 0 \\ 0 & -0.5240306 & 0.5235851 & 0 & 0 \\ 0 & 0 & -0.5240366 & 0.5235851 & 0 \\ 0 & 0 & 0 & -0.5240518 & 0.5235851 \\ 0 & 0 & 0 & 0 & -0.5240218 \end{bmatrix}, \ \boldsymbol{h}_b = \begin{bmatrix} 0.0004428392 \\ 0.0004455372 \\ 0.0004515255 \\ 0.0004667659 \\ 0.5240218392 \end{bmatrix}.$$

It can be verified graphically that the above two PTAMs have equivalent p.d.f.'s with a tolerance level $\epsilon < 0.00001$ in the numerical method. In addition, it can be verified that this example corresponds to Case 2.2, difference being that Algorithm 2 went through.

2.4 Discussion

It can be seen that the identifiability of the PTAM improves as m increases. When $m = 2$, countless non-identifiability constructions can be achieved. When $m = 3$, countless non-identifiability constructions can be achieved subject to certain constraints on λ. When $m = 4$ or $m = 5$, the non-identifiability construction becomes more difficult as more constraints are involved. In this instance, although non-identifiable candidates can be established in theory, it does not affect much the reliability of parameters as the parameters are still rather close. All of these conclusions are consistent with intuition: as the parameter constraints are increasingly strong as m increases, the model will move closer and closer towards being identifiable. Until when $m \geq 6$, the model will then become identifiable.

Last but not least, the threshold $m = 6$ exactly explains the balance between the two forces discussed earlier in this chapter: the non-identifiability concern brought about by the freedom of the last state and the identifiability improvement associated with the parameter constraints in the PTAM. It is now clear that when $2 \leq m < 6$, the former dominates whereas the latter dominates when $m \geq 6$.

2.5 Conclusion

In this chapter, we have thoroughly investigated the identifiability of the PTAM. The PTAM is identifiable when $m \geq 6$, but possibly non-identifiable when $2 \leq m < 6$. However, identifiability is only the first topic being investigated since other noisy measurements will hinder the reliability of parameter estimation when it comes to practical implementations. Such issues rest with the concept of estimability which will be discussed in Chapter 3.

Chapter 3

An Objective Measure of Estimability for Statistical Models

In this chapter, a novel definition of estimability is proposed in order to objectively quantify estimability in the context of statistical models. More specifically, this objectivity is achieved via a carefully designed c.d.f. sensitivity measure, under which the threshold will be tailored to the empirical c.d.f. and therefore become an experiment-based quantity. The proposed definition which is validated to be innately sound, will then be applied to assess and improve the estimability of the PTAM.

3.1 Motivation

In Chapter 2, we have established the identifiability of the PTAM for $m \geq 6$, where m denotes the number of states. We also provided illustrative examples of non-identifiable PTAM for $2 \leq m < 6$ to clarify the concept. The identifiability, or mathematical uniqueness, guarantees that no other c.d.f.-equivalent representations of the model exists and that the likelihood function has a unique global maximum.

However, identifiability does not imply estimability. Although the model representation is unique, parameter estimates can still be unreliable when the profile likelihood functions are extremely flat, as a large range of different estimates can produce nearly the same likelihood.[Raue et al. (2009)]. According to Cheng (2021), the profile likelihood functions of h_1, h_m and s of the PTAM turn out to be flat, which gives rise to this estimability issue.

3.2 Literature review on estimability

3.2.1 Methods for estimability assessment

Estimability is also referred to as practical identifiability in the literature. The non-estimability issue will arise in the case of flat likelihood functions [Raue et al. (2009)], or equivalently,

the insensitivity of the model c.d.f. with respect to its parameters [McLean and McAuley (2012)], which may be due to the following two reasons:

(i) The model c.d.f. is insensitive with respect to parameter changes. Accordingly, this aspect involves the model sensitivity.

(ii) The effect of one parameter on the c.d.f. might be offset by that of one of the other parameters. This is defined as parameter correlation.

If the model is identifiable, then unreliable parameter estimates, if any, will be caused by non-estimability issues that may be due to experimental error including data quality (too few or too noisy), algorithm approximation or other noisy measurements [Miao et al. (2011)]. Unlike identifiability, estimability is less well-defined and its characterization has remained an open problem. While it is straightforward to think qualitatively that parameters can be "loosely estimated" under noisy measurements, one would need to define quantitatively what this really means [Raue et al. (2009); Gontier and Pfister (2020)].

Estimability has been widely studied in system biology where an ODE model is utilized to model dynamic biological systems. For instance, it is assumed that

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = f(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}), \tag{3.1}$$

$$\boldsymbol{y}(t) = h(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) + \boldsymbol{\epsilon}(t), \tag{3.2}$$

where

$\boldsymbol{x}(t) \in \mathbb{R}^m$ is a vector of state variables,
$\boldsymbol{y}(t) \in \mathbb{R}^d$ is the output vector,
$\boldsymbol{u}(t) \in \mathbb{R}^p$ is the known system input vector,
$\boldsymbol{\theta} \in \mathbb{R}^q$ is the parameter vector,
$\boldsymbol{\epsilon}(t) \sim N(\mathbf{0}, \sigma^2(t)\boldsymbol{I_d})$ is the measurement noise.

There are four broad types of methodologies for investigating estimability. We briefly describe them next.

The Monte Carlo method

The first method involves repeated parameter estimation from a large number of data sets simulated by the Monte Carlo method. To apply this method, threshold values for parameter uncertainty levels are required to distinguish estimable and non-estimable parameters. Let $\boldsymbol{\theta_0}$ be the nominal parameters obtained from either fitting the model to original data or from prior knowledge [Miao et al. (2011)]. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ be the parameter estimates at the i^{th} trial which are based on data simulated from the model having $\boldsymbol{\theta_0}$ as its parameters. Then, the average

relative estimation error (ARE) associated with $\hat{\theta}_i^{(k)}$, the k^{th} element of $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$, is defined as

$$ARE_k := \frac{1}{N}\sum_{i=1}^{N}\frac{|\theta_0^{(k)} - \hat{\theta}_i^{(k)}|}{|\theta_0^{(k)}|}, \tag{3.3}$$

where $\theta_0^{(k)}$ is the k^{th} element $\boldsymbol{\theta_0}$ and $k = 1, 2, \ldots, q$. Then, Miao et al. (2011) defined non-estimability as occurring when the ARE of a parameter is sufficiently high, or equivalently, exceeds a pre-selected threshold Δ.

Methods based on the correlation matrix or the Fisher information matrix

According to Petersen et al. (2001), the Fisher information matrix (FIM) associated with the ODE model is given by

$$FIM = \sum_{i=1}^{N}\left(\frac{\partial\hat{\boldsymbol{y}}_i}{\partial\hat{\boldsymbol{\theta}}}\right)^T \boldsymbol{V}^{-1}\left(\frac{\partial\hat{\boldsymbol{y}}_i}{\partial\hat{\boldsymbol{\theta}}}\right), \tag{3.4}$$

where

$\left(\frac{\partial\hat{\boldsymbol{y}}_i}{\partial\hat{\boldsymbol{\theta}}}\right)$ is defined as the sensitivity matrix, and

$\boldsymbol{V}$ is a known positive definite matrix of weights on the variances.

Rodriguez-Fernandez et al. (2006) proposed a correlation matrix approach for analyzing the estimability of the ODE model. By the Cramér-Rao Theorem, the covariance matrix can be obtained as

$$\boldsymbol{C} \approx FIM^{-1}, \tag{3.5}$$

the correlation between θ_i and θ_j being

$$r_{ij} = \frac{\boldsymbol{C}_{ij}}{\sqrt{\boldsymbol{C}_{ii}\boldsymbol{C}_{jj}}},\ i \neq j, 1 \leq i, j \leq q.$$

Similarly, Quaiser and Mönnigmann (2009) proposed a total correlation measure, and θ_i and θ_j are deemed non-estimable if their correlation is sufficiently high, or equivalently, exceeds a certain threshold Δ.

There exist several methods focusing on the FIM. Dochain and Vanrolleghem (2001) proposed that the condition number, which is defined as the ratio of the largest eigenvalue to the smallest eigenvalue of the FIM, can also be used to assess estimability. The larger the condition number, the more correlated the parameters, and the less estimable the parameters will be. The model will then be non-estimable if the condition number is sufficiently high, or equivalently, exceeds a certain threshold Δ. In addition, Brun et al. (2001) proposed

a collinearity index to measure the parameter correlations. The model will be more non-estimable if the collinearity index is relatively large. A threshold Δ is also needed in that case.

Methods based on the model sensitivity

The third approach is based on the model sensitivity. As seen from (3.4), the FIM is obtained in terms of the sensitivity matrix. Thus, the sensitivity matrix may be extracted from the FIM and analyzed specifically. The sensitivity matrix $\boldsymbol{S}$ with observation times $(t_1, \ldots, t_N)$ is defined as

$$\boldsymbol{S}_{dN\times q}(t_1, \cdots, t_N) := \begin{bmatrix} \frac{\partial y_1(t_1;\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial y_1(t_1;\boldsymbol{\theta})}{\partial \theta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_d(t_1;\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial y_d(t_1;\boldsymbol{\theta})}{\partial \theta_q} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \frac{\partial y_1(t_N;\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial y_1(t_N;\boldsymbol{\theta})}{\partial \theta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_d(t_N;\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial y_d(t_N;\boldsymbol{\theta})}{\partial \theta_q} \end{bmatrix}.$$

Several methods are based on the sensitivity matrix. Jacquez and Greif (1985) calculated the sample correlation between the matrix columns. If $\rho(\boldsymbol{S}_{\cdot i}, \boldsymbol{S}_{\cdot j})$ is close to one within a threshold Δ, then θ_i and θ_j are considered non-estimable. Other methods exist such as the principal components analysis (PCA) technique [Degenring et al. (2004)], the orthogonal method [Yao et al. (2003)] and the eigenvalue method [Vajda et al. (1989)]. They all rely on a subjective threshold Δ.

If the model is sufficiently simplified and involves fewer parameters, the sensitivity function $\frac{\partial \boldsymbol{S}(t)}{\partial \boldsymbol{\theta}}$ can be solved analytically, in which case the sensitivity matrix is not needed. Holmberg (1982) proposed a visual inspection approach on the sensitivity function. The larger the sensitivity measure of one parameter, the greater the change in the model c.d.f. with respect to the change of that parameter. If the sensitivity functions of certain parameters are linearly dependent, then those parameters are functionally related. The drawback of this approach is that correlation cannot be quantified based on graphs. Moreover, subjective assessment is needed when resorting to visual inspection. Determining whether the graphs of the sensitivity functions are linearly dependent will depend on the experimenter's assessment.

Methods based on profile likelihood

Raue et al. (2009) proposed an explicit definition of estimability that is based on the profile likelihood function. They defined the profile likelihood confidence interval for parameter θ_i

as

$$C_{i,\Delta} := \{\theta_i | PL(\boldsymbol{x}; \hat{\theta}_i) - PL(\boldsymbol{x}; \theta_i) < \Delta\}, \tag{3.6}$$

where $PL(\cdot)$ is the profile likelihood function of θ_i and Δ is a subjectively chosen threshold. Then, θ_i is said to be non-estimable if $C_{i,\Delta}$ is infinite. In other words, given a certain threshold Δ, there exists a $\delta_i > 0$ such that for all $|\theta_i| > \delta_i$, $PL(\boldsymbol{x}; \hat{\theta}_i) - PL(\boldsymbol{x}; \theta_i) < \Delta$ holds true. This definition is mathematically clear as it relies on a binary event: whether $C_{i,\Delta}$ is infinite or not. However, a subjective threshold Δ is still required as in the case of other methods.

3.2.2 Relationships between identifiability, estimability and sensitivity

Informally, sensitivity refers to the degree to which a model will be affected by its parameter values. Graphically, the more sensitive a model is with respect to one parameter, the more noticeably the model's c.d.f. will be affected by changes in that parameter. In this chapter, we quantify this concept by the c.d.f. sensitivity measure to be defined in (3.7).

The concept of sensitivity can bridge identifiability and estimability. From the perspective of sensitivity, if a statistical model $f(x; \boldsymbol{\theta})$ is non-identifiable with the non-identifiable set being

$$\boldsymbol{\mathcal{A}} := \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \boldsymbol{\Theta} | \forall x, f(x; \boldsymbol{\theta_1}) = f(x; \boldsymbol{\theta_2})\},$$

then f will have zero sensitivity if the parameter changes within $\mathcal{A}$ since the model c.d.f. will not change. Thus, the non-identifiability issue cannot be overcome by improving the experimental design because the model output will be the same for all x.

On the other hand, if the sensitivity is not zero, then different parameters will produce different model c.d.f.'s. In that case, different x will produce different model outputs, and the experiment will contribute to providing some information towards parameter inference. The more sensitive the model with respect to one parameter, the more estimable that parameter will be. Although, as previously mentioned, different measures such as those based on correlations, condition numbers and eigenvalues may be employed, in each case a threshold in connection with sensitivity must be set.

It should be emphasized that identifiability and estimability are equally important. If a model is non-identifiable, then the likelihood function will have multiple global maxima. In that case, although we are certain that a numerical algorithm is based on maximizing the likelihood function, we may question the reliability of the parameter estimates produced by that algorithm since other maxima may potentially exist. In the case of estimability, although we are certain that the MLEs are unique, we may as well question the reliability of the parameter estimates produced by that algorithm since a wide variety of estimates can produce nearly the same likelihood values.

3.3 Proposed methodology

Based on a thorough review of the literature, very little research on estimability appears to have been conducted in connection with statistical models. To the best of our knowledge, there is only one paper, namely Gontier and Pfister (2020), that studies estimability on a statistical model, wherein a new definition of estimability based on a model selection perspective is proposed. One of their contributions is the elimination of the subjective threshold in (3.6) by introducing a Bayes factor into the new definition. Our contribution also aims at addressing the problem of having to set a subjective threshold. Rather than eliminating it, we shall make the threshold an objective, experiment-based quantity.

In order to do so, a methodology needs to be established, which relates the confidence region to the experimental protocol. To achieve this, we rely on following two considerations:

(i) The curvature of the likelihood function reflects the sensitivity of the model (c.d.f.) with respect to the parameters [McLean and McAuley (2012)].

(ii) The non-estimability issue is defined as occurring when the likelihood-based confidence interval is infinite [Raue et al. (2009)].

With respect to the first consideration, we replace the likelihood-based confidence region in Raue et al. (2009) with an innovative confidence region that based on a carefully designed c.d.f. sensitivity measure. The purpose of defining such a c.d.f. sensitivity measure is to relate the confidence region to the experimental error by quantifying them into single numbers. Then, by comparing the quantified numbers under the same measure, one may indirectly achieve the comparison between the confidence region and the experimental error. In that case, the threshold will be tailored to the experimental protocol and then, become objective. Additionally, in light of the second consideration, we define the non-estimability issue as occurring when this innovative confidence region is infinite.

Several preliminary definitions are needed before defining estimability. They are presented as Definitions 3.3.1–3.3.8.

Definition 3.3.1. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space and* $F(x;\boldsymbol{\theta})$*, the associated c.d.f. Then, for* $\boldsymbol{\theta_1},\ \boldsymbol{\theta_2} \in \boldsymbol{\Theta}$*, the c.d.f. sensitivity between* $f(x;\boldsymbol{\theta_1})$ *and* $f(x;\boldsymbol{\theta_2})$ *with respect to data* $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ *is defined as*

$$\max_{x_i \in \boldsymbol{x}} \Big| F(x_i;\boldsymbol{\theta_1}) - F(x_i;\boldsymbol{\theta_2}) \Big|. \tag{3.7}$$

Definition 3.3.2. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space and* $F(x;\boldsymbol{\theta})$*, the associated c.d.f. Then, for all real number* $e > 0$*,* $f(x;\boldsymbol{\theta_1})$ *and* $f(x;\boldsymbol{\theta_2})$ *are said to be indistinguishable with respect to* e *if their c.d.f. sensitivity with respect to data* $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ *is no greater than* e*. That is, for* $\boldsymbol{\theta_1},\ \boldsymbol{\theta_2} \in \Theta,\ e > 0$*,*

$$\max_{x_i \in \boldsymbol{x}} \Big| F(x_i;\boldsymbol{\theta_1}) - F(x_i;\boldsymbol{\theta_2}) \Big| \leq e. \tag{3.8}$$

Definition 3.3.3. *For a given statistical model $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Consider the procedure utilized for obtaining parameter estimates for $\boldsymbol{\theta}$ based on the data $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ and the numerical algorithm being implemented. We define such a procedure as experiment Φ.*

Definition 3.3.4. *Consider an identifiable statistical model $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space and $F(x;\boldsymbol{\theta})$, the associated c.d.f. Let $f(x;\hat{\boldsymbol{\theta}})$ be an estimated model of f with respect to experiment Φ, and $\hat{F}_n(t)$ be the empirical c.d.f. (ECDF) obtained from Φ. Then, the experimental error of Φ is defined as the c.d.f. sensitivity between the estimated model and the ECDF. That is,*

$$\epsilon := \max_{x_i \in \boldsymbol{x}} \left| F(x_i;\hat{\boldsymbol{\theta}}) - \hat{F}_n(x_i) \right|. \tag{3.9}$$

Definition 3.3.5. *Consider an identifiable statistical model $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space and $F(x;\boldsymbol{\theta})$, the associated c.d.f. Let $f(x;\hat{\boldsymbol{\theta}})$ be an estimated model of f with respect to experiment Φ. Then, $f(x;\boldsymbol{\theta})$ is said to be indistinguishable with respect to Φ if $f(x;\boldsymbol{\theta})$ and $f(x;\hat{\boldsymbol{\theta}})$ are indistinguishable with respect to the experimental error ϵ as defined in (3.9). That is, for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\max_{x_i \in \boldsymbol{x}} \left| F(x_i;\hat{\boldsymbol{\theta}}) - F(x_i;\boldsymbol{\theta}) \right| \leq \epsilon. \tag{3.10}$$

Definition 3.3.6. *Consider an identifiable statistical model $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space and $F(x;\boldsymbol{\theta})$, the associated c.d.f. Let $f(x;\hat{\boldsymbol{\theta}})$ be an estimated model of f with respect to experiment Φ; then the set $\boldsymbol{\mathcal{N}}(f,\Phi) \subset \boldsymbol{\Theta}$ is called a c.d.f. sensitivity-based confidence region (CSCR) with respect to Φ if*

$$\boldsymbol{\mathcal{N}} := \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| \max_{x_i \in \boldsymbol{x}} \left| F(x_i;\hat{\boldsymbol{\theta}}) - F(x_i;\boldsymbol{\theta}) \right| \leq \epsilon \right\}. \tag{3.11}$$

where ϵ is as defined in (3.9).

Definition 3.3.7. *Consider an identifiable statistical model $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $dim(\boldsymbol{\Theta}) = d_1$ is the parameter space. Define a sub-space $\boldsymbol{\mathcal{P}} \subset \boldsymbol{\Theta}$ with $dim(\boldsymbol{\mathcal{P}}) = d_2 < d_1$. Define $\boldsymbol{\mathcal{R}} := \boldsymbol{\Theta} \setminus \boldsymbol{\mathcal{P}}$ so that $dim(\boldsymbol{\mathcal{R}}) = d_1 - d_2$, and let the parameters in $\boldsymbol{\mathcal{R}}$ be $\boldsymbol{r} = \{r_1, r_2, \ldots, r_{d_1-d_2}\}$. Let $\boldsymbol{\theta}_B$ denote the boundary of the domain in the parameter space $\boldsymbol{\Theta}$. Then, a statistical model $\mathcal{M}_1 = \{g(x;\boldsymbol{p}) : \boldsymbol{p} \in \boldsymbol{\mathcal{P}}\}$ with parameter space $\boldsymbol{\mathcal{P}}$ is said to be a sub-model of $\mathcal{M}$ if*

$$\lim_{\boldsymbol{r} \to \boldsymbol{r}_B} F(x;\boldsymbol{\theta}) = G(x;\boldsymbol{p}), \tag{3.12}$$

where $\boldsymbol{r}_B$ are the elements in $\boldsymbol{\theta}_B$ corresponding to $\boldsymbol{r}$, and G is the associated c.d.f.

Definition 3.3.8. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space. Then,* $\mathcal{F}$ *is said to be the sub-model family of* $\mathcal{M}$ *if it comprises all the sub-models of* $\mathcal{M}$*. Namely,*

$$\mathcal{F} := \bigcup_{i=1}^{n} \mathcal{M}_i, \tag{3.13}$$

where $\mathcal{M}_i$ *is a sub-model of* $\mathcal{M}$ *and* $n \geq 1$ *is the number of sub-models.*

The principal contribution in this chapter is the definition of the estimability of a statistical model that follows.

Definition 3.3.9. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *with parameter space* $\boldsymbol{\Theta}$*. Then,* $\mathcal{M}$ *is said to be non-estimable with respect to experiment* Φ *if its CSCR with respect to experiment* Φ *is infinite. Accordingly,* $\mathcal{M}$ *is said to be estimable if its CSCR with respect to experiment* Φ *is bounded.*

Observe that expression (3.9) quantifies the experimental error as ϵ. It can also be interpreted as the tolerance level within which the estimated model c.d.f. can vary. The CSCR essentially includes all possible parameters such that the c.d.f. sensitivity of model having those parameters is less than the experimental error. Clearly, the smaller the experimental error, the smaller the CSCR, this being due to the fact that the experimental error is set as an upper bound in (3.11).

The next step is to make Definition 3.3.9 applicable in practice, as this definition may not be of practical use if utilized directly. Theorem 3.3.10 addresses this issue.

Theorem 3.3.10. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space and* $F(x;\boldsymbol{\theta})$*, the associated c.d.f. Assume that* $\mathcal{M}$ *has a sub-model family* $\mathcal{F} := \bigcup_{k=1}^{n} \mathcal{M}_k$*. Then,* $\mathcal{M}$ *is non-estimable if there exists at least one sub-model* $\mathcal{M}_k = \{g(x;\boldsymbol{p}) : \boldsymbol{p} \in \boldsymbol{\mathcal{P}}\}$ *which satisfies both of the following conditions:*

(i) $\max\limits_{x_i \in \boldsymbol{x}} \left| F(x_i;\hat{\boldsymbol{\theta}}) - G(x_i;\hat{\boldsymbol{p}}) \right| \leq \epsilon$*, where* ϵ *is as defined in (3.9), and* F *and* G *are the associated c.d.f.'s.*

(ii) Let $\boldsymbol{r}_B$ *be the boundary associated with* M_k*, that is,* $\lim_{\boldsymbol{r} \to \boldsymbol{r}_B} F(x;\boldsymbol{\theta}) = G(x;\boldsymbol{p})$*, then* $\boldsymbol{r}_B$ *consists of either* ∞ *or* $-\infty$*.*

Proof. By Definition 3.3.7, we have

$$\lim_{\boldsymbol{r} \to \boldsymbol{r}_B} F(x;\boldsymbol{\theta}) = G(x;\boldsymbol{p}). \tag{3.14}$$

Let $\tilde{\boldsymbol{p}}(\boldsymbol{r})$ be the estimates of $\boldsymbol{p}$ for given values of $\boldsymbol{r}$. Assuming the parameters are estimated under the same numerical algorithm, then for all x, we have

$$\lim_{\boldsymbol{r} \to \boldsymbol{r}_B} F(x;\tilde{\boldsymbol{p}}(\boldsymbol{r}) \cup \boldsymbol{r}) = G(x;\hat{\boldsymbol{p}}). \tag{3.15}$$

Subsequently, based on (3.15), we have

$$\lim_{\boldsymbol{r}\to\boldsymbol{r}_B}\max_{x_i\in\boldsymbol{x}}\left|F(x_i;\hat{\boldsymbol{\theta}})-F(x_i;\tilde{\boldsymbol{p}}(\boldsymbol{r})\cup\boldsymbol{r})\right|=\max_{x_i\in\boldsymbol{x}}\left|F(x_i;\hat{\boldsymbol{\theta}})-G(x_i;\hat{\boldsymbol{p}})\right|. \tag{3.16}$$

Applying conditions (i) and (ii), (3.16) then implies that,

$$\exists\delta_j>0 \text{ such that } \forall|r_j|>\delta_j, \max_{x_i\in\boldsymbol{x}}\left|F(x_i;\hat{\boldsymbol{\theta}})-F(x_i;\tilde{\boldsymbol{p}}(\boldsymbol{r})\cup\boldsymbol{r})\right|\leq\epsilon,$$

where $j=1,2,\ldots,d_1-d_2$.

Thus, $\tilde{\boldsymbol{p}}(\boldsymbol{r})\cup\boldsymbol{r}$ will be in the CSCR by Definition 3.3.6. In that case, the CSCR is infinite in $\boldsymbol{\mathcal{R}}$. Then, by Definition 3.3.9, $\mathcal{M}$ will be non-estimable with respect to experiment Φ. □

Consequently, Corollary 3.3.11 can be obtained as the complement of Theorem 3.3.10:

Corollary 3.3.11. *Consider an identifiable statistical model $\mathcal{M}=\{f(x;\boldsymbol{\theta}):\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space and $F(x;\boldsymbol{\theta})$, the associated c.d.f. Let $f(x;\hat{\boldsymbol{\theta}})$ be an estimated model to experiment Φ. Then, $\mathcal{M}$ is estimable if one of the following statements holds true:*

(i) $\mathcal{M}$ does not have the sub-model family $\mathcal{F}$.

(ii) As $\mathcal{M}$ has the sub-model family $\mathcal{F}:=\bigcup_{k=1}^{n}\mathcal{M}_k$, then for all $M_k\in\mathcal{F}$, (i) and (ii) specified in Theorem 3.3.10 cannot be simultaneously satisfied.

Interestingly, as can been seen from the next result, the existence of sub-models also brings some insights into the non-estimability issue in terms of flat profile likelihood surface.

Theorem 3.3.12. *Consider an identifiable statistical model $\mathcal{M}=\{f(x;\boldsymbol{\theta}):\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space and $F(x;\boldsymbol{\theta})$, the associated c.d.f. If $\mathcal{M}$ has a sub-model family $\mathcal{F}:=\bigcup_{k=1}^{n}\mathcal{M}_k$, then at least one profile likelihood surface has a finite limit.*

Proof. Let $\mathcal{M}_k=\{g(x;\boldsymbol{p}):\boldsymbol{p}\in\boldsymbol{\mathcal{P}}\}$ satisfy (i) and (ii) as specified in Theorem 3.3.10, where $k\in\{1,2,\ldots,n\}$. In other words, it is $\mathcal{M}_k$ that makes $\mathcal{M}$ non-estimable. Moreover, let $h(\cdot)$ be the profile likelihood surface of $\boldsymbol{r}$. Then, we have

$$\lim_{\boldsymbol{r}\to\boldsymbol{r}_B}h(\boldsymbol{r})=\lim_{\boldsymbol{r}\to\boldsymbol{r}_B}\max_{\boldsymbol{p}}\prod_{i=1}^{N}f(x_i;\boldsymbol{\theta})=\max_{\boldsymbol{p}}\prod_{i=1}^{N}g(x_i;\boldsymbol{p})=:C.$$

Let $\tilde{\boldsymbol{p}}(\boldsymbol{r})$ be the parameter estimates of $\boldsymbol{p}$ given certain values of $\boldsymbol{r}$, under a numerical maximum-likelihood algorithm. Then, we have[1]

$$\lim_{\boldsymbol{r}\to\boldsymbol{r}_B} h(\boldsymbol{r}) = \lim_{\boldsymbol{r}\to\boldsymbol{r}_B} \prod_{i=1}^{N} f(x_i;\tilde{\boldsymbol{p}}(\boldsymbol{r})\cup\boldsymbol{r}) = \prod_{i=1}^{N} g(x_i;\hat{\boldsymbol{p}}) =: \hat{C}.$$

Without any loss of generality, the same logic applies to other sub-models. Therefore, the profile likelihood surface will have finite limits corresponding to each one of the sub-models. □

Therefore, given the k^{th} sub-model $\mathcal{M}_k$, the profile likelihood surface of $\boldsymbol{r}$ will converge to a finite limit of $\hat{C}$. This is analogous to results available in the literature in that the non-estimability issue arises from flat profile likelihood functions. The only difference is that, based on the proposed definition, this flatness is not only obtained by visual inspection, but is rigorously proved to converge to finite limit(s).

3.4 Validation of the proposed definition

In this section, we will validate the proposed definition and establish that it is innately sound. The main idea is to show consistencies between the theoretical results and common sense. This will be achieved by considering multiple perspectives, as discussed in Sections 3.4.1, 3.4.2 and 3.4.3. Finally, these consistencies are further supported by the illustrative examples presented in Section 3.4.4.

3.4.1 Validation of the data noise, the algorithm noise and the experimental error

We first validate the data noise, the algorithm noise and the experimental error as specified in Definitions 3.4.1 and 3.3.4. Without any loss of generality, it is assumed that the experimental error comes from the data noise and the algorithm noise [Chis et al. (2011)]. The data noise and the algorithm noise can be defined as the following c.d.f. sensitivity measures:

Definition 3.4.1. *Consider an identifiable statistical model* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space and* $F(x;\boldsymbol{\theta})$*, the associated c.d.f. Let* $f(x;\hat{\boldsymbol{\theta}})$ *be an estimated model of* f *with respect to experiment* Φ*. Denoting by* $\hat{\boldsymbol{\theta}}^T$ *the true value of the parameter estimate, then the data noise* ϵ_d *and the algorithm noise* ϵ_{AL} *of* $\mathcal{M}$ *with respect to experiment* Φ *can*

[1] Due to algorithm noise, the exact value of $\max_{\boldsymbol{p}} \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta})$, will slightly deviate from $\prod_{i=1}^{N} f(x_i;\tilde{\boldsymbol{p}}(\boldsymbol{r})\cup\boldsymbol{r})$. This explains the difference between C and $\hat{C}$.

be respectively defined as the following c.d.f. sensitivities:

$$\epsilon_d := \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - \hat{F}_n(x_i) \right|, \tag{3.17}$$

$$\epsilon_{AL} := \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - F_X(\hat{\boldsymbol{\theta}}; x_i) \right|. \tag{3.18}$$

We now validate the data noise and the algorithm noise. As the sample size goes to infinity,

$$\begin{aligned} \epsilon_d &:= \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - \hat{F}_n(x_i) \right| \\ &\to \max_{x \in \boldsymbol{\Omega}} \left| F_X(\boldsymbol{\theta}^*; x) - F_X(\boldsymbol{\theta}^*; x) \right| \\ &= 0, \end{aligned}$$

where $\boldsymbol{\theta}^*$ is the true parameter of the model (unknown of course) and $\boldsymbol{\Omega}$ is the support of F_X. Notice that the limits of $F_X(\hat{\boldsymbol{\theta}}^T; x_i)$ and $\hat{F}_n(x_i)$ are based on asymptotic theories of consistent estimator $\hat{\boldsymbol{\theta}}^T$ and ECDF $\hat{F}_n(t)$, respectively. Thus, ϵ_d is a valid measure of the data noise because it tends to zero as the number of observations increases to infinity. This agrees with common sense.

Similarly, for the algorithm noise, as the accuracy of the algorithm tends to perfection,

$$\begin{aligned} \epsilon_{AL} &:= \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - F_X(\hat{\boldsymbol{\theta}}; x_i) \right| \\ &\to \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - F_X(\hat{\boldsymbol{\theta}}^T; x_i) \right| \\ &= 0. \end{aligned}$$

This also agrees with common sense as the more accurate the algorithm is, the better the approximation of $\hat{\boldsymbol{\theta}}^T$ will be. Thus, ϵ_{AL} is a valid measure of the algorithm noise as well. After validating the data noise and the algorithm noise, we may now validate the experimental error. Notice that

$$\begin{aligned} \epsilon :=& \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}; x_i) - \hat{F}_n(x_i) \right| \\ =& \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}; x_i) - F_X(\hat{\boldsymbol{\theta}}^T; x_i) + F_X(\hat{\boldsymbol{\theta}}^T; x_i) - \hat{F}_n(x_i) \right| \\ \leq& \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}^T; x_i) - \hat{F}_n(x_i) \right| + \max_{x_i \in \boldsymbol{x}} \left| F_X(\hat{\boldsymbol{\theta}}; x_i) - F_X(\hat{\boldsymbol{\theta}}^T; x_i) \right| \\ =& \epsilon_d + \epsilon_{AL}. \end{aligned} \tag{3.19}$$

Then, as the experimental design tends to perfection, both the data noise and the algorithm noise will tend to zero. In that case, ϵ will tend to zero by the Squeeze Theorem. Thus, ϵ is a valid measure of the experimental error.

It is worth emphasizing that ϵ_d and ϵ_{AL} will be unknown to the experimenter because one will never[2] know the true value of the parameter estimate, $\hat{\boldsymbol{\theta}}^T$. Instead, one only knows $\hat{\boldsymbol{\theta}}$ which is the output of the numerical algorithm. However, this does not prevent us from validating the proposed measures by applying the Squeeze Theorem.

3.4.2 Validation of the c.d.f. sensitivity-based confidence region

We next validate the definition of the CSCR. Observe that the experimental error, ϵ, is an upper bound in (3.11). Thus, as the experimental error decreases, the CSCR will shrink, making the model more estimable, which is consistent with intuition. Now, as was done in Section 3.4.1, consider the argument that the experimental design tends to perfection. Then,

$$\begin{aligned}
\boldsymbol{\mathcal{N}} &:= \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| \max_{x_i \in \boldsymbol{x}} \left| F(x_i; \hat{\boldsymbol{\theta}}) - F(x_i; \boldsymbol{\theta}) \right| \leq \epsilon \right\} \\
&\to \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| \max_{x \in \boldsymbol{\Omega}} |F(x; \boldsymbol{\theta}^*) - F(x; \boldsymbol{\theta})| \leq 0 \right\} \\
&= \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| \max_{x \in \boldsymbol{\Omega}} |F(x; \boldsymbol{\theta}^*) - F(x; \boldsymbol{\theta})| = 0 \right\} \\
&= \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| F(x; \boldsymbol{\theta}^*) - F(x; \boldsymbol{\theta}) = 0, \forall x \in \boldsymbol{\Omega} \right\} \\
&= \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \middle| F(x; \boldsymbol{\theta}) = F(x; \boldsymbol{\theta}^*), \forall x \in \boldsymbol{\Omega} \right\} \\
&=: \boldsymbol{\mathcal{A}},
\end{aligned}$$

where $\boldsymbol{\theta}^*$ is the true (unknown) parameter of the model.

The CSCR, $\boldsymbol{\mathcal{N}}$, will then collapse to $\boldsymbol{\mathcal{A}}$ which is exactly the parameter set for which $\mathcal{M}$ is non-identifiable. Therefore, if the experimental error goes to zero, then any remaining unreliability associated with the parameter estimates must originate from the non-identifiability issue. This agrees with the statement found in the literature to the effect that it is unnecessary to analyze estimability if the model is non-identifiable, since the non-identifiability issue cannot be overcome by improvements in the experimental design. Thus, the consistency regarding the CSCR further validates the proposed definition of estimability.

3.4.3 Validation by known methods for improving estimability

The proposed definition can also be validated by making use of known methods for improving estimability. These are discussed next.

[2]Otherwise, one would be able to calculate $\hat{\boldsymbol{\theta}}^T$ analytically without resorting to numerical algorithms.

Increasing the sample size

Increasing the sample size can decrease both ϵ_d and ϵ_{AL}. More observations will not only decrease the data noise, but also decrease the algorithm noise as the likelihood functions will then become more and more concave. Then, in light of the arguments presented in Sections 3.4.1 and 3.4.2, ϵ will decrease and the CSCR will shrink, which will improve estimability. As the sample size tends to infinity, ϵ will tend to zero and the experiment will tend to perfection. Thus, the proposed definition is consistent with an increase in the sample size.

Increasing the convexity of the log-likelihood function

Another way to increase the sample size is to clone the simulated data multiple times. This is called the data cloning method [Lele et al. (2007, 2010)]. With this approach, the likelihood functions become more and more concave, which makes the algorithm approximation more accurate. Accordingly, ϵ_{AL} will decrease. However, it only reduces the algorithm noise. It cannot improve ϵ_d as $\hat{\boldsymbol{\theta}}^T$ will not change. Thus, the proposed definition is consistent with the data cloning method.

Improving the algorithm design

As with the data cloning method, we can also decrease ϵ_{AL} by improving the quality of the algorithm approximation. However, the data noise will remain unchanged, which is consistent with intuition as the improvement pertains to the algorithm only.

Securing more complete information

Estimability can also be enhanced if more information is secured. As the data imparts more complete information, ϵ_d will decrease. Then, in light of the arguments presented in Sections 3.4.1 and 3.4.2, ϵ will decrease and the CSCR will shrink, which will improve estimability. Thus, the proposed definition is consistent with securing more complete information.

Using Bayesian inference with sound prior information

The non-estimability issue could also be improved by applying Bayesian methodologies. Since the likelihood function is flat, the posterior distribution will be highly dependent on the prior. In that case, sound prior information, say from an expert opinion for example, will improve parameter estimability as the posterior distribution will then produce narrower credible intervals. However, it turns out that Bayesian inference on the PTAM involves the Markov chain Monte Carlo theory which is another big field of study. Thus, we will specifically investigate this method in Chapter 4.

3.4.4 Illustrative examples

In Sections 3.4.1, 3.4.2 and 3.4.3, we have theoretically validated the proposed definition by establishing its consistencies with common sense. This will be further illustrated via two examples in this section. These examples respectively consider the application of the proposed definition on discrete and continuous statistical models.

Example A

Consider a constrained binomial model

$$\mathcal{M} = \left\{ f(k; m, p) = \binom{m}{k} p^k (1-p)^{m-k} : mp = \lambda > 0,\ m \in \mathbb{Z}^+,\ 0 < p < 1 \right\}.$$

First, the binomial distribution is identifiable, which makes it eligible for estimability assessment. According to Definition 3.3.8, $\boldsymbol{\Theta} = \{m, \lambda\}$, $\mathcal{P} = \{\lambda\}$, $\mathcal{R} = \{m\}$, $\boldsymbol{\theta}_B = \{0, 1, \infty\}$, $r_B = \{\infty\}$. Then, the sub-model family of $\mathcal{M}$ based on r_B is

$$\mathcal{F} = \mathcal{M}_1 := \left\{ g(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!} : \lambda > 0 \right\}$$

since the Poisson distribution is the limiting distribution of the binomial distribution as $m \to \infty$ with the constraint $mp = \lambda$.

Let the underlying model have as its parameters $m = 10$ and $p = 0.07$, and consider following three experiments:

- Experiment 1: The sample size is 50.
- Experiment 2: The sample size is 500.
- Experiment 3: The sample size is 1000.

The estimability assessment results are presented in Table 3.1, Figures 3.1 and 3.2. Table 3.1 compares the experimental error with the c.d.f. sensitivity between the fitted binomial distribution and its sub-model, which is a direct application of Theorem 3.3.10.

Sample Size	Sub-model	$\max_{k_i \in \boldsymbol{x}} \left\{ \left\lvert F(k_i; \hat{m}, \hat{p}) - G(k_i; \hat{\lambda}) \right\rvert \right\}$	ϵ
50	$\mathcal{M}_1$	0.004260893*	0.05132765
500	$\mathcal{M}_1$	0.004489167*	0.02269388
1000	$\mathcal{M}_1$	0.01275535	0.004800656

Table 3.1: Comparison between the c.d.f. sensitivity and the experimental error for the constrained binomial model - Experiments 1 to 3. The asterisk indicates that the binomial model is non-estimable.

It can be observed that the experimental error decreases as the sample size increases, which again validates the definition of the experimental error. According to Theorem 3.3.10, $\mathcal{M}$ is non-estimable with respect to Experiments 1 and 2 because the c.d.f. sensitivity measure is less than the experimental error. However, $\mathcal{M}$ becomes estimable, with respect to Experiment 3.

The above conclusions are further supported by Figure 3.1 where the CSCRs for Experiments 1, 2 and 3 are visualized. In line with Theorem 3.3.10, the CSCR for Experiments 1 and 2 are both infinite indeed. As the sample size increases, the CSCR shrinks, until it becomes bounded in Experiment 3 where $\mathcal{M}$ becomes estimable.

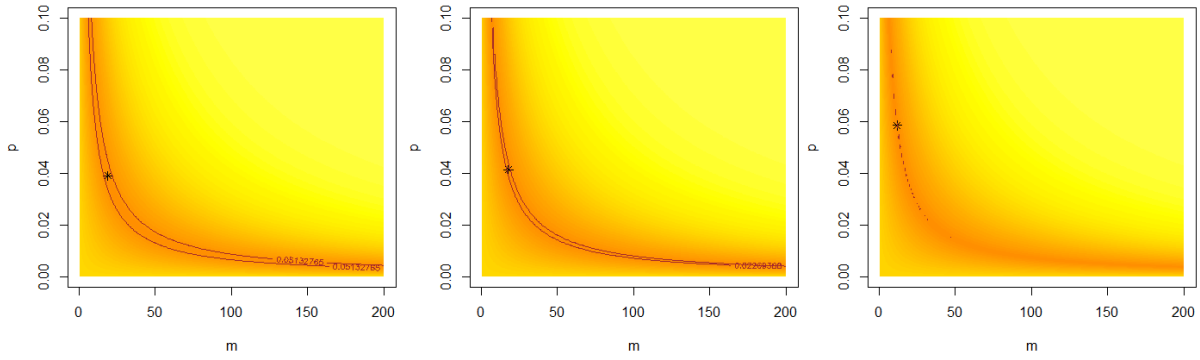

Figure 3.1: Contour plots of the CSCR on the parameter space of $\mathcal{M}$. Shades from yellow to orange correspond to high and low values of the c.d.f. sensitivity measure. The contour lines display the boundary of the CSCR and the asterisk indicates the optimal parameter estimates $\hat{m}$ and $\hat{p}$. Left panel: Experiment 1 - non-estimable. Middle panel: Experiment 2 - non-estimable but with some improvements. Right panel: Experiment 3 - estimable.

A more intuitive way of interpreting the concept of estimability is displayed in Figure 3.2. In Experiments 1 and 2, the model is assessed to be non-estimable. This can be intuitively interpreted as, the inferential power from the data displayed in the histogram is not sufficiently noticeable to tell apart the estimated p.m.f.s of the underlying model and its sub-model. This is exactly why estimability is also referred to as "practical identifiability" since the shape of the histogram cannot "practically identify" the estimated p.m.f.'s of the underlying model and its sub-model. On the other hand, in Experiment 3, the model is assessed to be estimable. Thus, the inferential power from the data displayed in the histogram is sufficiently noticeable to tell the estimated p.m.f.'s apart. In this case, the shape of the histogram will be sufficient to favor the underlying model and negate its sub-model. The histogram then "practically identifies" the underlying model.

However, these conclusions cannot be reached only by eyeballing Figure 3.2, which is why the CSCR associated with the proposed definition is crucial.

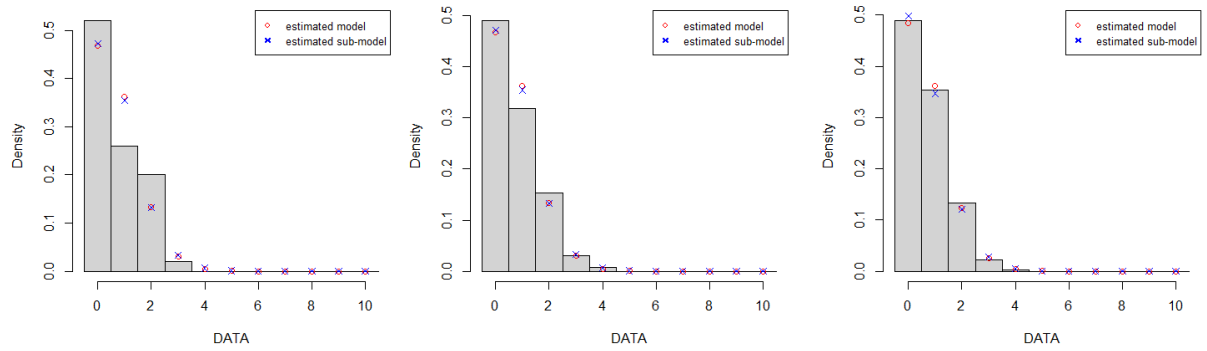

Figure 3.2: Simulated data, estimated model and estimated sub-model for the constrained binomial model. Left panel: Experiment 1 - non-estimable. Middle panel: Experiment 2 - non-estimable but with some improvements. Right panel: Experiment 3 - estimable.

Example B

Consider a constrained Pareto model

$$\mathcal{M} = \left\{ f(x;\alpha,\theta) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}} : \frac{\alpha}{\theta} = \lambda > 0, \alpha > 0,\ \theta > 0 \right\}.$$

First, the Pareto distribution is identifiable, which makes it eligible for estimability assessment. According to Definition 3.3.8, $\boldsymbol{\Theta} = \{\alpha, \lambda\}$, $\mathcal{P} = \{\lambda\}$, $\mathcal{R} = \{\alpha\}$, $\boldsymbol{\theta}_B = \{0, \infty\}$, $r_B = \{\infty\}$. Then, the sub-model family of $\mathcal{M}$ based on r_B is

$$\mathcal{F} = \mathcal{M}_1 := \left\{ g(x;\lambda) = \lambda e^{-\lambda x} : \lambda > 0 \right\}$$

since the exponential distribution is the limiting distribution of the Pareto distribution as $\alpha \to \infty$ with the constraint $\frac{\alpha}{\theta} = \lambda$.

Let the underlying model have as its parameters $\alpha = 3$ and $\theta = 30$, and consider following three experiments:

- Experiment 1: The sample size is 100.
- Experiment 2: The sample size is 250.
- Experiment 3: The sample size is 500.

The estimability assessment results are presented in Table 3.2, Figures 3.3 and 3.4. Table 3.2 compares the experimental error with the c.d.f. sensitivity between the fitted Pareto distribution and its sub-model, which is a direct application of Theorem 3.3.10. It can be seen that the results yield the same conclusions as those in Example A.

Sample Size	Sub-model	$\max_{x_i \in \boldsymbol{x}} \left\{ \left\lvert F(x_i; \hat{\alpha}, \hat{\theta}) - G(x_i; \hat{\lambda}) \right\rvert \right\}$	ϵ
100	$\mathcal{M}_1$	0.02081047*	0.05438388
200	$\mathcal{M}_1$	0.0262806*	0.02685047
500	$\mathcal{M}_1$	0.03654049	0.02500097

Table 3.2: Comparison between the c.d.f. sensitivity and the experimental error for the constrained Pareto model - experiments 1 to 3. The asterisk indicates that the Pareto model is non-estimable.

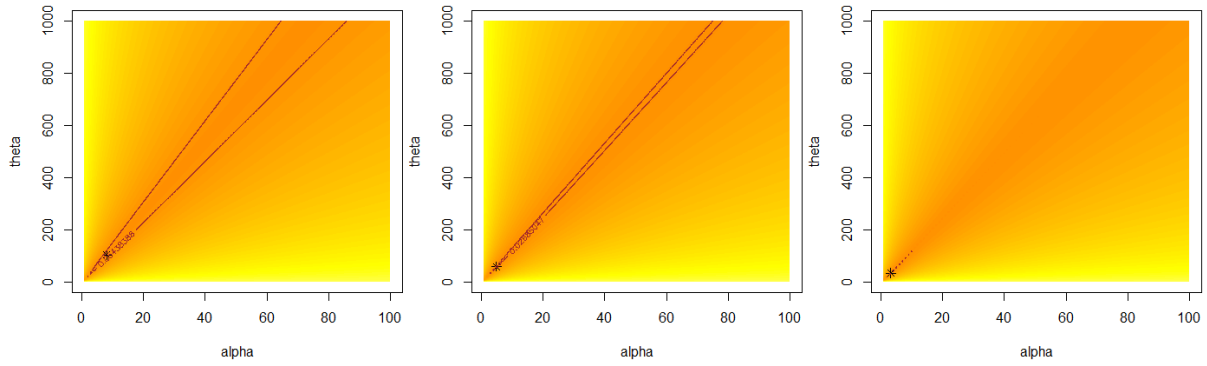

Figure 3.3: Contour plots of the CSCR on the parameter space of $\mathcal{M}$. Shades from yellow to orange correspond to high and low values of the c.d.f. sensitivity measure. The contour lines display the boundary of the CSCR and the asterisk indicates the optimal parameter estimates $\hat{\alpha}$ and $\hat{\theta}$. Left panel: Experiment 1 - non-estimable. Middle panel: Experiment 2 - non-estimable but with some improvements. Right panel: Experiment 3 - estimable.

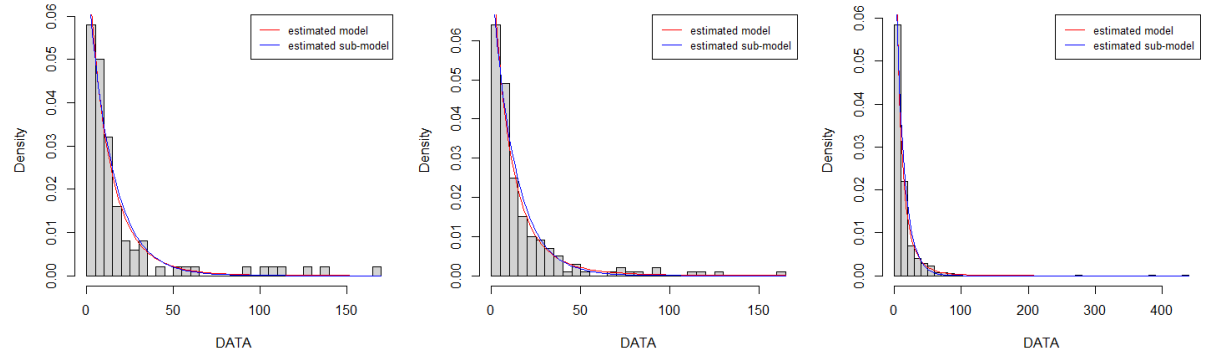

Figure 3.4: Simulated data, estimated model and estimated sub-model for the constrained Pareto model. Left panel: Experiment 1 - non-estimable. Middle panel: Experiment 2 - non-estimable but with some improvements. Right panel: Experiment 3 - estimable.

It is crucial to emphasize that the visualization of the CSCR displayed in Figures 3.1 and 3.3 will not be achievable if the parameter space extends to more than three dimensions, such as the PTAM which will be assessed next. In that case, we will have to fully rely on Theorem 3.3.10 and compare the c.d.f. sensitivity with the experimental error. The two-dimensional illustrative examples presented in the previous subsections provide supporting evidence corroborating the validity of Theorem 3.3.10.

Another crucial aspect is that the algorithm utilized to obtain parameter estimates follows the algorithm recommended in Section 3.6.1. We will investigate this further in the remaining part of this chapter.

3.5 Estimability of the PTAM

In Section 3.4, we have established that the proposed definition of estimability is innately sound. In this section, we will apply this definition to assess the estimability of the PTAM.

3.5.1 Identifiability of the PTAM

The PTAM was proved to be identifiable when the number of states is greater or equal to six in Chapter 2. Therefore, we may proceed to assess its estimability.

3.5.2 Sub-models of the PTAM

Based on Theorem 3.3.10, one must first investigate and obtain all sub-models of the PTAM in order to investigate parameter estimability. The results are presented in the following propositions, the proofs being provided in Appendix B.

Proposition 3.5.1. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$*, the associated p.d.f. Given* m*, the limiting distribution as* $s \to \infty$ *is Coxian of order* 2 *with*

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ 0 & -h_m \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_m \end{bmatrix}.$$

Denote this as sub-model $\mathcal{M}_1$*.*

Proposition 3.5.2. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$, the associated p.d.f. Given m, the limiting distribution as $s \to -\infty$ is Coxian of order m with

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda+h_1) & \lambda & 0 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda+h_1) & \lambda & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -(\lambda+h_1) & \lambda \\ 0 & 0 & 0 & 0 & \dots & 0 & -h_m \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_1 \\ \vdots \\ h_1 \\ h_m \end{bmatrix}.$$

Denote this as sub-model $\mathcal{M}_2$.

Proposition 3.5.3. *Consider the PTAM $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$, the associated p.d.f. Given m, the limiting distribution as $h_m \to \infty$ and $s \to -\infty$ is Coxian of order $m-1$ with

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda+h_1) & \lambda & 0 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda+h_1) & \lambda & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -(\lambda+h_1) & \lambda \\ 0 & 0 & 0 & 0 & \dots & 0 & -(\lambda+h_1) \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_1 \\ \vdots \\ h_1 \\ h_1+\lambda \end{bmatrix}.$$

Denote this as sub-model $\mathcal{M}_3$.

Proposition 3.5.4. *Consider the PTAM $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ wherein $\boldsymbol{\Theta}$ is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\psi, h_1, h_m, s, m) | h_m > h_1 > 0, \psi > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$, the associated p.d.f. According to Cheng et al. (2021), the limiting distribution as $m \to \infty$ is

$$f(t; h_1, h_m, s, \psi) = e^{-\int_0^t h(u;h_1,h_m,s,\psi)\mathrm{d}u} h(t; h_1, h_m, s, \psi),$$

where

$$h(t; h_1, h_m, s, \psi) = \begin{cases} \left((h_m^s - h_1^s)\frac{t}{\psi} + h_1^s\right)^{\frac{1}{s}}, & s \neq 0, \\ h_1^{1-\frac{t}{\psi}} h_m^{\frac{t}{\psi}}, & s = 0 \end{cases}$$

is the hazard rate of the limiting distribution. Denote this as sub-model $\mathcal{M}_4$.

It is worth pointing out that h_m, even though depending on m, still remains in the expression of the p.d.f. and the hazard rate. This is due to the fact that, while the number of interpolated dying rates in Figure 2.3 increases as m increases, the value of h_m remains unchanged. In other words, the value of m only controls the number of interpolated rates between h_1 and h_m, instead of the actual values of h_1 and h_m.

Interestingly, as m goes to infinity under the constraint (2.4), the hazard rates under $s \neq 0$ and $s = 0$, respectively correspond to the generalized Weibull distribution and the Gompertz law of mortality, which are well-known mortality models [Pham and Lai (2007); Gompertz (1825)].

Proposition 3.5.5. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$*, the associated p.d.f. The limiting distribution as* $h_m \to \infty$ *and* $s \to \infty$ *is exponential with rate* $h_1 + \lambda$*. Denote this as sub-model* $\mathcal{M}_5$*.*

Proposition 3.5.6. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$ *the associated p.d.f. The limiting distribution as* $m \to \infty$ *and* $s \to \infty$ *is exponential with rate* h_m*. This sub-model is again* $\mathcal{M}_5$*, with a different rate parameter for the exponential distribution.*

Proposition 3.5.7. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$*, the associated p.d.f. The limiting distribution as* $m \to \infty$ *and* $s \to -\infty$ *is exponential with rate* h_1*. This sub-model is again* $\mathcal{M}_5$*, with a different rate parameter for the exponential distribution.*

Proposition 3.5.8. *Consider the PTAM* $\mathcal{M} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *wherein* $\boldsymbol{\Theta}$ *is the parameter space*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\lambda, h_1, h_m, s, m) | h_m > h_1 > 0, \lambda > 0, s \in \mathbb{R}, m \geq 6\}.$$

and $f(x;\boldsymbol{\theta})$*, the associated p.d.f. The limiting distribution as* $m \to \infty$*,* $h_m \to \infty$ *and* $s \to -\infty$ *is exponential with rate* h_1*. This sub-model is again* $\mathcal{M}_5$*, with a different rate parameter for the exponential distribution.*

Combining all these sub-models, the sub-model family of the PTAM becomes

$$\mathcal{F} := \bigcup_{i=1}^{5} \mathcal{M}_i.$$

3.5.3 Simulation studies

In this section, the proposed definition of estimability will be applied to assess the estimability of the PTAM via simulation studies.

We consider an underlying PTAM with $h_1 = 0.0007$, $h_m = 0.1542$, $s = -2.2152, \lambda = 0.3122$ and $m = 10$. This underlying PTAM is utilized in Chapter 4 to model the aging process of individuals residing in Channing House - a retirement community in Palo Alto, California [Hyde (1980)].

In Table 3.3, the proposed definition is applied to assess model estimability with respect to six experiments:

- Experiment 1: A sample size of 50, 100000 initial values in the algorithm.
- Experiment 2: A sample size of 50, 200000 initial values in the algorithm.
- Experiment 3: A sample size of 500, 100000 initial values in the algorithm.
- Experiment 4: A sample size of 500, 200000 initial values in the algorithm.
- Experiment 5: A sample size of 2000, 100000 initial values in the algorithm.
- Experiment 6: A sample size of 2000, 200000 initial values in the algorithm.

It can be observed from Table 3.3 that

(i) According the Theorem 3.3.10, the PTAM is non-estimable with respect to Experiments 1 to 5, whereas it is estimable with respect to Experiment 6.

(ii) One potential threat to the estimability of the PTAM occurs when $s \to -\infty$, which corresponds to $\mathcal{M}_2$ and $\mathcal{M}_3$.

(iii) The algorithm noise decreases as the number of initial values increases. This is again a validation of the proposed definition. However, this trend does not always hold, unless the algorithm is selected to be that recommended in Section 3.6.1. We will discuss this shortly in Section 3.6.1.

In conclusion, the non-estimability issue of the PTAM can be thoroughly investigated utilizing the proposed definition. One may then arrive at an experimental design that makes the PTAM estimable such as that utilized Experiment 6.

Sample Size	Sub-model	100000 IVs		200000 IVs	
		$\max_{x_i \in \boldsymbol{x}} \left\lvert F(x_i;\hat{\boldsymbol{\theta}}) - G(x_i;\hat{\boldsymbol{p}}) \right\rvert$	ϵ	$\max_{x_i \in \boldsymbol{x}} \left\lvert F(x_i;\hat{\boldsymbol{\theta}}) - G(x_i;\hat{\boldsymbol{p}}) \right\rvert$	ϵ
50	$\mathcal{M}_1$	0.23261	0.05807	0.22661	0.05310
	$\mathcal{M}_2$	0.02414*	0.05807	0.01698*	0.05310
	$\mathcal{M}_3$	0.04271*	0.05807	0.01863*	0.05310
	$\mathcal{M}_4$	0.16902	0.05807	0.12708	0.05310
	$\mathcal{M}_5$	0.36318	0.05807	0.35797	0.05310
500	$\mathcal{M}_1$	0.32839	0.04051	0.31173	0.03845
	$\mathcal{M}_2$	0.02024*	0.04051	0.01520*	0.03845
	$\mathcal{M}_3$	0.02352*	0.04051	0.02048*	0.03845
	$\mathcal{M}_4$	0.30614	0.04051	0.155581	0.03845
	$\mathcal{M}_5$	0.33164	0.04051	0.34158	0.03845
2000	$\mathcal{M}_1$	0.20239	0.02972	0.26365	0.01824
	$\mathcal{M}_2$	0.02612*	0.02972	0.02270	0.01824
	$\mathcal{M}_3$	0.03645	0.02972	0.02312	0.01824
	$\mathcal{M}_4$	0.11637	0.02972	0.15976	0.01824
	$\mathcal{M}_5$	0.34331	0.02972	0.36079	0.01824

Table 3.3: Comparison between the c.d.f. sensitivity and the experimental error for the PTAM - experiment 1 to 6. The asterisk indicates that the PTAM is non-estimable.

3.6 Discussion

3.6.1 A recommendation regarding the algorithm design

A recommended though not mandatory design for the algorithm consists of searching a parameter estimate $\hat{\boldsymbol{\theta}}^T$ such that the following function g is minimized:

$$g(\boldsymbol{\theta}) = \max_{x_i \in \boldsymbol{x}} \left| F_X(\boldsymbol{\theta}; x_i) - \hat{F}_n(x_i) \right|, \tag{3.20}$$

where $F_X, \hat{F}_n$ and $\boldsymbol{x}$ are as defined in Definition 3.3.4. Otherwise, there will possibly exist counter-intuitive situations where the experimental error actually increases when the algorithm accuracy improves. To verify this, first notice that

$$\epsilon = g(\hat{\boldsymbol{\theta}}), \tag{3.21}$$

$$\epsilon_d = g(\hat{\boldsymbol{\theta}}^T). \tag{3.22}$$

Accordingly, if $\hat{\boldsymbol{\theta}}^T$ is based on other optimization criterion instead of (3.20), for instance the maximization of likelihood, then it is plausible to have $\epsilon = g(\hat{\boldsymbol{\theta}}) < g(\hat{\boldsymbol{\theta}}^T) = \epsilon_d$ before improving the algorithm accuracy. This is due to the fact that ϵ_d is not the minimum of g. Now, let ϵ' be the experimental error after the algorithm accuracy has improved. As explained in Section 3.4.3, increasing the algorithm accuracy will not change ϵ_d but the experimental error will converge to ϵ_d, so that it would be possible to have $\epsilon < \epsilon' < \epsilon_d$. This will result in a counter-intuitive situation. Therefore, the only way to eliminate this concern is to base $\hat{\boldsymbol{\theta}}^T$ on minimizing g, in which case ϵ_d will be the minimum of g.

3.6.2 Other potential definitions of estimability

Interestingly, minimizing g is essentially equivalent to fitting a regression model to the ECDF $\hat{F}_n$, treating g as the loss function. This perspective then suggests additional possibilities for g to define estimability. However, another valid g, if it exists, must also pass all the validations presented in Sections 3.4.1, 3.4.2 and 3.4.3. For example, consider the mean absolute deviation, that is,

$$g(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left|F_X(\boldsymbol{\theta}; x_i) - \hat{F}_n(x_i)\right|, \tag{3.23}$$

where $F_X, \hat{F}_n$ and $\boldsymbol{x}$ are as defined in Definition 3.3.4. Indeed, we believe that there possibly exist other alternative choices for g in addition to (3.20) and (3.23), which points to a possible avenue of research.

Different loss functions will lead to different extents to which the inequality is scaled in (3.19). The more stringent the inequality scaling (or the further ϵ to $\epsilon_d + \epsilon_{AL}$), the less stringent the estimability assessment will be. This is due to the fact that a smaller ϵ will more likely produce a narrower CSCR, which makes the model more likely to be assessed as estimable. Therefore, different loss functions might help experimenters to further investigate the stringency of the proposed definition.

3.6.3 Estimability versus density approximation

Estimability also relates to density approximation because they both involve measuring how close models are relative to each other, as displayed in Figures 3.2 and 3.4. Interestingly, estimability can be converted into a density approximation problem if one switches perspectives. For example, Experiments 1 and 2 in Figures 3.2 and 3.4, while considered to be disadvantaged from an estimability perspective, can be favored from a density approximation perspective. This is the case because we can alternatively interpret the results as a good approximation of the underlying model with its sub-models, since the histogram cannot tell them apart. Accordingly, we believe that the proposed definition has the potential to be extended to density approximation problems. In this context, the experimental error will then provide an objective threshold that determines how well the proposed density approximates the target density.

3.6.4 Caveat

A potential limitation of the proposed definition pertains to how one can define the non-estimability issue on the basis of characteristics of the CSCR. Similarly to Raue et al. (2009), we define the non-estimability issue as occurring when the CSCR is infinite. However, for sufficiently wide but bounded confidence region, the region is almost as deficient as when it

is infinite[3]. This is due to the fact that, as was done in Raue et al. (2009), we also relied on a binary event: whether the confidence region is infinite or not. There might exists another mathematical concept pertaining to multidimensional regions which relies on another binary event instead of finite versus infinite. If such a concept exists, it would be worth attempting to adapt the threshold to it.

3.7 Conclusion

A novel definition of estimability was proposed to objectively quantify this concept in the context of statistical models. The proposed definition is similar to an existing method in that the non-estimability is defined as occurring when the confidence region is infinite. However, unlike the existing method which utilizes a likelihood-based confidence region, a c.d.f. sensitivity-based confidence region was proposed with a view to tailor the confidence region to the experimental protocol. Under that setting, the threshold becomes objective as the experimental error becomes an experiment-based quantity under the proposed methodology. Multiple arguments showed that the proposed definition was innately sound as the corresponding theoretical results agree with common sense. The validated definition was then applied to assess the estimability of the PTAM, which solved the potential non-estimability issue.

[3]This is confirmed by Raue et al. (2009).

Chapter 4

Markov Chain Monte Carlo for Bayesian Inference on the Phase-Type Aging Model

In this chapter, we will investigate and implement Bayesian inference with sound prior information on the PTAM as one of the methods for improving the estimability of the PTAM presented in Chapter 3, Section 3.4.3. The proposed method provides two methodological extensions based on an existing MCMC inference method. First, we propose a two-level MCMC sampling scheme that makes the method applicable to situations where the posterior distributions do not assume simple forms after data augmentation. Secondly, an existing data augmentation technique for Bayesian inference on continuous phase-type distributions is further developed in order to incorporate left-truncated data. While numerical experimental results indicate that the proposed methodology improves parameter estimability for the PTAM as opposed to the MLE method, this approach may also be utilized as a standalone model fitting technique.

4.1 Motivation

As presented in Chapter 3, two sources of noises are currently assumed in connection with estimability. The data noise ϵ_d and the algorithm noise ϵ_{AL}. We begin this chapter by further elaborating on ϵ_{AL} for the PTAM since it is essentially more complicated than expected. According to Cheng et al. (2021), the likelihood function of the PTAM with representation $(\boldsymbol{\pi}, \boldsymbol{S})$ and order m, given data set $\boldsymbol{y} = (y_1, y_2, ..., y_M)$, can be simplified using the Kolmogorov forward equation. This is due to the fact that only the first row of $e^{\boldsymbol{S}t}$ is utilized to calculate the density of the PTAM. The likelihood function of the PTAM is then:

$$L(h_1, h_m, s, \lambda, m; \boldsymbol{t}) = \prod_{i=1}^{M} \boldsymbol{\pi}' e^{\boldsymbol{S}t_i} \boldsymbol{h} = \prod_{i=1}^{M} \sum_{k=1}^{m} P_{1k}(t_i) h_k, \tag{4.1}$$

where

$$P_{1k}(t) = \begin{cases} e^{-(\lambda_1+h_1)t}, & k=1, \\ \sum_{j=1}^{k} \frac{(-1)^{k-1}\lambda^{k-1}}{\prod_{\substack{i=1 \\ i\neq j}}^{k}(h_j-h_i)} e^{-(\lambda+h_j)t}, & k=2,3,...,m-1, \\ \sum_{j=1}^{m-1} \frac{(-1)^{m-1}\lambda^{m-1}}{\left(\prod_{\substack{i=1 \\ i\neq j}}^{m-1}(h_j-h_i)\right)(\lambda+h_j-h_i)} e^{-h_m t} + \frac{(-1)^{m-1}\lambda^{m-1}}{\prod_{i}^{m-1}(h_m-\lambda-h_i)} e^{-h_m t}, & k=m. \end{cases} \tag{4.2}$$

The gradient and Hessian matrix of the likelihood function are clearly intractable as the expression of $P_{1k}(t)$ is very complicated, in the sense that it is impossible to find a closed form representation of the MLEs at which the likelihood function is maximized; nor is it possible to mathematically manipulate the gradient and the Hessian matrix. As a result, certain analytical properties of the likelihood function of the PTAM cannot be determined, which affects parameter estimation. For instance, any hill-climbing algorithm may not lead to the correct MLEs, because the search might get stuck into local maxima (if any). Such algorithms only guarantee that the final output approximately ends at one of the local maxima, but it does not guarantee that it is the global maximum. To circumvent this problem, the scatter search method used in Cheng (2021) repeatedly randomizes the initial values of the optimization algorithm. In this chapter, we will develop an alternative approach based on a Bayesian perspective.

The Bayesian method can be adopted to improve the parameter estimability issue of the PTAM. It is a powerful approach for improving estimability owing to three main reasons:

(i) If the convexity of the likelihood remains extremely small over the entire parameter space, then graphically the likelihood function will exhibit a very flat pattern, which makes the posterior distributions very close to the prior distributions. Therefore, sound prior information can improve the reliability of the posterior estimates.

(ii) Due to the nature of the Bayesian approach, parameters are inherently random. This will eliminate the risk of getting stuck into local maxima. Actually, global optimization algorithms can also be created by incorporating stochastic components into hill-climbing algorithms used in frequentist settings. Examples include the scatter search method where initial values are randomized [Burke et al. (2014)] and the simulated annealing method which is essentially an MCMC algorithm [Michiels et al. (2007)].

(iii) As an approximate inference technique, the MCMC method changed our traditional way of thinking from "analytical solution" to "algorithmic approximation". This trend is fostered by the advent of the 'Big Data Era'.

Therefore, employing a Bayesian approach not only improves the parameter estimability issue brought about by flat likelihood functions, but also eliminates the risk of getting stuck into local maxima which is a further complication in ϵ_{AL}. In the next section, we will provide a thorough literature review on the MCMC-based Bayesian methodology.

4.2 Literature review on MCMC

4.2.1 The MCMC method's role in Bayesian statistics development

The debate between frequentist and Bayesian proponents has been well-known and enduring. Unlike frequentist inference where model parameters are assumed to be unknown constants, Bayesian inference stands from the view of the experimenter and incorporates the experimenter's initial belief via prior distributions. The prior distribution represents relevant information apart from experimental data and is based on the experimenter's subjective judgment before the experiment. Different experimenters will assume different prior distributions. The concept of prior distribution has always been attacked by frequentists because they believe that a subjective component has no place in scientific methods. This caused Bayesian statistics to nearly stagnate during the first half of 20^{th} century [Tanner and Wong (2010)].

In Bayesian statistics, the experimenter's belief will be updated as more experimental data are collected. The updated belief is depicted by a posterior distribution, taking into account the information from prior knowledge and experimental data. Therefore, the inference is made on the posterior distribution that contains all the relevant information. In many problems, we often obtain the posterior density up to a constant of proportionality, that is, the posterior kernel. Traditional methods rely on obtaining conjugate priors after inspecting the structure of a posterior kernel. However, some posterior kernels are analytically intractable, especially for high-dimensional distributions. In that case, the normalizing constant of the posterior has to be found via numerical integration in order to obtain the posterior distribution. Several approaches are proposed such as Gaussian quadrature [(Naylor and Smith, 1982; Smith et al., 1985)], importance sampling [(Kloek and Van Dijk, 1978; Van Dijk and Kloek, 1980)], and Laplace approximation [Tierney and Kadane (1986)]. However, these approaches only work well when the parameter space has a moderate dimension. Thus, numerical integration in high-dimensional parameter space has long been the bottleneck in the development of Bayesian statistics, until the MCMC method was introduced.

Modern Bayesian statistics with the MCMC method have following advantages:

(i) Pertinent prior knowledge (from experts) can be introduced as model input.

(ii) Using a prior distribution can counteract the non-identifiability issue to some degree, as it is more likely to converge to global optimum.

(iii) The MCMC method makes it possible to sample from analytically intractable posterior distribution, especially in high dimensions.

(iv) As an approximate inference technique, the MCMC method changed our traditional way of thinking from "analytical solution" to "algorithmic approximation".

4.2.2 The MCMC method

The MCMC method is a powerful approach that overcomes the bottleneck associated with sampling from high-dimensional posterior distributions. This technique was first proposed by Metropolis et al. (1953), with a view to address high-dimensional particle state calculations in nuclear physics. Later on, Hastings (1970) generalized the method as a statistical sampling tool and proposed the Metropolis-Hasting algorithm.

At the heart of the MCMC method is the construction of a Markov chain whose stationary distribution is the target distribution $\pi(\theta)$. Under this setting, the parameter samples, $\theta_1, \theta_2, \ldots, \theta_N$, are no longer directly sampled from $\pi(\theta)$ as in the traditional Monte Carlo method. Instead, the samples are generated by a carefully designed Markov chain based on the detailed balance condition. This circumvents having to deal with any analytical difficulties associated with the mathematical expression of $\pi(\theta)$. Finally, as the sample size, N, becomes large, the samples will converge to the stationary distribution [Brooks et al. (2011)].

Needless to say, there must be a relationship between the Markov chain and its stationary distribution. Such a relationship is the so-called detailed balance condition.

Theorem 4.2.1. *Consider an ergodic continuous-time Markov chain (CTMC) of size n with transition kernel* $p_{ij} = k\left(\theta^{(j)}|\theta^{(i)}\right), 1 \leq i,\ j \leq n$ *with initial distribution* $\pi(\theta)$ *having density* $\pi_i := \pi\left(\theta^{(i)}\right), 1 \leq i \leq n$. *Then, a sufficient condition for* π *to be the stationary distribution of the CTMC (continuous time Markov chain) is*

$$\pi_i p_{ij} = \pi_j p_{ji}, \forall 1 \leq i,\ j \leq n. \tag{4.3}$$

Equation (4.3) is called the detailed balance condition. If the constructed CTMC satisfies the detailed balance condition, then the CTMC will converge to the stationary distribution π.

The Metropolis-Hasting algorithm proposed by Hastings (1970) is widely used as one of the MCMC methods to sample from a probability distribution. In the algorithm, the samples are generated from a proposal density $q(\theta)$ chosen at the discretion of the experimenter, then the samples are accepted or rejected based on an acceptance probability α determined at each iteration. Algorithm 3 presents the Metropolis-Hasting algorithm for sampling from $\pi(\theta)$.

Algorithm 3 The Metropolis-Hasting algorithm

1: **initialization** Sample $\theta^{(i)}$ from the proposal distribution $q(\theta)$, where $i \in \{1, 2, \ldots, n\}$.
2: **for** $k = 2 : N$ **do**
3: Sample $\theta^{(j)}$ from the proposal distribution $\theta^{(j)} \sim q(\theta|\theta^{(i)})$.
4: Calculate the acceptance probability for $\theta^{(j)}$ as

$$\alpha_{ij} = \min\left(1, \frac{\pi\left(\theta^{(j)}\right) q\left(\theta^{(i)}|\theta^{(j)}\right)}{\pi\left(\theta^{(i)}\right) q\left(\theta^{(j)}|\theta^{(i)}\right)}\right).$$

5: Simulate $u \sim U(0, 1)$.
6: If $u < \alpha_{ij}$, then set $\theta^{(i)} = \theta^{(j)}$.
7: **end for**

Suppose p_{ij} is the transition kernel of the CTMC constructed from the Metropolis-Hasting algorithm, and Q_{ij} is the transition kernel for the proposal distribution. Then,

$$\begin{aligned}
\pi_i p_{ij} &= \pi\left(\theta^{(i)}\right) \alpha_{ij} Q_{ij} \\
&= \pi\left(\theta^{(i)}\right) \min\left(1, \frac{\pi\left(\theta^{(j)}\right) q\left(\theta^{(i)}|\theta^{(j)}\right)}{\pi\left(\theta^{(i)}\right) q\left(\theta^{(j)}|\theta^{(i)}\right)}\right) q\left(\theta^{(j)}|\theta^{(i)}\right) \\
&= \min\left(\pi\left(\theta^{(i)}\right) q\left(\theta^{(j)}|\theta^{(i)}\right), \pi\left(\theta^{(j)}\right) q\left(\theta^{(i)}|\theta^{(j)}\right)\right) \\
&= \pi\left(\theta^{(j)}\right) \min\left(1, \frac{\pi\left(\theta^{(i)}\right) q\left(\theta^{(j)}|\theta^{(i)}\right)}{\pi\left(\theta^{(j)}\right) q\left(\theta^{(i)}|\theta^{(j)}\right)}\right) q\left(\theta^{(i)}|\theta^{(j)}\right) \\
&= \pi\left(\theta^{(j)}\right) \alpha_{ji} Q_{ji} \\
&= \pi_j p_{ji}.
\end{aligned}$$

The detailed balance condition is satisfied. Therefore, $\pi(\theta)$ is the stationary distribution of the CTMC constructed in the Metropolis-Hasting algorithm.

An intuitive way of looking at the Metropolis-Hasting algorithm is the balance between two driving forces: (1) the proposal distribution $q(.)$ and (2) the Markov relationship between the previous and next sample point. While the samples are randomly sampled from a subjectively defined proposal distribution via the traditional Monte Carlo method, the Markov chain controls the trajectory of sample paths so that the samples are not totally independent of each other. As a balance between these two forces, the simulated samples will converge to $\pi(\theta)$.

Gibbs sampling is a special case of the Metropolis-Hasting algorithm. It is used for sampling from high-dimensional distributions $\pi(\boldsymbol{\theta})$, where $dim(\boldsymbol{\theta}) = n > 1$. When applying Gibbs sampling, one does not need to specify other types of proposal distributions. Instead, the proposal distribution will be the conditional distributions obtained from the target distribution $\pi(\boldsymbol{\theta})$. Algorithm 4 presents the Gibbs algorithm for sampling from $\pi(\boldsymbol{\theta})$.

Algorithm 4 The Gibbs algorithm

1: **initialization** Sample $(\theta_1, \theta_2, \ldots, \theta_n)$ from the proposal distribution $q(\boldsymbol{\theta})$.
2: **for** $k = 2 : N$ **do**
3: Sample θ_1^* from $\pi_1\left(\theta_1|\theta_2, \theta_3, \theta_4, \ldots, \theta_{n-1}, \theta_n\right)$.
4: Sample θ_2^* from $\pi_2\left(\theta_2|\theta_1^*, \theta_3, \theta_4, \ldots, \theta_{n-1}, \theta_n\right)$.
5: $\vdots$
6: Sample θ_{n-1}^* from $\pi_{n-1}\left(\theta_{n-1}|\theta_1^*, \theta_2^*, \theta_3^*, ..., \theta_{n-2}^*, \theta_n\right)$.
7: Sample θ_n^* from $\pi_{n-1}\left(\theta_n|\theta_1^*, \theta_2^*, \theta_3^*, ..., \theta_{n-2}^*, \theta_{n-1}^*\right)$.
8: **end for**

Without any loss of generality, in each step of the loop in Algorithm 4, the acceptance probability becomes

$$\begin{aligned}\alpha = p(\boldsymbol{\theta}^*|\boldsymbol{\theta}) &= \min\left(1, \frac{\pi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^*\right)}{\pi\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}\right)}\right) \\ &= \min\left(1, \frac{\pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) \pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right)}{\pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}\right) \pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}\right)}\right),\end{aligned}$$

where $i = 1, 2, \ldots, n$. Since the remaining dimensions do not change, $\boldsymbol{\theta}_{-\boldsymbol{i}} = \boldsymbol{\theta}_{-\boldsymbol{i}}^*$. Then, we have

$$\begin{aligned}\alpha = p(\boldsymbol{\theta}^*|\boldsymbol{\theta}) &= \min\left(1, \frac{\pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) \pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right)}{\pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}\right) \pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}\right)}\right) \\ &= \min\left(1, \frac{\pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) \pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right)}{\pi_i\left(\theta_i|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) p\left(\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right) \pi_i\left(\theta_i^*|\boldsymbol{\theta}_{-\boldsymbol{i}}^*\right)}\right) \\ &= \min(1, 1) \\ &= 1,\end{aligned}$$

where $i = 1, 2, \ldots, n$. Thus, Gibbs sampling is a special case of the Metropolis-Hasting algorithm with an acceptance rate equal to one [Gianola (2007)].

Geman and Geman (1984) applied Gibbs sampling to the Bayesian restoration of images. Tanner and Wong (1987) applied Gibbs sampling to data augmentation. Gelfand and Smith (1990) illustrated the essence of Gibbs sampling more deeply and comprehensively, and established a relatively mature theory of the MCMC method.

Data augmentation Gibbs sampler

Gibbs sampling is dramatically improved when the data augmentation technique is further applied. The idea behind data augmentation is to augment the original data $\boldsymbol{y}$ with new data, $\boldsymbol{x}$, so that the density function after data augmentation, that is $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$, will have a more tractable form [Tanner and Wong (2010)]. Figure 4.1 and Algorithm 5 present the

iterative framework and Gibbs algorithm in connection with data augmentation. The technique consists of a data augmentation step and a posterior sampling step, which respectively correspond to sampling from $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$.

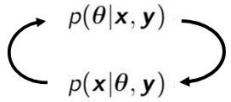

Figure 4.1: General framework of the data augmentation algorithm.

Algorithm 5 The Gibbs algorithm in conjunction with the data augmentation

1: **initialization** Sample $\boldsymbol{\theta}^{(i)}$ from the proposal distribution.
2: **for** $k = 1 : N$ **do**
3: Sample $\boldsymbol{x}$ from $p(\boldsymbol{x}|\boldsymbol{\theta}^{(i)}, y)$.
4: Sample $\boldsymbol{\theta}^{(j)}$ from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$.
5: **end for**

Proof of convergence [Hobert (2011)]

Proof. The logic behind data augmentation is essentially identical to that justifying the Gibbs algorithm:

$$\begin{aligned} p\left(\boldsymbol{\theta}^{(i)}\right) k\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(i)}\right) &= p\left(\boldsymbol{\theta}^{(i)}\right) \int p_1\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{x}\right) p_2\left(\boldsymbol{x}|\boldsymbol{\theta}^{(i)}\right) \mathrm{d}\boldsymbol{x} \\ &= \int \frac{p_1\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{x}\right) p_2\left(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\right)}{p_3(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \int \frac{p_2\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{x}\right) p_1\left(\boldsymbol{x}, \boldsymbol{\theta}^{(j)}\right)}{p_3(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= p\left(\boldsymbol{\theta}^{(j)}\right) \int p_2\left(\boldsymbol{\theta}^{(i)}|\boldsymbol{x}\right) p_1\left(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}\right) \mathrm{d}\boldsymbol{x} \\ &= p\left(\boldsymbol{\theta}^{(j)}\right) k\left(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(j)}\right), \end{aligned}$$

so that the detailed balance condition is satisfied. □

In connection with the proof, one has that the transition kernel of data augmentation will become

$$k\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{y}\right) = \int p_1\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{x}, \boldsymbol{y}\right) p_2\left(\boldsymbol{x}|\boldsymbol{\theta}^{(i)}, \boldsymbol{y}\right) \mathrm{d}\boldsymbol{x}.$$

Burn-in period and thinning

Although the constructed Markov chain will converge to the target distribution in theory, one may wonder how many iterations ought to be performed before convergence to the target distribution is assured. To address this question, a burn-in period is often used. The idea consists of discarding the samples from the first iteration up to a certain iteration, the final value being then utilized as the starting point for the samples to be used in the analysis. According to Lynch (2007), one way to select the burn-in period sample size consists of visually inspecting the cumulative standard deviation (CSD) plot. The threshold is the point beyond which the CSD plot does not show significant fluctuations.

Thinning is a common technique to reduce the autocorrelation among the samples. Then, only one sample point is collected for every L samples. This will reduce the autocorrelation and makes the samples similar to those obtained from traditional Monte Carlo sampling. According to Lynch (2007), L can be chosen in terms of the autocorrelation functions (ACFs) on the original samples, and should be such that the ACFs fall within the tolerance range after one or two lags. If the ACFs are not decreasing as lags increase, this will indicate that the constructed Markov chain does not converge.

The mixing time of the MCMC method

A well-known drawback of the MCMC method pertains to the mixing time of the Markov chain. Although the convergence to the stationary distribution is guaranteed based on the theory of Markov chain, to the best of our knowledge, there is no theoretical proof that specifies the threshold beyond which we can reasonably accept the samples and believe that they are "good enough" approximations of the stationary distribution. To address this issue, scientific tools such as trace plot and ergodic means are often utilized. However, the interpretation of the plot results relies on subjective assessment.

On the other hand, when the stationary distribution is multimodal, much longer chain may be needed to allow the sampling to encompass the entire parameter space. Moreover, the trace plots might be far more irregular than a level-off pattern. For example, consider using simulated annealing algorithm to find the global minimum of a multimodal function. If the initial temperature utilized in the simulated annealing algorithm is too low, then it will take a significantly long time for the search to escape from one of the troughs, though the MCMC samples do converge to the multimodal stationary distribution in theory. In that case, the MCMC samples might appear to converge to a distribution but the search is still getting stuck in the trough. This phenomenon is defined as "pseudo-convergence" in Brooks et al. (2011).

4.3 Literature review on MCMC-based Bayesian inference applied to the CPH distributions

In this section, existing MCMC algorithms for the CPH distributions are briefly reviewed. Bladt et al. (2003) first proposed the Gibbs algorithm for the CPH distributions. Later on, Aslett and Wilson (2011) enhanced the algorithm in terms of data augmentation and special parameter structures.

4.3.1 Data augmentation for the CPH distributions

The data augmentation scheme with respect to the CPH distributions was proposed by Asmussen et al. (1996). It is widely applied to EM, MCMC, and variational Bayes algorithms [Asmussen et al. (1996); Bladt et al. (2003); Aslett and Wilson (2011); Watanabe et al. (2012); Okamura et al. (2014)]. Consider a CPH distribution of order m with $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{S})$. Its likelihood function, given the data set $\boldsymbol{y} = (y_1, y_2, ..., y_M)$, is

$$L(\boldsymbol{\pi}, \boldsymbol{S}; \boldsymbol{y}) = \prod_{i=1}^{M} \boldsymbol{\pi}' e^{\boldsymbol{S} y_i} \boldsymbol{h}, \tag{4.4}$$

where $\boldsymbol{h} = -\boldsymbol{S}\boldsymbol{e}$.

According to Asmussen et al. (1996), a sample path associated with a CPH distribution can be characterized by the initial state, the transitions among states and the sojourn time at each state. Let $\boldsymbol{X} = \left\{X(t)^{(k)}\right\}_{t\geq 0},\ k = 1, 2, \ldots, M$, be M independent sample paths augmented from observed absorption time data $\boldsymbol{y} = \left\{y^{(k)}\right\},\ k = 1, 2, \ldots, M$. Each sample path is generated by a CPH distribution $(\boldsymbol{\pi}, \boldsymbol{S})$ of order m. Then, the likelihood function for the augmented data, $(\boldsymbol{x}, \boldsymbol{y})$, is given by

$$L(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) = \Big(\prod_{i=1}^{m} \pi_i^{B_i}\Big)\Big(\prod_{i=1}^{m}\prod_{j\neq i}^{m} \lambda_{ij}^{N_{ij}} e^{-\lambda_{ij} Z_i}\Big)\Big(\prod_{i=1}^{m} h_i^{N_{i,m+1}} e^{-h_i Z_i}\Big), \tag{4.5}$$

where B_i is the number of sample paths starting at state i among M individuals, N_{ij} is the total number of transitions from state i to state j among M individuals, and Z_i is the total sojourn time in state i among M individuals. In that case, the absorption time for an individual is equal to the sum of the sojourn times of its corresponding sample path.

4.3.2 Sampling from $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$

Sampling the latent sample path given the absorption time proves to be a difficult problem. To solve it, Bladt et al. (2003) suggested to make use of the Metropolis-Hasting algorithm with proposal distribution $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{Y} \geq \boldsymbol{y})$. Rejection sampling is utilized to draw the latent sample path from the proposal distribution, and the distribution will eventually converge to $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{Y} = \boldsymbol{y})$. However, this method is time-consuming because many sample paths will

get rejected if some data points in $\boldsymbol{y}$ is large. This will hinder computational efficiency.

Later on, Aslett and Wilson (2011) further improved the methodology by making two principal contributions:

(i) A faster and more efficient algorithm was developed to simulate the latent sample path from $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. The algorithm is named the exact conditional sampling (ECS) algorithm.

(ii) Unlike the CPH distributions considered by Bladt et al. (2003) which assumed full and unstructured parameters, Aslett and Wilson (2011) took into account the special parameter structure as indicated by the context of the experiment. Using a reliability model as an example, they discussed situations where parameters might be zero or have identical values. For identical parameters, they argued to combine all the relevant terms in (4.5) so that parameters with same values be sampled from one single distribution. The zero-valued parameters were ignored when constructing the likelihood function.

Watanabe et al. (2012) proposed another efficient MCMC algorithm to sample the latent sample path. The method is based on the uniformization technique and backward likelihood computation. In this chapter, we elected to adopt the ECS algorithm in Aslett and Wilson (2011) as part of the proposed methodology.

We now conveniently present the ECS algorithms applied to the PTAM in Algorithms 6 and 7, respectively. Since the PTAM is a Coxian distribution whose underlying Markov process is irreversible, Algorithms 6 and 7 turn out to be somewhat simpler than the original ECS algorithm that was introduced in Aslett and Wilson (2011).

Algorithm 6 The ECS algorithm [Aslett and Wilson (2011)] applied to the PTAM given absorption times

1: Sample a starting state i from the probability mass function:

$$\mathbb{P}(X(0) = i|\boldsymbol{\pi}, \boldsymbol{S}, Y = y) = \frac{\left(\boldsymbol{e}_i' e^{\boldsymbol{S}y}\boldsymbol{h}\right)\pi_i}{\boldsymbol{\pi}' e^{\boldsymbol{S}y}\boldsymbol{h}}$$

and set $t = 0$.

2: With probability

$$\mathbb{P}(X[t, y) = i \cap Y\{y\} = m + 1|\boldsymbol{S}, Y = y, X(t) = i) = \frac{e^{S_{ii}(y-t)}h_i}{\boldsymbol{e}_{\boldsymbol{i}}' e^{\boldsymbol{S}(y-t)}\boldsymbol{h}}$$

set $X[t, y) = i$ and $X(y) = m + 1$ and **end** the algorithm; **else continue**.

3: Sample the sojourn time d from

$$\begin{aligned} &p(\delta = d|\boldsymbol{S}, Y = y, X[t, t + \delta) = i, X(t + \delta) \in \{1, 2, ..., m\} \setminus i) \\ =&\frac{\boldsymbol{p}_{i.}' e^{\boldsymbol{S}(y-t-d)}\boldsymbol{s}(-S_{ii})e^{S_{ii}d}}{\int_0^{y-t} \boldsymbol{p}_{i.}' e^{\boldsymbol{S}(y-t-d)}\boldsymbol{s}(-S_{ii})e^{S_{ii}d}\mathrm{d}\delta} \end{aligned}$$

and set $X[t, t + d) = i$.

4: Update $t = t + d$ and $i = i + 1$, then go to Step 2.

Algorithm 7 The ECS algorithm [Aslett and Wilson (2011)] applied to the PTAM given right-censored times

1: Sample a starting state i from the probability mass function

$$\mathbb{P}(X(0) = i|\boldsymbol{\pi}, \boldsymbol{S}, Y \geq y) = \frac{\left(\boldsymbol{e}_i' e^{\boldsymbol{S}y}\boldsymbol{e}\right)\pi_i}{\boldsymbol{\pi}' e^{\boldsymbol{S}y}\boldsymbol{e}}$$

and set $t = 0$.

2: With probability $\min\left\{1, e^{S_{ii}(y-t)}\right\}$, the sojourn time is $d = \max\{y-t, 0\} + T$ where $T \sim exp(S_{ii})$. Else, the sojourn time is a sample from the finitely supported density on $[0, y-t)$, that is,

$$p(\delta = d|\boldsymbol{S}, Y \geq y, X(t) = i) \propto \boldsymbol{p}_{i\cdot}' e^{\boldsymbol{S}(y-t-d)}\boldsymbol{e}(-S_{ii})e^{S_{ii}d}$$

and set $X[t, t+d) = i$.

3: $j = i + 1$.

4: If $j = m + 1$, **end** the algorithm; **else**, update $t = t + d$ and $i = j$, go to Step 2.

4.3.3 Sampling from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$

The next step consists of simulating the posterior distribution of the parameter $\boldsymbol{\theta}$ from the augmented data. Fortunately, this step is quite straightforward for the CPH distributions. The likelihood function consisting of kernels of Dirichlet and Gamma distributions provides an indication to utilize Dirichlet and Gamma distributions as the conjugate prior distributions. According to Bladt et al. (2003), the prior distributions are

$$\boldsymbol{\pi} \sim Dirichlet(\beta_1, \beta_2, \ldots, \beta_m), \tag{4.6}$$
$$\lambda_{ij} \sim Gamma(v_{ij},\ \xi_{ij}), \tag{4.7}$$
$$h_i \sim Gamma(v_{i,m+1},\ \xi_{i,m+1}), \tag{4.8}$$

and posterior distributions after data augmentation are

$$\boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{y} \sim Dirichlet(\beta_1 + B_1,\ \beta_2 + B_2,\ \ldots\ ,\ \beta_m + B_m), \tag{4.9}$$
$$\lambda_{ij}|\boldsymbol{x}, \boldsymbol{y} \sim Gamma(v_{ij} + N_{ij},\ \xi_{ij} + Z_{ij}), \tag{4.10}$$
$$h_i|\boldsymbol{x}, \boldsymbol{y} \sim Gamma(v_{i,m+1} + N_{i,m+1},\ \xi_{i,m+1} + Z_{i,m+1}). \tag{4.11}$$

The p.d.f.'s of Dirichlet and Gamma distributions introduced above are as follows:

$$f(\boldsymbol{\pi}) := \frac{\prod_{i=1}^m \Gamma(\pi_i)}{\Gamma(\sum_{i=1}^m \pi_i)} \prod_{i=1}^m \pi_i^{\beta_i - 1},\ \text{where } \sum_{i=1}^m \pi_i = 1 \text{ and } \beta_i > 0 \text{ for all } i, \tag{4.12}$$

$$f(\lambda_{ij}) := \frac{\xi_{ij}^{v_{ij}} \lambda_{ij}^{v_{ij}-1} e^{-\xi_{ij}\lambda_{ij}}}{\Gamma(v_{ij})},\ \text{where } \xi_{ij} > 0, v_{ij} > 0, \tag{4.13}$$

$$f(h_i) := \frac{\xi_{i,m+1}^{v_{i,m+1}} \lambda_{i,m+1}^{v_{i,m+1}-1} e^{-\xi_{i,m+1}\lambda_{i,m+1}}}{\Gamma(v_{i,m+1})},\ \text{where } \xi_{i,m+1} > 0, v_{i,m+1} > 0. \tag{4.14}$$

4.3.4 The MCMC algorithm for the CPH distributions

The MCMC algorithm (or Gibbs sampler) for the CPH distributions can finally be constructed. It is presented in Algorithm 8.

Algorithm 8 The MCMC algorithm for the CPH distributions [Bladt et al. (2003); Aslett and Wilson (2011)]

1: **initialization** $\boldsymbol{\theta}^{(\mathbf{0})}$
2: **for** $k = 1 : N$ **do**
3: Sample $\boldsymbol{x}$ from $p(\boldsymbol{x}|\boldsymbol{\theta}^{(\boldsymbol{k-1})}, \boldsymbol{y})$, based on the ECS algorithms for full and right-censored data.
4: Sample $\boldsymbol{\theta}^{(\boldsymbol{k})}$ from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$, based on the Dirichlet and Gamma distributions.
5: **end for**

Note that $\boldsymbol{x}$ is as previously defined, the M sample paths augmented from the data $\boldsymbol{y} = \{y^{(k)}\}, k = 1, 2, \ldots, M$.

4.4 MCMC for Bayesian inference on the PTAM

As mentioned earlier, there are convincing reasons for applying the Bayesian approach on the PTAM due to its potential contributions. The MCMC algorithm for Bayesian inference on the PTAM introduced in this section constitutes the principal contribution of this chapter. This contribution involves two aspects. Firstly, the proposed MCMC algorithm can be considered as a methodological extension of the existing algorithm in terms of sampling from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$. This is due to the fact that the likelihood function of the PTAM is so involved that no simple conjugate prior distributions such as the Dirichlet and Gamma distributions are adequate. Although special parameter structures such as zero-valued and identical parameters are considered in Aslett and Wilson (2011), the prior conjugacy still holds as it simply involves deleting and regrouping parameters. However, further extensions are required in the case of the PTAM, since its parameters exhibit more complicated functional relationships as a result of the constraint specified in (2.4). Secondly, similarly to Olsson (1996) where the EM algorithm was developed for censored data from the CPH, we have developed the MCMC-based Bayesian approach for left-truncated data from the PTAM. This development is crucial for the estimation of the PTAM parameters based on real-life data, since it is unlikely that, in practice, each individual will enter the study at the same physiological age. Thus, there exists additional difficulty in analyzing left-truncated data.

With these contributions, a methodologically extended MCMC algorithm is proposed in order to carry out the sampling from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$, so that an MCMC-based Bayesian inference on the PTAM could be achieved.

4.4.1 Likelihood function of the PTAM with left-truncated data

Taking into account left-truncated data, the likelihood function for the PTAM after data augmentation becomes the following:

$$L(\lambda, h_1, h_m, s; \boldsymbol{x}, \boldsymbol{y}) = \frac{\left(\prod_{i=1}^{m-1} \lambda^{N_{i,i+1}-Q_{i,i+1}} e^{-\lambda Z_i^{\mathcal{A}}}\right)\left(\prod_{i=1}^{m} h_i^{N_{i,m+1}} e^{-h_i G_i}\right)}{\left(\prod_{i\in\mathcal{A}} e^{-\lambda d_i}\right)}, \tag{4.15}$$

where d_i is the time at which individual i enters the study, Q_{ij} is the total number of transitions from state i to j which occurred before the entry times, G_i is the total sojourn time in state i for the portions of the sample paths after the entry times, $Z_i^{\mathcal{A}}$ is the total sojourn time in state i for the sample paths in $\mathcal{A}$, and N_{ij} is as specified in Section 4.3. Finally, $\mathcal{A}$ is defined as

$$\mathcal{A} := \left\{k \in \mathbb{Z}^+ \middle| \text{the } k^{th} \text{ sample path enters the study before reaching state } m\right\}, \tag{4.16}$$

where $t_j^{(k)}$ is the sojourn time at state j for the k^{th} sample path.

The likelihood function (4.15) can be regarded as a generalized version of the likelihood function given in Asmussen et al. (1996), taking into account left-truncated data. To verify this, if the data do not involve left truncation, then the Q_{ij}'s and d_i's will be reduced to zero for all i and j, $\mathcal{A}$ will be reduced to the set of indices of all sample paths, and both the G_i's and $Z_i^{\mathcal{A}}$'s will be reduced to Z_i's for all i. Thus, the likelihood function in (4.15) will boil down to (4.5). The details of the derivation of the likelihood function (4.15) are presented in Appendix C.

4.4.2 Characteristics of the posterior distribution of the PTAM

In the PTAM, the posterior distribution of the model parameters is no longer a product of independent kernels. To verify this, we start by substituting (2.4) into the likelihood function (4.15):

$$\begin{aligned} L(\lambda, h_1, h_m, s; \boldsymbol{x}, \boldsymbol{y}) = &\frac{\left(\prod_{i=1}^{m-1} \lambda^{N_{i,i+1}-Q_{i,i+1}} e^{-\lambda Z_i^{\mathcal{A}}}\right)}{\left(\prod_{i\in\mathcal{A}} e^{-\lambda d_i}\right)} \\ &\times \left(\prod_{i=1}^{m}\left(\frac{m-i}{m-1}h_1^s + \frac{i-1}{m-1}h_m^s\right)^{\frac{N_{i,m+1}}{s}} e^{-\left(\frac{m-i}{m-1}h_1^s + \frac{i-1}{m-1}h_m^s\right)^{\frac{1}{s}} G_i}\right), \end{aligned} \tag{4.17}$$

where $s \neq 0$.

Then, the posterior distribution $p(\lambda, h_1, h_m, s|\boldsymbol{x}, \boldsymbol{y})$ can be written as

$$p(\lambda, h_1, h_m, s|\boldsymbol{x}, \boldsymbol{y}) \propto \Big(\pi_1(\lambda) L_1(\lambda; \boldsymbol{x}, \boldsymbol{y})\Big)\Big(\pi_2(h_1, h_m, s) L_2(h_1, h_m, s; \boldsymbol{x}, \boldsymbol{y})\Big), \tag{4.18}$$

where

$$\pi_1(\lambda)L_1(\lambda;\boldsymbol{x},\boldsymbol{y}) = \pi_1(\lambda)\Big(\prod_{i=1}^{m-1}\lambda^{N_{i,i+1}-Q_{i,i+1}}e^{-\lambda Z_i^{\mathcal{A}}}\Big)\Big(\prod_{i\in\mathcal{A}}e^{\lambda d_i}\Big), \tag{4.19}$$

$$\begin{aligned}\pi_2(h_1,h_m,s)L_2(h_1,h_m,s;\boldsymbol{x},\boldsymbol{y}) &= \pi_2(h_1,h_m,s)\\ &\times\Big(\prod_{i=1}^{m}\Big(\frac{m-i}{m-1}h_1^s+\frac{i-1}{m-1}h_m^s\Big)^{\frac{N_{i,m+1}}{s}}e^{-\left(\frac{m-i}{m-1}h_1^s+\frac{i-1}{m-1}h_m^s\right)^{\frac{1}{s}}G_i}\Big),\end{aligned} \tag{4.20}$$

with π_i and L_i denoting the respective prior distributions and likelihood functions, for $i = 1, 2$.

Based on (4.18), one has that the posterior distribution of the PTAM parameters can be decomposed into two independent posterior distributions: $p(\lambda|\boldsymbol{x},\boldsymbol{y})$ and a joint posterior distribution $p(h_1,h_m,s|\boldsymbol{x},\boldsymbol{y})$ where

$$p(\lambda|\boldsymbol{x},\boldsymbol{y}) \propto \pi_1(\lambda)L_1(\lambda;\boldsymbol{x},\boldsymbol{y}), \tag{4.21}$$

$$p(h_1,h_m,s|\boldsymbol{x},\boldsymbol{y}) \propto \pi_2(h_1,h_m,s)L_2(h_1,h_m,s;\boldsymbol{x},\boldsymbol{y}). \tag{4.22}$$

Thus, we can evaluate the posterior distribution for λ separately using as conjugate prior the gamma distribution specified in (4.23), which will produce the posterior distribution given in (4.24) of the same class.

$$\lambda \sim Gamma\left(v_\lambda,\xi_\lambda\right), \tag{4.23}$$

$$\lambda|\boldsymbol{x},\boldsymbol{y} \sim Gamma\Big(v_\lambda+\sum_{i=1}^{m-1}N_{i,i+1}-\sum_{i=1}^{m-1}Q_{i,i+1},\ \xi_\lambda+\sum_{i=1}^{m-1}Z_i^{\mathcal{A}}-\sum_{i\in\mathcal{A}}d_i\Big). \tag{4.24}$$

However, the likelihood function of (h_1,h_m,s) does not consist of independent kernels, which prevents one from determining conjugate priors. The prior distributions for h_1,h_m and s which are then subjectively determined, are taken to be $\pi_{H_1}(h_1)$, $\pi_{H_m}(h_m)$ and $\pi_S(s)$. We assume for simplicity that h_1,h_m and s are independently distributed. Accordingly, their joint prior distribution, $\pi_2(h_1,h_m,s)$, will be the product of $\pi_{H_1}(h_1)$, $\pi_{H_m}(h_m)$ and $\pi_S(s)$.

4.4.3 The proposed methodology for sampling from $p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y})$

Next, a methodology is developed for sampling from the joint posterior distribution given in (4.20). The Gibbs algorithm can be utilized again, further taking advantage of the MCMC method. In that case, the proposed algorithm will become a nested MCMC algorithm. The nested Gibbs algorithm samples from the joint posterior distribution given the augmented data. The algorithm framework is presented in Figure 4.2, for a p-dimensional posterior distribution. Since general notations are adopted in Figure 4.2, we believe that the algorithm is applicable to other models whose posterior distributions are complicated after data

augmentation. The PTAM considered in this paper is but one of its applications, where $p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = p(h_1, h_m, s, \lambda|\boldsymbol{x},\boldsymbol{y})$.

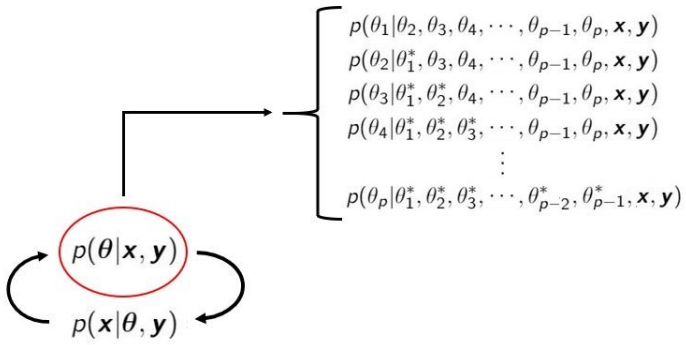

Figure 4.2: The MCMC algorithm framework for the proposed methodology.

As in the Gibbs sampling algorithm, in order to sample from $p(h_1, h_m, s|\boldsymbol{x},\boldsymbol{y})$ in the $(k+1)^{th}$ iteration for example, we need to sample from the corresponding conditional distributions. These are also the transition kernels of the Gibbs algorithm, that is,

$$p\Big(\big(h_1^{(k+1)}, h_m^{(k)}, s^{(k)}\big)\Big|\big(h_1^{(k)}, h_m^{(k)}, s^{(k)}\big)\Big) := p\big(h_1\big|h_m^{(k)}, s^{(k)}, \boldsymbol{x},\boldsymbol{y}\big), \tag{4.25}$$

$$p\Big(\big(h_1^{(k+1)}, h_m^{(k+1)}, s^{(k)}\big)\Big|\big(h_1^{(k+1)}, h_m^{(k)}, s^{(k)}\big)\Big) := p\big(h_m\big|h_1^{(k+1)}, s^{(k)}, \boldsymbol{x},\boldsymbol{y}\big), \tag{4.26}$$

$$p\Big(\big(h_1^{(k+1)}, h_m^{(k+1)}, s^{(k+1)}\big)\Big|\big(h_1^{(k+1)}, h_m^{(k+1)}, s^{(k)}\big)\Big) := p\big(s\big|h_1^{(k+1)}, h_m^{(k+1)}, \boldsymbol{x},\boldsymbol{y}\big). \tag{4.27}$$

Define $g(h_1, h_m, s) := \pi_2(h_1, h_m, s)L_2(h_1, h_m, s; \boldsymbol{x},\boldsymbol{y})$.

First, we introduce a sampling scheme for $p\big(h_1\big|h_m^{(k)}, s^{(k)}, \boldsymbol{x},\boldsymbol{y}\big)$, the conditional distribution of h_1. We know that

$$p\big(h_1\big|h_m^{(k)}, s^{(k)}, \boldsymbol{x},\boldsymbol{y}\big) = \frac{g\big(h_1, h_m^{(k)}, s^{(k)}\big)}{\int_0^{h_m^{(k)}} g\big(h_1, h_m^{(k)}, s^{(k)}\big)\mathrm{dh}_1} \propto g\big(h_1, h_m^{(k)}, s^{(k)}\big). \tag{4.28}$$

Rejection sampling can then be utilized in conjunction with $g\big(h_1, h_m^{(k)}, s^{(k)}\big)$ as described in Algorithm 9.

Algorithm 9 The rejection sampling algorithm for $p\big(h_1\big|h_m^{(k)}, s^{(k)}, \boldsymbol{x}, \boldsymbol{y}\big)$

1: Calculate the maximum value of $ln\left(g\big(h_1, h_m^{(k)}, s^{(k)}\big)\right)$ on $(0, h_m^{(k)})$. Denote it by $ln.m_{h1}$.
2: Draw a pair of samples $(x, ln(y))$. $X \sim Unif(0, h_m^{(k)})$ and $ln(Y) = ln(U) + ln.m_{h_1}$, where $U \sim Unif(0,1)$.
3: **while** $ln\left(g\big(x, h_m^{(k)}, s^{(k)}\big)\right) \leq ln(y)$ **do**
4: **repeat** Step 2
5: **end while**
6: Take $h_1^{(k+1)} = x$.

Secondly, we consider the sampling scheme for $p\big(h_m\big|h_1^{(k+1)}, s^{(k)}, \boldsymbol{x}, \boldsymbol{y}\big)$, the marginal distribution of h_m. We know that

$$p\big(h_m\big|h_1^{(k+1)}, s^{(k)}, \boldsymbol{x}, \boldsymbol{y}\big) = \frac{g\big(h_m, h_1^{(k+1)}, s^{(k)}\big)}{\int_{h_1^{(k+1)}}^{\infty} g\big(h_m, h_1^{(k+1)}, s^{(k)}\big)\mathrm{dh_m}} \propto g\big(h_m, h_1^{(k+1)}, s^{(k)}\big). \tag{4.29}$$

Rejection sampling can be utilized in conjunction with $g\big(h_m, h_1^{(k+1)}, s^{(k)}\big)$ as described in Algorithm 10.

Algorithm 10 The rejection sampling algorithm for $p\big(h_m\big|h_1^{(k+1)}, s^{(k)}, \boldsymbol{x}, \boldsymbol{y}\big)$

1: Calculate the maximum value of $ln\left(g\big(h_m, h_1^{(k+1)}, s^{(k)}\big)\right)$ on $(h_1^{(k+1)}, a)$. Denote it by $ln.m_{hm}$. In this case, a is a large enough truncation point.
2: Draw a pair of samples $(x, ln(y))$. $X \sim Unif(h_1^{(k+1)}, a)$ and $ln(Y) = ln(U) + ln.m_{hm}$, where $U \sim Unif(0,1)$.
3: **while** $ln\left(g\big(x, h_1^{(k+1)}, s^{(k)}\big)\right) \leq ln(y)$ **do**
4: **repeat** Step 2
5: **end while**
6: Take $h_m^{(k+1)} = x$.

Thirdly, we consider the sampling scheme for $p\big(s\big|h_1^{(k+1)}, h_m^{(k+1)}, \boldsymbol{x}, \boldsymbol{y}\big)$, the marginal distribution of s. We know that

$$p\big(s\big|h_1^{(k+1)}, h_m^{(k+1)}, \boldsymbol{x}, \boldsymbol{y}\big) = \frac{g\big(s, h_1^{(k+1)}, h_m^{(k+1)}\big)}{\int_{-\infty}^{\infty} g\big(s, h_1^{(k+1)}, h_m^{(k+1)}\big)\mathrm{ds}} \propto g\big(s, h_1^{(k+1)}, h_m^{(k+1)}\big). \tag{4.30}$$

Rejection sampling can be utilized in conjunction with $g\big(s, h_1^{(k+1)}, h_m^{(k+1)}\big)$ as described in Algorithm 11.

Algorithm 11 The rejection sampling algorithm for $p\big(s\big|h_1^{(k+1)}, h_m^{(k+1)}, \boldsymbol{x}, \boldsymbol{y}\big)$

1: Calculate the maximum value of $ln\left(g\big(s, h_1^{(k+1)}, h_m^{(k+1)}\big)\right)$ on (b, c). Denote it by $ln.m_s$. In this case, $|b|, c$ are large enough truncation points.
2: Draw a pair of samples $(x, ln(y))$. $X \sim Unif(b, c)$ and $ln(Y) = ln(U) + ln.m_s$, where $U \sim Unif(0,1)$.
3: **while** $ln\left(g\big(x, h_1^{(k+1)}, h_m^{(k+1)}\big)\right) \leq ln(y)$ **do**
4: **repeat** Step 2
5: **end while**
6: Take $s^{(k+1)} = x$.

The rejection sampling schemes presented in Algorithms 9, 10 and 11 constitute original contributions among others that are made in this chapter. Unlike traditional rejection sampling where a proposal function is chosen to fully cover the target density, the proposed rejection sampling transforms them to a logarithmic scale. This is due to the fact that the values of the posterior kernels are often too small to be handled by making use of the likelihood functions. In fact, sampling on a logarithmic scale is analogous to taking the logarithm of likelihood functions in order to find MLEs, since both frequentist and Bayesian will face the same problem caused by small likelihood function values. However, they deal with this problem differently. According to Bishop and Nasrabadi (2006), frequentist inference is essentially a numerical optimization problem, corresponding to the well-known MLE method where log-likelihood functions are maximized. On the other hand, Bayesian inference is essentially a numerical integration problem, where the output is a (posterior) distribution rather than a point estimate. In this context, it will involve random sampling techniques instead of optimization techniques, for instance the rejection sampling on a logarithmic scale presented in Algorithms 9, 10 and 11. Therefore, the proposed rejection sampling scheme on a logarithmic scale is parallel to its frequentist counterpart, we then believe it should as well be as widely applicable as the maximization of log-likelihood functions. Technical details regarding rejection sampling on a logarithmic scale are elaborated on in Appendix D.

4.4.4 The MCMC for the PTAM

Combining all these building blocks, Algorithm 12 presents the proposed steps for Bayesian inference on the PTAM:

Algorithm 12 The MCMC algorithm for Bayesian inference on the PTAM

Require: The number of states, m, based on prior knowledge or subjective judgment.
Input:

1. The data observations $\boldsymbol{y}$.
2. The hyper-parameters for posterior distributions: (v_λ, ξ_λ), (v_{h_1}, ξ_{h_1}), (v_{h_m}, ξ_{h_m}) and β.
3. The number of states m.
4. The number of inner iterations w_1.
5. The number of outer iterations w_2.
6. The size of the burn-in period.

Output: The posterior samples for h_1, h_m, s and λ, each of which has w_2 sample points.

1: **Initialization** $\left(\lambda^{(1)}, h_1^{(1)}, h_m^{(1)}, s^{(1)}\right)$
2: **Initialization** $\left(\lambda_{Gibbs}^{(1)}, h_{1,Gibbs}^{(1)}, h_{m,Gibbs}^{(1)}, s_{Gibbs}^{(1)}\right) = \left(\lambda^{(1)}, h_1^{(1)}, h_m^{(1)}, s^{(1)}\right)$
3: **for** $k = 2 : w_2$ **do**
4: Draw sample paths $\boldsymbol{x}$ from $p\left(\boldsymbol{x}\middle|\lambda^{(k-1)}, h_1^{(k-1)}, h_m^{(k-1)}, s^{(k-1)}, \boldsymbol{y}\right)$, based on Algorithm 6 (or Algorithm 7 for right-censored data).
5: Based on $\boldsymbol{x}$, calculate $(\boldsymbol{N}, \boldsymbol{Q}, \boldsymbol{Z}^{\mathcal{A}}, \mathcal{A}, \boldsymbol{G})$.
6: **for** $j = 2 : w_1$ **do**
7: Sample $h_{1,Gibbs}^{(j)}$ from $p\left(h_1\middle|h_{m,Gibbs}^{(j-1)}, s_{Gibbs}^{(j-1)}, \boldsymbol{N}, \boldsymbol{G}\right)$, based on Algorithm 9.
8: Sample $h_{m,Gibbs}^{(j)}$ from $p\left(h_m\middle|h_{1,Gibbs}^{(j)}, s_{Gibbs}^{(j-1)}, \boldsymbol{N}, \boldsymbol{G}\right)$, based on Algorithm 10.
9: Sample $s_{Gibbs}^{(j)}$ from $p\left(s\middle|h_{1,Gibbs}^{(j)}, h_{m,Gibbs}^{(j)}, \boldsymbol{N}, \boldsymbol{G}\right)$, based on Algorithm 11.
10: **end for**
11: Sample $\lambda^{(k)}$ from $p\left(\lambda|\boldsymbol{N}, \boldsymbol{Q}, \boldsymbol{Z}^{\mathcal{A}}\right)$.
12: $\left(h_1^{(k)}, h_m^{(k)}, s^{(k)}\right) = \left(h_{1,Gibbs}^{(N)}, h_{m,Gibbs}^{(N)}, s_{Gibbs}^{(N)}\right)$
13: Reset the inner Gibbs sampling vector to zeros.
14: $\left(\lambda_{Gibbs}^{(1)}, h_{1,Gibbs}^{(1)}, h_{m,Gibbs}^{(1)}, s_{Gibbs}^{(1)}\right) = \left(\lambda^{(k)}, h_1^{(k)}, h_m^{(k)}, s^{(k)}\right)$
15: **end for**

It is worth noting that, for the inner Gibbs sampling in each iteration, the initial values are selected to be the parameter outputs in the previous iteration, as indicated in Step 14 of Algorithm 12. Because the parameter outputs themselves also become increasingly accurate as they converge to the true posterior distribution, using the parameter outputs in previous iterations as initial values is then believed to be more reasonable and objective than random selection. In that case, we can make the most of Algorithm 12.

4.5 Simulation study

In this section, the proposed algorithm is implemented via a simulation study. We then determine whether parameter estimability is improved. Consider the following experimental conditions:

- The underlying parameters are $m = 10, \lambda = 1.99908, h_1 = 0.0008, h_m = 1.65349, s = -0.11118$. The parameter values were taken from the simulation study on the Le Bras limiting distribution that was carried out in Cheng et al. (2021), except that m is assumed to take a moderate value of ten.
- The sample size is 50.
- There are 4500 iterations of the Gibbs sampler for data augmentation.
- There are 500 iterations of the inner Gibbs sampling for the posterior distribution.
- The first 500 iterations are taken as burn-in, based on the CSD plots.
- A thinning rate of 10 is adopted, based on the ACFs.
- The prior distributions are:

$$\begin{aligned}
\pi_{H_1}(h_1) &= \text{p.d.f. of } Gamma(v_{h_1} = 1,\ \xi_{h_1} = 1000),\\
\pi_{H_m}(h_m) &= \text{p.d.f. of } Gamma(v_{h_m} = 30,\ \xi_{h_m} = 18),\\
\pi_S(s) &= 8e^{8s},\ s < 0,\\
\pi_1(\lambda) &= \text{p.d.f. of } Gamma(v_\lambda = 24,\ \xi_\lambda = 16).
\end{aligned}$$

After implementing Algorithm 12, the results are listed in Table 4.1, and illustrated in Figures 4.3 and 4.4:

Parameter	True	Posterior Mean	95% Credible Interval
h_1	0.00080	0.001326039	(0.00006395895, 0.00353304747)
h_m	1.65349	1.770213467	(1.230238, 2.478763)
s	-0.11118	-0.085397339	$(-0.347967221, -0.001345516)$
λ	1.99908	1.977085034	(1.698050, 2.258719)

Table 4.1: Posterior means and 95% credible intervals obtained from the MCMC algorithm and the true parameters.

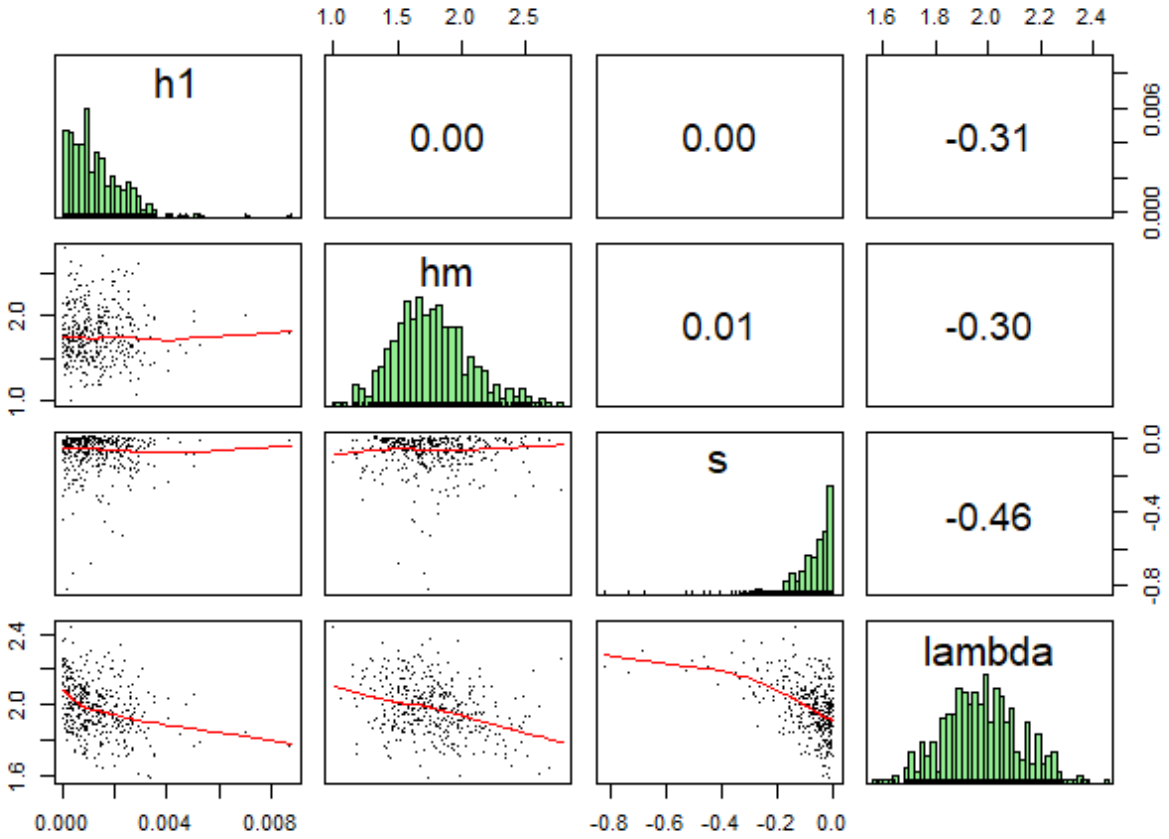

Figure 4.3: Posterior distributions and parameter correlations obtained from the MCMC samples.

In Table 4.1, the Bayesian estimates, taken as the posterior means, are all within their corresponding 95% credible intervals. This indicates that the proposed MCMC algorithm for Bayesian inference is quite satisfactory. It can be seen from Figure 4.3 that the correlations between h_1, h_m and s are minimal. This indicates that the likelihood function has little effect on the shape of the posterior distributions, so that h_1, h_m and s are still nearly independent as was assumed in the prior distributions. This observation suggests that the estimability of h_1, h_m and s could be poor. In fact, same conclusion can as well be reached by observing the diagonal panels in Figure 4.3, which show the shapes of their posterior distributions. According to Lynch (2007), posterior distributions should tend to be normal. Clearly, in Figure 4.3, the posterior distribution for λ demonstrates a more bell-shaped behaviour than the posterior distributions of h_1, h_m and s do (particularly h_1 and s). This suggests that the posterior distributions of h_1, h_m and s are less responsive to data so that the prior effects are to some degree still preserved in the behaviour of their posterior distributions.

This indicates a weaker inferential power and therefore poorer estimability for h_1, h_m and s. In contrast, the estimability for λ is better. Therefore, the role of prior distributions is crucial for estimating h_1, h_m and s. Sound prior information can improve the accuracy of the parameter estimates as the posterior distributions are highly dependent on the priors. Beautifully, the poor estimability of h_m and s is also consistent with the conclusions included in Chapter 3, Section 3.5.3.

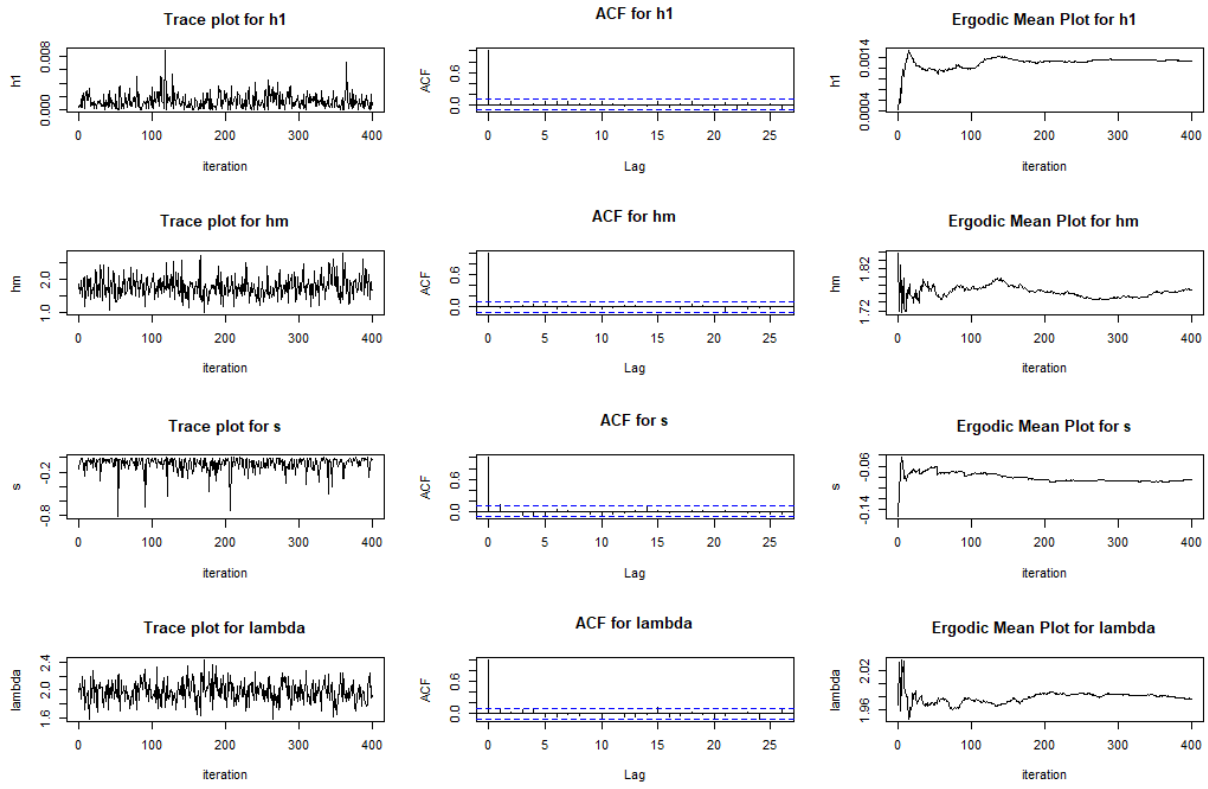

Figure 4.4: Diagnostics plots of the MCMC samples.

In Figure 4.4, the convergence of the proposed MCMC algorithm is being assessed by means of trace plots, ACFs and ergodic mean plots. First, the trace plots demonstrate the stationarity of the MCMC samples in terms of level-off patterns, though there are occasionally a few spikes for h_1 and s. However, such spikes are a normal phenomenon as the shapes of their posterior densities still remain close to their skewed prior densities due to poor estimability. Secondly, the ACFs for all parameters are within the tolerance range after the second lag. This indicates that the thinning rate effectively reduces the ACFs between the MCMC samples. Thirdly, the ergodic means all converge as the number of iterations increases. This suggests that the number of iterations, that is, 4500, is sufficient to believe that the simulated MCMC samples were approximately generated from the stationary distributions which are the target posterior distributions.

4.5.1 Estimability improvement

We compare the Bayesian estimates, the MLEs and the true values. The MLEs are calculated based on 500000 initial values, employing the scatter search method utilized in Cheng (2021). The results are presented in Figure 4.5:

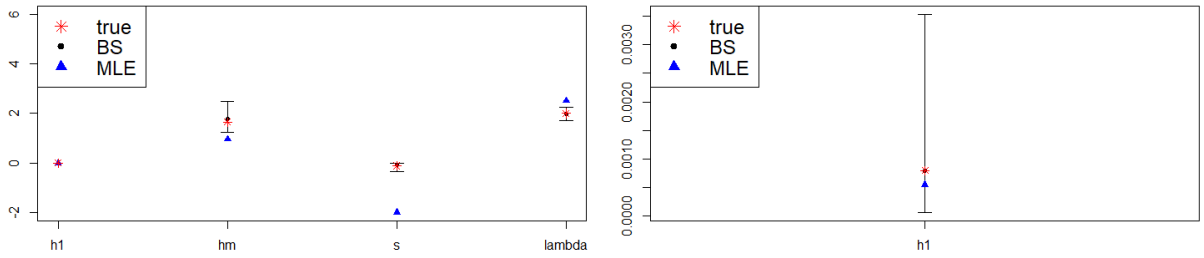

Figure 4.5: Left panel: Parameter estimates and 95% credible intervals. Right panel: Enlarged plot for h_1.

It is seen that Bayesian inference indeed performs better than the MLE approach as it improves parameter estimability. This is particularly manifest for h_m, s and λ whose MLEs lie outside their credible intervals, whereas the intervals still successfully cover the true values. In that case, their Bayesian estimates are more accurate than the MLEs in terms of narrower credible intervals. This is due to the advantage of making use of the Bayesian methodology which incorporates sound prior information.

4.5.2 Prior sensitivity analysis

To further validate the vital role of sound prior information in terms of estimability improvement, we now conduct a prior sensitivity analysis. Two alternative types of priors are tested. The first type is taken to be falsely informative,

where the prior means deviate noticeably from the true parameter values with low variances. The second type is taken to be non-informative, where parameters are uniformly distributed. The results are listed in Table 4.2 and illustrated in Figures 4.6 and 4.7:

Parameter	True	MLE	Falsely informative priors	Non-informative priors
h_1	0.00080	0.001210155	(0.01122067, 0.02576926)	(0.0006370074, 0.0514890336)
h_m	1.65349	0.957246758	(3.936409, 6.170863)	0.5202541, 21.9348132)
s	-0.11118	-1.989832312	$(-4.8561343, -0.5556046)$	$(-49.4620012, -0.7391115)$
λ	1.99908	2.514277429	(1.749293, 2.170121)	(1.819338, 3.221812)

Table 4.2: 95% credible intervals obtained from falsely informative and non-informative priors, MLEs and true parameters.

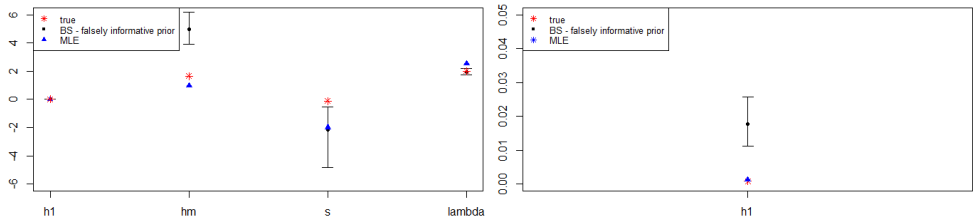

Figure 4.6: Left panel: Parameter estimates and 95% credible intervals for falsely informative priors. Right panel: Enlarged plot for h_1.

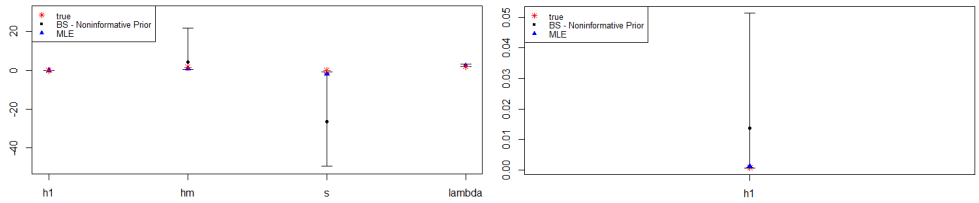

Figure 4.7: Left panel: Parameter estimates and 95% credible intervals for non-informative priors. Right panel: Enlarged plot for h_1.

It can be seen from Table 4.2 and Figure 4.6 that, when priors are taken to be falsely informative, the 95% credible intervals for h_1, h_m and s all failed to cover the true values. This is as expected because their likelihood functions are flat due to poor estimability. Then, the posterior distributions will be highly dependent on the prior distributions. On the other hand, the interval for λ remains narrow and covers the true value, which corroborates a better estimability than that of h_1, h_m and s.

Next, when the priors are taken to be non-informative, the shape of posterior density will be totally determined by the shape of the likelihood function. It can be seen from Table 4.2 and Figure 4.7 that the 95% credible intervals for h_1, h_m and s, while covering their MLEs as expected, are extremely wide. This further corroborates the flatness of their likelihood functions and therefore poor estimability. On the other hand, the interval for λ still remains narrow while covering its MLE, which corroborates a better estimability.

Upon completing this prior sensitivity analysis, all conclusions are consistent with each other throughout this simulation study. The poor estimability of h_1, h_m, s and the better estimability of λ have been supported with solid evidence. The significant prior sensitivity on h_1, h_m and s indicates that sound prior information indeed plays a significant role on improving their estimability. Therefore, it is crucial to select priors that are as sound as possible when making Bayesian inference. Otherwise, deficient priors might yield unreliable parameter estimates, particularly when their estimability is poor or unknown.

4.6 Data analysis

In Section 4.5, we have shown that the proposed Bayesian approach can improve parameter estimability for the PTAM by making use of sound prior information. In this section, we will demonstrate that, in addition to its advantage on improving estimability, the proposed Bayesian approach can also be utilized to adapt the PTAM to real-life data as a model fitting methodology.

Consider data collected from the Channing House - a retirement community in Palo Alto, California. The data consist of entry ages, ages at death and ages at study end for 462 people (97 males and 365 females) who resided in the facility between January 1964 to July 1975 [Hyde (1980)]. The Channing House data is chosen because all the residents in the community are approximately subject to the same circumstances, so that relatively speaking the aging process is the most significant factor that contributes to the variability in their lifetimes, which is the process we intend to model using the PTAM. What is more, the female data is chosen to preclude the effects of gender differences.

Only 361 records are considered, as three records have equal entry and exit ages and one record has a typo that causes the entry age to be greater than the exit age. Of the 361 females, 129 died while residing in the Channing House, whereas the other 232 survived until the end of the study.

In practice, residents join a retirement community at various physiological ages. According to the Channing House data, the youngest entry age is 61. Thus, for modelling purposes, it will be assumed that the aging process starts at calendar age 50 for all residents. Under that setting, residents will then be expected to have variability in their physiological ages at the time of entering the study. Moreover, we continue to assume that $m = 20$ for the PTAM. We currently are limited to using moderate values of m, a restriction that will be further discussed in Section 4.7.

As opposed to Section 4.5, there does not exist an underlying model. In that case, the prior distributions are surmised to be as follows:

$$\begin{aligned}
\pi_{H_1}(h_1) &= \text{p.d.f. of } Gamma(v_{h_1} = 0.002,\ \xi_{h_1} = 2),\\
\pi_{H_m}(h_m) &= \text{p.d.f. of } Gamma(v_{h_m} = 12.5,\ \xi_{h_m} = 5),\\
\pi_S(s) &= e^s,\ s < 0,\\
\pi_1(\lambda) &= \text{p.d.f. of } Gamma(v_\lambda = 1.5,\ \xi_\lambda = 5).
\end{aligned}$$

It is worth stressing that the priors are deliberately chosen such that the model with parameters taken as the prior means is far away from the Kaplan-Meier survival function estimates, as displayed in Figure 4.8. The purpose of this is to more persuasively demonstrate that the proposed Bayesian approach is sound. This is only done for the purpose of this study. In practice, of course, one should assume the priors to be such that the model with its parameters taken as the prior means is as close to the Kaplan-Meier survival function estimates as possible.

Using the proposed Bayesian approach, the parameter estimation results are displayed in Table 4.3.

Parameter	Posterior Mean	95% Credible Interval
h_1	0.0045658	(0.00006130736, 0.00923589434)
h_m	2.475408	(1.459456, 3.422970)
s	-1.085645	$(-1.8089289, -0.1331294)$
λ	0.4906715	(0.4353424, 0.5284059)

Table 4.3: Posterior means and 95% credible intervals obtained from the MCMC algorithm for the Channing House female data.

In Figure 4.8, we illustrate the goodness of fit of the PTAM to the Channing House female data by plotting the fitted survival function along with the nonparametric Kaplan-Meier survival function estimates. In addition, for comparison purposes, we also plotted the model with parameters taken as the prior mean, the fitted model using the MLE method and the fitted model obtained in Cheng et al. (2021). It can be observed that the PTAM fits the Channing House female data very well as the associated fitted survival function stays within the 95% confidence limits of the Kaplan-Meier estimates. The significant difference between the fitted model and the model with parameters taken as the prior mean, as mentioned earlier, very persuasively validates the proposed Bayesian approach. This difference clearly shows that the prior distributions are actually updated to the corresponding posterior distributions for the Channing House female data.

Furthermore, the fitted models with $m = 20$, whether estimated based on the MLEs or the proposed Bayesian method, are in very close agreement with the fitted model in Cheng et al. (2021) where $m = 100$. In fact, the fitted model with $m = 20$ fits the data even better for older ages between 91 and 101.

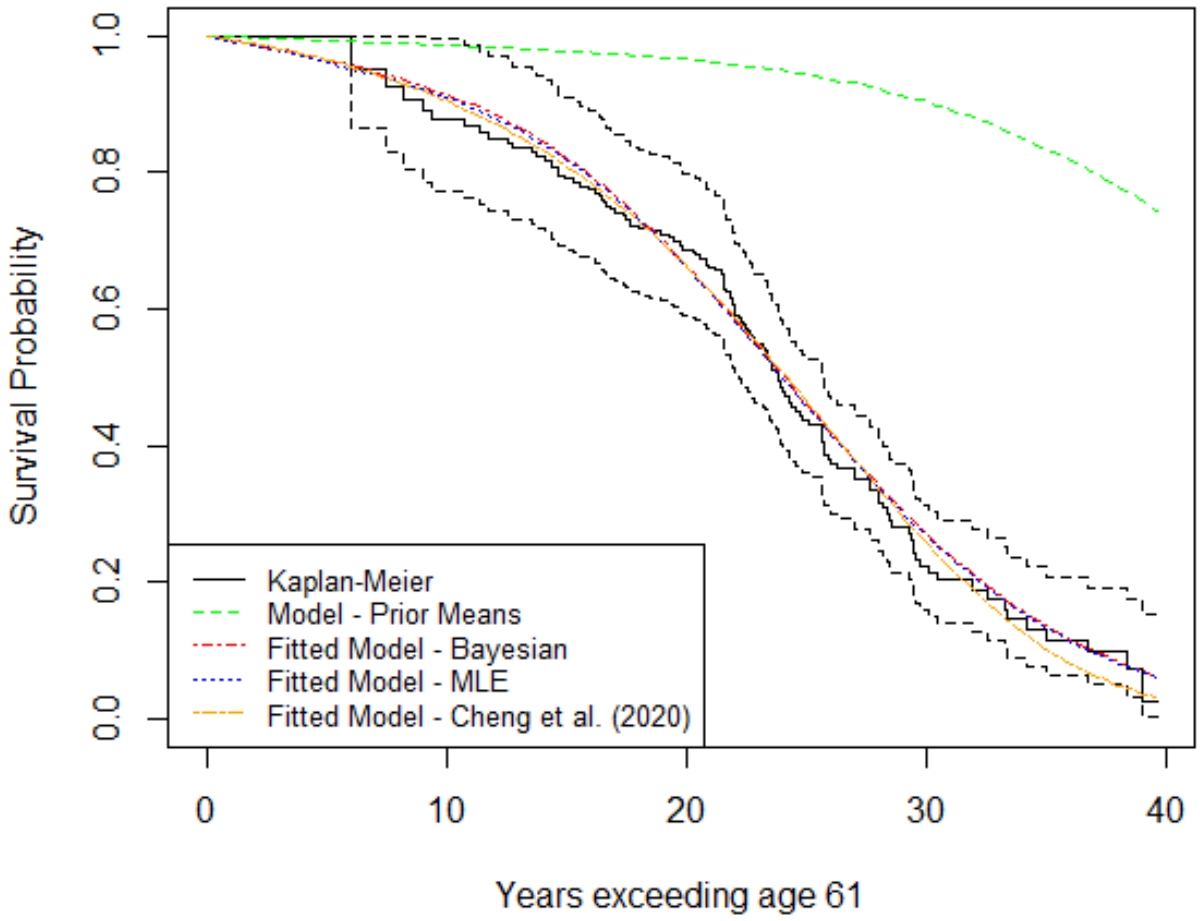

Figure 4.8: Survival functions of the PTAM calibrated to the Channing House female data using maximum likelihood estimates and the proposed Bayesian approach; the calibrated survival function with parameters taken as the prior mean; the calibrated survival function obtained in Cheng et al. (2021) and the Kaplan-Meier estimates of the survival function and corresponding 95% confidence limits.

4.7 Discussion

4.7.1 Computing matrix exponentials

The calculation of matrix exponentials is the principal drawback that affects the computational efficiency of the proposed Bayesian approach. The computing time will increase significantly when m is large. It almost takes 10 minutes for a single iteration when m is greater than 100. Thus, we are currently somewhat limited to using moderate values of m for Bayesian inference on the PTAM. Actually, MCMC-based Bayesian inference in connection with continuous phase-type distributions with a large transition intensity matrix has never been discussed in the literature. Aslett and Wilson (2011) made use of a simple reliability model with $m = 4$ and Watanabe et al. (2012) presented several examples with m no greater than 5. However, for large m, the required matrix exponential calculation will compound the computing time requested by the MCMC method. The calculation of matrix exponentials has been an open problem for years.

4.7.2 Number of states

Currently, the number of states m is treated as being fixed throughout the proposed MCMC-based Bayesian approach. Inference on m has rarely been discussed as it was not defined as a parameter according to the definition of phase-type distributions. However, it deserves attention for the PTAM as it determines the variability of the physiological age [Cheng (2021); Cheng et al. (2021)]. To the best of our knowledge, only Bladt et al. (2003) suggested that the reversible jump MCMC (RJMCMC) might constitute a plausible way to incorporate m into an MCMC-based Bayesian inference framework. This aspect could be further investigated.

4.8 Conclusion

An MCMC algorithm for Bayesian inference on the PTAM was proposed. Two contributions were made on the basis of existing MCMC algorithms for Bayesian inference on continuous phase-type distributions. First, a sampling scheme was proposed for posterior sampling after data augmentation. Secondly, existing data augmentation technique was further developed to incorporate left-truncated data. In the simulation study, the proposed approach was applied to a twenty-state PTAM. The results showed that, with sound prior information, the proposed approach indeed improved parameter estimability by producing narrower credible intervals which captured the true values. Then, it was also applied to calibrate the PTAM to aging-related mortality data from a retirement community, which produced reasonable results that are comparable to those obtained in previous research. All in all, while numerical experimental results indicate that the proposed methodology improves parameter estimability for the PTAM as opposed to the MLE method, this approach may also be utilized as a standalone model-fitting technique.

Chapter 5

Combining the Markov Chain Monte Carlo Procedure with Data Cloning to Make Inferences on the Discrete Multivariate Phase-Type Model

The discrete multivariate phase-type model (DMPTM) is a class of discrete phase-type distributions based on discrete time Markov chains with marked transitions. With respect to its parameter estimation, He and Ren (2016b) proposed an EM algorithm which is classified as a deterministic approximation in the field of approximate inference. However, parallel to deterministic approximation, the other methodology in the field of approximate inference is stochastic approximation which utilizes stochastic techniques. This approach remains unexplored with respect to the DMPTM. In this chapter, we address this gap by developing an MCMC algorithm for estimating the parameters of the DMPTM. Once combined with the data cloning method, the proposed approach can be regarded as an alternative for determining the MLEs of the DMPTM. The design of the proposed algorithm is inspired by the ECS algorithm [Aslett and Wilson (2011)] presented in Chapter 4. Numerical experiments show that the proposed MCMC algorithm combined with data cloning achieves results that are comparable to those obtained by applying the EM algorithm.

5.1 Motivation

The discrete multivariate phase-type model (DMPTM) was proposed in He and Ren (2016a). It belongs to a class of discrete phase-type distributions that is based on discrete time Markov chains with marked transitions. It can be viewed as a generalization of the discrete univariate phase-type distributions. In He and Ren (2016a), the DMPTM was utilized to model multivariate insurance claim processes in risk analysis, with claims allowed to arrive in batches. Later on, He and Ren (2016b) developed an EM algorithm to estimate the parameters of the DMPTM.

The theoretical background of the EM approach traces back to the field of approximate inference. There are two parallel, exhaustive classes of approximation schemes in this: deterministic approximation and stochastic approximation [Bishop and Nasrabadi (2006)]. In the context of deterministic approximation, a variational inference (VI) algorithm is applied to maximize the evidence lower bound (ELBO) at each iteration. The EM algorithm, as a special case of the VI algorithm, then belongs to the class of deterministic approximations[1].

Unlike deterministic approximation, stochastic approximation utilizes stochastic techniques such as the MCMC procedure to make inferences on the model parameters, which was reviewed in Section 4.2.

Thus, from the perspective of approximate inference, it is manifest that, while the application of a deterministic approximation to the DMPTM has been investigated via the EM algorithm proposed in He and Ren (2016b), the stochastic approximation counterpart remains to this day unexplored. This gap is addressed in this chapter by developing an MCMC algorithm that is applicable to the DMPTM. Contrary to the approach employed in Chapter 4, a frequentist perspective is held in this chapter. The application of the MCMC in connection with the frequentist view can be realized via the data cloning method, which will be briefly reviewed in Section 5.3.

5.2 Discrete multivariate phase-type model

5.2.1 Preliminaries

Definition 5.2.1. *Let $\{J_t\}_{t=0,1,\dots}$ be a discrete time Markov chain (DTMC) defined on a finite state space $\mathcal{S} = \mathcal{E} \cup \Delta = \{1, 2, \dots, m\} \cup \Delta$, where $\Delta = \{m+1\}$ is the absorbing state and $\mathcal{E}$ is the set of transient states. Let $\{J_t\}_{t=0,1,\dots}$ have initial distribution $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}'\boldsymbol{e} = 1$, and let the transition probability matrix be*

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{b_0} \\ \boldsymbol{0} & 0 \end{bmatrix}, \tag{5.1}$$

where $\boldsymbol{b_0} = (\boldsymbol{I_m} - \boldsymbol{B})\boldsymbol{e}$ and $\boldsymbol{e}$ is the column vector of ones. Define T as the number of transitions before absorption (the absorbing transition is not included). Then, T is said to follow a discrete phase-type (DPH) distribution denoted by $DPH(\boldsymbol{\beta}, \boldsymbol{B})$ of order m. The exit vector is denoted by $\boldsymbol{h}$.

Result 2. *Given $T \sim DPH(\boldsymbol{\beta}, \boldsymbol{B})$ of order m,*

(i) The p.m.f. of T is $p_T(t) = \boldsymbol{\beta}'\boldsymbol{B}^t\boldsymbol{h}$.

(ii) The c.d.f. of T is $F_T(t) = 1 - \boldsymbol{\beta}'\boldsymbol{B}^t\boldsymbol{e}$.

[1]Some familiarity with the EM and VI approaches is assumed. Details are available in Chapters 9 and 10 of Bishop and Nasrabadi (2006).

Definition 5.2.2. *Consider a $DPH(\boldsymbol{\beta}, \boldsymbol{B})$ of order m whose underlying DTMC includes batch event arrivals. Define the collection of batch events as $\mathcal{C}_0$ with $|\mathcal{C}_0|$ denoting the number of batches. Then, the vector $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ follows a multivariate phase-type distribution if $X_{\boldsymbol{h}}$ is the number of arrivals of batch $\boldsymbol{h}$ accumulated before the underlying DTMC enters the absorbing state. Denote it by $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$ of order m.*

Result 3. *Consider a $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$ of order m. Then $\boldsymbol{B}$ can be decomposed into $\{\boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0\}$, where*

$$\boldsymbol{B} = \boldsymbol{B_0} + \sum_{\boldsymbol{h}:\boldsymbol{h}\in\mathcal{C}_0} \boldsymbol{B_h}. \tag{5.2}$$

Throughout this thesis, we assume that no batches are associated with the absorbing transition $\boldsymbol{b_0}$. This assumption was made in He and Ren (2016b) and does not involve any loss of generality.

To facilitate the application of Definition 5.2.2, Algorithm 13 presents the algorithm for simulating $X_{\boldsymbol{h}}$ from a $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$:

Algorithm 13 Simulation of $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ from the DMPTM

1: **initialization** $X_{\boldsymbol{h},0} = 0, \forall h \in \mathcal{C}_0, k = 0$.
2: Let $I_k = i$ be simulated from initial distribution $\boldsymbol{\beta}$.
3: **while** $I_k < m + 1$ **do**
4: Simulate I_{k+1} based on $\boldsymbol{B_0}$ and $\boldsymbol{B_h}$.
5: **for** $\boldsymbol{h} \in \mathcal{C}_0$ **do**
6: If $\boldsymbol{h}$ is associated with the state move in Step 4, then let $X_{\boldsymbol{h},k+1} = X_{\boldsymbol{h},k} + 1$. Otherwise, $X_{\boldsymbol{h},k+1} = X_{\boldsymbol{h},k}$.
7: $k = k + 1$.
8: **end for**
9: **end while**
10: $\forall \boldsymbol{h} \in \mathcal{C}_0$, let $X_{\boldsymbol{h}} = X_{\boldsymbol{h},k}$.
11: **end** the algorithm.

Let Y_k be the total number of type k items that arrived before absorption. That is,

$$Y_k := \sum_{\boldsymbol{h}\in\mathcal{C}_0} h_k X_{\boldsymbol{h}}, \tag{5.3}$$

where h_k is the k^{th} element of $\boldsymbol{h}$ and K is the number of elements in the batches. For example, letting $\mathcal{C}_0 = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \boldsymbol{h_3}\} = \{(0,0,1,1,0), (0,2,1,0,1), (1,0,0,1,1)\}$ and $X_{\boldsymbol{h_1}} = 2, X_{\boldsymbol{h_2}} = 1, X_{\boldsymbol{h_3}} = 1$, we have $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5) = (1, 2, 3, 3, 2)$. Moreover, $|\mathcal{C}_0| = 3$ and $K = 5$.

The vectors $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ and $\{Y_k, 1 \le k \le K\}$ can be useful in insurance risk and other areas. For example, an organization could be subject to several types of perils and a loss event can result in claims of more than one type. Consider two types of claims in workers' compensation insurance, medical and income replacement. A single loss event could give rise to any type of claims or both. So, $K = 2$ and $\mathcal{C}_0 = \{\boldsymbol{h_1} = (1,0), \boldsymbol{h_2} = (0,1), \boldsymbol{h_3} = (1,1)\}$, Y_1 and Y_2 being the total number of the two types of claims.

5.2.2 Mathematical properties in connection with the DMPTM

He and Ren (2016a,b) established recursive relationships for state probabilities of $\boldsymbol{Y}$ for full and right-censored observations, which then underlies the EM algorithm. In this chapter, the recursive relationships for these state probabilities will as well underpin the proposed MCMC algorithm. Thus, we now conveniently present them in Results 4, 5 and 6 which correspond to full, partially right-censored and fully right-censored observations, respectively. Relevant derivations and proofs can be found in He and Ren (2016b).

Result 4. *Consider* $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ *following a* $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$ *of order* m *and* Y_k *as defined in (5.3). Define* $\boldsymbol{p_Y(y)} := \Big(p_{\boldsymbol{Y},1}(\boldsymbol{y}), \ldots, p_{\boldsymbol{Y},m}(\boldsymbol{y})\Big)'$ *as an* $m \times 1$ *vector with*

$$p_{\boldsymbol{Y},i}(\boldsymbol{y}) := \mathbb{P}(Y_1 = y_1, \ldots, Y_K = y_K | J_t = i), \tag{5.4}$$

where $i = 1, 2, \ldots, m$, $t = 0, 1, \ldots,$ *and* $k = 1, 2, \ldots, K$. *Then,*

$$\boldsymbol{p_Y(0)} = (\boldsymbol{I_m} - \boldsymbol{B_0})^{-1}(\boldsymbol{I_m} - \boldsymbol{B})\boldsymbol{e}, \tag{5.5}$$

$$\boldsymbol{p_Y}(\boldsymbol{y}) = (\boldsymbol{I_m} - \boldsymbol{B_0})^{-1}\Big(\sum_{\boldsymbol{h}:\boldsymbol{y}\geq\boldsymbol{h}} \boldsymbol{B_h}\boldsymbol{p_Y}(\boldsymbol{y} - \boldsymbol{h})\Big), \tag{5.6}$$

where $\boldsymbol{y} \geq \boldsymbol{h}$ *means* $y_k \geq h_k, \forall k \in \{1, \ldots, K\}$.

Result 5. *Consider* $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ *following a* $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$ *of order* m *and* Y_k *as defined in (5.3). Define* $\boldsymbol{p^{(0,1)}_{Y,\geq}(y^{(1)}, y^{(2)})} := \Big(p^{(0,1)}_{\boldsymbol{Y},\geq,1}(\boldsymbol{y^{(1)}, y^{(2)}}), \ldots, p^{(0,1)}_{\boldsymbol{Y},\geq,m}(\boldsymbol{y^{(1)}, y^{(2)}})\Big)'$ *as an* $m \times 1$ *vector with*

$$p^{(0,1)}_{\boldsymbol{Y},\geq,i}(\boldsymbol{y^{(1)}, y^{(2)}}) := \mathbb{P}(Y_1 = y_1, \ldots, Y_L = y_L, Y_{L+1} \geq y_{L+1}, \ldots, Y_K \geq y_K | J_t = i), \tag{5.7}$$

where

$$\begin{aligned}
\boldsymbol{y^{(1)}} &:= (y_1, \ldots, y_L)', \\
\boldsymbol{y^{(2)}} &:= (y_{L+1}, \ldots, y_K)', \\
\boldsymbol{h^{(1)}} &:= (h_1, \ldots, h_L)', \\
\boldsymbol{h^{(2)}} &:= (h_{L+1}, \ldots, h_K)',
\end{aligned}$$

and $i = 1, 2, \ldots, m$, $t = 0, 1, \ldots,$ *and* $k = 1, 2, \ldots, K$. *Then,*

$$\boldsymbol{p}^{(0,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{0^{(1)}}, \boldsymbol{0^{(2)}}) = \Big(\boldsymbol{I_m} - \boldsymbol{B_0} - \sum_{\boldsymbol{h}:\boldsymbol{h^{(1)}}=\boldsymbol{0}} \boldsymbol{B_h}\Big)^{-1}(\boldsymbol{I_m} - \boldsymbol{B})\boldsymbol{e}, \tag{5.8}$$

$$\begin{aligned}\boldsymbol{p}^{(0,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}) = &\Big(\boldsymbol{I_m} - \boldsymbol{B_0} - \sum_{\boldsymbol{h}:\boldsymbol{y^{(1)}}\geq\boldsymbol{h^{(1)}},(\boldsymbol{y^{(1)}}-\boldsymbol{h^{(1)}},(\boldsymbol{y^{(2)}}-\boldsymbol{h^{(2)}})^+)=\boldsymbol{y}} \boldsymbol{B_h}\Big)^{-1} \\ &\times \Big(\sum_{\boldsymbol{h}:\boldsymbol{y^{(1)}}\geq\boldsymbol{h^{(1)}},(\boldsymbol{y^{(1)}}-\boldsymbol{h^{(1)}},(\boldsymbol{y^{(2)}}-\boldsymbol{h^{(2)}})^+)\neq\boldsymbol{y}} \boldsymbol{B_h}\boldsymbol{p}^{(0,1)}_{\boldsymbol{Y}}(\boldsymbol{y^{(1)}} - \boldsymbol{h^{(1)}}, (\boldsymbol{y^{(2)}} - \boldsymbol{h^{(2)}})^+)\Big),\end{aligned} \tag{5.9}$$

where $\boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}$ *means* $y_k \geq h_k$, $k = 1, \ldots, L$; *and* $(\boldsymbol{y^{(2)}} - \boldsymbol{h^{(2)}})^+ := \{\max(y_k - h_k, 0)\}$, $k = L+1, \ldots, K$.

Similarly, one can derive a recursion relationship for $\boldsymbol{p}^{(1,0)}_{\boldsymbol{Y},\geq}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})$ with

$$p^{(1,0)}_{\boldsymbol{Y},\geq,i}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}) := \mathbb{P}(Y_1 \geq y_1, \ldots, Y_L \geq y_L, Y_{L+1} = y_{L+1}, \ldots, Y_K = y_K | J_t = i). \tag{5.10}$$

However, this is not necessary because the expression of $\boldsymbol{p}^{(0,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})$ does not involve any loss of generality. One can obtain this expression by rearranging all the right-censored variables in $\boldsymbol{p}^{(1,0)}_{\boldsymbol{Y},\geq}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})$ last. Result 5 is then sufficient for partially right-censored data.

Result 6. *Consider* $\{X_{\boldsymbol{h}}, \boldsymbol{h} \in \mathcal{C}_0\}$ *following a* $DMPTM(\boldsymbol{\beta}, \boldsymbol{B_0}, \boldsymbol{B_h}, \boldsymbol{h} \in \mathcal{C}_0)$ *of order* m *and* Y_k *as defined in (5.3). Define* $\boldsymbol{p}^{(1,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{y}) := \Big(p^{(1,1)}_{\boldsymbol{Y},\geq,1}(\boldsymbol{y}), \ldots, p^{(1,1)}_{\boldsymbol{Y},\geq,m}(\boldsymbol{y})\Big)'$ *as an* $m \times 1$ *vector with*

$$p^{(1,1)}_{\boldsymbol{Y},\geq,1}(\boldsymbol{y}) := \mathbb{P}(Y_1 \geq y_1, \ldots, Y_K \geq y_K | J_t = i), \tag{5.11}$$

where $i = 1, 2, \ldots, m$, $t = 0, 1, \ldots,$ *and* $k = 1, 2, \ldots, K$. *Then,*

$$\boldsymbol{p}^{(1,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{0}) = \Big(\boldsymbol{I_m} - \boldsymbol{B_0} - \sum_{\boldsymbol{h}} \boldsymbol{B_h}\Big)^{-1}(\boldsymbol{I_m} - \boldsymbol{B})\boldsymbol{e} = \boldsymbol{e}, \tag{5.12}$$

$$\boldsymbol{p}^{(1,1)}_{\boldsymbol{Y},\geq}(\boldsymbol{y}) = \Big(\boldsymbol{I_m} - \boldsymbol{B_0} - \sum_{\boldsymbol{h}:(\boldsymbol{y}-\boldsymbol{h})^+=\boldsymbol{y}} \boldsymbol{B_h}\Big)^{-1}\Big(\sum_{\boldsymbol{h}:(\boldsymbol{y}-\boldsymbol{h})^+\neq\boldsymbol{y}} \boldsymbol{B_h}\boldsymbol{p}^{(1,1)}_{\boldsymbol{Y}}((\boldsymbol{y} - \boldsymbol{h})^+)\Big), \tag{5.13}$$

where $(\boldsymbol{y} - \boldsymbol{h})^+ := \{\max(y_k - h_k, 0)\}$, $k = 1, \ldots, K$.

5.3 Literature review on data cloning

Since the MCMC output is a distribution rather than a point estimate, it is usually applied in a Bayesian context (just as in Chapter 4). However, MCMC, as a stochastic technique alone, can also be applied in connection with the frequentist view. In this context, the MCMC method is utilized as an alternative approach for approximating MLEs, indicating that the contributions of the prior choices then become negligible.

The main idea of combining MCMC with the frequentist view is to make the target distribution narrowly spread out so that it will have significant peaks around the MLEs. In that case, the parameters simulated from the target distribution will be sufficiently close to the MLEs. In fact, this is exactly the mechanism behind the simulated annealing algorithm [Michiels et al. (2007); Burke et al. (2014)] which is a well-known application of MCMC in connection with the frequentist view. In the simulated annealing algorithm, the target distribution is the Boltzmann distribution which naturally exists in physics. As the temperature parameter of the Boltzmann distribution decreases, the distribution will produce significant peaks around the optimal values.

As for the data cloning method, it is completely analogous to the simulated annealing algorithm as it also manipulates the peaks of the target distribution, the only difference being that, rather than being related to physics and using the Boltzmann distribution, one starts with MCMC-based Bayesian inference and then clones the data numerous times. In that case, the posterior density will exhibit statistical features similar to those of the likelihood function such as the convexity, variance, mean and mode. The shape of the posterior density will then become very similar to the likelihood function and will have significant peaks around the MLEs. This method is called the data cloning method [Lele et al. (2007, 2010)].

Algorithm 14 presents the data cloning algorithm. It is essentially Algorithm 5 with cloned data.

Algorithm 14 The data cloning algorithm [Lele et al. (2007, 2010)]

1: **initialization** $\boldsymbol{\theta^{(0)}}$
2: Clone the data w times. Let the cloned data be $\boldsymbol{y^c}$.
3: **for** $k = 1 : N$ **do**
4: Sample $\boldsymbol{x}$ from $p(\boldsymbol{x}|\boldsymbol{\theta^{(k-1)}}, \boldsymbol{y^c})$.
5: Sample $\boldsymbol{\theta^{(k)}}$ from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y^c})$.
6: **end for**

In line with Section 3.4.3, data cloning can only decrease the algorithm noise by making the posterior estimates close to their MLEs. However, the values of MLEs will remain unchanged. In other words, data cloning cannot improve the values of MLEs produced by small data size. The larger the data size, the less necessary data cloning is, as both the data noise and algorithm noise will be reduced.

5.4 MCMC-based Bayesian inference for the DMPTM

As pointed out in Section 5.3, the starting point of the data cloning method is the MCMC-based Bayesian inference. We then start with the likelihood function (5.14) specified in He and Ren (2016b):

$$L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) = \Big(\prod_{i=1}^{m}\beta_i^{D_i}\Big)\Big(\prod_{\boldsymbol{h}\in\mathcal{C}_0\cup\{\phi\}}\prod_{i=1}^{m}\prod_{j=1}^{m}(B_{\boldsymbol{h},ij})^{N_{(i,j),\boldsymbol{h}}}\Big)\Big(\prod_{i=1}^{m}(b_{0,i})^{N_{(i,m+1)}}\Big), \tag{5.14}$$

where D_i is the number of the sample path starting at state i, $N_{(i,j),\boldsymbol{h}}$ is the number of transitions from state i and j which includes the batch event $\boldsymbol{h}$ (if $\boldsymbol{h}=\phi$, then no batch event is associated with the state move) and $N_{(i,m+1)}$ is the number of transitions entering the absorbing state from state i.

As was done in Chapter 4, a data augmentation step and a posterior sampling step need to be developed for MCMC-based Bayesian inference on the DMPTM, which will be presented in subsequent sections. We first introduce the sampling from the posterior sampling step in Section 5.4.1. Sampling from the data augmentation step, which is far more involved, will be expounded upon in Section 5.4.2. Finally, the proposed MCMC algorithm with data cloning on the DMPTM is presented in Section 5.4.3.

5.4.1 The posterior sampling step - sampling from $p(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{b_0}|\boldsymbol{x},\boldsymbol{y})$

The likelihood appearing in (5.14) consists of Dirichlet kernels, which immediately implies a Dirichlet distribution as the conjugate prior. Then, the conjugate prior for $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}' = (\beta_1,\beta_2,\ldots,\beta_m) \sim Dirichlet(v_1,v_2,\ldots,v_m). \tag{5.15}$$

Denote the $m\times 1$ vector corresponding to the i^{th} row of matrix $\boldsymbol{B_h}$ by $\boldsymbol{B_{h,i.}}$, where $\boldsymbol{h}\in\mathcal{C}_0\cup\{\phi\}$. The conjugate prior for $\boldsymbol{B}$ can be expressed in terms of its row vectors. Namely,

$$\Big((\boldsymbol{B'_{0,i.}}),(\boldsymbol{B'_{h_1,i.}}),\ldots,(\boldsymbol{B'_{h_{|\mathcal{C}_0|},i.}}),(b_{0,i})\Big) \sim Dirichlet\Big(\boldsymbol{\eta_{B_0}^{(i)\prime}},\boldsymbol{\eta_{B_{h_1}}^{(i)\prime}},\ldots,\boldsymbol{\eta_{B_{h_{|\mathcal{C}_0|}}}^{(i)\prime}},\eta_{b_0}^{(i)}\Big), \tag{5.16}$$

where

$$\begin{aligned}
\boldsymbol{\eta_{B_0}^{(i)}} &= \Big(\eta_1^{(i)},\eta_2^{(i)},\ldots,\eta_m^{(i)}\Big)',\\
\boldsymbol{\eta_{B_{h_1}}^{(i)}} &= \Big(\eta_{m+1}^{(i)},\eta_{m+2}^{(i)},\ldots,\eta_{2m}^{(i)}\Big)',\\
&\vdots\\
\boldsymbol{\eta_{B_{h_{|\mathcal{C}_0|}}}^{(i)}} &= \Big(\eta_{|\mathcal{C}_0|m+1}^{(i)},\eta_{|\mathcal{C}_0|m+2}^{(i)},\ldots,\eta_{(|\mathcal{C}_0|+1)m}^{(i)}\Big)',\\
\eta_{b_0}^{(i)} &= \eta_{(|\mathcal{C}_0|+1)m+1}^{(i)},
\end{aligned}$$

and $1 \leq i \leq m$.

Notice that there are $m(|\mathcal{C}_0| + 1) + 1$ parameters in (5.16) for a given i. They simply correspond to elements in each of the i^{th} row vector of $\boldsymbol{B_0}$ and $\boldsymbol{B'_h}s$, plus one more parameter as the i^{th} element of $\boldsymbol{b_0}$.

The posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{B}$ then become:

$$(\beta_1, \beta_2, \ldots, \beta_m | \boldsymbol{x}, \boldsymbol{y}) \sim Dirichlet(v_1 + D_1, v_2 + D_2, \ldots, v_m + D_m) \tag{5.17}$$

and

$$\Big((\boldsymbol{B'_{0,i\cdot}}), \ldots, (\boldsymbol{B'_{h_{|\mathcal{C}_0|},i\cdot}}), (b_{0,i}) \Big| \boldsymbol{x}, \boldsymbol{y}\Big) \sim \ Dirichlet\Big(\boldsymbol{\eta_{B_0}^{(i)\prime}} + \boldsymbol{N_{B_0}^{(i)\prime}}, \ldots, \eta_{\boldsymbol{b_0}}^{(i)} + N_{\boldsymbol{b_0}}^{(i)}\Big), \tag{5.18}$$

where

$$\begin{aligned}
\boldsymbol{\eta_{B_0}^{(i)}} + \boldsymbol{N_{B_0}^{(i)}} &= \Big(\eta_1^{(i)} + N_{(i,1),\phi}, \eta_2^{(i)} + N_{(i,2),\phi}, \ldots, \eta_m^{(i)} + N_{(i,m),\phi}\Big)', \\
\boldsymbol{\eta_{B_{h_1}}^{(i)}} + \boldsymbol{N_{B_{h_1}}^{(i)}} &= \Big(\eta_{m+1}^{(i)} + N_{(i,1),\boldsymbol{h_1}}, \eta_{m+2}^{(i)} + N_{(i,2),\boldsymbol{h_1}}, \ldots, \eta_{2m}^{(i)} + N_{(i,m),\boldsymbol{h_1}}\Big)', \\
&\vdots \\
\boldsymbol{\eta_{B_{h_{|\mathcal{C}_0|}}}^{(i)}} + \boldsymbol{N_{B_{h_{|\mathcal{C}_0|}}}^{(i)}} &= \Big(\eta_{|\mathcal{C}_0|m+1}^{(i)} + N_{(i,1),\boldsymbol{h_{|\mathcal{C}_0|}}}, \eta_{|\mathcal{C}_0|m+2}^{(i)} + N_{(i,2),\boldsymbol{h_{|\mathcal{C}_0|}}}, \ldots, \eta_{(|\mathcal{C}_0|+1)m}^{(i)} + N_{(i,m),\boldsymbol{h_{|\mathcal{C}_0|}}}\Big)', \\
\eta_{\boldsymbol{b_0}}^{(i)} + N_{\boldsymbol{b_0}}^{(i)} &= \eta_{m(|\mathcal{C}_0|+1)+1}^{(i)} + N_{(i,m+1)},
\end{aligned}$$

and $1 \leq i \leq m$.

5.4.2 The data augmentation step - sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{y})$

The data augmentation algorithm on the DMPTM can be developed from the ECS algorithm given in Aslett and Wilson (2011) (Algorithms 6 and 7). They are presented as Algorithms 15, 16 and 17 which correspond to full, partially right-censored and fully right-censored observations, respectively. The technical details are available in Appendix E.

First, let the data size be $|\boldsymbol{y}| = M$, let $\{J^{(k)}(t)\}_{t\geq 0}$ and $\{N_{\boldsymbol{h}}^{(k)}(t)\}_{t\geq 0}$ be the augmented sample paths for the k^{th} data, and let T_k be the absorption time for the k^{th} data. Then,

$$D_i = \sum_{k=1}^{M} \mathbf{1}_{\left\{J_0^{(k)} = i\right\}}, \tag{5.19}$$

$$N_{(i,j),\boldsymbol{h}} = \sum_{k=1}^{M} \sum_{t=0}^{T_k} \mathbf{1}_{\left\{N_{\boldsymbol{h}}^{(k)}(t+1) = N_{\boldsymbol{h}}^{(k)}(t)+1, J_t^{(k)} = i, J_{t+1}^{(k)} = j\right\}}, \tag{5.20}$$

$$N_{(i,m+1)} = \sum_{k=1}^{M} \mathbf{1}_{\left\{J_{T_k-1}^{(k)} = i\right\}}. \tag{5.21}$$

Therefore, in order to complete the data augmentation step, one would need to generate the latent Markov process and batch arrival process, namely $\{J_t\}_{t\geq 0}$ and $\{N_{\boldsymbol{h}}(t)\}_{t\geq 0}, \forall \boldsymbol{h} \in \mathcal{C}_0$. Algorithms 15, 16 and 17 presented in the following subsections will achieve this goal.

5.4.2.1 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} = \boldsymbol{y})$

Algorithm 15 presents the data augmentation algorithm for full observations $\boldsymbol{y}$.

Algorithm 15 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} = \boldsymbol{y})$

Output: $\{J_t\}_{t\geq 0}, \{N_{\boldsymbol{h}}(t)\}_{t\geq 0}, \forall \boldsymbol{h} \in \mathcal{C}_0$.

1: **Initialization** Set $t = 0$, and set $N_{\boldsymbol{h}}(t) = 0$, $\forall \boldsymbol{h} \in \mathcal{C}_0$.
2: Simulate $J_t = i$ from $p(J_t|\boldsymbol{Y} = \boldsymbol{y})$.
3: **if** $\boldsymbol{y} = \boldsymbol{0}$ **then**
4: Go to Step 17.
5: **end if**
6: Simulate $(J_{t+1} = j,\ I)$ from $p(J_{t+1} = j, I|\boldsymbol{Y} = \boldsymbol{y}, J_t = i)$, where $I \in \{\boldsymbol{h}|\boldsymbol{y} \geq \boldsymbol{h}\} \cup \{\phi\}$.
7: **if** $I = \phi$ **then**
8: $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$,
9: **else if** $I = \boldsymbol{h^*}$ **then**
10: $N_{\boldsymbol{h^*}}(t+1) = N_{\boldsymbol{h^*}}(t) + 1$.
11: $N_{\boldsymbol{u}}(t+1) = N_{\boldsymbol{u}}(t), \forall \boldsymbol{u} \in \mathcal{C}_0 \setminus \{\boldsymbol{h^*}\}$.
12: $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{h^*}$.
13: **end if**
14: $i = j$.
15: Delete j.
16: Go to Step 3.
17: Simulate $J_{t+1} = j$ from $p(J_{t+1} = j|\boldsymbol{Y} = \boldsymbol{0}, J_t = i)$.
18: **if** $j = m + 1$ **then**
19: **end** the algorithm.
20: **end if**
21: $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$.
22: $t = t + 1$; $i = j$.
23: Go to Step 17.

The followings are specific features of Algorithm 15 which are analogous to the ECS algorithm for full observations (Algorithm 6).

(i) Step 2 is analogous to Step 1 of Algorithm 6. That is, simulating the initial state of the sample path.

(ii) Steps 6 and 17 are analogous to Step 2 of Algorithm 6. That is, simulating the next state conditioning on the previous information.

(iii) Steps 7–12 are analogous to Step 3 of Algorithm 6. That is, subtracting the simulated data from the original data. Then, the simulated data is accumulated forward as the algorithm proceeds and less data will remain until all the data is used up, at which point the algorithm will end.

As the DMPTM involves both the underlying Markov process and the batch arrival process, what is unique in the case of the DMPTM is that two types of sample paths need to be simulated: $\{J_t\}_{t\geq 0}$ and $\{N_{\boldsymbol{h}}(t)\}_{t\geq 0}$.

5.4.2.2 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}})$

Consider the data where the last $K - L$ data are right-censored. That is, $\boldsymbol{Y} = \left(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}\right)$, where $|\boldsymbol{y^{(1)}}| = L$ and $|\boldsymbol{y^{(2)}}| = K - L$. The data augmentation algorithm for partially right-censored data $\boldsymbol{y}$ is given in Algorithm 16:

Algorithm 16 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}})$

Output: $\{J_t\}_{t\geq 0}, \{N_{\boldsymbol{h}}(t)\}_{t\geq 0}, \forall \boldsymbol{h} \in \mathcal{C}_0$.

Initialization Set $t = 0$, and set $N_{\boldsymbol{h}}(t) = 0$, $\forall \boldsymbol{h} \in \mathcal{C}_0$.
Simulate $J_t = i$ from $p(\cdot|\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}})$.
if $(\boldsymbol{y_1}, \boldsymbol{y_2}) = \boldsymbol{0}$ **then**
  Go to Step 17.
end if
Simulate $(J_{t+1} = j, I)$ from $p(J_{t+1} = j, I|\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}, J_t = i)$, where $I \in \{\boldsymbol{h}|\boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}\} \cup \{\phi\}$.
if $I = \phi$ **then**
  $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$.
else if $I = \boldsymbol{h^*} = (\boldsymbol{h^{(1)*}}, \boldsymbol{h^{(2)*}})$ **then**
  $N_{\boldsymbol{h^*}}(t+1) = N_{\boldsymbol{h^*}}(t) + 1$.
  $N_{\boldsymbol{u}}(t+1) = N_{\boldsymbol{u}}(t), \forall \boldsymbol{u} \in \mathcal{C}_0 \setminus \{\boldsymbol{h^*}\}$.
  $\boldsymbol{y^{(1)}} = \boldsymbol{y^{(1)}} - \boldsymbol{h^{(1)*}}$; $\boldsymbol{y^{(2)}} = (\boldsymbol{y^{(2)}} - \boldsymbol{h^{(2)*}})^+$.
end if
$i = j$.
Delete j.
Go to Step 3.
Simulate $(J_{t+1} = j, I)$ from $p(J_{t+1} = j, I|\boldsymbol{Y^{(1)}} = \boldsymbol{0}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0}, J_t = i)$, where $I \in \{\boldsymbol{h}|\boldsymbol{h^{(1)}} = \boldsymbol{0}\} \cup \{\phi\}$.
if $j = m + 1$ **then**
  end the algorithm.
end if
if $I = \phi$ **then**
  $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$.
else if $I = \boldsymbol{h^*} = (\boldsymbol{0^{(1)*}}, \boldsymbol{h^{(2)*}})$ **then**
  $N_{\boldsymbol{h^*}}(t+1) = N_{\boldsymbol{h^*}}(t) + 1$.
  $N_{\boldsymbol{u}}(t+1) = N_{\boldsymbol{u}}(t), \forall \boldsymbol{u} \in \mathcal{C}_0 \setminus \{\boldsymbol{h^*}\}$.
  $\boldsymbol{y^{(1)}} = \boldsymbol{0^{(1)}}$; $\boldsymbol{y^{(2)}} = (\boldsymbol{0^{(2)}} - \boldsymbol{h^{(2)*}})^+ = \boldsymbol{0^{(2)}}$.
end if
$t = t + 1$; $i = j$.
Go to Step 17.

The main idea of the ECS algorithm for right-censored data (Algorithm 7) is applicable to specific steps in Algorithms 16 and 17 as well. As they are quite similar, they will be omitted.

5.4.2.3 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} \geq \boldsymbol{y})$

Consider the data where all the data are right-censored. The data augmentation algorithm for fully right-censored data $\boldsymbol{y}$ is given in Algorithm 17:

Algorithm 17 Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} \geq \boldsymbol{y})$

Output: $\{J_t\}_{t\geq 0}, \{N_{\boldsymbol{h}}(t)\}_{t\geq 0}, \forall \boldsymbol{h} \in \mathcal{C}_0$.

Initialization Set $t = 0$, and set $N_{\boldsymbol{h}}(t) = 0$, $\forall \boldsymbol{h} \in \mathcal{C}_0$.
Simulate $J_t = i$ from $p(\cdot|\boldsymbol{Y} \geq \boldsymbol{y})$.
if $\boldsymbol{y} = \boldsymbol{0}$ **then**
    Go to Step 17.
end if
Simulate $(J_{t+1} = j, I)$ from $p(J_{t+1} = j, I|\boldsymbol{Y} \geq \boldsymbol{y}, J_t = i)$, where $I \in \{\phi, \mathcal{C}_0\}$.
if $I = \phi$ **then**
    $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$.
else if $I = \boldsymbol{h^*}$ **then**
    $N_{\boldsymbol{h^*}}(t+1) = N_{\boldsymbol{h^*}}(t) + 1$.
    $N_{\boldsymbol{u}}(t+1) = N_{\boldsymbol{u}}(t), \forall \boldsymbol{u} \in \mathcal{C}_0 \setminus \{\boldsymbol{h^*}\}$.
    $\boldsymbol{y} = (\boldsymbol{y} - \boldsymbol{h^*})^+$.
end if
$i = j$.
Delete j.
Go to Step 3.
Simulate $(J_{t+1} = j, I)$ from $p(J_{t+1} = j, I|\boldsymbol{Y} \geq \boldsymbol{0}, J_t = i)$, where $I := \{\phi, \mathcal{C}_0\}$.
if $j = m + 1$ **then**
    end the algorithm.
end if
if $I = \phi$ **then**
    $N_{\boldsymbol{h}}(t+1) = N_{\boldsymbol{h}}(t), \forall \boldsymbol{h} \in \mathcal{C}_0$.
else if $I = \boldsymbol{h^*}$ **then**
    $N_{\boldsymbol{h^*}}(t+1) = N_{\boldsymbol{h^*}}(t) + 1$.
    $N_{\boldsymbol{u}}(t+1) = N_{\boldsymbol{u}}(t), \forall \boldsymbol{u} \in \mathcal{C}_0 \setminus \{\boldsymbol{h^*}\}$.
    $\boldsymbol{y} = (\boldsymbol{0} - \boldsymbol{h^*})^+ = \boldsymbol{0}$.
end if
$t = t + 1$; $i = j$.
Go to Step 17.

5.4.3 The MCMC algorithm for the DMPTM

With the building blocks provided in Algorithms 15, 16 and 17, Algorithm 18 presents the MCMC algorithm for inference on the DMPTM using data cloning, which is the principal contribution of this chapter. Since the DMPTM is discrete, one does not need to sample sojourn times, which requires computing matrix exponentials. This fact, combined with the recursive relationships specified in Results 4, 5 and 6, will make the implementation of Algorithm 18 more efficient than Algorithm 12 for the PTAM.

Algorithm 18 The MCMC algorithm for inference on the DMPTM using data cloning

Require: m, $\mathcal{C}_0$, the number of data cloning w_1, the number of iterations w_2.
Input:

1. The data observations $\boldsymbol{y}$.
2. The hyper-parameters for posterior distributions: $\boldsymbol{v}, \boldsymbol{\eta}$.
3. The number of states m.
4. The number of MCMC iterations w_2.
5. The size of burn-in period.

Output: The posterior samples for the model parameters $\boldsymbol{\beta}, \boldsymbol{B_h}, \boldsymbol{b_0}$ where $\boldsymbol{h} \in \mathcal{C}_0$.

Initialization $\left(\boldsymbol{\beta^{(1)}}, \boldsymbol{B_h^{(1)}}, \boldsymbol{b_0^{(1)}}\right)$, where $\boldsymbol{h} \in \mathcal{C}_0$.
Repeat the data w_1 times. Let the cloned data be $\boldsymbol{y^c}$.
for $k = 2 : w_2$ **do**
    Calculate all possible $\boldsymbol{p_Y}(\cdot)$ based on recursive relationships in Result 4. If there are right-censored data, then calculate all possible $\boldsymbol{p_{Y,\geq}^{(0,1)}}(\cdot)$ and $\boldsymbol{p_{Y,\geq}^{(1,1)}}(\cdot)$ based on recursive relationships in Results 5 and 6, respectively.
    Sample $\boldsymbol{x} := \left(\{J_t\}_{t\geq 0}, \{N_{\boldsymbol{h}(t)}\}_{t\geq 0}\right)$ from $p\left(\boldsymbol{x}\middle|\boldsymbol{\beta^{(k-1)}}, \boldsymbol{B^{(k-1)}}, \boldsymbol{b_0^{(k-1)}}, \boldsymbol{y^c}\right)$, based on Algorithm 15. If there are right-censored data, then Algorithms 16 and 17 will also be used.
    Obtain D_i and $N_{(i,j),\boldsymbol{h}}$ for all $i, j, \boldsymbol{h}$.
    Sample $\left(\boldsymbol{\beta^{(k)}}, \boldsymbol{B^{(k)}}, \boldsymbol{b_0^{(k)}}\right)$ from $p\left(\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}\middle|\boldsymbol{x}, \boldsymbol{y^c}\right)$.
end for

5.5 Simulation studies

As pointed out earlier, the proposed approach has two components: the MCMC-based Bayesian inference and the data cloning. Thus, we will consider two simulation studies which are presented in Sections 5.5.1 and 5.5.2 to validate each component, respectively.

5.5.1 Example 2.1 in He and Ren (2016b)

Example 2.1 in He and Ren (2016b) provides one of the simplest form of the DMPTM, where $K = 2$, $|\mathcal{C}_0| = 3$, $\mathcal{C}_0 = \{\boldsymbol{h_1} = (1,0), \boldsymbol{h_2} = (0,1), \boldsymbol{h_3} = (1,1)\}$, $m = 1$, $\boldsymbol{\beta} = 1$, $B_0 = 0$, $B_{\boldsymbol{h_1}} > 0$, $B_{\boldsymbol{h_2}} > 0$, $B_{\boldsymbol{h_3}} > 0$ and $b_0 = 1 - B_{\boldsymbol{h_1}} - B_{\boldsymbol{h_2}} - B_{\boldsymbol{h_3}} > 0$. The DMPTM of this type is also called the trivariate geometric distribution.

We now perform a simulation study on this example. Data are simulated from a pre-specified underlying trivariate geometric distribution based on Algorithm 13. Then, the

proposed MCMC algorithm is applied to estimate the parameters which are then compared with the true underlying values.

Consider following experimental conditions:

- Let the underlying model be the trivariate geometric distribution with $B_{\boldsymbol{h_1}} = 0.3$, $B_{\boldsymbol{h_2}} = 0.2$, $B_{\boldsymbol{h_3}} = 0.1$ and $b_0 = 1 - B_{\boldsymbol{h_1}} - B_{\boldsymbol{h_2}} - B_{\boldsymbol{h_3}} = 0.4$.
- 500 sample points are simulated from the underlying model. They are not cloned in this example
- The conjugate prior distributions of the parameters is

$$(B_{\boldsymbol{h_1}}, B_{\boldsymbol{h_2}}, B_{\boldsymbol{h_3}}, b_0) \sim Dirichlet(1, 3, 5, 0.5).$$

  The prior is deliberately chosen to be informative yet false, similar to experiments in Lele et al. (2007, 2010). The purpose of this choice is to exhibit more significantly the effect of Bayesian inference.

- 10000 samples are simulated with the MCMC method.
- The first 1000 iterations are taken as burn-in, based on the CSD plot.
- A thinning rate of 20 is adopted, based on the ACFs.

The results are included in Table 5.1 and Figure 5.1:

Parameter	True	Posterior Mean	95% Credible Interval
$B_{\boldsymbol{h_1}}$	0.3	0.2922798	(0.2588544, 0.3238664)
$B_{\boldsymbol{h_2}}$	0.2	0.1820329	(0.1525789, 0.2119733)
$B_{\boldsymbol{h_3}}$	0.1	0.1097666	(0.07463419, 0.14450728)
b_0	0.4	0.4159207	(0.3836012, 0.4460370)

Table 5.1: Bayesian estimates and credible intervals for the trivariate geometric model.

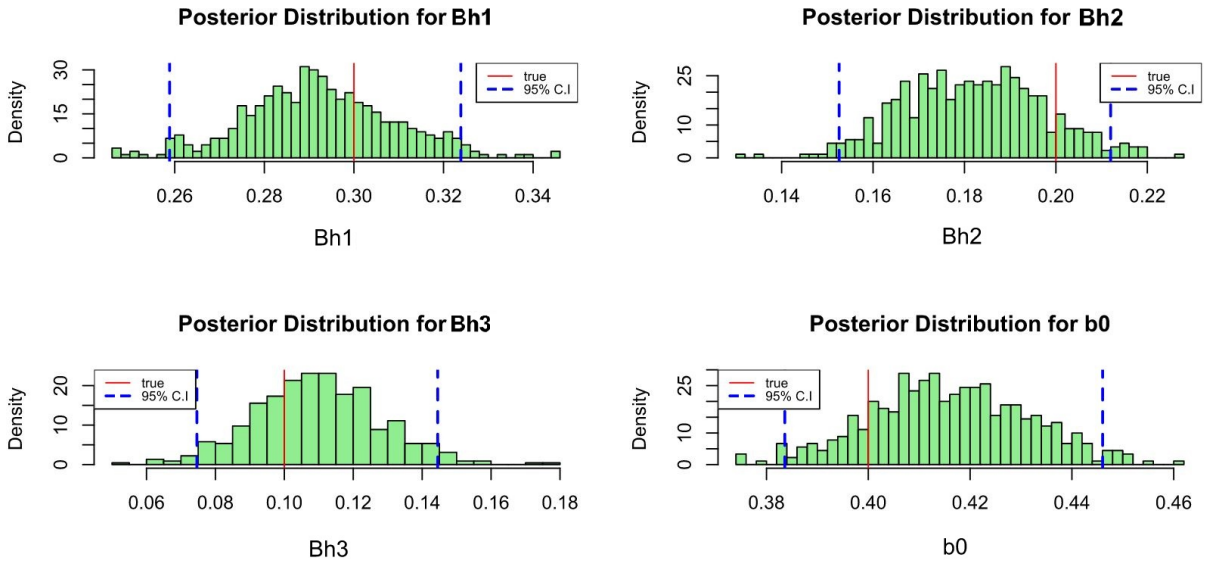

Figure 5.1: Posterior distributions for the trivariate geometric model.

In Table 5.1 and Figure 5.1, the Bayesian estimates, taken as the posterior means, are all within their corresponding 95% credible intervals. This is the initial indication that the proposed MCMC algorithm for Bayesian inference is quite satisfactory. In addition, the posterior distributions in Figure 5.1 are all bell-shaped, which indicates a desirable degree of estimability.

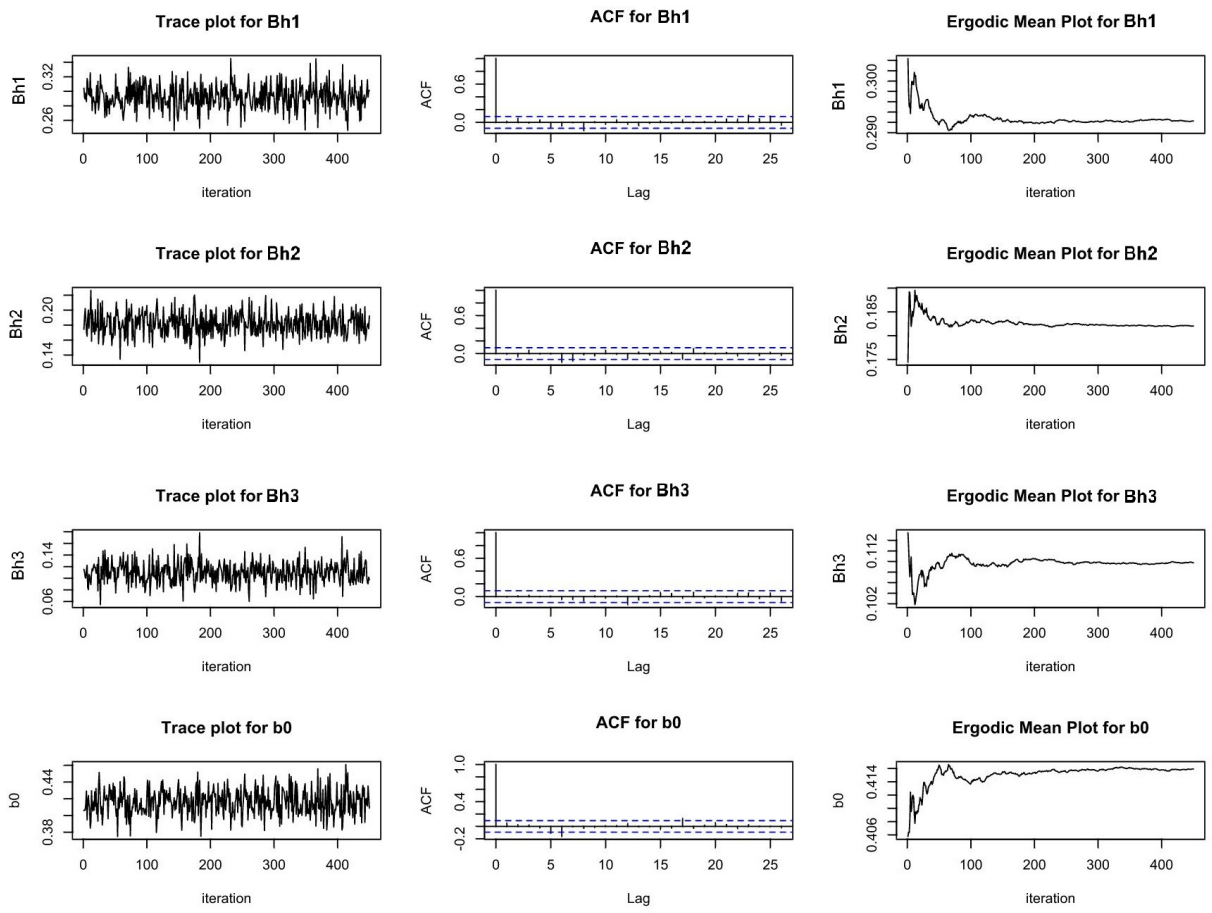

Figure 5.2: Diagnostic plots of the MCMC samples.

In Figure 5.2, the convergence of the proposed MCMC algorithm is being examined by the trace plot, the ACFs and the ergodic mean plot. First, the trace plots all point to stationary and level-off patterns of the MCMC samples. Secondly, the ACFs for all parameters are within the tolerance range after the second lag. This demonstrates that the thinning rate significantly reduces the ACFs between the MCMC samples. Thirdly, the ergodic means all start to converge as iteration increases. This indicates that the number of iterations, that is, 10000, is sufficient to intimate that the MCMC samples that were obtained, were approximately generated from the target posterior distributions.

5.5.2 Example 2.1 in He and Ren (2016b) with data cloning

Following Section 5.5.1, consider another simulation study with the following experimental conditions:

- 20 sample points are simulated from the underlying model. The sample size is deliberately chosen to be small. That way, the effect of data cloning on the posterior distributions is more noticeable.
- Data are cloned 0 (status quo), 10 and 50 times.
- All other conditions remain identical to those assumed in the simulation study carried out in Section 5.5.1.

The results are included in Figure 5.3. It can be seen that, without data cloning, the posterior distributions deviate markedly from the MLEs. This is due to the fact that the prior is chosen to be falsely informative and the data size is small, in which case the shape of the posterior density is still dominated by the shape of the prior density. However, as the number of cloned data increases, the shape of likelihood function starts to dominate the shape of posterior density, tilting it towards the MLEs. This is particularly obvious for $B_{\boldsymbol{h_1}}$ and $B_{\boldsymbol{h_3}}$. This validates the theory of data cloning as applied to the DMPTM.

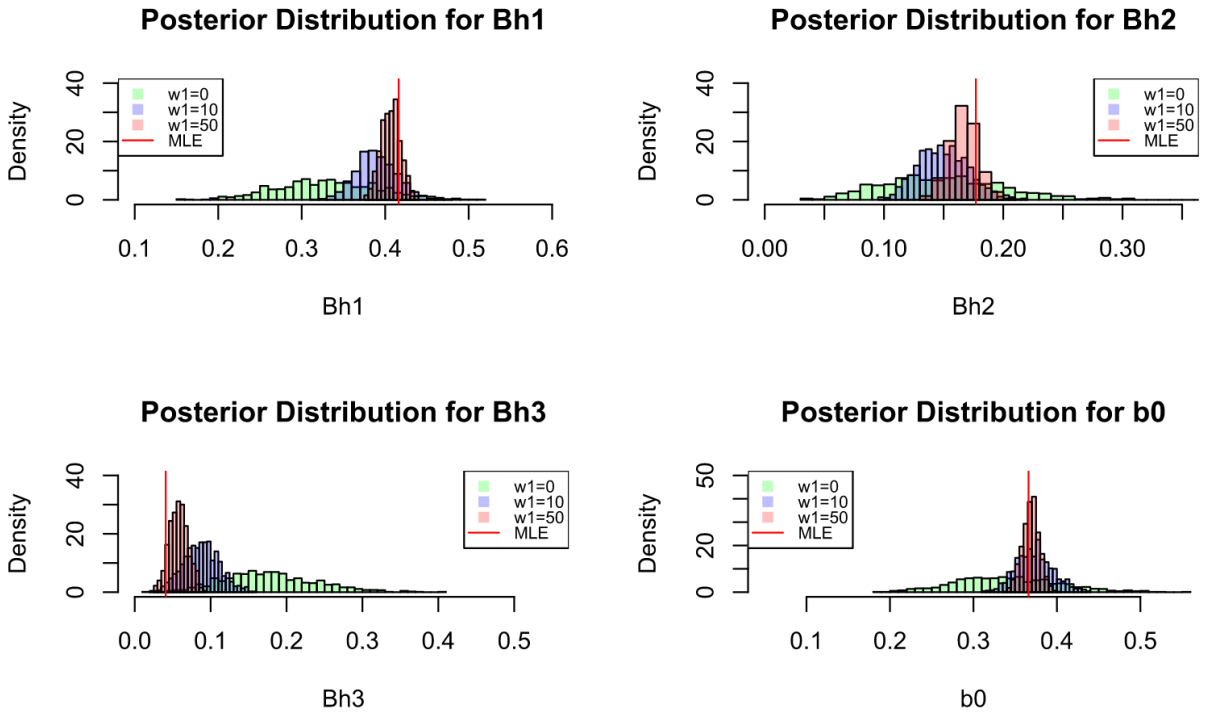

Figure 5.3: Comparison between the MLEs and the posterior distributions for the trivariate geometric model using data cloning. The numbers of times data is cloned denoted by w_1 are 0, 10 and 50.

Another crucial aspect to point out based on Figure 5.3 is that increasing the number of cloned data can only make the stochastic approximation closer to the MLEs, which decreases the algorithm noise of the MCMC algorithm. However, the MLEs themselves do not depend on data cloning as they relate to the data noise. They can only be improved upon if more data (or more information) is secured. This agrees beautifully with the contributions discussed in Section 3.4.3.

It is also worth emphasizing that the precondition of comparing the parameter estimates with the true parameter is that the underlying model must be identifiable, which is why we chose the trivariate geometric distribution in this section. However, the DMPTM with $m > 1$ will suffer from the non-identifiability issue such as in the experiment to be presented in the next section. In that case, comparing parameters will be meaningless. Instead, one should compare the prediction results.

5.6 Data analysis

In the simulation studies presented in Section 5.5, we have validated the proposed MCMC algorithm for the DMPTM using data cloning. In this section, we will apply the proposed approach to fit the DMPTM to real-life data. Following the research carried out in He and Ren (2016b), we continue to investigate the auto insurance property damage and bodily injury claim data discussed in Cummins and Wiltbank (1983). The data are listed in Table 5.2 with $K = 2$ types of claims.

	Property damage events					
	(y_1, y_2)	0	1	2	$\geq$3	Totals
Bodily injury events	0	44	49	2	1	96
	1	10	20	2	1	33
	2	2	6	1	1	10
	$\geq$3	0	4	5	1	10
	Totals	56	79	10	4	149

Table 5.2: Property damage data in Cummins and Wiltbank (1983).

As assumed in He and Ren (2016b), there are $|\mathcal{C}_0| = 4$ batches, that is,

$$\mathcal{C}_0 = \{\boldsymbol{h_1} = (1,0), \boldsymbol{h_2} = (0,1), \boldsymbol{h_3} = (1,1), \boldsymbol{h_4} = (1,2)\}.$$

This implies that the possible combinations of claims are:

- an accident causing one property damage claim only,
- an accident causing one bodily injury claim only,
- an accident causing one property damage claim and one bodily injury claim,
- an accident causing one property damage claim and two bodily injury claims.

The matrices $\boldsymbol{B_0}$ and $\boldsymbol{B_h}$ were assumed to have full entries, in which case the DMPTM will have $m - 1 + (|\mathcal{C}_0| + 1)m^2$ parameters to be estimated. The data is cloned 100 times and the prior is chosen arbitrarily. Finally, after the fitted DMPTM is obtained using the proposed approach, the prediction results are compared with the EM approach proposed in He and Ren (2016b). The results are presented in Tables 5.3, 5.4, 5.5, 5.6 and 5.7.

			Property damage events				
			0	1	2	≥3	m
Bodily injury events	0	Data	(44)	(49)	(2)	(1)	
		EM	66.153227	21.202277	6.795384	3.205218	1
		MCMC	64.232440	20.160677	6.327844	2.894679	1
	1	Data	(10)	(20)	(2)	(1)	
		EM	15.491349	9.979265	4.845027	3.428309	1
		MCMC	15.332570	10.672893	5.190542	3.601593	1
	2	Data	(2)	(6)	(1)	(1)	
		EM	3.627667	3.511087	2.276583	2.280428	1
		MCMC	3.659953	3.946583	2.652148	2.675831	1
	≥3	Data	(0)	(4)	(5)	(1)	
		EM	1.109265	1.542454	1.350268	2.202193	1
		MCMC	1.147579	1.813615	1.702878	2.988172	1

Table 5.3: Prediction results for the auto insurance claim data when $m = 1$ - EM vs MCMC with data cloned 100 times.

			Property damage events				
			0	1	2	≥3	m
Bodily injury events	0	Data	(44)	(49)	(2)	(1)	
		EM	44.6210500	45.697456	1.598340	0.9100688	2
		MCMC	43.5525635	45.648331	1.696490	0.7812590	2
	1	Data	(10)	(20)	(2)	(1)	
		EM	8.8677200	22.238564	3.182727	1.0731121	2
		MCMC	9.7269901	22.661608	3.507644	1.1024578	2
	2	Data	(2)	(6)	(1)	(1)	
		EM	1.7866334	8.049978	2.434963	0.8776059	2
		MCMC	2.1904608	7.498144	2.622216	0.9434692	2
	≥3	Data	(0)	(4)	(5)	(1)	
		EM	0.4628932	3.653967	2.261937	1.2829853	2
		MCMC	0.6416705	2.976259	2.141847	1.3085899	2

Table 5.4: Prediction results for the auto insurance claim data when $m = 2$ - EM vs MCMC with data cloned 100 times.

			Property damage events				
			0	1	2	≥3	m
Bodily injury events	0	Data	(44)	(49)	(2)	(1)	
		EM	44.8169137	49.1444185	2.0426025	0.8733413	3
		MCMC	44.0650724	49.106799	2.054026	0.8344180	3
	1	Data	(10)	(20)	(2)	(1)	
		EM	8.9692214	20.1817881	1.8492699	1.1942515	3
		MCMC	9.8338358	20.538636	1.861705	1.0203017	3
	2	Data	(2)	(6)	(1)	(1)	
		EM	1.7950128	5.9619534	1.0210211	0.7024357	3
		MCMC	1.7110203	5.319394	1.713639	0.7155012	3
	≥3	Data	(0)	(4)	(5)	(1)	
		EM	0.4491188	3.9719636	4.6113153	1.4153726	3
		MCMC	0.3791378	4.180922	3.944238	1.7213526	3

Table 5.5: Prediction results for the auto insurance claim data when $m = 3$ - EM vs MCMC with data cloned 100 times.

			Property damage events				
			0	1	2	≥3	m
Bodily injury events	0	Data	(44)	(49)	(2)	(1)	
		EM	44.0032477	49.0950878	1.9765702	0.9458369	4
		MCMC	44.1232817	48.500675	1.987970	0.8602734	4
	1	Data	(10)	(20)	(2)	(1)	
		EM	10.0440489	20.0356399	2.0202334	1.0235015	4
		MCMC	10.1761654	19.758579	2.266369	1.0812184	4
	2	Data	(2)	(6)	(1)	(1)	
		EM	1.9513996	5.9933529	0.9916737	0.9066508	4
		MCMC	1.8085572	5.448871	1.717686	0.8067970	4
	≥3	Data	(0)	(4)	(5)	(1)	
		EM	0.0000015	3.9758322	4.8757383	1.1611848	4
		MCMC	0.3374578	4.074688	4.303132	1.7482812	4

Table 5.6: Prediction results for the auto insurance claim data when $m = 4$ - EM vs MCMC with data cloned 100 times.

m	Methods	Log-likelihood	AIC
1	EM	-312.9074	635.8148
	MCMC	-312.3437	634.6874
2	EM	-280.1488	602.2976
	MCMC	-280.7239	603.4478
3	EM	-277.6401	649.2802
	MCMC	-276.9631	647.9262
4	EM	-276.9631	719.9262
	MCMC	-277.7945	721.5890

Table 5.7: Log-likelihood and AIC for the fitted DMPTM with $m = 1, 2, 3$ and 4 using the EM and the proposed MCMC algorithms.

It can be seen from Tables 5.3, 5.4, 5.5 and 5.6 that the prediction results obtained from the proposed MCMC approach and the EM approach are in close agreement for $m = 1, 2, 3$ and 4, which validates the proposed approach as an alternative way of determining MLEs. Moreover, based on the AIC values included in Table 5.7, we may arrive at the conclusion reached by He and Ren (2016b) to the effect that the DMPTM with $m = 2$ turns out to be the best model which produces the lowest AIC. This further validates the proposed MCMC algorithm with data cloning as an alternative method to obtain the MLEs of the DMPTM.

Then, we may utilize the fitted DMPTM with $m = 2$ as specified below:

$$\boldsymbol{\beta} = (0.6769916, 0.3230084)', \ \boldsymbol{B_{h_1}} = \begin{bmatrix} 0.1838569 & 0.009865069 \\ 0.0560322 & 0.143507819 \end{bmatrix},$$
$$\boldsymbol{B_{h_2}} = \begin{bmatrix} 0.005560552 & 0.5929008383 \\ 0.020528796 & 0.0001368295 \end{bmatrix}, \ \boldsymbol{B_{h_3}} = \begin{bmatrix} 0.001467756 & 0.04908077 \\ 0.001217004 & 0.00178014 \end{bmatrix},$$
$$\boldsymbol{B_{h_4}} = \begin{bmatrix} 0.0006199654 & 0.0004056259 \\ 0.0004124564 & 0.0003874211 \end{bmatrix}, \ \boldsymbol{b_0} = \begin{bmatrix} 0.02939067 \\ 0.60564915 \end{bmatrix}.$$

This fitted DMPTM is quite distinct from that obtained in He and Ren (2016b). This is due the fact that for $m > 1$, the DMPTM with full parameters is non-identifiable. Actually, the DMPTM is very likely to suffer from the non-identifiability issue. This is discussed in the next section.

5.7 Discussion

5.7.1 Identifiability of the DMPTM

At this moment, we have not yet fully investigated the identifiability of the DMPTM. However, based on preliminary results, it is certain that the DMPTM with full parameters will be beset by the non-identifiability issue for $m > 1$. In particular, one may simply rearrange the parameters and relabel the states in a consistent manner so that the overall process will remain unchanged. To verify this, consider the following simple example. Let the batch set be $\mathcal{C}_0 = \{\boldsymbol{h_1}, \boldsymbol{h_2}\}$, $m = 2$ and $\boldsymbol{B_0} = \boldsymbol{0}$; then two DMPTMs, namely, $DMPTM(\boldsymbol{\beta^{(1)}}, \boldsymbol{B_0^{(1)}} = \boldsymbol{0}, \boldsymbol{B_h^{(1)}}, \mathcal{C}_0)$ and $DMPTM(\boldsymbol{\beta^{(2)}}, \boldsymbol{B_0^{(2)}} = \boldsymbol{0}, \boldsymbol{B_h^{(2)}}, \mathcal{C}_0)$, will be non-identifiable if

$$\boldsymbol{\beta^{(1)}} = (0.3, 0.7)', \ \boldsymbol{B_{h_1}^{(1)}} = \begin{bmatrix} 0.04 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}, \ \boldsymbol{B_{h_2}^{(1)}} = \begin{bmatrix} 0.3 & 0.4 \\ 0.05 & 0.2 \end{bmatrix}, \ \boldsymbol{b_0^{(1)}} = \begin{bmatrix} 0.06 \\ 0.35 \end{bmatrix}, \tag{5.22}$$
$$\boldsymbol{\beta^{(2)}} = (0.7, 0.3)', \ \boldsymbol{B_{h_1}^{(2)}} = \begin{bmatrix} 0.3 & 0.1 \\ 0.2 & 0.04 \end{bmatrix}, \ \boldsymbol{B_{h_2}^{(2)}} = \begin{bmatrix} 0.2 & 0.05 \\ 0.4 & 0.3 \end{bmatrix}, \ \boldsymbol{b_0^{(2)}} = \begin{bmatrix} 0.35 \\ 0.06 \end{bmatrix}. \tag{5.23}$$

However, as the objective is to eventually fit the DMPTM to real-life data and to achieve sound prediction results, then the non-identifiability issue of the DMPTM does not seem to be the most crucial problem to tackle. Nonetheless, the identifiability of the DMPTM constitutes a worthwhile topic to investigate further.

5.7.2 Applicability

Unlike the case of univariate phase-type distributions, the observed data $\boldsymbol{Y} = \boldsymbol{y}$ of the DMPTM does provide at least some clues about the latent information $\boldsymbol{X_h}$, although they might not be sufficiently compelling for us to fully determine $\boldsymbol{X_h}$. To illustrate this, consider the following simple example. Let an underlying DMPTM have batches $\boldsymbol{h_1} = (2, 0)$ and $\boldsymbol{h_2} = (1, 1)$. Then, there is no possible way to generate the data point $\boldsymbol{y} = (3, 0)$, which eliminates the possibility of using this DMPTM when the data includes $(3, 0)$. In other words, for a specific underlying DMPTM with its batch types, the model cannot be utilized to arbitrarily fit any given data points. Thus, if we are given the observed data, then the first step must be to determine an eligible batch set $\mathcal{C}_0$. This might require subjective assessments or prior knowledge depending on the context of the experiment.

5.8 Conclusion

An MCMC algorithm was proposed for inference on the DMPTM using data cloning. The algorithm was constructed on the basis of existing MCMC algorithms for Bayesian inference on continuous phase-type distributions. While the known EM algorithm yields MLEs for the DMPTM based on a deterministic approximation, the proposed algorithm provides an alternative way to obtain the MLEs of the DMPTM based on a stochastic approximation, which directly contributes to the field of approximate inference. Two simulation studies validated the proposed MCMC algorithm combined with data cloning as applied to the DMPTM. Then, the proposed MCMC algorithm with data cloning was applied to calibrate the DMPTM to a real-life insurance claim frequency data set. It was observed that, once enhanced with data cloning, the proposed MCMC algorithm produced results that are as sound as those secured with the existing EM algorithm.

Chapter 6

Summary and Future Research Topics

6.1 Summary

In this thesis, we made contributions in two areas. In one area, the concept of estimability was objectively defined in the context of statistical models. In the other area, the Markov chain Monte Carlo procedure was applied to the PTAM and the DMPTM.

In Chapter 2, the identifiability of the PTAM was investigated, which paved the way for investigating its estimability. In Chapter 3, a novel definition of estimability was introduced to objectively quantify estimability for statistical models, and more particularly that of the PTAM, in order to solve the non-estimability issue. In Chapter 4, an MCMC algorithm was developed for Bayesian inference on the PTAM, which improved estimability via sound prior information. The algorithm was also utilized as a standalone model fitting technique. In Chapter 5, an MCMC algorithm combined with data cloning was developed for inference on the DMPTM. The algorithm which provides an alternative approach to determining MLEs achieved model fitting results that were comparable to those obtained by applying the EM algorithm.

Here are principal research contributions of this thesis:

(i) It is established that the PTAM is identifiable when the number of states is greater or equal to six; it is otherwise possibly non-identifiable.

(ii) Unlike existing methods utilized for estimability assessments which require a subjectively specified threshold, the proposed definition of estimability is objective. This objectivity is achieved via a carefully designed c.d.f. sensitivity measure which relates the confidence region to the experimental error. Under that setting, the threshold becomes objective as the experimental error becomes an experiment-based quantity. The proposed definition not only solves the issue of subjective thresholds in existing methods, but also extends the concept of estimability in the context of statistical models.

(iii) An MCMC algorithm is developed for Bayesian inference on the PTAM. The proposed method provides two methodological extensions based on an existing MCMC inference method. First, a two-level MCMC sampling scheme is proposed to make the method applicable to situations where the posterior distributions are complicated after data augmentation. Secondly, the data augmentation technique is further developed in order to incorporate left-truncated data.

(iv) An MCMC algorithm is developed for inference on the DMPTM using data cloning. From the perspective of approximate inference, while the application of a deterministic approximation to the DMPTM has been investigated via the EM algorithm, the stochastic approximation counterpart had remained hitherto unexplored. The proposed MCMC algorithm therefore fills this gap and provides another way of determining the MLEs of the DMPTM based on stochastic approximation.

6.2 Future research topics

The results obtained in Chapter 3 could be further investigated as follows:

(i) Other potential definitions of estimability can be envisaged by utilizing different loss functions involving the ECDF. This might help experimenters investigate the stringency of the proposed definition.

(ii) Other potential definitions of estimability might be put forward by relying on a binary event other than "the confidence region being finite or infinite". This might provide a more realistic criterion regarding the size of the confidence region. As confirmed by Raue et al. (2009), an extremely wide but bounded confidence region is almost as deficient as an infinite region, which makes the proposed definition somewhat less practical. Thus, it is a worthwhile direction to seek other mathematical concepts regarding the size of a multidimensional region. In that case, the criterion can be improved upon while retaining its objectivity.

(iii) The application of the proposed definition can be extended to density approximation problems. In that case, the experimental error will be interpreted as an objective threshold that determines how well the proposed density approximates the target density.

Possible future research undertakings that are related to Chapter 4 are suggested below.

(i) RJMCMC can be applied to make inferences on the number of states of phase-type distributions, which was suggested by Bladt et al. (2003). To the best of our knowledge, treating the number of states as one of the parameters of phase-type distributions has seldom been considered in the literature. However, since the number of states conveys biological meanings in the context of the PTAM, this aspect deserves attention.

(ii) Finding efficient ways of calculating large matrix exponentials has been an open problem for years. If large matrix exponentials can be calculated more efficiently, this will directly facilitate the application of the proposed MCMC on the PTAM when the number of states is large.

In Chapter 5, a possible future research topic is the identifiability of the DMPTM. Canonical forms such as those presented in Cumani (1982) might be expected. If it turns out that such canonical forms are indeed available for the DMPTM, then one may consider utilizing them rather than resorting to assuming full entries in the transition probability matrices.

Appendix A

Properties of the $h_i's$

A.1 Preliminaries

Proposition A.1.1. *Let the sequence* $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ *be generated by (2.4); then any sub-sequence of* $\boldsymbol{h}$ *with* n *elements* $\{h_a, h_{a+1}, \ldots, h_{a+n-1}\}$, *where* $1 \leq a < a+n-1 \leq m$ *and* $n \geq 3$, *can be described by the same structures as (2.4) with*

$$h_1^* = h_a, \tag{A.1}$$
$$h_n^* = h_{a+n-1}, \tag{A.2}$$
$$s* = s. \tag{A.3}$$

Proof. It suffices to prove that $h_j^* = h_{a-1+j},\ j = 1, 2, \ldots, n$. When $s^* = 0$, we have

$$\begin{aligned} h_j^* &= {h_1^*}^{\frac{n-j}{n-1}} {h_n^*}^{\frac{j-1}{n-1}} \\ &= h_a^{\frac{n-j}{n-1}} h_{a+n-1}^{\frac{j-1}{n-1}} \\ &= \left(h_1^{\frac{m-a}{m-1}} h_m^{\frac{a-1}{m-1}}\right)^{\frac{n-j}{n-1}} \left(h_1^{\frac{m-a-n+1}{m-1}} h_m^{\frac{a+n-2}{m-1}}\right)^{\frac{j-1}{n-1}} \\ &= h_1^{\frac{m-a-j+1}{m-1}} h_m^{\frac{a+j-2}{m-1}} \\ &= h_{a-1+j}. \end{aligned} \tag{A.4}$$

When $s^* \neq 0$, we have

$$\begin{aligned}
h_j^* &= \left(\frac{n-j}{n-1}h_1^{*s^*} + \frac{j-1}{n-1}h_n^{*s^*}\right)^{\frac{1}{s^*}} \\
&= \left(\frac{n-j}{n-1}h_a^s + \frac{j-1}{n-1}h_{a+n-1}^s\right)^{\frac{1}{s}} \\
&= \left(\frac{n-j}{n-1}\left(\frac{m-a}{m-1}h_1^s + \frac{a-1}{m-1}h_m^s\right) + \frac{j-1}{n-1}\left(\frac{m-a-n+1}{m-1}h_1^s + \frac{a+n-2}{m-1}h_m^s\right)\right)^{\frac{1}{s}} \\
&= \left(\frac{m-a-j+1}{m-1}h_a^s + \frac{a+j-2}{m-1}h_{a+n-1}^s\right)^{\frac{1}{s}} \\
&= h_{a-1+j}. \qquad \text{(A.5)}
\end{aligned}$$

□

The next proposition follows immediately.

Proposition A.1.2. *Let the sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ be generated by (2.4); then any new sequence with extrapolated points of $\boldsymbol{h}$, namely, $\{a_1, a_2, \ldots, a_M, h_1, h_2, \ldots, h_m, b_1, b_2, \ldots, b_N\}$, where $M \geq 1, N \geq 1$, can be described by the same structures as (2.4) with*

Case 1: $s* = s \neq 0$

$$h_1^* = a_1 = \left(\frac{M+m-1}{m-1}h_1^s - \frac{M}{m-1}h_m^s\right)^{\frac{1}{s}}, \qquad \text{(A.6)}$$

$$h_{m+M+N*} = b_N = \left(-\frac{N}{m-1}h_1^s + \frac{N+m-1}{m-1}h_m^s\right)^{\frac{1}{s}}, \qquad \text{(A.7)}$$

Case 2: $s* = s = 0$

$$h_1^* = a_1 = h_1^{\frac{M+m-1}{m-1}}h_m^{-\frac{M}{m-1}}, \qquad \text{(A.8)}$$

$$h_{m+M+N*} = b_N = h_1^{-\frac{N}{m-1}}h_m^{\frac{N+m-1}{m-1}}. \qquad \text{(A.9)}$$

Proof. From Proposition A.1.1, we know that $\boldsymbol{h}$ is a sub-sequence of the larger extrapolated sequence, and we have

$$h_{M+1}^* = h_1, \qquad \text{(A.10)}$$

$$h_{M+m}^* = h_m, \qquad \text{(A.11)}$$

$$s* = s. \qquad \text{(A.12)}$$

We can then solve for a_1 and b_N. When $s \neq 0$, we have

$$h_1^* = a_1 = \left(\frac{M+m-1}{m-1} h_1^s - \frac{M}{m-1} h_m^s \right)^{\frac{1}{s}}, \tag{A.13}$$

$$h_{m+M+N*} = b_N = \left(-\frac{N}{m-1} h_1^s + \frac{N+m-1}{m-1} h_m^s \right)^{\frac{1}{s}}. \tag{A.14}$$

When $s = 0$, we have

$$h_1^* = a_1 = h_1^{\frac{M+m-1}{m-1}} h_m^{-\frac{M}{m-1}}, \tag{A.15}$$

$$h_{m+M+N*} = b_N = h_1^{-\frac{N}{m-1}} h_m^{\frac{N+m-1}{m-1}}. \tag{A.16}$$

□

Therefore, this proposition provides manageable formulas to calculate extrapolated points of $\boldsymbol{h}$. In simple words, the two previous propositions establish that a given sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$, whether being extrapolated or a given subset, the resulting new sequence can be described by the same structure with new starting and ending points and the same s.

Proposition A.1.3. *Let the sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ be generated by (2.4) where $m \geq 6$. If a point h^* is inserted between h_i and h_{i+1} where $i = 1, 2, \ldots, m-1$, then there will be no structure of (2.4) that can describe the new sequence.*

Proof. **Case 1**

Suppose h^* is inserted between h_i and h_{i+1} where $1 \leq i \leq m-3$, then there are at least three points in $h_{i+1}, \ldots, h_m$. Assuming structure of (2.4) can describe the new sequence, then based on propositions A.1.1 and A.1.2, we know that we must have $h^* = h_i$; however, we cannot have $h^* = h_i$ because the sequence described by (2.4) must be increasing. Therefore, there will be no structure of (2.4) that can describe the new sequence $\{h_1, h_2, \ldots, h_i, h^*, h_{i+1}, \ldots, h_m\}$.

Case 2

Similarly, suppose h^* is inserted between h_i and h_{i+1} where $3 \leq i \leq m-1$, then there are at least three points in $h_1, \ldots, h_i$. Assuming structure of (2.4) can describe the new sequence, then based on propositions A.1.1 and A.1.2, we know that we must have $h^* = h_{i+1}$; however, we cannot have $h^* = h_{i+1}$ because the sequence described by (2.4) must be increasing. Therefore, there will be no structure of (2.4) that can describe the new sequence $\{h_1, h_2, \ldots, h_i, h^*, h_{i+1}, \ldots, h_m\}$. □

Proposition A.1.4. *Let the sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ be generated by (2.4) where $m \geq 6$. If a point h_i is removed where $2 \leq i \leq m-1$, then there will be no structure of (2.4) that can describe the new sequence.*

Proof. **Case 1**

Suppose h_i is removed where $2 \leq i \leq m-3$, then there are at least three points in $h_{i+1}, \ldots, h_m$. Assuming structure of (2.4) can describe the new sequence, then based on Propositions A.1.1 and A.1.2, we know that the first to $(i-1)^{th}$ points must be $\{h_2, h_3, \ldots, h_i\}$. However, they are in fact $\{h_1, h_2, \ldots, h_{i-1}\}$, which is certainly different because the sequence described by (2.4) must be increasing. Therefore, there will be no structure of (2.4) that can describe the new sequence $\{h_1, h_2, \ldots, h_{i-1}, h_{i+1}, \ldots, h_m\}$.

Case 2

Similarly, suppose h_i is removed where $4 \leq i \leq m-1$, then there are at least three points in $h_1, ..., h_i$. Assuming structure of (2.4) can describe the new sequence, then based on Propositions A.1.1 and A.1.2, we know that the i^{th} to $(m-1)^{th}$ points must be the set $\{h_i, h_{i+1}, \ldots, h_{m-1}\}$. However, they are in fact $\{h_{i+1}, h_{i+2}, \ldots, h_m\}$, which is surely different because the sequence described by (2.4) must be increasing. Therefore, there will be no structure of (2.4) that can describe the new sequence $\{h_1, h_2, \ldots, h_{i-1}, h_{i+1}, \ldots, h_m\}$. □

A.2 Inserting or removing an element in $\boldsymbol{h}$

Proposition A.2.1. *Let the sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ be generated by (2.4) where $m \geq 6$, if we insert or remove a point, then there will be no structure of (2.4) that can describe the new sequence, provided that the inserted or removed point is not the starting or ending point*

Proof. A combination of Propositions A.1.3 and A.1.4 will establish this result. □

A.3 Vertically shifting $\boldsymbol{h}$

Proposition A.3.1. *Let the sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ be generated by (2.4) where $m \geq 6$ and $s \neq 0$; if we vertically shifted the components by ϵ (where $\epsilon > 0$), then the shifted curve $\{h_1 + \epsilon, h_2 + \epsilon, \ldots, h_m + \epsilon\}$ cannot be described by the structure of (2.4), except for $s = 1$.*

Proof. For the new sequence we obviously need to choose

$$h_1^* = h_1 + \epsilon, \tag{A.17}$$

$$h_m^* = h_m + \epsilon. \tag{A.18}$$

Therefore, it suffices to prove that there is no such s^* that can produce

$$h_i^* = h_i + \epsilon \text{ for all } i. \tag{A.19}$$

Since $s \neq 0$, we have

$$h_i^s = \frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s \tag{A.20}$$

and

$$h_i^{*s*} = \frac{m-i}{m-1} (h_1 + \epsilon)^{s*} + \frac{i-1}{m-1} (h_m + \epsilon)^{s*}. \tag{A.21}$$

Since the two previous equations can be seen as two sequences exhibiting a linear pattern, for convenience, we define

$$a = h_1^s, \tag{A.22}$$

$$d = \frac{h_m^s - h_1^s}{m-1}, \tag{A.23}$$

$$a + \delta = (h_1 + \epsilon)^{s*}, \tag{A.24}$$

$$c = \frac{(h_m + \epsilon)^{s*} - (h_1 + \epsilon)^{s*}}{m-1}. \tag{A.25}$$

Thus, (A.20) and (A.21) become

$$h_i^s = a + (i-1)d, \tag{A.26}$$

$$h_i^{*s*} = a + \delta + (i-1)c, \tag{A.27}$$

and the problem then becomes determining whether one can find an $s*$ such that

$$[a + \delta + (i-1)c]^{\frac{1}{s*}} = [a + (i-1)d]^{\frac{1}{s}} + \epsilon, \text{ for all } i. \tag{A.28}$$

Solving for s^*, we have

$$s^* = \frac{ln(a + \delta + (i-1)c)}{ln\left([a + (i-1)d]^{\frac{1}{s}} + \epsilon\right)}. \tag{A.29}$$

We now can see that the value of s^* is a function of i, which means there is no constant value of s^* that can satisfy (A.28) for all i, unless $s = 1$. In fact, for $s = 1$, one will have $s* = 1$, in which case $\delta = \epsilon$ and $c = d$. □

Appendix B

Proofs in connection with the Sub-models of the PTAM

B.1 Proof of Proposition 3.5.1

Proof. Given m,

$$\begin{aligned}
\lim_{s\to\infty} h_i &= \lim_{s\to\infty}\left(\frac{m-i}{m-1}h_1^s + \frac{i-1}{m-1}h_m^s\right)^{\frac{1}{s}} \\
&= \max(h_1,h_m)\lim_{s\to\infty}\left(\frac{m-i}{m-1}\left(\frac{h_1}{\max(h_1,h_m)}\right)^s + \frac{i-1}{m-1}\left(\frac{h_m}{\max(h_1,h_m)}\right)^s\right)^{\frac{1}{s}} \\
&= h_m\lim_{s\to\infty}\left(\frac{m-i}{m-1}\left(\frac{h_1}{h_m}\right)^s + \frac{i-1}{m-1}\right)^{\frac{1}{s}}, \text{ since } h_m > h_1 > 0.
\end{aligned}$$

Let $L := \lim_{s\to\infty}\left(\frac{m-i}{m-1}\left(\frac{h_1}{h_m}\right)^s + \frac{i-1}{m-1}\right)^{\frac{1}{s}}$, then

$$\begin{aligned}
lnL &= ln\lim_{s\to\infty}\left(\frac{m-i}{m-1}\left(\frac{h_1}{h_m}\right)^s + \frac{i-1}{m-1}\right)^{\frac{1}{s}} \\
&= \lim_{s\to\infty} ln\left(\frac{m-i}{m-1}\left(\frac{h_1}{h_m}\right)^s + \frac{i-1}{m-1}\right)^{\frac{1}{s}} \\
&= \lim_{s\to\infty}\frac{ln\left(\frac{m-i}{m-1}\left(\frac{h_1}{h_m}\right)^s + \frac{i-1}{m-1}\right)}{s} \\
&= \frac{ln\left(\frac{i-1}{m-1}\right)}{\infty} \\
&= 0.
\end{aligned}$$

Thus,

$$\lim_{s\to\infty} h_i = h_m L = h_m e^0 = h_m$$

for $i = 2, 3, \ldots, m-1$.

Therefore, the limiting distribution as $s \to \infty$ is Coxian with

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda+h_1) & \lambda & 0 & 0 & \ldots & 0 & 0 \\ 0 & -(\lambda+h_m) & \lambda & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -(\lambda+h_m) & \lambda \\ 0 & 0 & 0 & 0 & \ldots & 0 & -h_m \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_m \\ \vdots \\ h_m \\ h_m \end{bmatrix}.$$

Since the last $m-1$ states have the dying rate h_m, they can be merged into one state. That is,

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda+h_1) & \lambda \\ 0 & -h_m \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_m \end{bmatrix}.$$

□

B.2 Proof of Proposition 3.5.2

Proof. Given m,

$$\begin{aligned} \lim_{s\to-\infty} h_i &= \lim_{s\to-\infty} \left(\frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s \right)^{\frac{1}{s}} \\ &= \min(h_1, h_m) \lim_{s\to-\infty} \left(\frac{m-i}{m-1} \left(\frac{h_1}{\min(h_1,h_m)} \right)^s + \frac{i-1}{m-1} \left(\frac{h_m}{\min(h_1,h_m)} \right)^s \right)^{\frac{1}{s}} \\ &= h_1 \lim_{s\to-\infty} \left(\frac{m-i}{m-1} + \frac{i-1}{m-1} \left(\frac{h_m}{h_1} \right)^s \right)^{\frac{1}{s}}, \text{ since } h_m > h_1 > 0. \end{aligned}$$

Let $L := \lim_{s\to-\infty}\left(\frac{m-i}{m-1}+\frac{i-1}{m-1}\left(\frac{h_m}{h_1}\right)^s\right)^{\frac{1}{s}}$, then

$$\begin{aligned}
lnL &= ln\lim_{s\to-\infty}\left(\frac{m-i}{m-1}+\frac{i-1}{m-1}\left(\frac{h_m}{h_1}\right)^s\right)^{\frac{1}{s}}\\
&= \lim_{s\to-\infty} ln\left(\frac{m-i}{m-1}+\frac{i-1}{m-1}\left(\frac{h_m}{h_1}\right)^s\right)^{\frac{1}{s}}\\
&= \lim_{s\to-\infty}\frac{ln\left(\frac{m-i}{m-1}+\frac{i-1}{m-1}\left(\frac{h_m}{h_1}\right)^s\right)}{s}\\
&= \frac{ln\left(\frac{m-i}{m-1}\right)}{-\infty}\\
&= 0.
\end{aligned}$$

Thus,

$$\lim_{s\to-\infty} h_i = h_1 L = h_1 e^0 = h_1$$

for $i = 2, 3, \ldots, m-1$.

Therefore, the limiting distribution as $s \to -\infty$ is Coxian with

$$\boldsymbol{S} = \begin{bmatrix}
-(\lambda+h_1) & \lambda & 0 & 0 & \ldots & 0 & 0\\
0 & -(\lambda+h_1) & \lambda & 0 & \ldots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & 0 & \ldots & -(\lambda+h_1) & \lambda\\
0 & 0 & 0 & 0 & \ldots & 0 & -h_m
\end{bmatrix},\ \boldsymbol{h} = \begin{bmatrix} h_1\\ h_1\\ \vdots\\ h_1\\ h_m \end{bmatrix}.$$

□

B.3 Proof of Proposition 3.5.3

Proof. Given m,

$$\begin{aligned}
\lim_{h_m\to\infty}\lim_{s\to-\infty} h_i &= \lim_{h_m\to\infty} h_1, \text{ by Proposition 3.5.2}\\
&= h_1,
\end{aligned}$$

where $i = 1, 2, 3, \ldots, m-1$. As h_m goes to infinity, the sojourn time in the last state will tend to zero. In other words, simply leaving the $(m-1)^{th}$ state is equivalent to absorption. The rate of absorption will be $h_1 + \lambda$ as there are two ways to leave the $(m-1)^{th}$ state.

Therefore, the limiting distribution as $h_m \to \infty$ and $s \to -\infty$ is Coxian of order $m-1$ with

$$\boldsymbol{S} = \begin{bmatrix} -(\lambda+h_1) & \lambda & 0 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda+h_1) & \lambda & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -(\lambda+h_1) & \lambda \\ 0 & 0 & 0 & 0 & \dots & 0 & -(\lambda+h_1) \end{bmatrix}, \ \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_1 \\ \vdots \\ h_1 \\ h_1+\lambda \end{bmatrix}.$$

□

B.4 Proof of Proposition 3.5.4

Proof. The limiting distribution of the PTAM as $m \to \infty$ is difficult to obtain. The result stated in Proposition 3.5.4 was proved in Cheng (2021) and mentioned in Cheng et al. (2021). Interested readers may refer to their work for more information. □

B.5 Proof of Proposition 3.5.5

Proof. Given m,

$$\begin{aligned} \lim_{h_m \to \infty} \lim_{s \to \infty} h_i &= \lim_{h_m \to \infty} h_m, \text{ by Proposition 3.5.1} \\ &= \infty, \end{aligned}$$

where $i = 2, 3, \dots, m$. Since the parameter space is continuous, the result can be obtained if the limits are switched. As h_m goes to infinity, the sojourn time goes to zero. In that instance, simply leaving the first state is equivalent to absorption. Therefore, the lifetime is exponentially distributed with rate $h_1 + \lambda$. □

B.6 Proof of Proposition 3.5.6

Proof. Let $h(t; h_1, h_m, s, \psi, m)$ be the hazard rate function associated with the PTAM. Then,

$$\begin{aligned}
&\lim_{s\to\infty}\lim_{m\to\infty} h(t; h_1, h_m, s, \psi, m) \\
&= \lim_{s\to\infty}\left((h_m^s - h_1^s)\frac{t}{\psi} + h_1^s\right)^{\frac{1}{s}}, \text{ by Cheng (2021)} \\
&= \max(h_1, h_m)\lim_{s\to\infty}\left(\left(\left(\frac{h_m}{\max(h_1, h_m)}\right)^s - \left(\frac{h_1}{\max(h_1, h_m)}\right)^s\right)\frac{t}{\psi} + \left(\frac{h_1}{\max(h_1, h_m)}\right)^s\right)^{\frac{1}{s}} \\
&= h_m \lim_{s\to\infty}\left(\frac{t}{\psi} + \left(1 - \frac{t}{\psi}\left(\frac{h_1}{h_m}\right)^s\right)\right)^{\frac{1}{s}}.
\end{aligned}$$

Let $L := \lim_{s\to\infty}\left(\frac{t}{\psi} + \left(1 - \frac{t}{\psi}\left(\frac{h_1}{h_m}\right)^s\right)\right)^{\frac{1}{s}}$, then

$$lnL = ln\lim_{s\to\infty} L = \lim_{s\to\infty} lnL = \lim_{s\to\infty}\frac{ln\left(\frac{t}{\psi} + \left(1 - \frac{t}{\psi}\left(\frac{h_1}{h_m}\right)^s\right)\right)}{s} = \frac{ln\left(\frac{t}{\psi}\right)}{\infty} = 0.$$

Therefore,

$$\lim_{s\to\infty}\lim_{m\to\infty} h(t; h_1, h_m, s, \psi, m) = h_m L = h_m e^0 = h_m,$$

which is the hazard function for an exponentially distributed random variable with rate h_m. □

B.7 Proof of Proposition 3.5.7

Proof. Let $h(t; h_1, h_m, s, \psi, m)$ be the hazard rate function associated with the PTAM. Then,

$$\begin{aligned}
&\lim_{s\to-\infty}\lim_{m\to\infty} h(t; h_1, h_m, s, \psi, m) \\
&= \lim_{s\to-\infty}\left((h_m^s - h_1^s)\frac{t}{\psi} + h_1^s\right)^{\frac{1}{s}}, \text{ by Cheng (2021)} \\
&= \min(h_1, h_m)\lim_{s\to-\infty}\left(\left(\left(\frac{h_m}{\min(h_1, h_m)}\right)^s - \left(\frac{h_1}{\min(h_1, h_m)}\right)^s\right)\frac{t}{\psi} + \left(\frac{h_1}{\min(h_1, h_m)}\right)^s\right)^{\frac{1}{s}} \\
&= h_1 \lim_{s\to-\infty}\left(\frac{t}{\psi}\left(\frac{h_m}{h_1}\right)^s + \left(1 - \frac{t}{\psi}\right)\right)^{\frac{1}{s}}.
\end{aligned}$$

Let $L := \lim_{s\to-\infty}\left(\frac{t}{\psi}\left(\frac{h_m}{h_1}\right)^s + \left(1-\frac{t}{\psi}\right)\right)^{\frac{1}{s}}$, then

$$lnL = ln\lim_{s\to-\infty} L = \lim_{s\to-\infty} lnL = \lim_{s\to-\infty}\frac{ln\left(\frac{t}{\psi}\left(\frac{h_m}{h_1}\right)^s + \left(1-\frac{t}{\psi}\right)\right)}{s} = \frac{ln\left(\frac{t}{\psi}\right)}{-\infty} = 0.$$

Therefore,

$$\lim_{s\to-\infty}\lim_{m\to\infty} h(t;h_1,h_m,s,\psi,m) = h_1L = h_1e^0 = h_1,$$

which is the hazard function for an exponentially distributed random variable with rate h_1. □

B.8 Proof of Proposition 3.5.8

Proof. Let $h(t;h_1,h_m,s,\psi,m)$ be the hazard rate function associated with the PTAM. Then,

$$\begin{aligned}\lim_{h_m\to\infty}\lim_{m\to\infty}\lim_{s\to-\infty} h(t;h_1,h_m,s,\psi,m) &= \lim_{h_m\to\infty} h_1, \text{by Proposition 3.5.7}\\ &= h_1,\end{aligned}$$

which is the hazard function for an exponentially distributed random variable with rate h_1. □

Appendix C

Data Augmentation with Left-truncated Data for the PTAM

C.1 Case 1 - entering the study before reaching state m

To begin with, consider a sample path of the underlying Markov process of the PTAM presented in Table C.1, where $m > 5$:

state	1	2	3	4	5	$m+1$
sojourn time	t_1	t_2	t_3	t_4	t_5	0

Table C.1: A PTAM sample path generated from data augmentation.

In that case, the augmented data is $\boldsymbol{x} = (t_1, t_2, t_3, t_4, t_5, 1, 2, 3, 4, 5)$ and the original data is the absorption time y, which is equal to the sum of all sojourn times. The likelihood function of this sample path is then

$$L(h_1, h_m, s, \lambda; \boldsymbol{x}, y) = \prod_{i=1}^{4} \left(\lambda e^{-(\lambda+h_i)t_i}\right) h_5 e^{-(\lambda+h_5)t_5}. \tag{C.1}$$

Now, suppose the individual enters the study at d where $t_1 < d < t_1 + t_2$ without any loss of generality; then the likelihood function for this left-truncated data is,

$$\begin{aligned} L(h_1, h_m, s, \lambda; \boldsymbol{x}, y) &= \frac{\prod_{i=1}^{4} \left(\lambda e^{-(\lambda+h_i)t_i}\right) h_5 e^{-(\lambda+h_5)t_5}}{\lambda e^{-(\lambda+h_1)t_1} e^{-(\lambda+h_2)(d-t_1)}} \\ &= \lambda e^{-(\lambda+h_2)(t_1+t_2-d)} \prod_{i=3}^{4} \left(\lambda e^{-(\lambda+h_i)t_i}\right) h_5 e^{-(\lambda+h_5)t_5}. \end{aligned} \tag{C.2}$$

C.2 Case 2 - entering the study after reaching state m

On the other hand, when the individual enters the study at the last physiological age, the likelihood function will be slightly different. To verify this, consider another case where the simulated sample path is as presented in Table C.2:

state	1	2	3	4	5	$\cdots$	m	$m+1$
sojourn time	t_1	t_2	t_3	t_4	t_5	$\cdots$	t_m	0

Table C.2: A PTAM sample path generated from data augmentation.

Accordingly, we have $\sum_{i=1}^{m-1} t_i < d < \sum_{i=1}^{m} t_i$. In that case, the likelihood function becomes

$$\begin{aligned} L(h_1, h_m, s, \lambda; \boldsymbol{x}, y) &= \frac{\prod_{i=1}^{m-1} \left(\lambda e^{-(\lambda+h_i)t_i}\right) h_m e^{-h_m t_m}}{\prod_{i=1}^{m-1} \left(\lambda e^{-(\lambda+h_i)t_i}\right) e^{-h_m (d - \sum_{i=1}^{m-1} t_i)}} \\ &= h_m e^{-h_m (\sum_{i=1}^{m} t_i - d)}. \end{aligned} \tag{C.3}$$

Clearly, what makes the two cases different is the rate in the exponent. For the previous $m-1$ states, the rate includes λ; however, for the last state, the rate does not, which is due to the PTAM definition. Thus, in order to construct the likelihood function for left-truncated data, one needs to consider these two cases separately, which explains why the set $\mathcal{A}$ needs to be defined.

C.3 Proof for likelihood function with left-truncated data

Now, let us finally consider sample paths from M individuals. To consider the two cases separately, let $\mathcal{A}$ is the set of indices of the sample paths occurring in the second case. Let $m^{(i)}$ be the state right before absorption for the i^{th} individual. Moreover, let $n^{(i)}$ be such

that $\sum_{j=1}^{n^{(i)}} t_j^{(i)} < d_i < \sum_{j=1}^{n^{(i)}+1} t_j^{(i)}$. Then, the likelihood function for M individual becomes

$$\begin{aligned}
L(\lambda, h_1, h_m, s; \boldsymbol{x}, \boldsymbol{y}) &= \prod_{i\in\mathcal{A}} \left(\lambda e^{-\left(\lambda + h_{n^{(i)}+1}\right)\left(\sum_{j=1}^{n^{(i)}+1} t_j^{(i)} - d_i\right)} \prod_{j=n^{(i)}+2}^{m^{(i)}-1} \left(\lambda e^{-(\lambda+h_j)t_j^{(i)}} \right) h_{m^{(i)}} e^{-\left(\lambda + h_{m^{(i)}}\right)t_{m^{(i)}}^{(i)}} \right) \\
&\quad \times \prod_{i\notin\mathcal{A}} \left(h_m e^{-h_m\left(\sum_{j=1}^{m} t_j^{(i)} - d_i\right)} \right) \\
&= \left(\lambda^{\sum_{i\in\mathcal{A}}(m^{(i)}-n^{(i)}-1)} \right) \left(e^{-\lambda \sum_{i\in\mathcal{A}} \sum_{j=1}^{m^{(i)}} t_j^{(i)}} \right) \left(\prod_{i=1}^{M} h_{m^{(i)}} \right) \\
&\quad \times \left(\prod_{i\in\mathcal{A}} e^{-h_{n^{(i)}+1}\left(\sum_{j=1}^{n^{(i)}+1} t_j^{(i)} - d_i\right) - \sum_{j=n^{(i)}+2}^{m^{(i)}} h_j t_j^{(i)}} \right) \left(\prod_{i\notin\mathcal{A}} e^{-h_m\left(\sum_{j=1}^{m} t_j^{(i)} - d_i\right)} \right) \\
&\quad \times \left(\prod_{i\in\mathcal{A}} e^{\lambda d_i} \right). \qquad \text{(C.4)}
\end{aligned}$$

According to our definitions of Q_{ij}, $Z_i^{\mathcal{A}}$ and M_i and definitions on N_{ij} and the definitions of Z_i in Asmussen et al. (1996), it can be observed from above that

$$\sum_{i\in\mathcal{A}} (m^{(i)} - n^{(i)} - 1) =: \sum_{i=1}^{m-1} \left(N_{i,i+1} - Q_{i,i+1} \right), \qquad \text{(C.5)}$$

$$\sum_{i\in\mathcal{A}} \sum_{j=1}^{m^{(i)}} t_j^{(i)} =: \sum_{i=1}^{m-1} Z_i^{\mathcal{A}}, \qquad \text{(C.6)}$$

$$\prod_{i=1}^{M} h_{m^{(i)}} =: \prod_{i=1}^{m} h_i^{N_{i,m+1}}, \qquad \text{(C.7)}$$

and

$$\left(\prod_{i\in\mathcal{A}} e^{-h_{n^{(i)}+1}\left(\sum_{j=1}^{n^{(i)}+1} t_j^{(i)} - d_i\right) - \sum_{j=n^{(i)}+2}^{m^{(i)}} h_j t_j^{(i)}} \right) \left(\prod_{i\notin\mathcal{A}} e^{-h_m\left(\sum_{j=1}^{m} t_j^{(i)} - d_i\right)} \right) =: \prod_{i=1}^{m} e^{-h_i G_i}. \qquad \text{(C.8)}$$

Substituting (C.5)–(C.8) into (C.4) yields the final representation given in (4.15).

Appendix D

Rejection Sampling on the Logarithmic Scale

We briefly recall the rejection sampling method, in order to sample from a given continuous p.d.f. $p(\theta)$ where $\theta \in (a, b)$, given that $f(\theta) \propto p(\theta)$:

Algorithm 19 The rejection sampling algorithm with a uniform proposal distribution

1: Calculate the global maximum of $f(\theta)$ on (a, b). Define it as w.
2: Draw a pair of uniformly distributed sample (θ, y); $\Theta \sim Unif(a, b)$ and $Y \sim Unif(0, w)$.
3: **while** $f(\theta) \leq y$ **do**
4: **repeat** Step 2
5: **end while**
6: Take θ as the sample.

Note that Algorithm 19 utilizes a uniform distribution of Θ as the proposal distribution, with p.d.f. defined as $v(\theta) = \frac{1}{b-a}$. Subsequently, in order to satisfy the requirement for rejection sampling, a constant $c = \frac{w}{(b-a)}$ is selected so that $cv(\theta) = w \geq f(\theta),\ \forall\theta \in (a, b)$. According to the theory on rejection sampling, the proposal distribution $v(\theta)$ does not have to be uniform, as long as the requirement $cv(\theta) \geq f(\theta)$ is satisfied. In this paper, it is taken as the uniform distribution as the implementation turns out to be simpler.

However, when f is a posterior kernel, its value is likely to be small by making use of the likelihood function. In fact, in the simulation study on the PTAM, its value is so small that it outputs a value of zero in R. Thus, in order to carry out the rejection sampling scheme, we have to transform the posterior kernel to a logarithmic scale. In other words, instead of comparing $f(\theta)$ and y, we compare $lnf(\theta)$ and $ln(y)$. Accordingly, we need to determine the distribution of $ln(Y)$.

Given that $x \in (-\infty,\ ln(w))$, we have

$$\begin{aligned} \mathbb{P}(ln(Y) \leq x) &= \mathbb{P}(Y \leq e^x) \\ &= \frac{e^x}{w}. \end{aligned} \tag{D.1}$$

Upon inverting the c.d.f., we may achieve the sampling of $ln(Y)$ via the relationship

$$ln(Y) = ln(w) + ln(U), \tag{D.2}$$

where $U \sim Unif(0,1)$. Therefore, one may sample on a logarithmic scale as in Algorithm 20. That allows one to work with $ln(w)$, when w is so small that it outputs a value of zero in R.

Algorithm 20 Algorithm 19 on a logarithmic scale

1: Calculate the global maximum of $lnf(\theta)$ on (a,b). Define it as $ln.w$.
2: Draw a pair of samples $(\theta, ln(y))$; $\Theta \sim Unif(a,b)$ and $ln(Y) = ln(U) + ln.w$, where $U \sim Unif(0,1)$.
3: **while** $lnf(\theta) \leq ln(y)$ **do**
4: **repeat** Step 2
5: **end while**
6: Take θ as the sample.

Then, Algorithms 9, 10 and 11 are direct applications of Algorithm 20.

Appendix E

Data Augmentation for the DMPTM

In this appendix, we will address the technical details in connection with certain specific steps of Algorithms 15, 16 and 17.

E.1 Algorithm 15: Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} = \boldsymbol{y})$

E.1.1 Step 2 of Algorithm 15

To simulate $J_t = i$ from $p(J_t|\boldsymbol{Y} = \boldsymbol{y})$, Bayes Theorem can be utilized:

$$\begin{aligned} p(J_t = i|\boldsymbol{Y} = \boldsymbol{y}) &= \frac{\pi_i \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|J_t = i)}{\sum_{j=1}^{m} \pi_j \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|J_t = i)} \\ &= \frac{\pi_i p_{\boldsymbol{Y},i}(\boldsymbol{y})}{\sum_{j=1}^{m} \pi_j p_{\boldsymbol{Y},j}(\boldsymbol{y})}, \ i = 1, 2, 3, \ldots, m, \end{aligned} \tag{E.1}$$

where $\boldsymbol{p_Y(y)}$ is obtained by applying the recursive relationship specified in Result 4.

E.1.2 Step 6 of Algorithm 15

Let ϕ be a fictitious batch corresponding to the event "no batches in $\mathcal{C}_0$ occurs". Letting $I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y} \geq \boldsymbol{h}\} \cup \{\phi\}$, we have

$$\begin{aligned} p(J_{t+1} = j, I|\boldsymbol{Y} = \boldsymbol{y}, J_t = i) &= \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y} = \boldsymbol{y}, J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}, J_t = i)} \\ &= \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y} = \boldsymbol{y}|J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|J_t = i)} \\ &= \frac{\mathbb{P}(I, \boldsymbol{Y} = \boldsymbol{y}|J_{t+1} = j)\mathbb{P}(J_{t+1} = j|J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|J_t = i)}, \end{aligned} \tag{E.2}$$

where $j = 1, 2, \ldots, m$.

Letting the numerator in expression (E.2) be denoted by R, we have

$$R = \begin{cases} (B_{0,ij}) p_{\boldsymbol{Y},j}(\boldsymbol{y}), & \text{if } I = \phi, \\ (B_{\boldsymbol{h},ij}) p_{\boldsymbol{Y},j}(\boldsymbol{y} - \boldsymbol{h}), & \text{if } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y} \geq \boldsymbol{h}\}, \end{cases} \tag{E.3}$$

$$=: \begin{cases} R_1, & \text{if } I = \phi, \\ R_2, & \text{if } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y} \geq \boldsymbol{h}\}, \end{cases} \tag{E.4}$$

where $\boldsymbol{p_Y(y)}$ and $\boldsymbol{p_Y}(\boldsymbol{y} - \boldsymbol{h})$ are obtained by applying the recursive relationship specified in Result 4.

Therefore,

$$\begin{aligned} p(J_{t+1} = j, I | \boldsymbol{Y} = \boldsymbol{y}, J_t = i) &= \frac{R}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | J_t = i)} \\ &= \frac{R}{\sum_{j:(B_0)_{ij} \neq 0} R_1 + \sum_{\boldsymbol{h} \in \mathcal{C}_0 : \boldsymbol{y} \geq \boldsymbol{h}} R_2}. \end{aligned}$$

E.1.3 Step 17 of Algorithm 15

Step 17 of Algorithm 15 is similar to Step 6 presented in Appendix E.1.2; however, there are two differences:

(i) At this stage, we know that $\boldsymbol{Y} = \boldsymbol{0}$ so that no batch in $\mathcal{C}_0$ should be generated.

(ii) $\boldsymbol{Y}$ being reduced to $\boldsymbol{0}$ means that no data remains, and the algorithm is ready to end. Thus, the absorbing transition should be included. That is, $j = m + 1$ is allowed.

Then, we have

$$p(J_{t+1} = j | \boldsymbol{Y} = \boldsymbol{0}, J_t = i) = p(J_{t+1} = j | \boldsymbol{Y} = \boldsymbol{0}, J_t = i, I = \phi) \tag{E.5}$$

$$\begin{aligned} &= \frac{\mathbb{P}(J_{t+1} = j, I = \phi, \boldsymbol{Y} = \boldsymbol{0}, J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}, J_t = i, I = \phi)} \\ &= \frac{\mathbb{P}(J_{t+1} = j, I = \phi, \boldsymbol{Y} = \boldsymbol{0} | J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}, I = \phi | J_t = i)} \\ &= \frac{\mathbb{P}(I = \phi, \boldsymbol{Y} = \boldsymbol{0} | J_{t+1} = j) \mathbb{P}(J_{t+1} = j | J_t = i)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}, I = \phi | J_t = i)}, \end{aligned} \tag{E.6}$$

where $j = 1, 2, \ldots, m+1$.

Letting the numerator in expression (E.6) be denoted by R, we have

$$R = \begin{cases} \mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}|J_{t+1} = j)\mathbb{P}(J_{t+1} = j|J_t = i), & \text{if } j \neq m+1, \\ \mathbb{P}(J_{t+1} = m+1|J_t = i), & \text{if } j = m+1 \text{ (absorbed)}, \end{cases} \tag{E.7}$$

$$= \begin{cases} (B_{0,ij})p_{\boldsymbol{Y},j}(\boldsymbol{0}), & \text{if } j \neq m+1, \\ b_{0,i}, & \text{if } j = m+1 \text{ (absorbed)}, \end{cases} \tag{E.8}$$

$$=: \begin{cases} R_1, & \text{if } j \neq m+1, \\ R_2, & \text{if } j = m+1 \text{ (absorbed)}, \end{cases} \tag{E.9}$$

where $\boldsymbol{p_Y(0)}$ is obtained by applying the recursive relationship specified in Result 4.

Therefore,

$$\begin{aligned} p(J_{t+1} = j|\boldsymbol{Y} = \boldsymbol{0}, J_t = i) &= \frac{R}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}|J_t = i)} \\ &= \frac{R}{\sum_{j:(B_0)_{ij} \neq 0} R_1 + R_2}. \end{aligned}$$

E.2 Algorithm 16: Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}})$

The technical details pertaining to Algorithm 16 are similar to those presented in Appendix E.1 for Algorithm 15, the only difference being that the recursive relationship specified in Result 5 is utilized rather than that given in Result 4.

E.2.1 Step 2 of Algorithm 16

To simulate $J_t = i$ from $p(\cdot|\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}})$, Bayes Theorem can be utilized:

$$\begin{aligned} p(J_t = i|\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}) &= \frac{\pi_i \mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}|J_t = i)}{\sum_{j=1}^m \pi_j \mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}|J_t = i)} \\ &= \frac{\pi_i p^{(0,1)}_{\boldsymbol{Y},\geq,i}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})}{\sum_{j=1}^m \pi_j p^{(0,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})}, \ i = 1, 2, 3, \ldots, m, \end{aligned} \tag{E.10}$$

where $\boldsymbol{p^{(0,1)}_{Y,\geq}(y^{(1)}, y^{(2)})}$ is obtained by applying the recursive relationship specified in Result 5.

E.2.2 Step 6 of Algorithm 16

Let ϕ be a fictitious batch corresponding to the event "no batches in $\mathcal{C}_0$ occurs". Let $I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}\} \cup \{\phi\}$. Define $Q := p(J_{t+1} = j, I | \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}, J_t = i)$. Then, we have

$$\begin{aligned} Q =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}, J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}, J_t = i)} \\ =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}} | J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}} | J_t = i)} \\ =& \frac{\mathbb{P}(I, \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}} | J_{t+1} = j) \mathbb{P}(J_{t+1} = j | J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}} | J_t = i)}, \end{aligned} \tag{E.11}$$

where $j = 1, 2, \ldots, m$.

Letting the numerator in expression (E.11) be denoted by R, we have

$$R = \begin{cases} (B_{0,ij}) p^{(0,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}), & \text{if } I = \phi, \\ (B_{\boldsymbol{h},ij}) p^{(0,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{y^{(1)}} - \boldsymbol{h^{(1)}}, (\boldsymbol{y^{(2)}} - \boldsymbol{h^{(2)}})^+), & \text{if } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}\}, \end{cases} \tag{E.12}$$

$$=: \begin{cases} R_1, & \text{if } I = \phi, \\ R_2, & \text{if } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}\}, \end{cases} \tag{E.13}$$

where $\boldsymbol{p^{(0,1)}_{Y,\geq}}(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}})$ and $\boldsymbol{p^{(0,1)}_{Y,\geq}}(\boldsymbol{y^{(1)}} - \boldsymbol{h^{(1)}}, (\boldsymbol{y^{(2)}} - \boldsymbol{h^{(2)}})^+)$ are obtained by applying the recursive relationship specified in Result 5.

Thus,

$$\begin{aligned} p(J_{t+1} = j, I | \boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}}, J_t = i) &= \frac{R}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{y^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{y^{(2)}} | J_t = i)} \\ &= \frac{R}{\sum_{j:(B_0)_{ij} \neq 0} R_1 + \sum_{\boldsymbol{h}:\boldsymbol{h} \in \mathcal{C}_0, \boldsymbol{y^{(1)}} \geq \boldsymbol{h^{(1)}}} R_2}. \end{aligned}$$

E.2.3 Step 17 of Algorithm 16

Step 17 of Algorithm 16 is similar to Step 6 presented in Appendix E.2.2; however, there are two differences:

(i) At this stage, we know that $\boldsymbol{Y} = \boldsymbol{0}$ so that any batch to be generated must have $\boldsymbol{y^{(1)}} = \boldsymbol{0}$.

(ii) $\boldsymbol{Y}$ being reduced to $\boldsymbol{0}$ means that no data remains, and the algorithm is ready to end. Thus, the absorbing transition should be included.

Define $Q_1 := p(J_{t+1} = j, I | \boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}}, J_t = i)$. Then, we have

$$\begin{aligned} Q_1 =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}}, J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}}, J_t = i)} \\ =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}} | J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}} | J_t = i)} \\ =& \frac{\mathbb{P}(I, \boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}} | J_{t+1} = j) \mathbb{P}(J_{t+1} = j | J_t = i)}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}} | J_t = i)}, \end{aligned} \tag{E.14}$$

where $j = 1, 2, \ldots, m+1$.

Letting the numerator in expression (E.14) be denoted by R, we have

$$R = \begin{cases} (B_{0,ij}) p^{(0,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{0^{(1)}}, \boldsymbol{0^{(2)}}), & \text{if } j \neq m+1 \text{ and } I = \phi, \\ (B_{\boldsymbol{h},ij}) p^{(0,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{0^{(1)}}, \boldsymbol{0^{(2)}}), & \text{if } j \neq m+1 \text{ and } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{h^{(1)}} = \boldsymbol{0^{(1)}}\}, \\ b_{0,i}, & \text{if } j = m+1 \text{ (absorbed)}, \end{cases} \tag{E.15}$$

$$=: \begin{cases} R_1, & \text{if } j \neq m+1 \text{ and } I = \phi, \\ R_2, & \text{if } j \neq m+1 \text{ and } I \in \{\boldsymbol{h} \in \mathcal{C}_0 | \boldsymbol{h^{(1)}} = \boldsymbol{0^{(1)}}\}, \\ R_3, & \text{if } j = m+1 \text{ (absorbed)}, \end{cases} \tag{E.16}$$

where $\boldsymbol{p^{(0,1)}_{Y,\geq}(0)}$ is obtained by applying the recursive relationship specified in Result 5.

Thus,

$$\begin{aligned} p(J_{t+1} = j, I | \boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}}, J_t = i) &= \frac{R}{\mathbb{P}(\boldsymbol{Y^{(1)}} = \boldsymbol{0^{(1)}}, \boldsymbol{Y^{(2)}} \geq \boldsymbol{0^{(2)}}, J_t = i)} \\ &= \frac{R}{\sum_{j:(B_0)_{ij} \neq 0} R_1 + \sum_{\boldsymbol{h} \in \mathcal{C}_0 : \boldsymbol{h^{(1)}} = \boldsymbol{0}} R_2 + R_3}. \end{aligned}$$

E.3 Algorithm 17: Sampling from $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{b_0}, \boldsymbol{Y} \geq \boldsymbol{y})$

The technical details pertaining to Algorithm 17 are similar to those presented in Appendix E.1 for Algorithm 15, the only difference being that the recursive relationship specified in Result 6 is utilized rather than that given in Result 4.

E.3.1 Step 2 of Algorithm 17

To simulate $J_t = i$ from $p(\cdot|\boldsymbol{Y} \geq \boldsymbol{y})$, Bayes Theorem can be utilized:

$$\begin{aligned} p(J_t = i|\boldsymbol{Y} \geq \boldsymbol{y}) &= \frac{\pi_i \mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y}|J_t = i)}{\sum_{j=1}^{m} \pi_j \mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y}|J_t = i)} \\ &= \frac{\pi_i p^{(1,1)}_{\boldsymbol{Y},\geq,i}(\boldsymbol{y})}{\sum_{j=1}^{m} \pi_j p^{(1,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{y})}, \ i = 1, 2, 3, \ldots, m, \end{aligned} \tag{E.17}$$

where $\boldsymbol{p^{(1,1)}_{Y,\geq}(y)}$ is obtained by applying the recursive relationship specified in Result 6.

E.3.2 Step 6 of Algorithm 17

Let ϕ be a fictitious batch corresponding to the event "no batches in $\mathcal{C}_0$ occurs". Letting $I \in \{\phi, \mathcal{C}_0\}$, we have

$$\begin{aligned} p(J_{t+1} = j, I|\boldsymbol{Y} \geq \boldsymbol{y}, J_t = i) =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y} \geq \boldsymbol{y}, J_t = i)}{\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y}, J_t = i)} \\ =& \frac{\mathbb{P}(J_{t+1} = j, I, \boldsymbol{Y} \geq \boldsymbol{y}|J_t = i)}{\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y}|J_t = i)} \\ =& \frac{\mathbb{P}(I, \boldsymbol{Y} \geq \boldsymbol{y}|J_{t+1} = j)\mathbb{P}(J_{t+1} = j|J_t = i)}{\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y}|J_t = i)}, \end{aligned} \tag{E.18}$$

where $j = 1, 2, \ldots, m$.

Letting the numerator in expression (E.18) be denoted by R, we have

$$R = \begin{cases} (B_{0,ij}) p^{(1,1)}_{\boldsymbol{Y},\geq,j}(\boldsymbol{y}), & \text{if } I = \phi, \\ (B_{\boldsymbol{h},ij}) p^{(1,1)}_{\boldsymbol{Y},\geq,j}((\boldsymbol{y} - \boldsymbol{h})^+), & \text{if } I \in \mathcal{C}_0, \end{cases} \tag{E.19}$$

$$=: \begin{cases} R_1, & \text{if } I = \phi, \\ R_2, & \text{if } I \in \mathcal{C}_0, \end{cases} \tag{E.20}$$

where $\boldsymbol{p^{(1,1)}_{Y,\geq}(y)}$ and $\boldsymbol{p^{(1,1)}_{Y,\geq}((y-h)^+)}$ are obtained by applying the recursive relationship specified in Result 6.

Finally,

$$p(J_{t+1}=j, I|\boldsymbol{Y}\geq\boldsymbol{y}, J_t=i) = \frac{R}{\mathbb{P}(\boldsymbol{Y}\geq\boldsymbol{y}|J_t=i)} = \frac{R}{\sum_{j:(B_0)_{ij}\neq 0} R_1 + \sum_{\boldsymbol{h}:\boldsymbol{h}\in\mathcal{C}_0} R_2}.$$

E.3.3 Step 17 of Algorithm 17

Step 17 of Algorithm 17 is similar to Step 6 presented in Appendix E.3.2. $\boldsymbol{Y}$ being reduced to $\mathbf{0}$ means that no data remains, and the algorithm is ready to end. Thus, the absorbing transition should be included. Then, we have

$$\begin{aligned} p(J_{t+1}=j, I|\boldsymbol{Y}\geq\mathbf{0}, J_t=i) =& \frac{\mathbb{P}(J_{t+1}=j, I, \boldsymbol{Y}\geq\mathbf{0}, J_t=i)}{\mathbb{P}(\boldsymbol{Y}\geq\mathbf{0}, J_t=i)} \\ =& \frac{\mathbb{P}(J_{t+1}=j, I, \boldsymbol{Y}\geq\mathbf{0}|J_t=i)}{\mathbb{P}(\boldsymbol{Y}\geq\mathbf{0}|J_t=i)} \\ =& \frac{\mathbb{P}(I, \boldsymbol{Y}\geq\mathbf{0}|J_{t+1}=j)\mathbb{P}(J_{t+1}=j|J_t=i)}{\mathbb{P}(\boldsymbol{Y}\geq\mathbf{0}|J_t=i)}, \end{aligned} \tag{E.21}$$

where $j=1,2,\ldots,m+1$.

Letting the numerator in expression (E.21) be denoted by R, we have

$$R = \begin{cases} (B_{0,ij})p^{(1,1)}_{\boldsymbol{Y},\geq,j}(\mathbf{0}), & \text{if } j\neq m+1 \text{ and } I=\phi, \\ (B_{\boldsymbol{h},ij})p^{(1,1)}_{\boldsymbol{Y},\geq,j}(\mathbf{0}), & \text{if } j\neq m+1 \text{ and } I\in\mathcal{C}_0, \\ b_{0,i}, & \text{if } j=m+1 \text{ (absorbed)}, \end{cases} \tag{E.22}$$

$$=: \begin{cases} R_1, & \text{if } j\neq m+1 \text{ and } I=\phi, \\ R_2, & \text{if } j\neq m+1 \text{ and } I\in\mathcal{C}_0, \\ R_3, & \text{if } j=m+1 \text{ (absorbed)}, \end{cases} \tag{E.23}$$

where $\boldsymbol{p}^{(1,1)}_{\boldsymbol{Y},\geq}(\mathbf{0})$ is obtained by recursive relationship specified in Result 6.

Therefore,

$$p(J_{t+1}=j, I|\boldsymbol{Y}\geq\mathbf{0}, J_t=i) = \frac{R}{\mathbb{P}(\boldsymbol{Y}\geq\mathbf{0}, J_t=i)} = \frac{R}{\sum_{j:(B_0)_{ij}\neq 0} R_1 + \sum_{\boldsymbol{h}:\boldsymbol{h}\in\mathcal{C}_0} R_2 + R_3}.$$

Bibliography

Aalen, O. O. (1995). Phase-type distributions in survival analysis. *Scandinavian Journal of Statistics*, pages 447–463.

Aslett, L. J. and Wilson, S. P. (2011). Markov chain Monte Carlo for inference on phase-type models. *Proceedings of the 2011 International Statistical Institute World Statistics Congress (ISI WSC)*, 120.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23: 419–441.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bladt, M., Gonzalez, A., and Lauritzen, S. L. (2003). The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal*, 2003(4): 280–300.

Bobbio, A. and Cumani, A. (1992). ML estimation of the parameters of a PH distribution in triangular canonical form. *Computer Performance Evaluation*, 22: 33–46.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2): 211–243.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press, New York.

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4): 1015–1030.

Burke, E. K., Burke, E. K., Kendall, G., and Kendall, G. (2014). *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Springer, New York.

Cheng, B. (2021). *A Class of Phase-Type Aging Models and their Lifetime Distributions*. PhD thesis, Western University.

Cheng, B., Jones, B., Liu, X., and Ren, J. (2021). The mathematical mechanism of biological aging. *North American Actuarial Journal*, 25(1): 73–93.

Chis, O.-T., Banga, J. R., and Balsa-Canto, E. (2011). Structural identifiability of systems biology models: a critical comparison of methods. *PloS One*, 6(11): e27755.

Cox, D. R. (1955a). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 433–441. Cambridge University Press.

Cox, D. R. (1955b). A use of complex probabilities in the theory of stochastic processes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 313–319. Cambridge University Press.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics Reliability*, 22(3): 583–602.

Cummins, J. D. and Wiltbank, L. J. (1983). Estimating the total claims distribution using multivariate frequency and severity distributions. *Journal of Risk and Insurance*, pages 377–403.

Degenring, D., Froemel, C., Dikta, G., and Takors, R. (2004). Sensitivity analysis for the reduction of complex metabolism models. *Journal of Process Control*, 14(7): 729–745.

Dochain, D. and Vanrolleghem, P. (2001). Dynamical modelling & estimation in wastewater treatment processes. *Water Intelligence Online*, 4.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409.

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

Gianola, D. (2007). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, volume 43. Springer, New York.

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.

Gontier, C. and Pfister, J.-P. (2020). Identifiability of a binomial synapse. *Frontiers in Computational Neuroscience*, 14: 86.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109.

He, Q.-M. and Ren, J. (2016a). Analysis of a multivariate claim process. *Methodology and Computing in Applied Probability*, 18(1): 257–273.

He, Q.-M. and Ren, J. (2016b). Parameter estimation of discrete multivariate phase-type distributions. *Methodology and Computing in Applied Probability*, 18(3): 629–651.

Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19): 2612–2618.

Hobert, J. P. (2011). The data augmentation algorithm: theory and methodology. *Handbook of Markov Chain Monte Carlo*, pages 253–293.

Holmberg, A. (1982). On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities. *Mathematical Biosciences*, 62(1): 23–43.

Hyde, J. (1980). Testing survival with incomplete observations. *Biostatistics Casebook*, pages 31–46.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2): 201–227.

Kloek, T. and Van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19.

Lehmann, E. and Casella, G. (1998). Unbiasedness. *Theory of Point Estimation*, pages 83–146.

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10(7): 551–563.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492): 1617–1625.

Lin, X. S. and Liu, X. (2007). Markov aging process and phase-type law of mortality. *North American Actuarial Journal*, 11(4): 92–109.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer Science & Business Media.

McLean, K. A. and McAuley, K. B. (2012). Mathematical modelling of chemical processes—obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *The Canadian Journal of Chemical Engineering*, 90(2): 351–366.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087–1092.

Miao, H., Xia, X., Perelson, A. S., and Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review*, 53(1): 3–39.

Michiels, W., Aarts, E. H., and Korst, J. (2007). *Theoretical Aspects of Local Search*. Springer, Berlin.

Naylor, J. C. and Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3): 214–225.

Okamura, H. and Dohi, T. (2016). PH fitting algorithm and its application to reliability engineering. *Journal of the Operations Research Society of Japan*, 59(1): 72–109.

Okamura, H., Watanabe, R., and Dohi, T. (2014). Variational Bayes for phase-type distribution. *Communications in Statistics-Simulation and Computation*, 43(8): 2031–2044.

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, pages 443–460.

Petersen, B., Gernaey, K., and Vanrolleghem, P. (2001). Practical identifiability of model parameters by combined respirometric-titrimetric measurements. *Water Science and Technology*, 43(7): 347–355.

Pham, H. and Lai, C.-D. (2007). On recent generalizations of the Weibull distribution. *IEEE Transactions on Reliability*, 56(3): 454–458.

Quaiser, T. and Mönnigmann, M. (2009). Systematic identifiability testing for unambiguous mechanistic modeling–application to jak-stat, map kinase, and nf-κ b signaling pathway models. *BMC Systems Biology*, 3(1): 1–21.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15): 1923–1929.

Rizk, J., Burke, K., and Walsh, C. (2019). On the non-uniqueness of representations of Coxian phase-type distributions. *arXiv preprint arXiv:1901.03849*.

Rizk, J., Walsh, C., and Burke, K. (2021). An alternative formulation of Coxian phase-type distributions with covariates: Application to emergency department length of stay. *Statistics in Medicine*, 40(6): 1574–1592.

Rodriguez-Fernandez, M., Egea, J. A., and Banga, J. R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics*, 7(1): 1–18.

Smith, A., Skene, A., Shaw, J., Naylor, J., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics-Theory and Methods*, 14(5): 1079–1102.

Su, S. and Sherris, M. (2012). Heterogeneity of Australian population mortality and implications for a viable life annuity market. *Insurance: Mathematics and Economics*, 51(2): 322–332.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398): 528–540.

Tanner, M. A. and Wong, W. H. (2010). From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. *Statistical Science*, 25(4): 506–516.

Telek, M. and Horváth, G. (2007). A minimal representation of Markov arrival processes and a moments matching method. *Performance Evaluation*, 64(9-12): 1153–1168.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393): 82–86.

Vajda, S., Rabitz, H., Walter, E., and Lecourtier, Y. (1989). Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83(1): 191–219.

Van Dijk, H. K. and Kloek, T. (1980). Further experience in Bayesian analysis using Monte Carlo integration. *Journal of Econometrics*, 14(3): 307–328.

Watanabe, R., Okamura, H., and Dohi, T. (2012). An efficient MCMC algorithm for continuous PH distributions. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEEE.

Yao, K. Z., Shaw, B. M., Kou, B., McAuley, K. B., and Bacon, D. (2003). Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Engineering*, 11(3): 563–588.

Curriculum Vitae

Name:	Cong Nie
Post-Secondary Education and Degrees:	University of Western Ontario, London, ON Ph.D. Statistics in Actuarial Science 2017 - 2019, 2020 - 2022
	University of Waterloo, Waterloo, ON MMath in Actuarial Science 2015 - 2016
	University of Western Ontario, London, ON BSc, Honour Specialization in Actuarial Science 2011 - 2015
Professional Designations	Associate of the Society of Actuaries (ASA) Associate of the Canadian Institute of Actuaries (ACIA) International Financial Reporting for Insurers (IFRI) Certificate
Honours and Awards:	Western Graduate Research Scholarship 2017 - 2019, 2020 - 2022
	Mathematics Graduate Experience Award Statistics & Actuarial Science Chair's Award 2015 - 2016
	Dean's Honour List 2011 - 2015
Related Work Experience:	Teaching and Research Assistant University of Western Ontario 2017 - 2019, 2020 - 2022

Publication:

Nie, C., Liu, X., Provost, S. B., and Ren, J. (2022). Markov chain Monte Carlo for Bayesian inference on the phase-type aging model. *North American Actuarial Journal*. Under revision.