Statistical modelling of queueing systems

Sarah Ellen James

Thesis submitted for the degree of

Doctor of Philosophy

in

Applied Mathematics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

February 5, 2022

Contents

Abstract

Queueing models are mathematical models used to describe queueing systems, such as healthcare systems or telecommunication systems. Standard queueing models such as the $M/M/\bullet$ and $M/PH/\bullet$ queueing models assume independence between the arrival process and the distribution of service times. For some queueing systems, this is a reasonable assumption. However, it is possible for some queueing systems to have dependence between how often customers arrive and how long each customer spends in service.

Intensive care units are generally described as complex queueing systems, in that a server is not clearly defined and patient admissions vary considerably and often depend on resource availability. In addition to this, studies have found evidence of a dependence between the patient admission process and the distribution of patient length of stay. Given that standard queueing models assume independence between the arrival process and the distribution of service times, such models are invalid for modelling the bed occupancy of an intensive care unit.

An alternative to modelling the bed occupancy of an intensive care unit is to use quasi-birth-and-death (QBD) processes, which not only allow for dependence between the patient admission process and the distribution of patient length of stay but also provide freedom in the distribution of time spent at each bed occupancy. However, limited research exists on the statistical modelling of queueing systems using QBD processes. Therefore, in this thesis we focus on developing statistical methods to fit various types of QBD processes to queueing system data, as well as a goodness of fit method to assess the fit of QBD processes to observed queueing system data.

Firstly, we develop two statistical fitting methods for level-dependent and level-independent QBD processes, respectively. These methods are based on the EM algorithm, since all that is observed while watching the evolution of a QBD process are the changes in level and the times at which those changes occurred. That is, the phase process remains hidden. We assess the accuracy of our methods by using simulated data from known QBD processes. In particular, we compare the stationary and transient behaviour of a known QBD process to what is expected under the fitted QBD process.

The statistical fitting of level-dependent QBD processes to queueing system data is advantageous in that we potentially gain valuable insight into the operation of a queueing system. However, such models can be over-parameterised and therefore cannot be used for prediction. We therefore develop a new class of QBD process called structured QBD processes which offer a reduction in the number of parameters through using observable behaviours of the queueing system. We then extend our statistical fitting method to structured QBD processes and assess the accuracy of our method by comparing the stationary and transient behaviour of several known structured QBD processes to what is expected under the respective fitted structured QBD process.

Since we often do not know the true QBD process for a queueing system, we develop a goodness of fit test which statistically determines if data observed from a queueing system is modelled as a realisation of a particular type of QBD process. We also develop methods to visually assess the fit of a QBD process to queueing system data, which is insightful in situations where the fitted QBD process does not capture the stationary and transient behaviour observed in the queueing system data. We then consider several numerical examples to demonstrate the application and performance of the goodness of fit test and usefulness of the diagnostic plots.

A benefit of modelling an intensive care unit using a level-dependent QBD process is that we gain valuable insight into the patient flow of the intensive care unit. However, structured QBD processes are more useful in that they have fewer parameters than a level-dependent QBD process, and hence can be used for prediction. Through the use of our goodness of fit test, we identify the best fitting structured QBD process which is then used to predict future behaviours under various scenarios.

The statistical methods developed in this thesis enable the fitting and analysis of QBD processes to provide meaningful insight and reliable predictions for queueing systems, including those with dependence between the arrival process and the distribution of service times. Therefore, our statistical methods for QBD processes provide an alternative to modelling intensive care units, particularly when there exists a dependence between the patient admission process and the distribution of patient length of stay.

Signed Statement

I, Sarah James, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

SIGNED: . . . . DATE: . . . .

Dedication

To my parents Ralda and Steven, for making everything possible and raising me to believe there are no limits to what you can achieve.

Acknowledgements

Firstly, I would like to thank my supervisors Simon Tuke, Nigel Bean, and Garique Glonek for encouraging me to pursue postgraduate study and providing endless support throughout my time at university.

To my parents, Ralda and Steven, my sister, Abbie, and my brother Stephen. Thank you for your constant support, unconditional love, and for always being there whenever I needed someone to chat to and have a laugh with. To my sister-in-law, Carrissa, and my nieces and nephew Madison, Hunter, and Jade. Thank you for always encouraging me to think about my next move and for sparking those little ideas to improve my work.

To my partner, Matt. Thank you for encouraging me to go further and reminding me that nothing is beyond my reach. I truly appreciate your love and endless support over the past few years.

To my closest friends, Bianca, Danielle, Kate, Laura, Max, and Shaniah. Thank you for all the fun and wonderful times we have had since we first met. I feel incredibly lucky to have such an amazing group of friends.

Lastly, thank you to all of my friends and colleagues who I have met at university and through the ARC Centre of Excellence for Mathematical and Statistical Frontiers. You have all shown me how supportive and vibrant the mathematics community can be, as well as the endless possibilities associated with a career in mathematics.

Chapter 1

Introduction

Waiting in a queue in a service facility of some kind is a common phenomenon. Mathematically speaking, service facilities are called queueing systems where customers or items arrive to receive a service and then leave the system once the service has been provided. Queueing systems may be simple in nature like a coffee shop or a supermarket, or complex like a telecommunication network or a surgery waiting list. Irrespective of complexity, the variation in arrivals and departures and the limited knowledge about the structure of a queueing system makes it difficult to predict behaviour when simply monitoring the system.

Analytical modelling of queueing systems involves detailed descriptions of various aspects of the system, such as the arrival process, the distribution of service times, the number of servers, and the capacity of the queueing system. The analysis of queueing systems aims to provide details on behaviour and performance, including the expected number of customers waiting in the queue and the proportion of time spent at full capacity.

Complex queueing systems such as intensive care units are particularly difficult to model, not only due to the unpredictable nature of patient admissions and the uncertainty of the definition of a server but also as a result of patient admissions depending on other patients' lengths of stay. Inherent with the current modelling process of intensive care units is the assumption of independence between the patient admission process and the distribution of patient length of stay, thus rendering standard queueing models invalid for queueing systems with dependence.

Quasi-birth-and-death (QBD) processes are Markov processes in two dimensions that provide a natural framework to model queueing systems with a dependence between the arrival process and the distribution of service times, as well as provide freedom in the distribution of sojourn time for each level. The modelling of the arrival process and the distribution of service times is a relatively straight-forward task when we can assume independence between the arrival process and distribution of service times. However, the modelling approach becomes less clear when dependence exists between the arrival process and distribution of service times. Therefore, we focus on the problem of statistically fitting arbitrary quasi-birth-and-death processes to queueing system data and developing methods to assess the fit of a fitted QBD process to queueing system data.

The data and work presented in this thesis is motivated by the limitations of the current modelling processes of intensive care units. Therefore, we begin with an introduction to intensive care units. In Chapter 2, we emphasise the importance of intensive care services within healthcare systems and describe the patient flow within an intensive care unit. Patient demand for intensive care services has particularly increased since the beginning of the COVID-19 outbreak, thus encouraging more mathematical research into the modelling of patient flow and bed occupancy in intensive care units using queueing theory.

Before we look further into modelling intensive care units, we need an understanding of random phenomena and queueing systems. In Chapter 3, we review the fundamentals of stochastic processes and Markov chains which form a basis for queueing theory. Stochastic processes are random processes that evolve over time, such as currency exchange rates or the number of calls to a telephone call centre. Under certain conditions, a stochastic process is called a Poisson process which is often used to model the number of arrivals to a queueing system.

Another important type of stochastic process is the Markov process which has the property that the future of the process is independent of the past, conditioned on the current state of the process. A special type of Markov process is the birth-and-death process which is often used to model the size of a population or the number of customers waiting in a queue. Birth-and-death processes are fundamental to the modelling of various types of queueing systems, such that they are used to derive the maximum likelihood estimates of elementary queueing models.

Queueing models are mathematical models used to describe queueing systems and provide a means of assessing the performance of a queueing system, such as how often the servers are busy and the amount of time a customer waits in the queue. In Chapter 4, we review Markovian queueing models which assume that customers arrive according to a Poisson arrival process and the amount of time each customer spends in service is exponentially distributed. These types of queueing models are examples of the general birth-and-death process, whereby the maximum likelihood estimates of the parameters are easily derived.

With a general understanding of the fundamentals of queueing theory, we then demonstrate how Markovian queueing models are used to model intensive care units. In Chapter 5, we model the bed occupancy of the intensive care unit at the Royal Adelaide Hospital using various types of Markovian queueing models. We draw attention to the invalidity of the assumptions of these queueing models by analysing the times between patient admissions and the distribution of patient length of stay, as well as discussing the dependence between patient admission and patient length of stay. As a result, standard queueing models are invalid in modelling the bed occupancy of intensive care units as they do not allow for a possible dependence between the arrival process and service times.

Alternatively, we can model queueing systems with dependence between the arrival process and the service time distribution using quasi-birth-and-death processes. In Chapter 6, we first review phase-type distributions which are often used to model service times in healthcare and telecommunication systems due to the balance between generality and tractability. Phase-type distributions also form a foundation for QBD processes, which are defined as a two-dimensional continuous-time Markov chain where the state of the process at any given time is represented by a level and phase. Similar to birth-and-death processes, a QBD process can move up or down a level, but the distribution of time between a change in level is explained by a phase process. However, we cannot fully observe the evolution of a QBD process, in that all that is observed are the changes in level and the times at which those changes occurred. That is, the phase process within a QBD process remains hidden.

Statistical methods have been developed to fit general-birth-and-death processes and phase-type distributions to queueing system data. However, limited research has been done on the statistical modelling of queueing systems using QBD processes. Therefore, in Chapter 7 we develop an EM algorithm to fit arbitrary QBD processes to queueing system data, where the unobserved data is assumed to be the phase transition process. In particular, we develop statistical modelling methods for both level-dependent and level-independent QBD processes. Using simulated queueing system data from a known QBD process, we assess the accuracy of each method by comparing the stationary and transient behaviour of the known QBD process to that of the QBD process fitted to the simulated data.

Level-dependent QBD processes have the ability to provide meaningful insight into the operation of a queueing system. However, such models can be over-parameterised and are thus not suitable for predicting future behaviour. Despite a level-independent QBD process having fewer parameters, a queueing system may have level-dependent behaviour which is therefore not captured by such a process. Therefore, in Chapter 8 we develop a new type of QBD process called the structured QBD process, such that we apply constraints to the block matrices in the infinitesimal generator matrix of the level-dependent QBD process. We then extend the EM algorithm developed in Chapter 7 to structured QBD processes. To assess the accuracy of this estimation method, we compare the stationary and transient behaviour of several known structured QBD processes to that of each respective structured QBD process fitted to data simulated from the known structured QBD process.

With the range of QBD processes now available to model a queueing system, the choice of the most appropriate QBD process becomes unclear. In Chapter 9, we develop a goodness of fit test that statistically determines whether observed queueing system data is modelled as a realisation of a particular type of QBD process. This goodness of fit test focuses on the stationary distribution of the observed queueing system data, as well as the transition probabilities between each level and the time spent in each level before moving up or down a level. Given the non-identifiability of QBD processes, the comparison of stationary and transient behaviour is per level, not by phase. Where discrepancies exist, diagnostic plots of the stationary and transient behaviour are used to determine where the QBD process fails to model the queueing system of interest. We then demonstrate the application and performance of our goodness of fit test through the use of several numerical examples and illustrate the statistical power and significance by means of a small simulation study.

In Chapter 10, we use the QBD process estimation methods developed in this thesis to model the bed occupancy of the Royal Adelaide Hospital intensive care unit. We consider various types of QBD processes and use our goodness of fit method to find the best fitting structured QBD process. We also demonstrate how we can use structured QBD processes to predict future bed occupancy behaviour of the Royal Adelaide Hospital intensive care unit under various scenarios.

In Chapter 11, we summarise our findings and discuss possible extensions to the work presented in this thesis.

Chapter 2

Intensive care units

2.1 Healthcare in Australia

Healthcare systems are an essential part of society that provide health care to assist in maintaining good health, prevent and manage disease, and reduce the risk of unnecessary disability and premature death. Healthcare in Australia has evolved remarkably over the years, as demonstrated by the various private and public health services located across the metropolitan and regional areas of Australia.

Prior to the outbreak of COVID-19, pressure on health services in Australia had increased due to depleted resources and an increase in demand for health care. Due to the complexity and self-organising nature of healthcare systems, attention has been drawn to examining the performance of health care services, particularly the intensive care services. Since the beginning of the COVID-19 outbreak, patient demand for intensive care surged across the world. This forced hospital staff and healthcare system modellers to improve workflow for diagnosis and isolation, management of disease and illness, and to increase the critical care bed capacity in intensive care units.

2.2 Intensive care units

An intensive care unit (ICU) is a crucial and limited resource in a hospital which provides high-level care to critically ill patients. This includes burns, cardiothoracic, general medicine, respiratory, surgical, and trauma cases, as well as patients who require life support. Generally speaking, patients are admitted to the ICU as either an elective patient or an emergency patient. Elective procedures are planned in advance and include heart surgery and neurological surgery. Emergency procedures are done in emergency situations, such as a road trauma, or a sudden and critical change in a person's health status, such as appendicitis or an aortic dissection.

Admission to the ICU depends on the state of the ICU and the status of the patient. The state of the ICU refers to the number of beds occupied, nurse and doctor availability, and any potential discharges from the ICU. The status of the patient refers to the age of the patient, the severity of the medical condition, the expected length of stay, and the expected outcome. Ideally, a patient is admitted to the ICU without delay. However, there are a finite number of resources in an ICU, so a decision regarding which patients are going to be admitted and when patients will be admitted needs to be made.

ICU management controls patient admissions and discharges, the treatment and length of stay of patients, re-admissions to the ICU, bed allocations, and bed capacity. In particular, ICU management carefully decides the order in which patients are admitted to the ICU so as to avoid worsening the health status of the patients waiting for a bed in the ICU.

ICUs often operate close to capacity but an increase in resource demand and uncertainty in the ICU can cause many negative effects both within the ICU and in other connected departments [2]. Hence, there is a strong interest in how to improve patient care with such a limited capacity.

2.3 Modelling intensive care units

An intensive care unit can be described as a queueing system, as illustrated in Figure 2.3.1. The main input into the intensive care units is patients and the service is the medical care provided to patients. Patients not only vary in health status, but also demographically and geographically which has been shown to influence the rate of patient admissions [46]. In the ICU, there are a finite number of resources which includes bed, monitors, staff, ventilators, and other medical equipment. The output from an ICU mostly includes treated patients who are being transferred to another unit or ward.

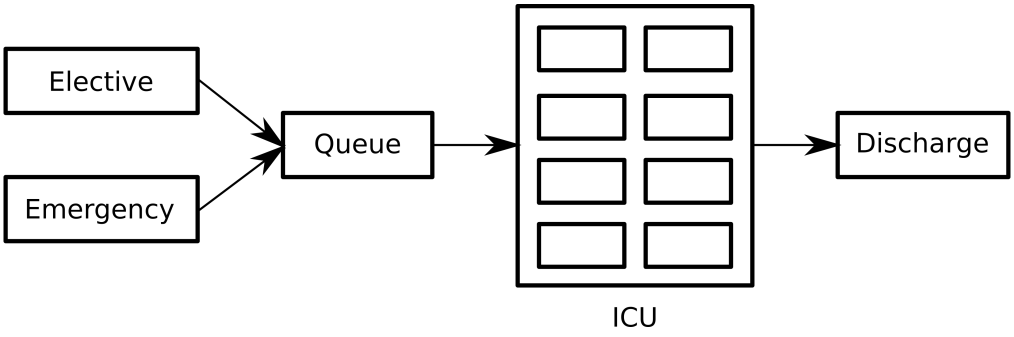


Figure 2.3.1: A general model of the intensive care unit.

The performance of health care services such as intensive care units is continuously monitored and focuses on waiting times, mortality rates, re-admission rates, and the throughput. Over the past few years, there has been an increasing amount of research into modelling patient flow in ICUs, with a particular focus on using statistical methods and queueing theory [2].

2.4 Summary

A considerable amount of planning and work is required for ICU management to ensure patients are provided with correct medical care, resources are in adequate supply and that the ICU is staffed with highly trained medical professionals. Admission rates into the ICU and the types of medical care provided to patients dynamically changes over time and may be a result of unpredictable disease outbreaks and the autonomous behaviour of patients. Additionally, the admission and discharge patterns of elective and emergency patients are considerably different, as well as the distributions of patient length of stay. Due to the unpredictable nature of bed occupancy in ICUs, researchers often use queueing theory to model the patient flow in ICUs to improve the operation of the ICU with such a limited capacity and limited resources.

Chapter 3

Stochastic processes and Markov chains

When aiming to understand the behaviour of various natural and artificial processes, researchers often consider a sequence or family of random variables to represent such phenomena over time. In mathematics, these processes are called stochastic processes. In this chapter, we provide a background to stochastic processes and Markov chains, which form a basis for queueing theory and the modelling of queueing systems discussed in the following chapters.

3.1 Stochastic processes

A stochastic process is defined as a family of random variables $\{X(t); t \in T\}$, where T is the index set. In many applications, the index t is used to model time. In other words, a stochastic process is a random process evolving over time. For example, if the index set such that $T = \{0, 1, 2, \ldots\}$, then $\{X(t); t \in T\}$ is a discrete-time stochastic process. If the index set is continuous, then $\{X(t); t \in T\}$ is a continuous-time stochastic process.

Furthermore, $\{X(t); t \in T\}$ is called a discrete-space stochastic process if the random variables $X(t)$ are discrete, and $\{X(t); t \in T\}$ is called a continuous-space stochastic process if the random variables $X(t)$ are continuous.

Characteristics of stochastic processes include independence, stationarity, and homogeneity.

- *Independence*: If there is no structure or dependence between the random variables, $\{X(t); t \in T\}$, then the stochastic process is called an independent stochastic process.

- *Stationarity*: Generally, if the evolution of a stochastic process depends on the time it started, then it is non-stationary. If we shift the time of origin and the stochastic process remains invariant, then it is called stationary. Mathematically speaking, a stochastic process is stationary if for any value $d \in \mathbb{R}$,

  $$P(X(t_1) \leq x_1, \dots, X(t_n) \leq x_n) = P(X(t_1 + d) \leq x_1, \dots, X(t_n + d) \leq x_n)$$

  for all values of $n \in \mathbb{N}$ and $t_1, t_2, \dots, t_n \in \mathbb{R}$.

- *Time homogeneity*: If the transitions between states are independent of time, then the stochastic process is called time-homogeneous. Alternatively, if the transitions between states depends on time, then the stochastic process is non-time-homogeneous.

3.1.1 Poisson process

A fundamental stochastic process used in queueing theory is the Poisson process. Poisson processes are often used to model the occurrence of events, such as the number of telephone calls to a call centre.

Let $\{N(t); t \geq 0\}$ be the total number of events by time $t \geq 0$ where

- $N(t) \geq 0$,
- $N(t)$ is integer,
- if $s < t$, then $N(s) \leq N(t)$ which means that $\{N(t); t \geq 0\}$ is a non-decreasing function of time, and
- $N(t) - N(s)$ represents the number of events that occurred during the time interval $[s, t]$.

Then the stochastic process $\{N(t); t \geq 0\}$ is called a counting process.

Furthermore, we say that a counting process $\{N(t); t \geq 0\}$ is a Poisson process if

- $N(0) = 0$,
- the number of events that occur in any time interval depends only on the length of the time interval,
- the events are mutually independent in non-overlapping time intervals, such that for any $s > t > u > v > 0$, $N(s) - N(t)$ and $N(u) - N(v)$ are independent, and
- for sufficiently small h and some positive constant λ,

  $$P(\text{one event in } (t, t+h]) = \lambda h + o(h),$$

  where $\lim_{h \to 0} \frac{o(h)}{h} = 0$.

For any Poisson process $\{N(t); t \geq 0\}$, $N(t_2) - N(t_1)$ and $N(t_2 + u) - N(t_1 + u)$ have the same distribution for any $u > 0$ and $t_2 > t_1$. That is, the Poisson process has stationary increments, although it is not stationary itself.

Researchers often describe a stochastic process by using the inter-arrival times (time between arrivals or events) defined as

$$\tau_i = t_{i+1} - t_i,$$

where t_i is the time of the i^{th} event and t_{i+1} is the time of the $i+1^{th}$ event. Notice that for a Poisson process $\{N(t); t \geq 0\}$,

$$P(\tau > t) = P(N(t) = 0) = \exp(-\lambda t).$$

By the properties of independent and stationary increments of a Poisson process, the inter-arrival times are therefore exponentially distributed with rate λ.

3.2 Markov processes

Another type of stochastic process is a Markov process. Markov processes are widely used in queueing theory and also in many other areas across theoretical and applied science.

A Markov process, $\{X(t); t \geq 0\}$, with state space $\mathcal{S}$, is a stochastic process with the property that given the state of the process at time t, $X(t)$, the value of $X(s)$ for $s > t$ is not affected by the value of $X(u)$ for $u < t$. That is,

$$P(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, \ldots, X(t_1) = x_1) = P(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}),$$

for any $0 \leq t_1 < t_2 < \ldots < t_n$ and any $x_n \in \mathcal{S}$. In other words, the future of the process is conditionally independent of the past, given the current state of the process.

3.2.1 Continuous-time Markov chains

A continuous-time Markov chain (CTMC) is a Markov process with a discrete state space, $\mathcal{S}$, which describes the state of the process at some time $t > 0$. More specifically, let $\{X(t); t \in T\}$ be a sequence of random variables with a state space $\mathcal{S}$. Then the continuous-time stochastic process $\{X(t); t \in T\}$ is a continuous-time Markov chain if for $u \leq s$,

$$P(X(t+s) = j | X(u) = i_u, X(s) = i_s) = P(X(t+s) = j | X(s) = i_s),$$

for all $s, t \in [0, \infty)$ and all $i_s, i_u, j \in \mathcal{S}$. So the probability of moving from state i_s to state j depends only on the present state $X(s)$. As before, this is known as the memoryless property.

Typically, the interactions between states in a continuous-time Markov chain are described by the rates at which the transitions occur. For example, the rate of moving from state $i \in \mathcal{S}$ to state $j \in \mathcal{S}$ at time $t > 0$ is denoted as $q_{i,j}(t)$. The transition rates are contained within the transition rate matrix, Q, otherwise known as the infinitesimal generator matrix, such that

$$\mathcal{Q} = \begin{bmatrix} q_{0,0} & q_{0,1} & q_{0,2} & \cdots & q_{0,j} & \cdots \\ q_{1,0} & q_{1,1} & q_{1,2} & \cdots & q_{1,j} & \cdots \\ q_{2,0} & q_{2,1} & q_{2,2} & \cdots & q_{2,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{i,0} & q_{i,1} & q_{i,2} & \cdots & q_{i,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Assuming that the continuous-time Markov chain, $X(t)$, is time-homogeneous such that

$$P(X(t+s)=j|X(s)=i)=P(X(t)=j|X(0)=i),$$

for all $i,j\in\mathcal{S}$ and $s,t\in[0,\infty)$, the entries of Q are defined as

$$q_{i,j}=\lim_{h\to 0^+}\frac{p_{i,j}(h)}{h},\quad i,j\in\mathcal{S}, j\neq i,\text{ and}$$

$$q_{i,i}=\lim_{h\to 0^+}\frac{p_{i,i}(h)-1}{h},\quad i\in\mathcal{S},$$

where $p_{i,j}(h)=P(X(t+h)=j|X(t)=i)$, and $q_{i,j}$ must satisfy the following three conditions:

- $q_{i,j}\geq 0$ for $i,j\in\mathcal{S}$, $j\neq i$,
- $q_{i,i}\leq 0$ for $i\in\mathcal{S}$, and
- $\displaystyle\sum_{j\in\mathcal{S}}q_{i,j}=0,$ for all $i\in\mathcal{S}$.

A physical interpretation of $q_{i,j}$ is that for small h and for $i\neq j$,

$$p_{i,j}(h)=q_{i,j}h+o(h),$$

where $o(h)$ denotes a function $f(h)$ that satisfies $\lim_{h\to 0}\frac{f(h)}{h}=0$. That is,

$$P(\text{moving from state } i \text{ to state } j \text{ in some small time } h)\approx q_{i,j}h.$$

So we interpret $q_{i,j}$ as the instantaneous rate of moving from state $i\in\mathcal{S}$ to state $j\in\mathcal{S}$.

Similarly, a physical interpretation of $q_{i,i}$ is that for small h and for $i \in \mathcal{S}$,

$$1 - p_{i,i}(h) = -q_{i,i}h + o(h) = \sum_{\substack{j \neq i \\ j \in \mathcal{S}}} q_{i,j}h + o(h).$$

Hence, we interpret $-q_{i,i}$ as the instantaneous rate of moving out of state $i \in \mathcal{S}$.

We can also think of $-q_{i,i}$ as the hazard rate for the time spent in state $i \in \mathcal{S}$, where the hazard rate is defined as

$$h(t) = \frac{f(t)}{1 - F(t)},$$

where $f(t)$ and $F(t)$ are the probability density and cumulative distribution functions representing the time spent in state $i \in \mathcal{S}$, respectively. The only distribution which has a constant hazard rate is the exponential distribution. Hence,

$$F(t) = 1 - e^{q_{i,i}t},$$

for $t \geq 0$, which indicates that the time spent in state $i \in \mathcal{S}$ is exponentially distributed with rate

$$-q_{i,i} = \sum_{\substack{j \neq i \\ j \in \mathcal{S}}} q_{i,j}.$$

Let's now consider the probability that the process moves from state $i \in \mathcal{S}$ to state $j \in \mathcal{S}$ in time $t > 0$,

$$p_{i,j}(t) = P(X(s+t) = j | X(s) = i), \quad \text{for all } i, j \in \mathcal{S},$$

which has the property that

$$p_{i,j}(t + \tau) = \sum_{k \in \mathcal{S}} p_{i,k}(t) p_{k,j}(\tau), \quad \text{for all } i, j \in \mathcal{S}. \tag{3.2.1}$$

The equations presented in Equation (3.2.1) are known as the Chapman-Kolmogorov equations and are interpreted as the probability of moving from state $i \in \mathcal{S}$ to state $j \in \mathcal{S}$ in time $t + \tau > 0$ is equal to the probability of moving from state $i \in \mathcal{S}$ to state $k \in \mathcal{S}$ in time $t > 0$ and then from state $k \in \mathcal{S}$ to state $j \in \mathcal{S}$ in time $\tau > 0$, for all $k \in \mathcal{S}$.

From the Chapman-Kolmogorov equations, we obtain the Kolmogorov backward and forward differential equations of a continuous-time Markov chain which are defined as

$$\frac{dp_{i,j}(t)}{dt} = \sum_{k \in \mathcal{S}} q_{i,k} p_{k,j}(t), \quad i, j \in \mathcal{S}, \text{ and}$$

$$\frac{dp_{i,j}(t)}{dt} = \sum_{k \in \mathcal{S}} p_{i,k}(t) q_{k,j}, \quad i, j \in \mathcal{S},$$

respectively.

Suppose $\mathcal{P}(t)$ is the transition function of a time-homogeneous continuous-time Markov chain where

$$p_{i,j}(t) = P(X(t) = j | X(0) = i), \quad \text{for all } i, j \in \mathcal{S},$$

and let $\pi(0)$ be the initial probability distribution. If the limit

$$\lim_{t \to \infty} P(t)$$

exists, then the limiting probabilities are defined as

$$\pi_j = \lim_{t \to \infty} p_{i,j}(t).$$

for all $i, j \in \mathcal{S}$. Assuming the limiting distribution $\boldsymbol{\pi}$ exists,

$$\lim_{t \to \infty} \frac{dp_{i,j}(t)}{dt} = 0.$$

Hence, the Kolmogorov forward differential equations imply that

$$\begin{aligned}
0 &= \lim_{t\to\infty} \sum_{k\in\mathcal{S}} p_{i,k}(t) q_{k,j} \\
&= \sum_{k\in\mathcal{S}} \lim_{t\to\infty} p_{i,k}(t) q_{k,j}, \text{ by Fatou's Lemma,} \\
&= \sum_{k\in\mathcal{S}} \pi_k q_{k,j} \\
&= \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_k q_{k,j} + \pi_j q_{j,j},
\end{aligned}$$

which leads to the global balance equations defined as

$$\sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_k q_{k,j} = \pi_j \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} q_{j,k}.$$

Hence, the equilibrium equations of a continuous-time Markov chain are defined as

$$\sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_k q_{k,j} = \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_j q_{j,k},$$

where the left-hand side represents the probability flux into state $j \in \mathcal{S}$ and the right-hand side represents the probability flux out of state $j \in \mathcal{S}$.

3.3 Birth-and-death processes

A particular type of continuous-time Markov chain used in queueing theory is a birth-and-death process. Birth-and-death processes are often used to model the size of a population, such that a birth is defined as an increase by one individual and a death is defined as a decrease by one individual. Other applications include modelling the number of customers in a supermarket or doctor's surgery, where a birth is considered as customer or patient arriving and a death is considered as a customer or patient leaving.

In birth-and-death processes, the birth rate when the population is of size $k \in \mathcal{S}$ is defined as λ_k and the death rate when the population is of size $k \in \mathcal{S}$ is defined as μ_k. A key property of birth-and-death processes is that transitions only occur between neighbouring states, as shown in the state transition diagram presented in Figure 3.3.1. That is, if $X(t)$ represents the size of the population at time $t \geq 0$ and is such that $X(t) = n$, then $X(s) = n+1$ or $X(s) = n-1$ where $s > t$ is the time of the next event.

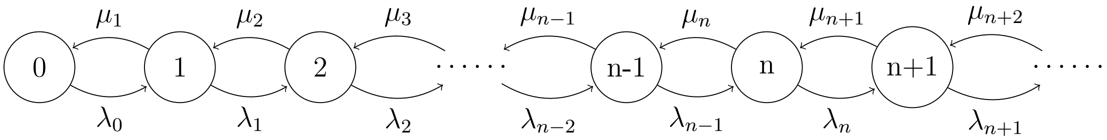


Figure 3.3.1: State transition diagram for the general birth-and-death process.

In terms of a CTMC, the birth rate is defined as

$$\lambda_k = q_{k,k+1}, \quad \text{for } k \in \mathcal{S},$$

and the death rate is defined as

$$\mu_k = q_{k,k-1}, \quad \text{for } k \in \mathcal{S}/\{0\}.$$

Given that transitions only occur between neighbouring states, we have that

$$q_{j,k} = 0, \quad \text{for } k \in \mathcal{S}/\{j-1, j, j+1\}, \text{ and } j \in \mathcal{S}.$$

Hence, the transition rates $-q_{k,k}$ are defined as

$$-q_{k,k} = \sum_{\substack{j \neq k \\ j \in \mathcal{S}}} q_{k,j} = \begin{cases} \lambda_k + \mu_k, & k \in \mathcal{S}/\{0\}, \\ \lambda_0, & k = 0. \end{cases}$$

So the infinitesimal generator matrix for a general birth-and-death process is written as

$$\mathcal{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Using the state transition diagram and the global balance equations, we see that the equilibrium equations are

$$\pi_j(\lambda_j+\mu_j) = \pi_{j+1}\mu_{j+1} + \pi_{j-1}\lambda_{j-1}, \quad j \geq 1,$$

$$\pi_0\lambda_0 = \pi_1\mu_1.$$

Re-writing, we obtain

$$\pi_{j+1}\mu_{j+1} - \pi_j\lambda_j = \pi_j\mu_j - \pi_{j-1}\lambda_{j-1},$$

which is of the form $A_{j+1} = A_j$, where $A_j = \pi_j\mu_j - \pi_{j-1}\lambda_{j-1}$, for $j \geq 1$.

We know that $A_1 = 0$ using $\pi_0\lambda_0 = \pi_1\mu_1$, so $A_j = 0$ for all $j \geq 1$. Hence,

$$\pi_j\mu_j = \pi_{j-1}\lambda_{j-1}, \text{ for all } j \geq 1.$$

Therefore, for $j \geq 1$,

$$\begin{aligned} \pi_j &= \pi_{j-1}\frac{\lambda_{j-1}}{\mu_j} \\ &= \pi_0\frac{\lambda_0}{\mu_1}\frac{\lambda_1}{\mu_2}\cdots\frac{\lambda_{j-1}}{\mu_j} \\ &= \pi_0\prod_{l=0}^{j-1}\frac{\lambda_l}{\mu_{l+1}}. \end{aligned}$$

Using the fact that $\sum_{i=0}^{\infty} \pi_i = 1$, we have that

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}.$$

3.3.1 Transition probabilities

Consider the time interval $(t, t + dt)$, for some small value dt. The probability of a birth in the time interval $(t, t + dt)$ is

$$P(X(t + dt) = k + 1 | X(t) = k) = \lambda_k dt + o(dt),$$

the probability of a death in the time interval $(t, t + dt)$ is

$$P(X(t + dt) = k - 1 | X(t) = k) = \mu_k dt + o(dt),$$

and the probability of no births or deaths in the time interval $(t, t + dt)$ is

$$P(X(t + dt) = k | X(t) = k) = 1 - (\lambda_k + \mu_k)dt + o(dt),$$

where $o(dt)$ denotes a function $f(dt)$ that satisfies $\lim_{dt \to 0} \frac{f(dt)}{dt} = 0$.

3.3.2 Waiting times

For a continuous-time Markov chain, the time spent waiting in state $k \in \mathcal{S}$ is exponentially distributed and is also independent of the next jump. So the waiting time, W, until the next event is exponentially distributed with rate $\lambda_k + \mu_k$.

3.3.3 Likelihood and parameter estimation

Suppose we continuously observe a general birth-and-death process over some finite time period, $[0, t]$. That is, suppose we observed N jumps at times $t_1, t_2, \ldots, t_N$, where $t_0 = 0$ and $t_N < t$. Various methods have been developed to estimate the parameters of the general birth-and-death process using maximum likelihood estimation [14, 31]. In this section, we present an overview of the likelihood and calculation of the maximum likelihood estimators.

We define $B_n = 1$ if there is a birth at the n^{th} jump, and $B_n = 0$ if there is a death at the n^{th} jump. Then the likelihood of the sequence of observations, $\{X(\tau); 0 < \tau < t\}$ is as follows.

$$\begin{aligned} L &= \prod_{i=1}^{N} \big(P(W_i = t_i - t_{i-1} | X(t_{i-1})) \times P(\text{birth}|X(t_{i-1}))^{B_i} P(\text{death}|X(t_{i-1}))^{1-B_i} \big) \\ &\qquad \times P(W_{N+1} \geq t - t_N | \text{no births or deaths}, X(t_N)) \\ &= \prod_{i=1}^{N} \left((\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})}) e^{-(\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})})(t_i - t_{i-1})} \times \frac{\lambda_{X(t_{i-1})}^{B_i} \mu_{X(t_{i-1})}^{1-B_i}}{\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})}} \right) \\ &\qquad \times e^{-(\lambda_{X(t_N)} + \mu_{X(t_N)})(t - t_N)} \\ &= \prod_{i=1}^{N} \left(\lambda_{X(t_{i-1})}^{B_i} \mu_{X(t_{i-1})}^{1-B_i} e^{-(\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})})(t_i - t_{i-1})} \right) \times e^{-(\lambda_{X(t_N)} + \mu_{X(t_N)})(t - t_N)}. \end{aligned}$$

Let T_k be the total amount of time spent in state $k \in \mathcal{S}$, U_k be the total number of births from state $k \in \mathcal{S}$, and D_k be the total number of deaths from state $k \in \mathcal{S}$. That is, for $k \in \mathcal{S}$,

$$T_k = \sum_{i=1}^{N} (t_i - t_{i-1}) I(X(t_{i-1}) = k),$$

$$T_{X(t_N)} = t - t_N,$$

$$U_k = \sum_{i=1}^{N} I(X(t_{i-1}) = k, B_i = 1), \text{ and}$$

$$D_k = \sum_{i=1}^{N} I(X(t_{i-1}) = k, B_i = 0),$$

respectively. Then the likelihood is simplified to the following:

$$L = \prod_{k=0}^{\infty} \lambda_k^{U_k} \mu_k^{D_k} e^{-(\lambda_k + \mu_k) T_k}.$$

Hence, the log-likelihood is written as

$$\ell = \sum_{k=0}^{\infty} U_k \log(\lambda_k) + \sum_{k=0}^{\infty} D_k \log(\mu_k) - \sum_{k=0}^{\infty} (\lambda_k + \mu_k) T_k.$$

Using the log-likelihood, we find that the maximum likelihood estimates $\hat{\theta} = (\{\widehat{\lambda}_k; k = 0, 1, \ldots\}, \{\widehat{\mu}_k; k = 1, 2, \ldots\})$ are defined as

$$\widehat{\lambda}_k = \frac{U_k}{T_k}, \quad k = 0, 1, \ldots, \text{ and}$$

$$\widehat{\mu}_k = \frac{D_k}{T_k}, \quad k = 1, 2, \ldots.$$

Crawford *et al.* [14] developed an EM algorithm for estimating the parameters of a general birth-and-death process. The reader is referred to the paper for further discussion of the parameter estimation method.

3.4 Summary

Stochastic processes and Markov chains are used to model many random phenomena and are fundamental to the development and understanding of queueing theory. In particular, the general birth-and-death process is related to various elementary queueing models and is used to derive the maximum likelihood estimates of the parameters of such queueing models. In the next chapter, we review queueing systems and discuss the theory behind the elementary modelling of queueing systems.

Chapter 4

Queueing theory

Queueing systems, such as call centres, production plants, and hospitals, are systems of servers or resources where customers or units queue up to receive a service and leave upon service completion. Queueing models are mathematical models that are used to describe queueing systems and are traditionally determined by understanding the structure of the queueing system and analysing data. Queueing theory is an area of research within applied probability and operations research that focuses on modelling the performance of queueing systems. In this chapter, we briefly review queueing theory and present an overview of elementary queueing models.

4.1 Queueing systems

A queueing system is a service facility where customers arrive to receive a service and leave upon service completion. Queueing systems are generally described by the arrival process, the distribution of service times, the number of servers in the queueing system, the capacity of the queueing system, and the queueing discipline.

An illustration of customer flow through a service facility is shown in Figure 4.1.1. The customers arrive to the service facility in a random manner. Most service facilities have a limited number of servers, so customers will join a queue and await service if there is no available server. When a server becomes available, the customer moves into service and upon completion, the customer will leave the service facility.

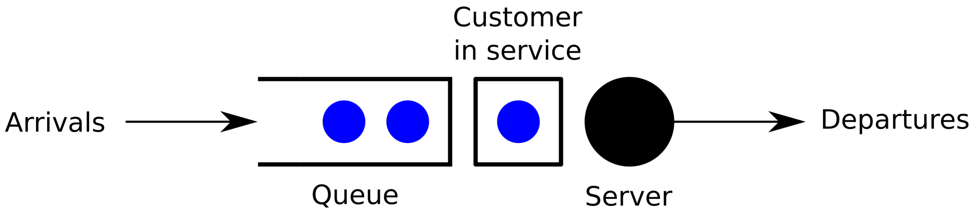


Figure 4.1.1: An illustration of a simple queueing system.

4.1.1 The arrival process

The arrival process for a queueing system is a stochastic process and is often described by the probability distribution of the times between customer arrivals, otherwise known as the inter-arrival times. However, there is no restriction to how many customers that can arrive at any point in time. That is, customers can arrive one by one or in batches. It is also possible for the arrival process to be time-dependent, such as the arrival rate depending on the hour of the day or the day of the week, which is known as a non-stationary arrival process. If the arrival process does not depend on time, then it is known as a stationary arrival process.

4.1.2 The service process

The service process of a queueing system describes the manner in which customers are served. The service process is typically characterised by the distribution of service times. The service time of customers can be affected by the type of service required and the service rate of the server. Depending on the queueing system, the service time may also depend on the number of customers waiting in the queue, which is known as a state-dependent queueing system. Similar to the arrival process, if the service times are time-dependent, then the service process is called a non-stationary service process. Otherwise, it is called a stationary service process.

4.1.3 Number of servers

Queueing systems may have any number of servers that provide a service of some kind. When the queueing system only has one server, it is called a single-server queueing system. When the queueing system has more than one server, it is known as a multi-server queueing system.

In a multi-server queueing system with parallel servers, several customers are served simultaneously. Intuitively, we see that the more servers in a queueing system, the lower the amount of time customers have to wait in the queue. In a queueing system where the servers are in a series, customers must complete the service at each server in a specified order before leaving the queueing system.

Another type of multi-server queueing system is the infinite server queueing system. The infinite server queueing system assumes an unlimited number of servers which means that all customers are immediately served upon arrival without waiting. The importance of an infinite server queueing system is demonstrated by its ability to provide insight into the service demand and performance of the queueing system.

4.1.4 Queue and system capacity

Many service facilities have a limited number of servers, and so customers will join a queue and await service if there is no available server. In addition to this, service facilities may have a finite queue capacity which are known as finite queueing systems. So when the number of customers in the queue reaches capacity, no further customers will be able to enter and instead those customers are lost to the system.

For service facilities with infinite capacity, such as a finite number of servers and unlimited queueing space, customers move directly into the queue and proceed to service when a server becomes available.

4.1.5 The queueing discipline

The queueing discipline of a queueing system refers to the order in which customers are chosen from the queue and taken into service. Generally, there are two types of queueing disciplines: static and dynamic.

Static queueing disciplines are based on the individual customer's status in the queue. Examples of static queueing disciplines include First In First Out (FIFO) and Last Come First Serve (LCFS).

- *FIFO*: The first customer at the head of the queue is the next customer to be taken into service.
- *LCFS*: The last customer at the tail of the queue is the next customer to be taken into service.

Dynamic queueing disciplines are based on the individual customer's attributes in the queue. Examples of dynamic queueing disciplines include Serve In Random Order (SIRO) and priority queueing.

- *SIRO*: A randomly selected customer in the queue is the next customer to be taken into service.
- *Priority*: The customer with the highest priority is the next customer to be taken into service. Priority queues can be divided into two types; non-preemptive priority queue, where a customer in service cannot be interrupted and a preemptive priority queue, where a customer in service can be interrupted by a higher priority customer.

In other queueing systems, the service may be shared equally between customers. This is known as the processor sharing queueing discipline.

4.2 Kendall's notation

Kendall's notation is the standard notation used to describe a queueing system. Kendall first introduced the form A/B/C but it has since been extended to A/B/C/D/E/F, where

- A describes the inter-arrival time distribution,
- B describes the service time distribution,
- C denotes the number of servers in the system,
- D denotes the system capacity,
- E denotes the size of the calling population, and
- F describes the queueing discipline.

The standard symbols used in Kendall's notation are given in Table 4.2.1. For example, a single-server queueing system with Poisson arrivals, exponential service times, infinite system capacity, infinite calling population, and a FIFO queueing discipline is written as M/M/1/∞/∞/FIFO, or M/M/1 using the shorthand notation.

Characteristic	Symbol	Description
Inter-arrival time distribution (A)	M	Markovian
Service time distribution (B)	D	Deterministic
	G	General
	GI	General Independent
	PH	Phase-type distribution
	E_k	Erlang with k phases
	H_k	Mixture of k exponentials
	C_k	Coxian distribution with k phases
Number of servers (C)	1, 2, $\ldots$, ∞	
System capacity (D)	1, 2, $\ldots$, ∞	
Size of the calling population (E)	1, 2, $\ldots$, ∞	
Queueing discipline (F)	FIFO	First In First Out
	LCFS	Last Come First Serve
	SIRO	Serve In Random Order
	Priority	Customer with the highest priority
	Processor sharing	Shared service

Table 4.2.1: Kendall's queueing notation.

4.3 Performance measures

Once a suitable queueing model has been identified for a queueing system, focus is then drawn to the performance of the queueing system under the selected queueing model. The analysis of the performance of a queueing system typically involves estimating performance measures, such as the server utilisation, the distribution of the number of customers in the queue/system, and the distribution of the amount of time a customer waits in the queue/system. These performance measures assist in quantifying potential improvements to the system, such as changing the capacity of the queueing system or changing the number of servers.

4.4 Little's law

An intuitive assessment of the efficiency of a queueing system is that the average number of customers in a queueing system depends on the average waiting time in the queue and the average arrival rate of customers to the queueing system.

Let $A(t)$ be the number of arrivals in the time interval $[0, t]$ and $D(t)$ be the number of departures from the system in the time interval $[0, t]$. Then the number of customers in the system at time $t \geq 0$ is defined as

$$N(t) = A(t) - D(t).$$

For the time interval $[0, t]$, the average arrival rate is defined as

$$\lambda(t) = \frac{A(t)}{t}, \quad t > 0.$$

Let $S(t)$ denote the total amount of time customers spent in the system during the time interval $[0, t]$, as illustrated in Figure 4.4.1. Then the average time spent in the system per customer in the interval $[0, t]$ is defined as

$$\bar{T}(t) = \frac{S(t)}{A(t)}.$$

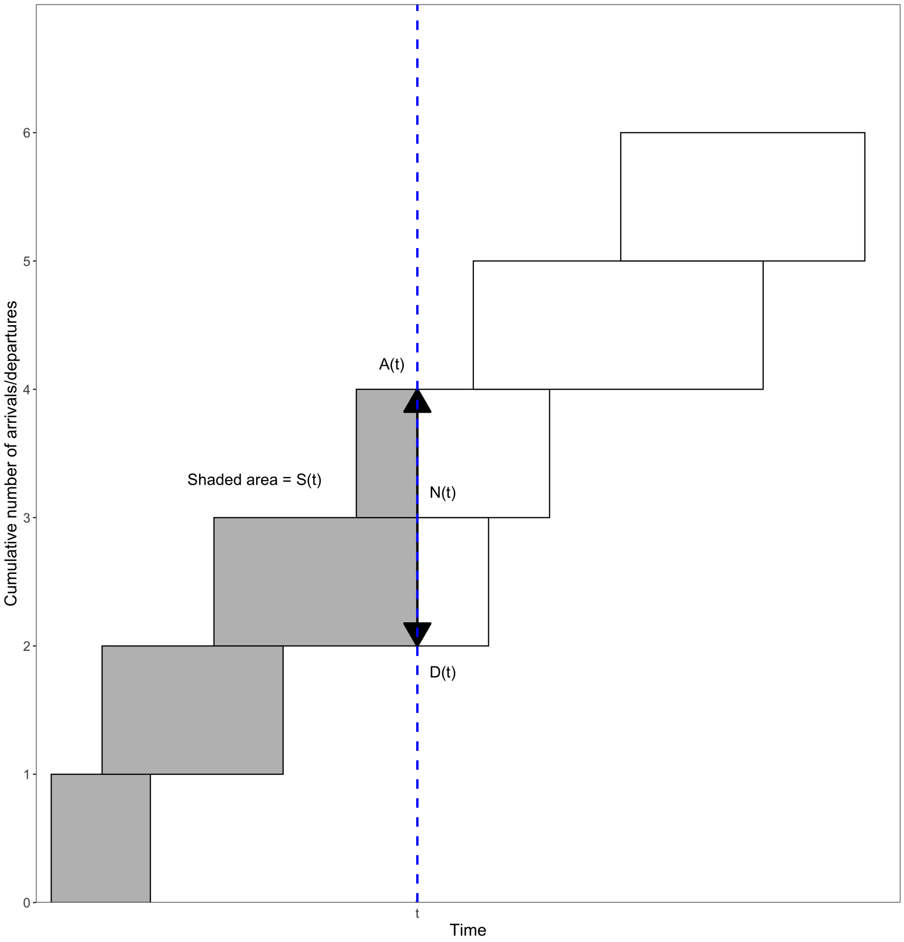


Figure 4.4.1: Graphical representation of the number of arrivals and departures over time.

Lastly, we define the average number of customers in the system in the interval $[0, t]$ as

$$\bar{N}(t) = \frac{S(t)}{t}.$$

It then follows that,

$$\begin{aligned}\bar{N}(t) &= \frac{S(t)}{t},\\ &= \frac{S(t)}{A(t)} \times \frac{A(t)}{t},\\ &= \lambda(t)\bar{T}(t).\end{aligned}$$

Taking the limit as $t \to \infty$, let

$$\bar{N} = \lim_{t\to\infty} \bar{N}(t),$$

$$\bar{T} = \lim_{t\to\infty} \bar{T}(t), \text{ and}$$

$$\lambda = \lim_{t\to\infty} \lambda(t).$$

Hence, we have that

$$\bar{N} = \lambda\bar{T},$$

where λ is the average arrival rate of the queueing system. This result is known as Little's Law. More explicitly, Little's Law states that the average number of customers in the system is equal to the average arrival rate of customers multiplied by the average time spent in the system. Given that Little's Law does not depend on the arrival process, the distribution of service times, the capacity of the system, or the queueing discipline, it can be used as a performance measure for any type of queueing system.

4.5 Elementary queueing models

The most studied queueing models are those with Poisson arrivals and exponential service times, which are examples of a general birth-and-death process. In this section, we review various Markovian queueing models of the form M/M/·/·.

4.5.1 M/M/1 queue

The first Markovian queueing system that we consider is a single-server queuing system with an infinite queue capacity and a FIFO queueing discipline, as illustrated in Figure 4.5.1.

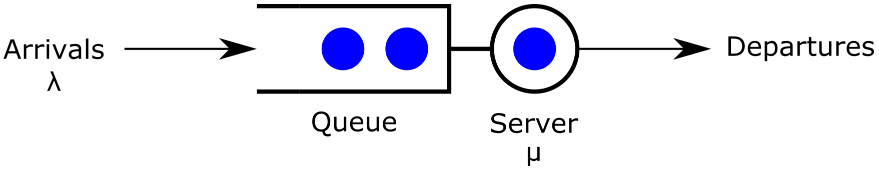


Figure 4.5.1: Example of a single-server queue. The blue dots represent customers.

The M/M/1 queueing system has a Poisson arrival process with parameter λ and exponentially distributed service times with parameter μ. The state transition diagram for the M/M/1 queue is shown in Figure 4.5.2.

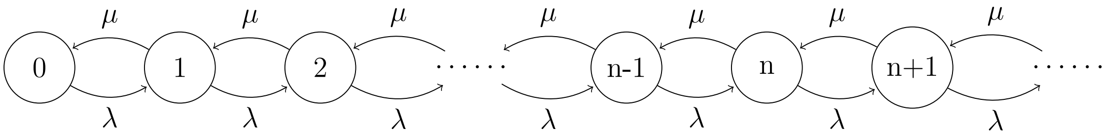


Figure 4.5.2: State transition diagram for the M/M/1 queue.

The transition rates for a M/M/1 queueing model with state space $\mathcal{S} = \{0, 1, 2, \ldots\}$ are

$$q_{i,i+1} = \lambda, \quad i \in \mathcal{S},$$

$$q_{i,i-1} = \mu, \quad i \in \mathcal{S}\backslash\{0\}, \text{ and}$$

$$q_{i,j} = 0, \quad i \in \mathcal{S}, j \in \mathcal{S}\backslash\{i, i+1, i-1\}.$$

So the infinitesimal generator matrix is

$$\mathcal{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Using the stationary distribution of a birth-and-death process from Section 3.3, we find that the stationary distribution for an M/M/1 queue is

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k, \quad k = 0, 1, \dots,$$

where

$$\pi_0 = \left(\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n\right)^{-1} = 1 - \frac{\lambda}{\mu}.$$

4.5.2 M/M/c queue

The second Markovian queueing system we consider is a system with c identical servers, infinite queue capacity, and a FIFO queueing discipline, as illustrated in Figure 4.5.3.

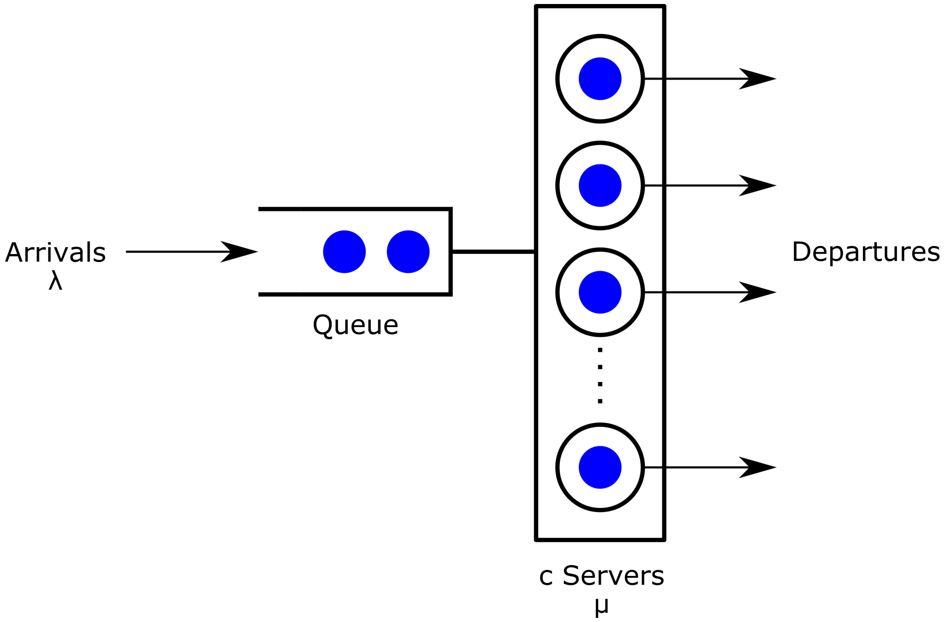


Figure 4.5.3: Example of a queue with c identical servers. The blue dots represent customers.

The M/M/c queueing system has a Poisson arrival process with parameter λ and each server provides independent and identically distributed exponential service at rate μ. If there are $n < c$ customers in service, then the effective service rate is equal to $n\mu$. If there are $n \geq c$ customers in the system, then the effective service rate is equal to $c\mu$.

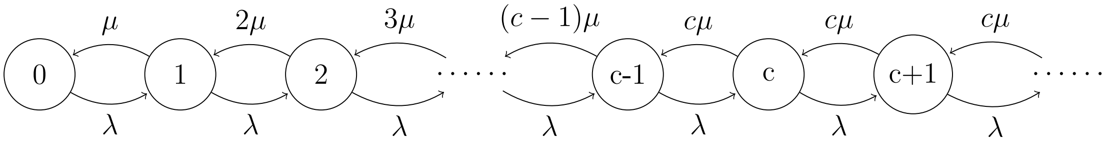


Figure 4.5.4: State transition diagram for the M/M/c queue.

The state transition diagram shown in Figure 4.5.4 illustrates that the transition rates are

$$q_{i,i+1} = \lambda, \quad i \in \mathcal{S},$$

$$q_{i,i-1} = min(i\mu, c\mu), \quad i \in \mathcal{S} \backslash \{0\}, \text{ and}$$

$$q_{i,j} = 0, \quad i \in \mathcal{S}, j \in \mathcal{S} \backslash \{i, i+1, i-1\}.$$

So the infinitesimal generator matrix is

$$\mathcal{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & c\mu & -(\lambda+c\mu) & \lambda & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Using the stationary distribution of a birth-and-death process from Section 3.3, we find that the stationary distribution for an M/M/c queue is

$$\pi_k = \begin{cases} \pi_0 \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k, & k \leq c, \\ \pi_0 \frac{1}{c^{k-c} c!} \left(\frac{\lambda}{\mu}\right)^k, & k > c, \end{cases}$$

where

$$\pi_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n! \mu^n} + \frac{\lambda^c}{c! \mu^c} \frac{1}{c - \frac{\lambda}{c\mu}} \right)^{-1}.$$

4.5.3 M/M/∞ queue

Next, we consider a Markovian queueing system with an infinite number of identical servers, infinite queue capacity, and a FIFO queueing discipline, as illustrated in Figure 4.5.5. Since there are an infinite number of servers in this queueing system, no customer will ever wait in a queue. Hence, the time spent in the queue and the number of customers in the queue will be zero by definition.

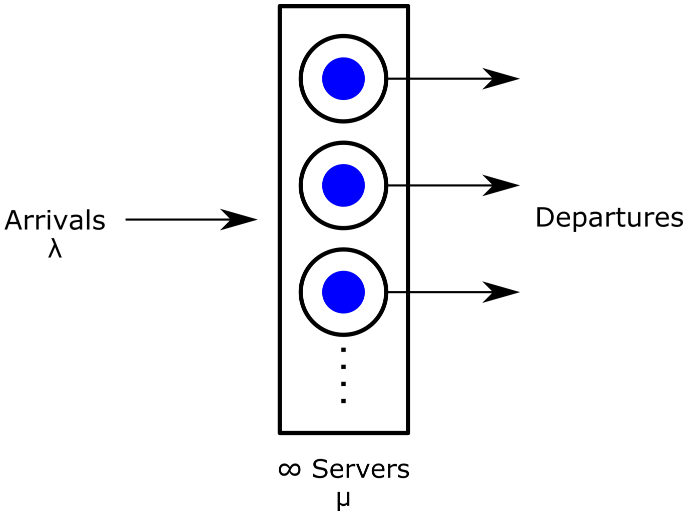


Figure 4.5.5: Example of a queue with an infinite number of identical servers. The blue dots represent customers in the system.

For the M/M/∞ queueing system, the arrival process is Poisson with parameter λ and each server provides independent and identically distributed exponential service at rate μ.

The state transition diagram for the M/M/∞ queue is shown in Figure 4.5.6.

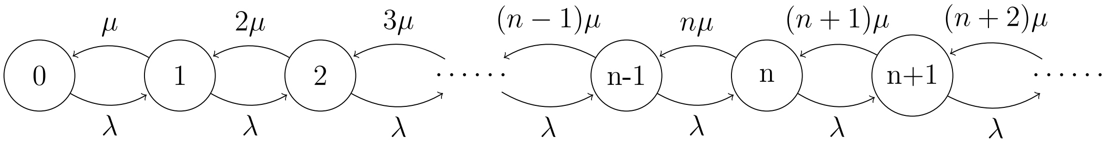


Figure 4.5.6: State transition diagram for the M/M/∞ queue.

From this, we see that the transition rates are

$$q_{i,i+1} = \lambda, \quad i \in \mathcal{S},$$

$$q_{i,i-1} = i\mu, \quad i \in \mathcal{S}\backslash\{0\}, \text{ and}$$

$$q_{i,j} = 0, \quad i \in \mathcal{S}, j \in \mathcal{S}\backslash\{i, i+1, i-1\}.$$

So the infinitesimal generator matrix is

$$\mathcal{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Using the stationary distribution of a birth-and-death process from Section 3.3, we find that the stationary distribution for an M/M/∞ queue is

$$\pi_k = \pi_0 \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n, \quad k = 0, 1, \dots,$$

where

$$\pi_0 = \left(\sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right)^{-1} = e^{-\frac{\lambda}{\mu}}.$$

4.5.4 M/M/c/K queue

Lastly, we consider a Markovian queueing system with c identical servers, c, a finite system capacity of K customers, and a FIFO queueing discipline, as illustrated in Figure 4.5.7. This queueing system is known as a loss system, as customers are lost if the system is at capacity.

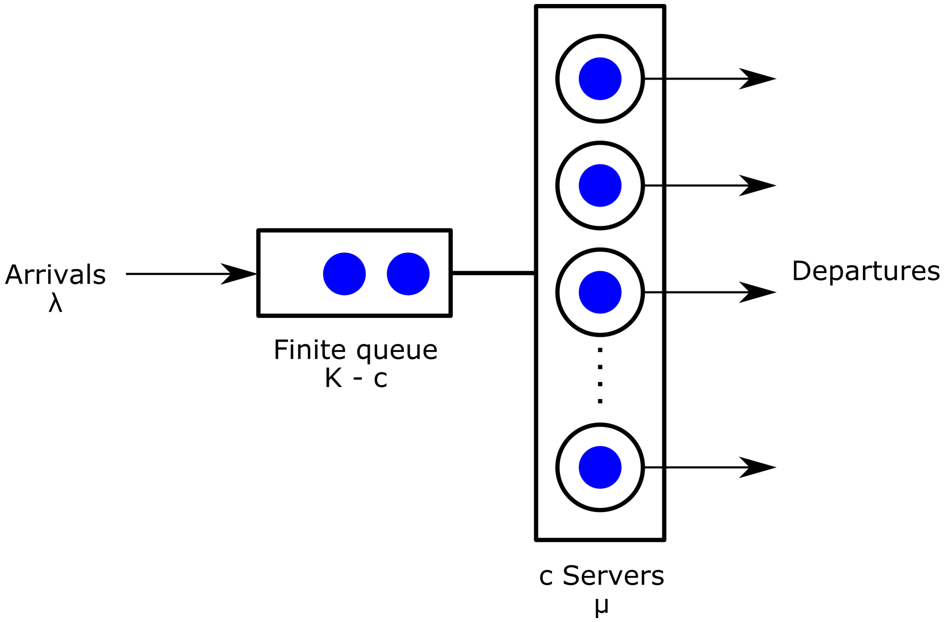


Figure 4.5.7: Example of a queue with c identical servers and a finite system capacity of K customers. The blue dots represent customers.

For the M/M/c/K queueing system, the arrival process is Poisson with parameter λ and each server provides independent and identically distributed exponential service at rate μ.

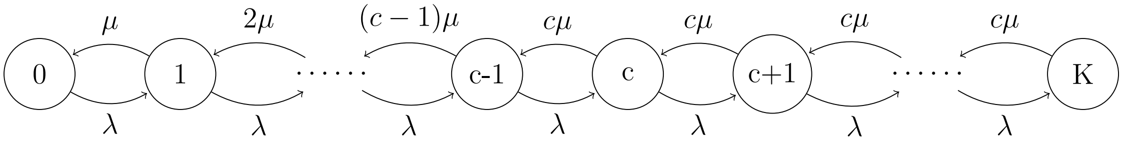


Figure 4.5.8: State transition diagram for the M/M/c/K queue.

From the state transition diagram shown in Figure 4.5.8, we see that the transition rates are

$$q_{i,i+1} = \begin{cases} \lambda, & i \in \mathcal{S}, i < K \\ 0, & i \geq K, \end{cases}$$

$$q_{i,i-1} = \begin{cases} i\mu, & i \in \mathcal{S}, 0 < i < c, \\ c\mu, & c \leq i \leq K, \end{cases}$$

and,

$$q_{i,j} = 0, \quad i \in \mathcal{S}, j \in \mathcal{S} \backslash \{i, i+1, i-1\}.$$

So the infinitesimal generator matrix is

$$\mathcal{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots & 0 \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & c\mu & -(\lambda+c\mu) & \lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & c\mu & -(\lambda+c\mu) & 0 \\ 0 & 0 & 0 & 0 & 0 & & c\mu & -c\mu \end{bmatrix}.$$

Using the stationary distribution of a birth-and-death process from Section 3.3, we find that the stationary distribution for an M/M/c/K queue is

$$\pi_k = \begin{cases} \pi_0 \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k, & k \leq c, \\ \pi_0 \frac{1}{c^{k-c} c!} \left(\frac{\lambda}{\mu}\right)^k, & c < k \leq K, \end{cases}$$

where

$$\pi_0 = \left(\sum_{n=0}^{c-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K} \frac{1}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n \right)^{-1}.$$

4.6 Summary

Queueing theory benefits the design, analysis, and modification of queueing systems in ways, such as adjusting the number of servers, adjusting the capacity, or making small changes to the flow of customers to improve the performance of the system. One of the many applications of queueing theory is healthcare system modelling. In the next chapter, we show how elementary queueing models are used to model the bed occupancy of intensive care units.

Chapter 5

Modelling intensive care units using elementary queueing models

The modelling of healthcare systems has become increasingly popular over the past few years as the demand for resources in health care services increases. In particular, there is a growing interest in improving the operation of intensive care units in terms of patient care and bed management.

Intensive care units have been extensively modelled using stochastic modelling techniques over the last few decades, with a particular focus on using queueing theory [2]. Queueing models, such as $M/M/\bullet$ queueing models, have been used to address bed occupancy management and capacity planning problems in ICUs [47]. These queueing models assume a Poisson arrival process and an exponential distribution for the service times, as well as independence between the arrival and service processes.

In the context of a queueing process, the customers are the patients being admitted to the ICU and the servers are the resources being modelled. For example, the servers could be the beds in the ICU, medical equipment, medical staff, or a combination of all three resources. The queueing discipline of the ICU is very complex as patient admission and treatment depends on the severity of a patient's condition and the resources required, as well as the availability of beds in the ICU.

In this section, we use data from the Royal Adelaide Hospital [47] to model the bed occupancy in the ICU using the $M/M/\infty$, $M/M/c$, and $M/M/c/K$ queueing models. For each of these models, we assume that the customers are the patients being admitted to the ICU, the servers are the beds in the ICU, and the queueing discipline is a combination of priority and processor sharing. We also assess the fit of each model by comparing the model distributions of the bed occupancy to the observed distribution of bed occupancy from the RAH ICU data set, as well as check the assumptions of the queueing models.

5.1 Royal Adelaide Hospital

The Royal Adelaide Hospital (RAH) is the largest hospital in Adelaide, South Australia. Prior to November 2017, the Royal Adelaide Hospital was located on North Terrace, Adelaide, containing 680 beds. The Royal Adelaide Hospital has since relocated to Port Road, Adelaide, where it now contains 800 beds. Among several improvements to the hospital, the treatment capacities in the intensive care unit increased to accomodate more patients with various conditions. The intensive care unit at the New Royal Adelaide Hospital (nRAH) contains 60 beds including quarantine rooms for infectious diseases, whereas the previous hospital contained only 42 beds.

5.1.1 Data set

The data set used in this thesis was obtained from the intensive care unit in the original Royal Adelaide Hospital. The data set contains de-identified information on 7124 patients who arrived to the ICU from 1 November 2014 to 31 October 2016, such as

- patient demographics (age, gender, etc.),
- the date and time of admission to the ICU,
- the source of patient admission (ambulance, private vehicle, etc.),
- the health status and reason for admission,
- the date and time of discharge from the ICU,
- the patient survival status (died or survived), and
- the discharge destination (home or another unit/ward).

5.1.2 Analysis

In this section, we analyse the patient admissions to the ICU, how long patients stay in the ICU, patient discharges from the ICU, as well as the number of patients in the ICU across the time period of interest. Note that the shifts defined in this thesis are as follows:

- day shift - 7:00 a.m. to 3:00 p.m,
- afternoon shift - 3:00 p.m. to 11:00 p.m, and
- night shift - 11:00 p.m. to 7:00 a.m.

Patient admissions

Patients generally arrive to the ICU following surgery or from another unit or ward within the hospital, as shown in Figure 5.1.1. The largest source of admission to the ICU is the operating theatre, which accounts for 48.7% of ICU admissions, followed by the emergency department, which accounts for 31.5%.

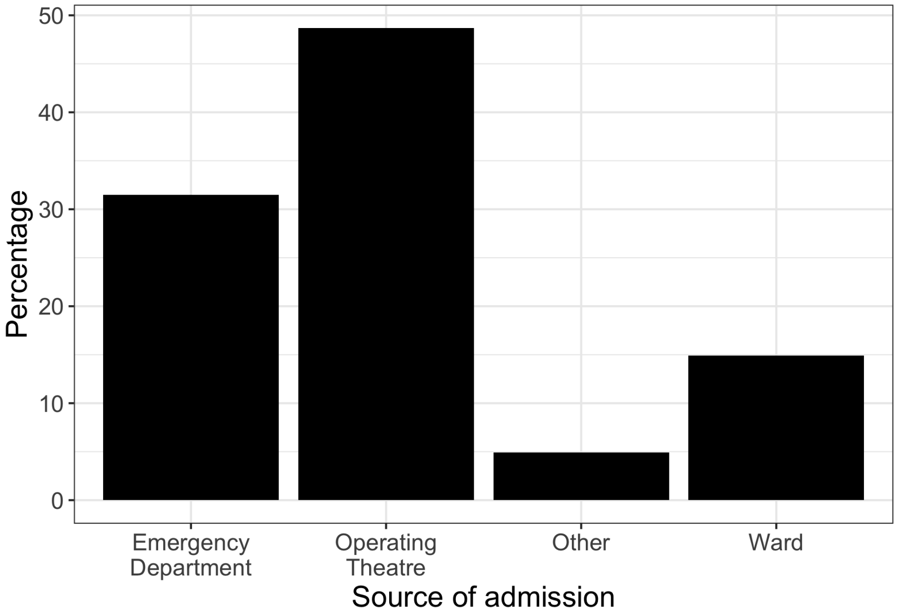


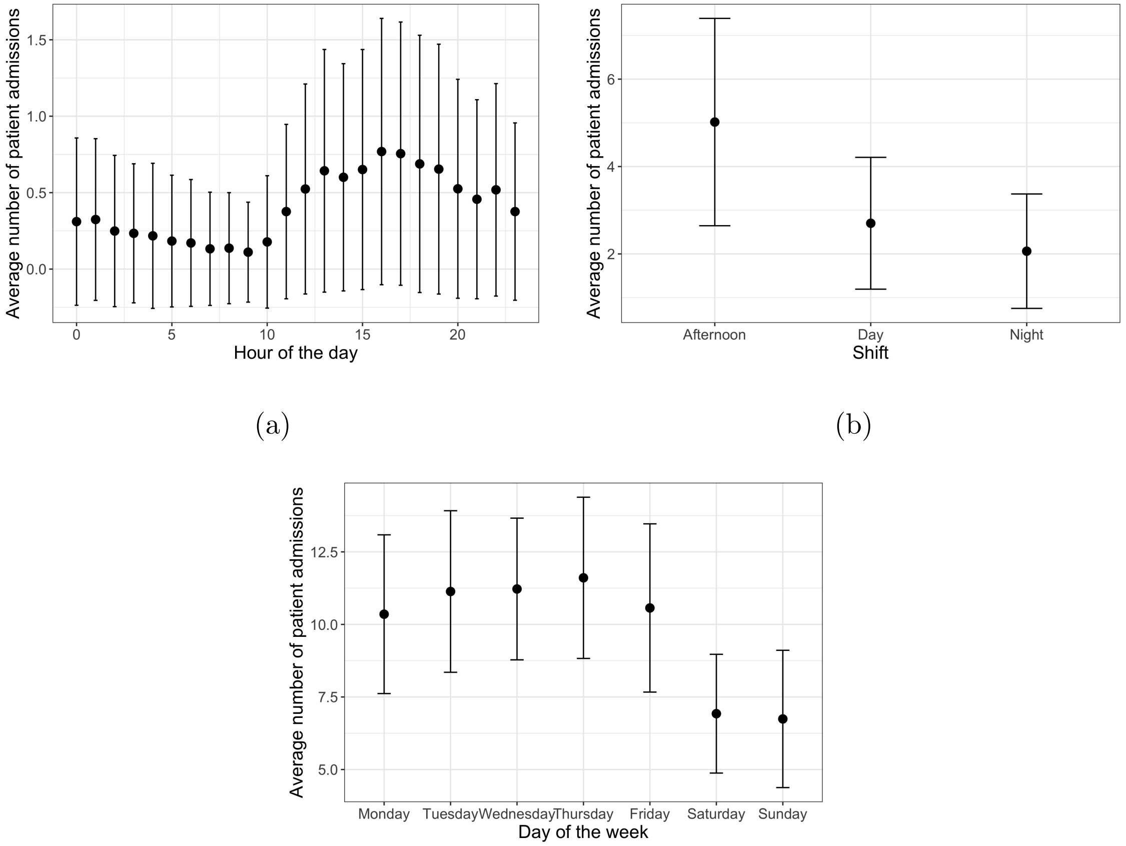
Figure 5.1.1: Sources of admission to the intensive care unit at the Royal Adelaide Hospital.

Between November 2014 and October 2016, the average admission rate of patients to the ICU was 9.785 patients per day, as shown in Table 5.1.1. The variability of patient admissions to the ICU is illustrated by the variance of 10.336 patients per day, as well as the minimum and maximum number of patient admissions being 1 and 18 respectively.

Mean	Median	Variance	IQR	Minimum	Maximum
9.785	10	10.336	5	1	18

Table 5.1.1: Summary statistics of the number of patient admissions to the RAH ICU per day.

The RAH ICU has a high influx of patient admissions between Monday and Friday due to more medical staff available during the week, as shown in Figure 5.1.2c. The increase in patient admissions to the ICU in the afternoon, shown in Figures 5.1.2a and 5.1.2b, is generally due to patients coming out of scheduled surgeries in the morning.



(a)

(b)

(c)

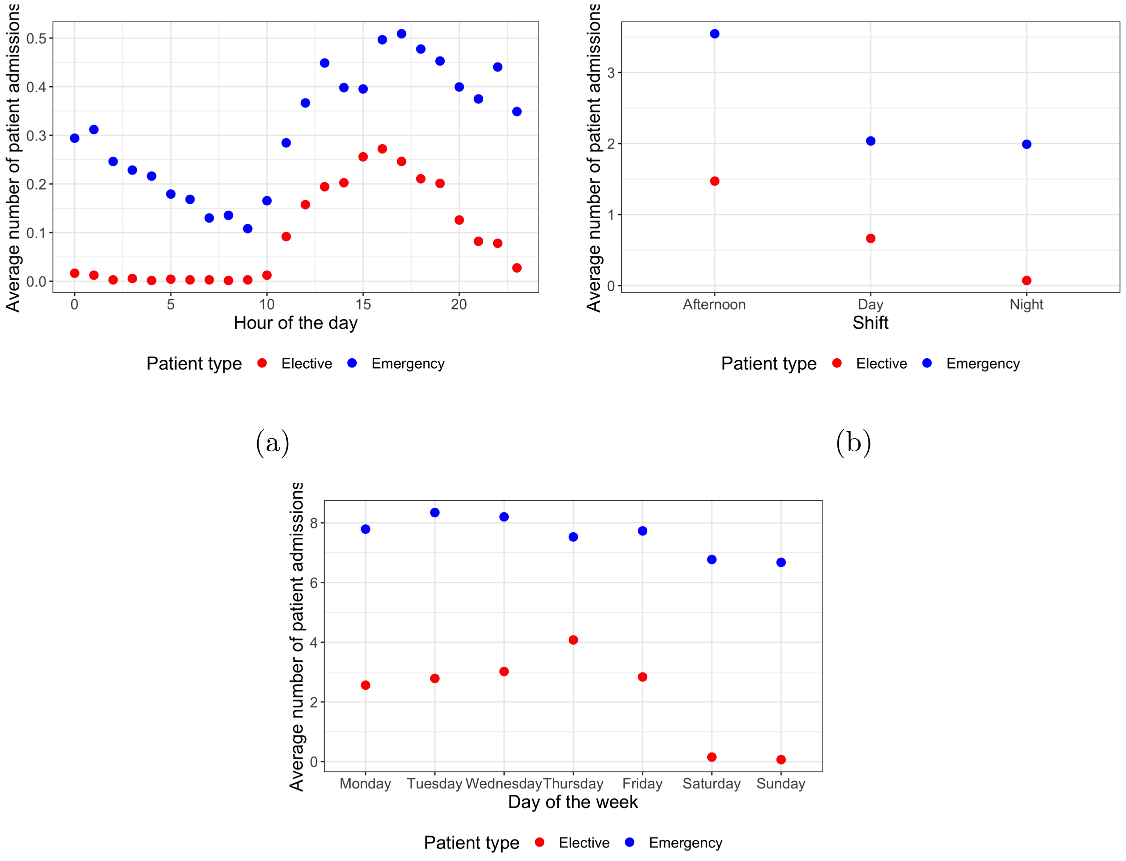
Figure 5.1.2: Average number of patient admissions to the RAH ICU (dot) for each hour of the day (a), for each shift (b), and for each day of the week (c). Note that the error bars represent mean ± standard deviation.

Emergency procedures are unplanned, and so admission to the ICU can occur on any day during the week and at any time of the day. Hence, the average admission rate of emergency patients to the ICU is higher than the average admission rate of elective patients to the ICU, as shown in Table 5.1.2. Also note the higher amount of variability in the number of emergency patient admissions per day compared to that of elective patients, which is reflective of the unplanned nature of emergency procedures.

Summary statistic	Elective	Emergency
Mean	2.209	7.576
Median	2	7
Variance	3.831	6.379
IQR	4	3
Minimum	0	1
Maximum	9	15

Table 5.1.2: Summary statistics of the number of elective and emergency patient admissions to the RAH ICU per day.

Elective procedures are planned in advanced and are generally scheduled on weekdays and during the day as there are more medical staff working during these hours. Hence, there are fewer elective patient admissions to the ICU on weekends and during night shifts, as illustrated in Figures 5.1.3c, 5.1.3a, and 5.1.3b.



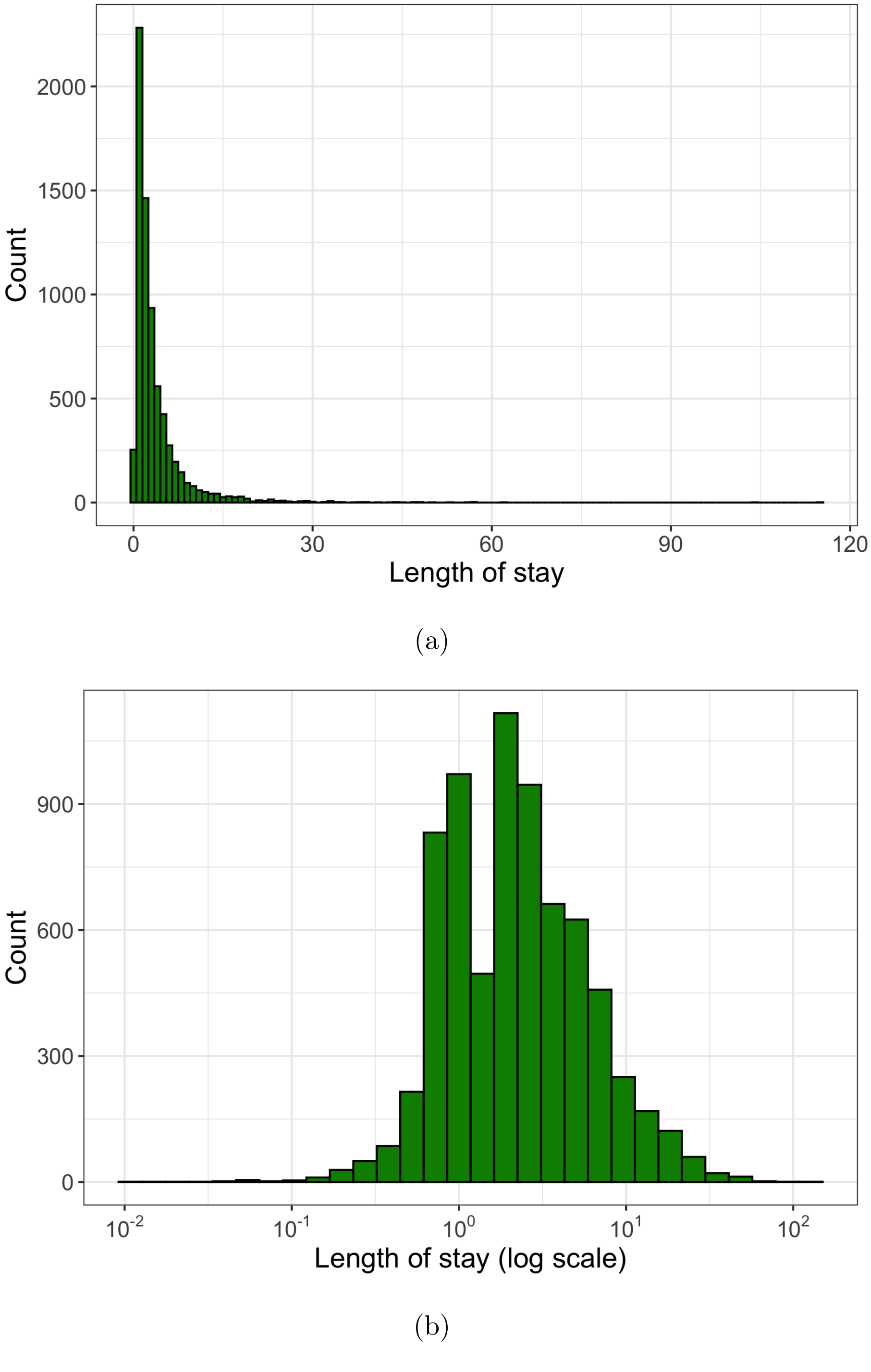
(a) (b)

(c)

Figure 5.1.3: Average number of elective and emergency patient admissions to the RAH ICU for each hour of the day (a), for each shift (b), and for each day of the week (c).

Length of stay

The length of stay of a patient is defined as the number of days spent in a bed. There is a considerable amount of variability in the length of stay which is due to the varying severity of medical conditions of patients presented to the ICU, as shown in Figure 5.1.4 and Table 5.1.3.



(a)

(b)

Figure 5.1.4: Histograms of patient length of stay (days) in the RAH ICU. Note that (b) plots the length of stay on a log base 10 scale.

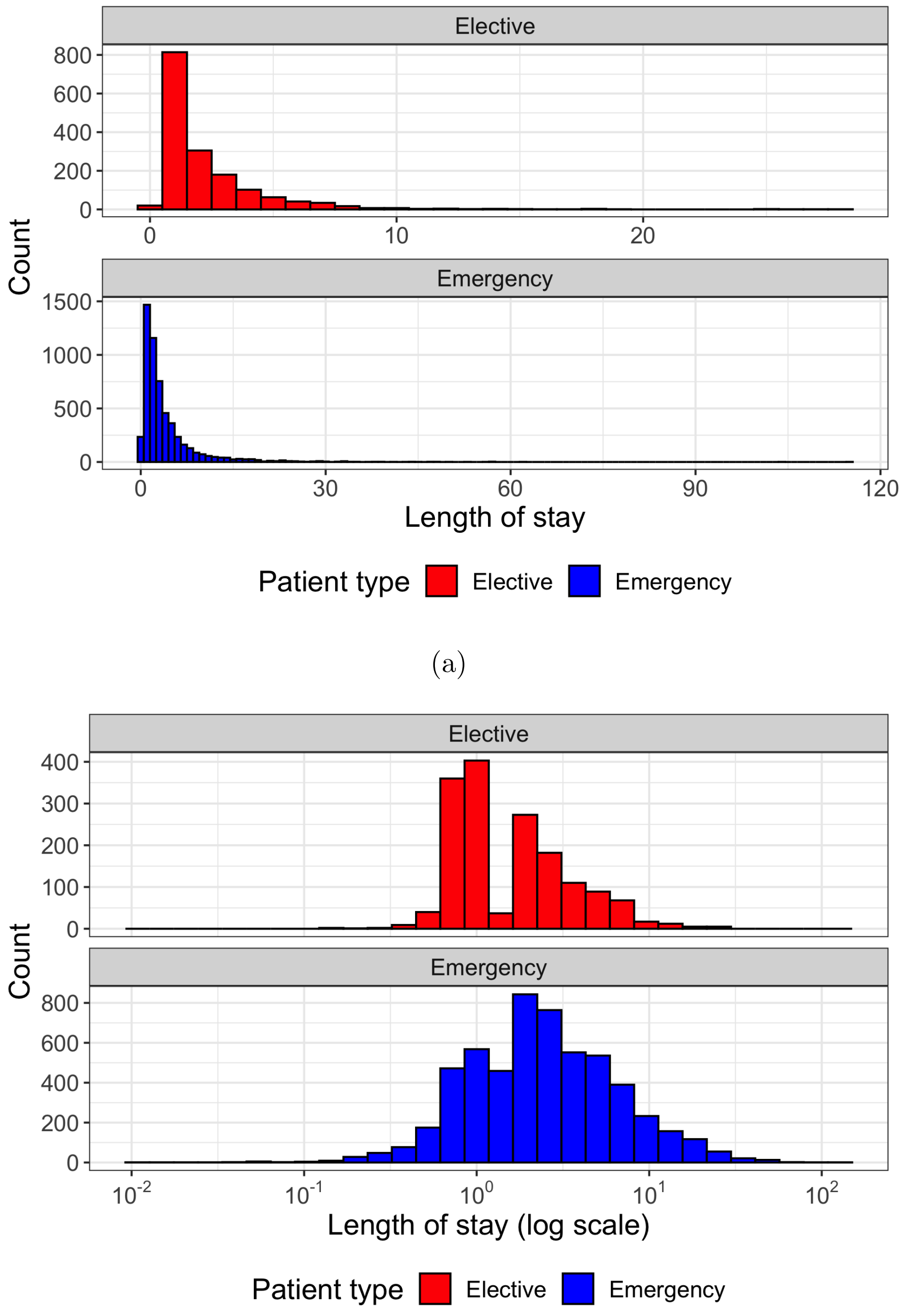
Summary statistic	Value
Mean	3.628
Median	2.001
Variance	26.100
IQR	3.106
Minimum	0.010
Maximum	114.712

Table 5.1.3: Summary statistics of patient length of stay (days) in the RAH ICU.

Elective procedures are extensively planned, so medical staff are quite efficient in treating elective patients. Whereas, emergency patients often require a lot of medical attention and complications can occur if the medical condition is severe. Hence, the length of stay of emergency patients is longer on average and more variable than the length of stay of elective patients, as shown in Table 5.1.4 and Figure 5.1.5.

Summary statistic	Elective	Emergency
Mean	2.257	4.027
Median	1.104	2.375
Variance	6.716	31.047
IQR	1.992	3.502
Minimum	0.160	0.010
Maximum	28.156	114.712

Table 5.1.4: Summary statistics of patient length of stay (days) of elective and emergency patients in the RAH ICU.



(a)

(b)

Figure 5.1.5: Histograms of patient length of stay (days) for elective and emergency patients in the RAH ICU. Note that (b) plots the length of stay on a log base 10 scale.

Patient discharges

Once a patient no longer requires intensive care, the patient is transferred to another unit or ward either within or externally to the hospital, as shown in Figure 5.1.6.

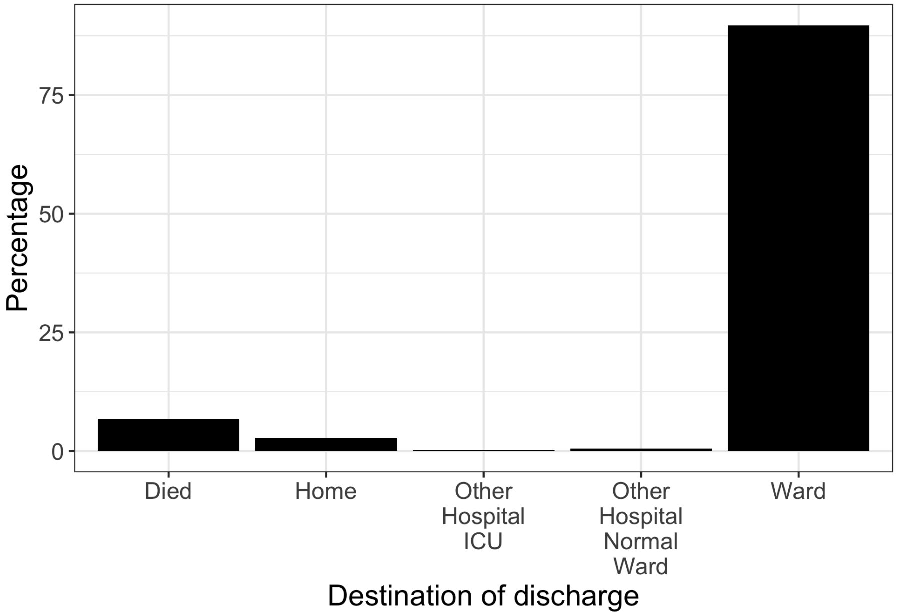


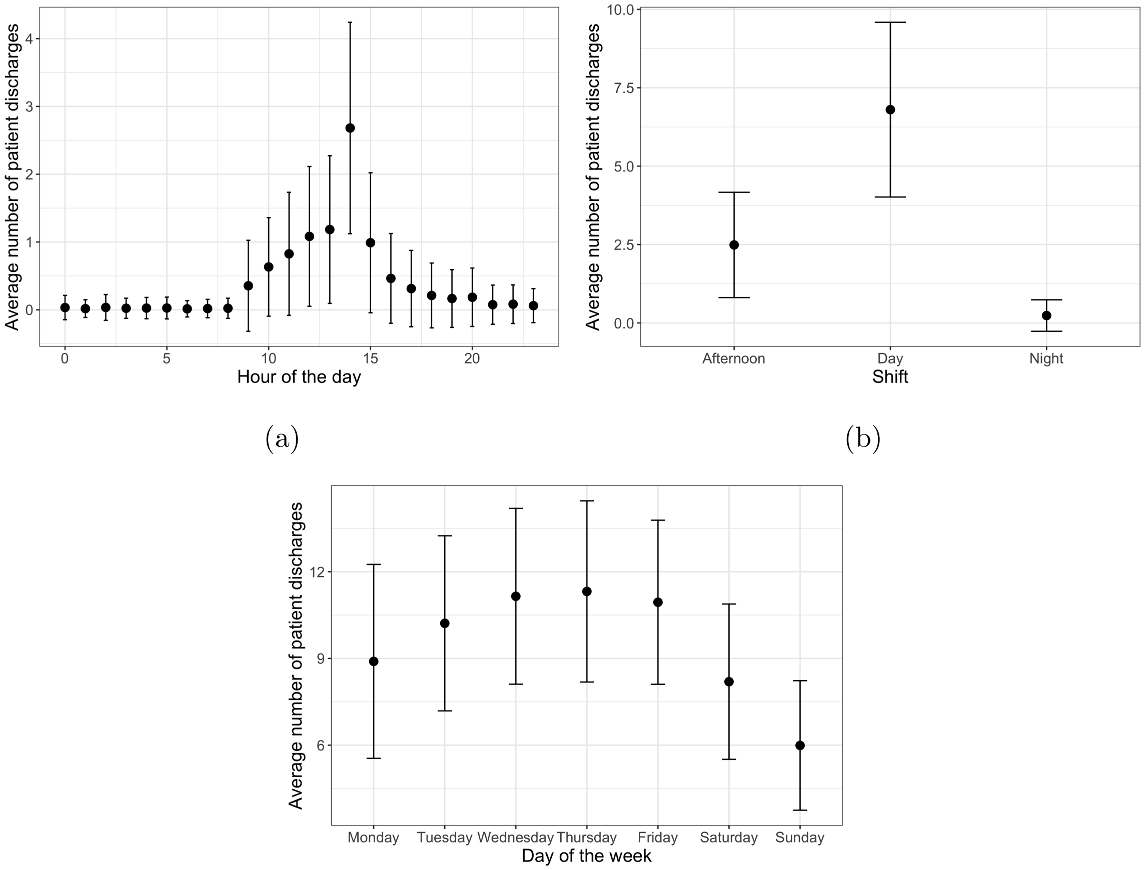
Figure 5.1.6: Destinations of discharges from the intensive care unit at the Royal Adelaide Hospital.

Between November 2014 and October 2016, the average discharge rate of patients from the ICU was 9.525 patients per day, as shown in Table 5.1.5.

Mean	Median	Variance	IQR	Minimum	Maximum
9.525	10	11.767	5	0	19

Table 5.1.5: Summary statistics of the number of patient discharges from the RAH ICU per day.

Given that there are limited staff in the hospital on weekends, there are more patient discharges from the ICU during the week, as shown in Figure 5.1.7c. In addition to this, ICU management tend to discharge patients from the ICU in the late morning or early afternoon as there is an influx of patient admissions to the ICU following morning surgery, as illustrated in Figures 5.1.7a and 5.1.7b.



(a)

(b)

(c)

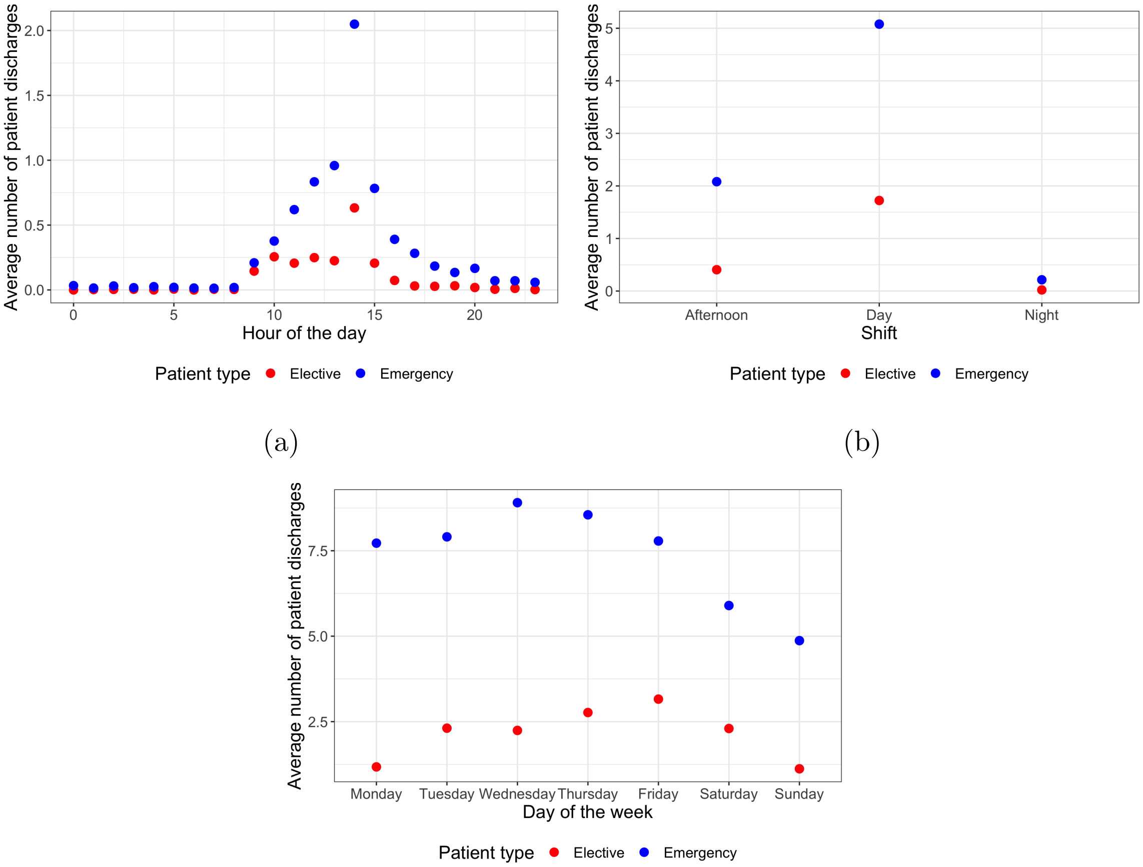
Figure 5.1.7: Average number of patient discharges from the RAH ICU (dot) for each hour of the day (a), for each shift (b), and for each day of the week (c). Note that the error bars represent mean ± standard deviation.

Given that the admission patterns of elective and emergency patients are different, the discharge patterns between elective and emergency patients are consequently different as well, as shown in Table 5.1.6.

Summary statistic	Elective	Emergency
Mean	2.150	7.374
Median	2	7
Variance	2.755	8.765
IQR	2	4
Minimum	0	0
Maximum	9	17

Table 5.1.6: Summary statistics of the number of elective and emergency patient discharges from the RAH ICU per day.

Furthermore, Figures 5.1.8a, 5.1.8b, and 5.1.8c illustrate that the average discharge rate of emergency patients is higher across the week than the average discharge rate of elective patients, particularly during the day shift.



(a)

(b)

(c)

Figure 5.1.8: Average number of elective and emergency patient discharges from the RAH ICU for each hour of the day (a), for each shift (b), and for each day of the week (c).

Bed occupancy

In an ICU, bed occupancy is defined as the number of patients occupying beds within the ICU. Between November 2014 and October 2016, the average bed occupancy was 34.85 patients and it varies considerably, as shown in Table 5.1.7. Figure 5.1.9 shows that the bed occupancy increases during the colder months and flu season and it decreases over the warmer months and Christmas holidays. The minimum observed bed occupancy during this time period was 20 patients and the ICU exceeded capacity several times, as shown in Figure 5.1.10.

Mean	Median	Variance	IQR	Minimum	Maximum
34.850	35	13.255	5	20	45

Table 5.1.7: Summary statistics of the ICU bed occupancy at the Royal Adelaide Hospital.

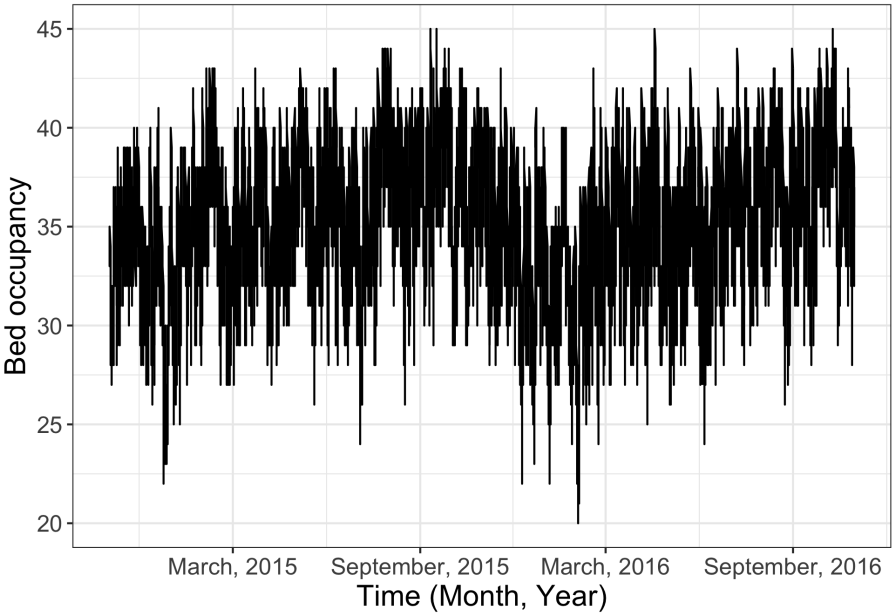


Figure 5.1.9: Plot of ICU bed occupancy at the Royal Adelaide Hospital.

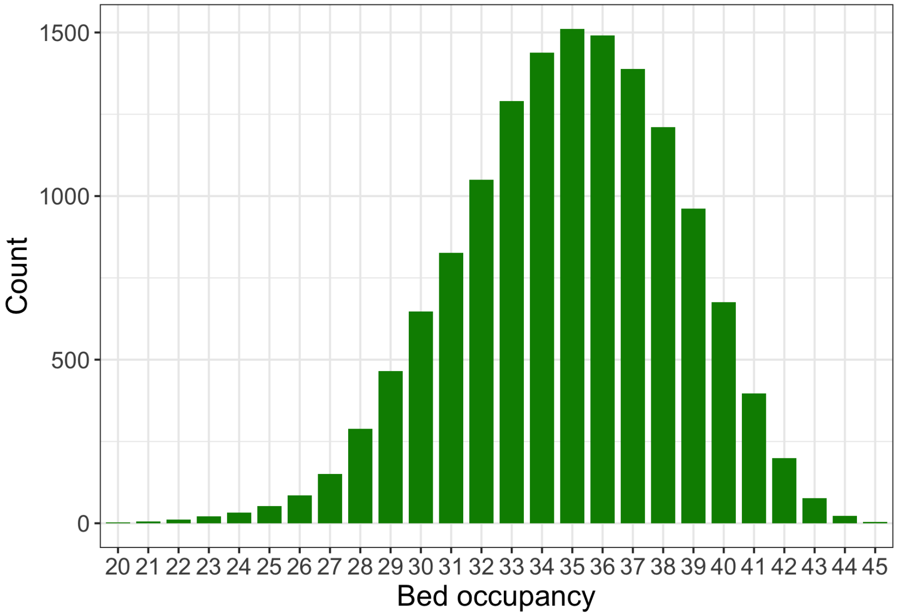


Figure 5.1.10: Bar chart of ICU bed occupancy at the Royal Adelaide Hospital.

Service priority

Given the nature of the ICU, patients are admitted based not only on the severity of their condition but also on the availability of a bed. In terms of treatment, the severity of their condition and availability of resources greatly determines how long a patient remains in the ICU. Hence, the service discipline of the ICU is very complex and tends to relate to priority and processor sharing queueing models.

However, in this chapter we only consider the queueing models that are commonly used to address bed occupancy management and capacity planning problems in ICUs. The reason behind this is to draw attention to the limitations of using these queueing models in terms of the arrival and service processes and not the queueing discipline.

Summary

From the analysis above, the admission and discharge patterns of patients are time dependent, meaning that admissions and discharges depend on the time of the day and the day of the week. There is also evidence of different admission, discharge, and length of stay patterns between elective and emergency patients.

Despite this, the bed occupancy of ICUs is often modelled using $M/M/\bullet$ queueing models due to their simplicity [47, 2]. In the following sections, we initially model the RAH ICU using an $M/M/\infty$ queueing model and then restrict the capacity by considering an $M/M/42$ queueing model and an $M/M/42/45$ queueing model.

5.2 $M/M/\infty$ queueing model

In the context of an ICU, we first consider an $M/M/\infty$ queueing model which suggests that

- patients arrive to the ICU according to a Poisson process,
- patient lengths of stay are exponentially distributed,
- there are an unlimited number of beds in the ICU, and
- there are an unlimited number of waiting spaces available.

Using the definition of the likelihood for a general birth-and-death process presented in Section 3.3.3, the likelihood for an $M/M/\infty$ queueing model is defined as

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=0}^{\infty} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right),$$

where u_k is the observed number of times the bed occupancy increased from $k \geq 0$ to $k+1$, d_k is the observed number of times the bed occupancy decreased from $k \geq 1$ to $k-1$, and t_k is the observed amount of time the ICU had a bed occupancy of $k \geq 0$.

Recall that the minimum observed occupancy was 20 and the maximum observed occupancy was 45. Hence, the likelihood becomes

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=20}^{45} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right).$$

Using the observed data, we estimate λ and μ using the maximum likelihood estimates,

$$\widehat{\lambda} = \frac{\displaystyle\sum_{k=20}^{45} u_k}{\displaystyle\sum_{k=20}^{45} t_k} = 9.786,$$

and,

$$\widehat{\mu} = \frac{\displaystyle\sum_{k=20}^{45} d_k}{\displaystyle\sum_{k=20}^{45} kt_k} = 0.276.$$

Hence, the transition rates of the $M/M/\infty$ queueing model are

$$\lambda_k = 9.786, \quad k \geq 20,$$

and

$$\mu_k = 0.276k, \quad k \geq 21.$$

Using the definition of the stationary distribution for an $M/M/\infty$ queueing model, the expected proportion of time spent at each bed occupancy is

$$\pi_k = \pi_{20} \frac{1}{\frac{k!}{20!}} \left(\frac{9.786}{0.276} \right)^{k-20}, \quad k \geq 21,$$

where

$$\pi_{20} = \left(1 + \sum_{n=21}^{\infty} \frac{1}{\frac{n!}{20!}} \left(\frac{9.786}{0.276} \right)^{n-20} \right)^{-1}.$$

Figure 5.2.1 compares the expected proportion of time spent at each bed occupancy assuming an $M/M/\infty$ queueing model to the observed proportion of time spent at each bed occupancy. Based on the significant difference between the distributions, the $M/M/\infty$ queueing model is clearly not a suitable model for the RAH ICU.

In addition to this, the $M/M/\infty$ queueing model allows transitions to bed occupancies greater than 45 which is an unreasonable assumption for the RAH ICU.

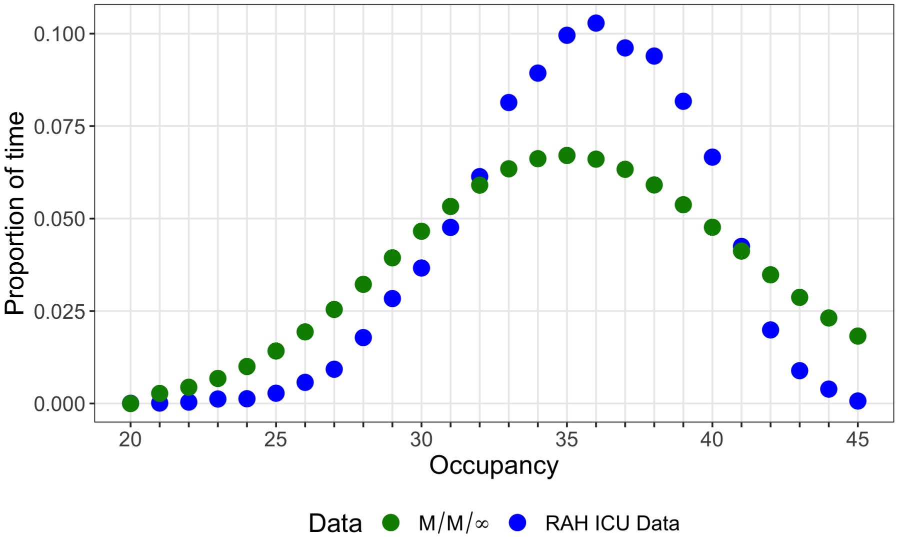


Figure 5.2.1: Observed proportion of time spent at each bed occupancy (blue) versus the expected proportion of time spent at each bed occupancy assuming an $M/M/\infty$ queueing model (green).

5.3 $M/M/42$ queueing model

Next, we consider an $M/M/42$ queueing model which now assumes a finite number of beds in the ICU. In particular, the $M/M/42$ queueing model suggests that

- patients arrive to the ICU according to a Poisson process,
- patient lengths of stay are exponentially distributed,
- there are 42 beds in the ICU, and
- there are an unlimited number of waiting spaces available.

Using the definition of the likelihood for a general birth-and-death process presented in Section 3.3.3, the likelihood for an $M/M/42$ queueing model is defined as

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=0}^{41} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right) \prod_{k=42}^{\infty} \lambda^{u_k} (42\mu)^{d_k} \exp\left(-\left(\lambda + 42\mu\right) t_k\right),$$

where u_k is the observed number of times the bed occupancy increased from $k \geq 0$ to $k+1$, d_k is the observed number of times the bed occupancy decreased from $k \geq 1$ to $k-1$, and t_k is the observed amount of time the ICU had a bed occupancy of $k \geq 0$.

Given that the minimum observed occupancy was 20 and the maximum observed occupancy was 45, the likelihood becomes

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=20}^{41} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right) \prod_{k=42}^{45} \lambda^{u_k} (42\mu)^{d_k} \exp\left(-\left(\lambda + 42\mu\right) t_k\right).$$

Using the observed data, we estimate λ and μ using the maximum likelihood estimates,

$$\widehat{\lambda} = \frac{\sum\limits_{k=20}^{45} u_k}{\sum\limits_{k=20}^{45} t_k} = 9.786,$$

and,

$$\widehat{\mu} = \frac{\sum\limits_{k=20}^{45} d_k}{\sum\limits_{k=20}^{41} kt_k + \sum\limits_{k=42}^{45} 42t_k} = 0.276,$$

which are the same as those obtained in the $M/M/\infty$ queueing model.

Hence, the transition rates of the $M/M/42$ queueing model are

$$\lambda_k = 9.786, \quad k \geq 20,$$

and

$$\mu_k = \begin{cases} 0.276k, & 21 \leq k \leq 41, \\ 42 \times 0.276, & k \geq 42. \end{cases}$$

Using the definition of the stationary distribution for an $M/M/42$ queueing model, the expected proportion of time spent at each bed occupancy is

$$\pi_k = \begin{cases} \pi_{20} \frac{1}{\frac{k!}{20!}} \left(\frac{9.786}{0.0127}\right)^{k-20}, & 20 \leq k \leq 41, \\ \pi_{20} \frac{1}{\frac{42^{k-42} 42!}{20!}} \left(\frac{9.786}{0.0127}\right)^{k-20}, & k \geq 42, \end{cases}$$

where

$$\pi_{20} = \left(1 + \sum_{n=21}^{41} \frac{1}{\frac{n!}{20!}} \left(\frac{9.786}{0.0127}\right)^{n-20} + \frac{1}{\frac{42!}{20!}} \left(\frac{9.786}{0.0127}\right)^{42-20} \frac{1}{1 - \frac{9.786}{42 \times 0.0127}}\right)^{-1}.$$

Based on the significant difference between the expected proportion of time spent at each bed occupancy assuming an $M/M/42$ queueing model and the observed proportion of time spent at each bed occupancy illustrated in Figure 5.3.1, the $M/M/42$ queueing model is again clearly not a suitable model for the RAH ICU.

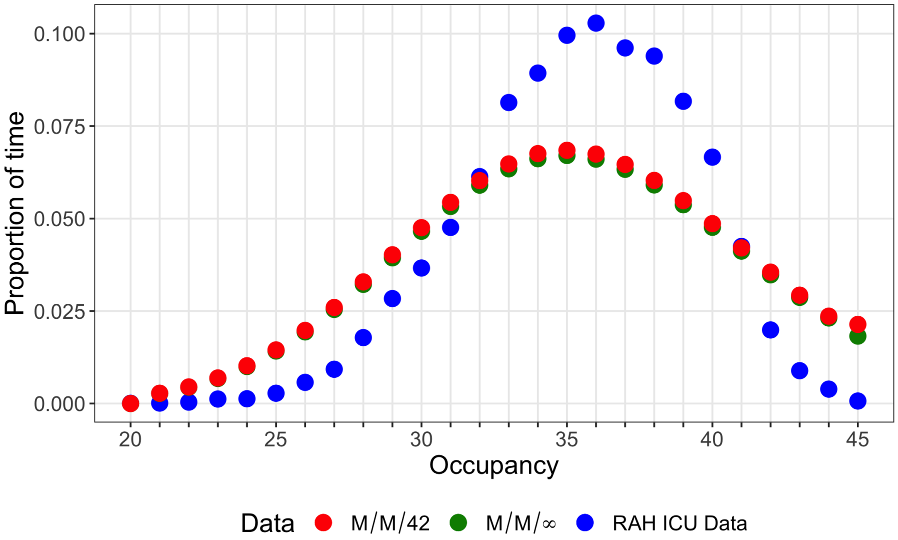


Figure 5.3.1: Observed proportion of time spent at each bed occupancy (blue) versus the expected proportion of time spent at each bed occupancy assuming an $M/M/42$ queueing model (red) and the expected proportion of time spent at each bed occupancy assuming an $M/M/\infty$ queueing model (green).

5.4 $M/M/42/45$ queueing model

Lastly, we consider an $M/M/42/45$ queueing model which further reduces the capacity of the ICU, such that there are now a finite number of beds in the ICU and a finite number of waiting spaces. That is, we consider an $M/M/42/45$ queueing model which suggests that

- patients arrive to the ICU according to a Poisson process,
- patient lengths of stay are exponentially distributed,
- there are 42 beds in the ICU, and
- there are 3 waiting spaces available.

Using the definition of the likelihood for a general birth-and-death process presented in Section 3.3.3, the likelihood for an $M/M/42/45$ queueing model is defined as

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=0}^{411} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right) \prod_{k=42}^{45} \lambda^{u_k} (42\mu)^{d_k} \exp\left(-\left(\lambda + 42\mu\right) t_k\right),$$

where u_k is the observed number of times the bed occupancy increased from $k \geq 0$ to $k+1$, d_k is the observed number of times the bed occupancy decreased from $k \geq 1$ to $k-1$, and t_k is the observed amount of time the ICU had a bed occupancy of $k \geq 0$.

Since the minimum observed occupancy was 20 and the maximum observed occupancy was 45, the likelihood becomes

$$L(u_k, d_k, t_k; \lambda, \mu) = \prod_{k=20}^{41} \lambda^{u_k} (k\mu)^{d_k} \exp\left(-\left(\lambda + k\mu\right) t_k\right) \prod_{k=42}^{45} \lambda^{u_k} (42\mu)^{d_k} \exp\left(-\left(\lambda + 42\mu\right) t_k\right).$$

Using the observed data, we estimate λ and μ using the maximum likelihood estimates,

$$\widehat{\lambda} = \frac{\displaystyle\sum_{k=20}^{45} u_k}{\displaystyle\sum_{k=20}^{45} t_k} = 9.786,$$

and,

$$\widehat{\mu} = \frac{\displaystyle\sum_{k=20}^{45} d_k}{\displaystyle\sum_{k=20}^{41} k t_k + \sum_{k=42}^{45} 42 t_k} = 0.276,$$

which are also the same as those obtained in the $M/M/\infty$ queueing model.

Hence, the transition rates of the $M/M/42/45$ queueing model are

$$\lambda_k = \begin{cases} 9.786, & 20 \leq k \leq 44, \\ 0, & k \geq 45 \end{cases}$$

and

$$\mu_k = \begin{cases} 0.276k, & 21 \leq k \leq 42, \\ 42 \times 0.276, & 43 \leq k \leq 45. \end{cases}$$

Using the definition of the stationary distribution for an $M/M/42/45$ queueing model, the expected proportion of time spent at each bed occupancy is

$$\pi_k = \begin{cases} \pi_{20} \frac{1}{\frac{k!}{20!}} \left(\frac{9.786}{0.0127}\right)^{k-20}, & 20 \leq k \leq 42, \\ \pi_{20} \frac{1}{\frac{42^{k-42} 42!}{20!}} \left(\frac{9.786}{0.0127}\right)^{k-20}, & 43 \leq k \leq 45, \\ 0, & k > 45, \end{cases}$$

where

$$\pi_{20} = \left(1 + \sum_{n=21}^{41} \frac{1}{\frac{n!}{20!}} \left(\frac{9.786}{0.0127}\right)^{n-20} + \sum_{n=42}^{45} \frac{1}{\frac{42^{n-42} 42!}{20!}} \left(\frac{9.786}{0.0127}\right)^{n-20} \right)^{-1}.$$

Figure 5.4.1 compares the expected proportion of time spent at each bed occupancy $M/M/42/45$ queueing model to the observed proportion of time spent at each bed occupancy. Again, there is a clear difference in the distributions and so the $M/M/42/45$ queueing model is not a suitable model for the RAH ICU.

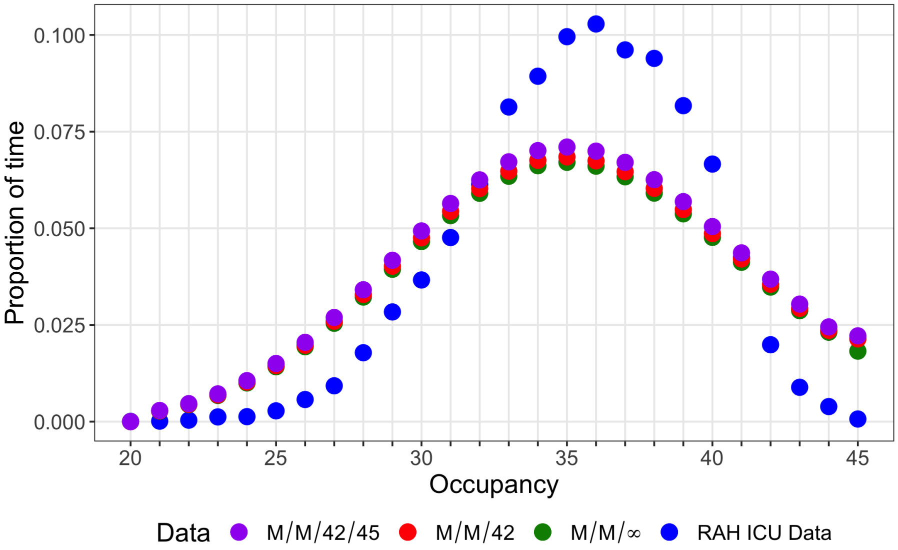


Figure 5.4.1: Observed proportion of time spent at each bed occupancy (blue) versus the expected proportion of time spent at each bed occupancy assuming an $M/M/42/45$ queueing model (purple), the expected proportion of time spent at each bed occupancy assuming an $M/M/42$ queueing model (red) and the expected proportion of time spent at each bed occupancy assuming an $M/M/\infty$ queueing model (green).

5.5 Discussion

The $M/M/\bullet$ queueing models have the following assumptions in common:

- the inter-arrival times are exponentially distributed,
- the service times are exponentially distributed, and
- the arrival process and the distribution of service times are independent.

In this section, we assess the validity of each assumption using statistical modelling techniques.

5.5.1 Patient inter-arrival times

In the literature, the patient admission process is often assumed to be a Poisson arrival process, leading to exponential inter-arrival times. Between November 2014 and October 2016, the average inter-arrival time was 2.46 hours with a variance of 8.26 hours. Figure 5.5.1 plots the distribution of the patient inter-arrival times, drawing attention to the drop in density for inter-arrivals between 0 and 1 hour.

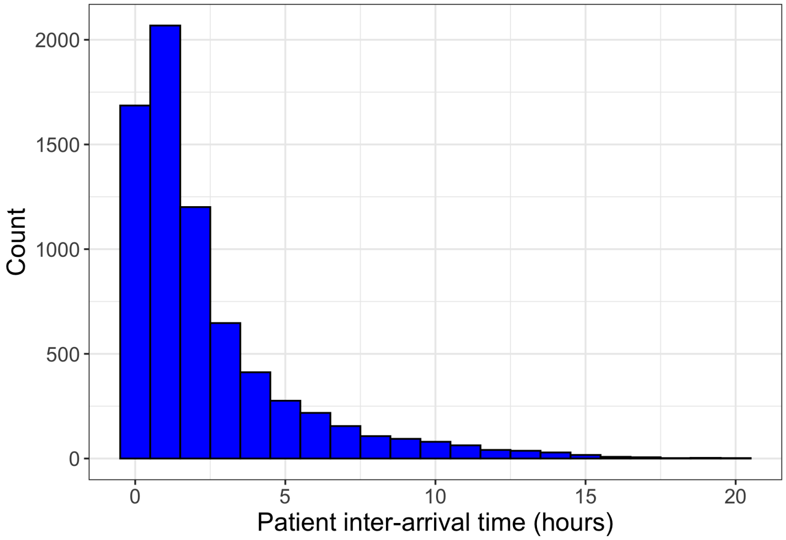


Figure 5.5.1: Histogram of patient inter-arrival times from the RAH ICU data set.

Using maximum likelihood techniques, the probability density function of the fitted exponential distribution is

$$f(x) = 0.407 \exp(-0.407x), \quad x \geq 0,$$

and is plotted in Figure 5.5.2. If the inter-arrival times were in fact exponentially distributed, then the points in the quantile-quantile plot in Figure 5.5.3 would fall on the solid line. However, there is a strong deviation away from the exponential distribution. Hence, a Poisson arrival process does not seem reasonable in this case.

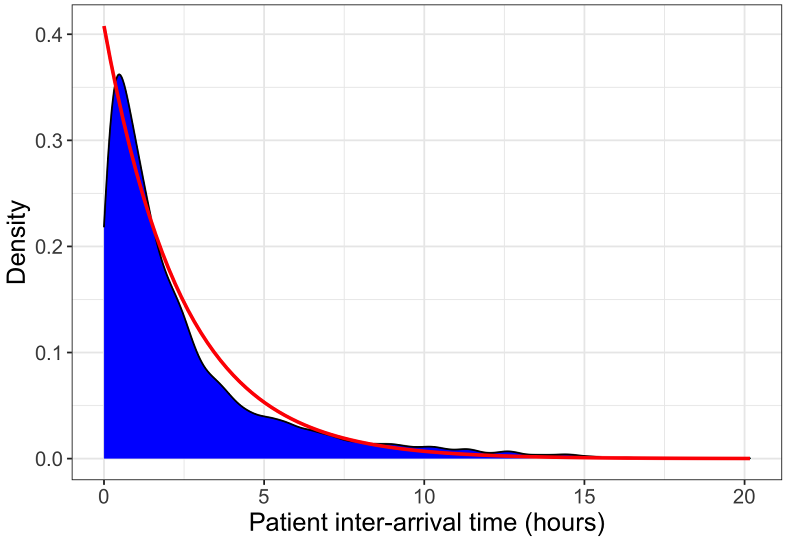


Figure 5.5.2: Observed density (blue) and fitted exponential density (red) of the patient inter-arrival times from the RAH ICU data set.

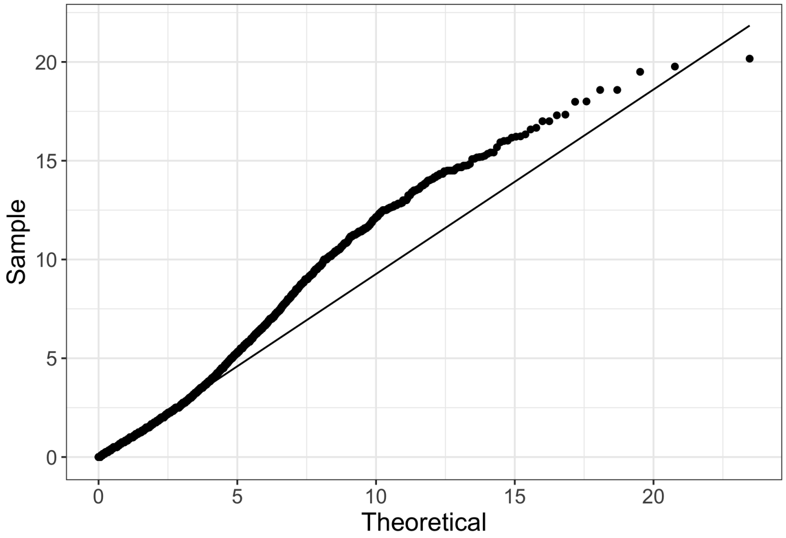


Figure 5.5.3: Quantile-quantile plot of the observed patient inter-arrival times from the RAH ICU data set compared to those expected assuming an exponential distribution.

The Poisson process also assumes stationary inter-arrival times. That is, the inter-arrival distribution should be the same, regardless of the time of the day or day of the week. Looking at Tables 5.5.1 and 5.5.2 and Figures 5.5.4 and 5.5.5, we see that the inter-arrival distribution varies between shift and day of the week, thus suggesting that the arrival process to the RAH ICU is not stationary. As a result of this analysis, it is clear that the assumption of a Poisson arrival process is not valid for the RAH ICU.

Shift	Mean	Variance
Day	3.677	14.832
Afternoon	1.523	3.749
Night	2.600	5.415

Table 5.5.1: Mean and variance of the patient inter-arrival times for each shift from the RAH ICU data set.

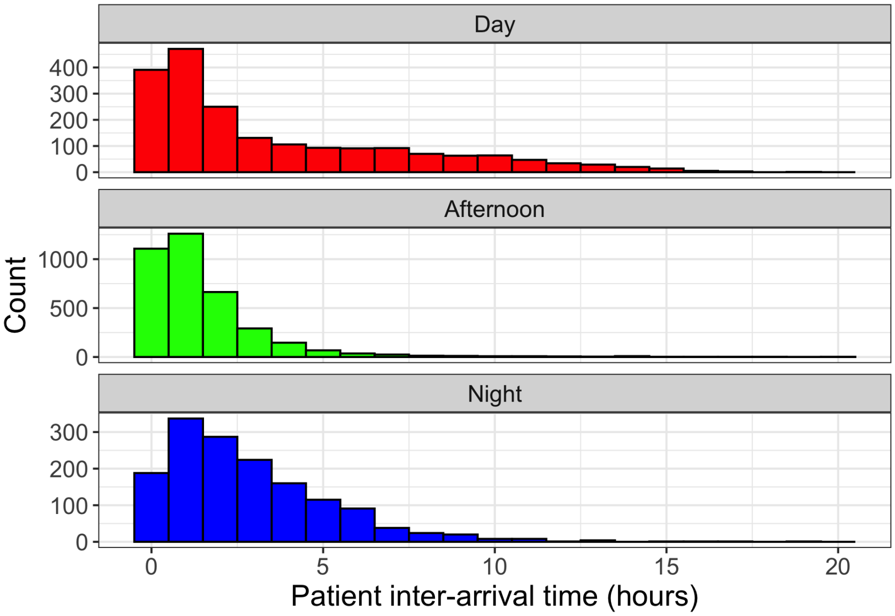


Figure 5.5.4: Histogram of patient inter-arrival times for each shift from the RAH ICU data set.

Day	Mean	Variance
Monday	2.379	8.890
Tuesday	2.164	6.715
Wednesday	2.136	6.784
Thursday	2.068	6.204
Friday	2.251	6.595
Saturday	3.414	11.042
Sunday	3.539	11.222

Table 5.5.2: Mean and variance of the patient inter-arrival times for each day of the week from the RAH ICU data set.

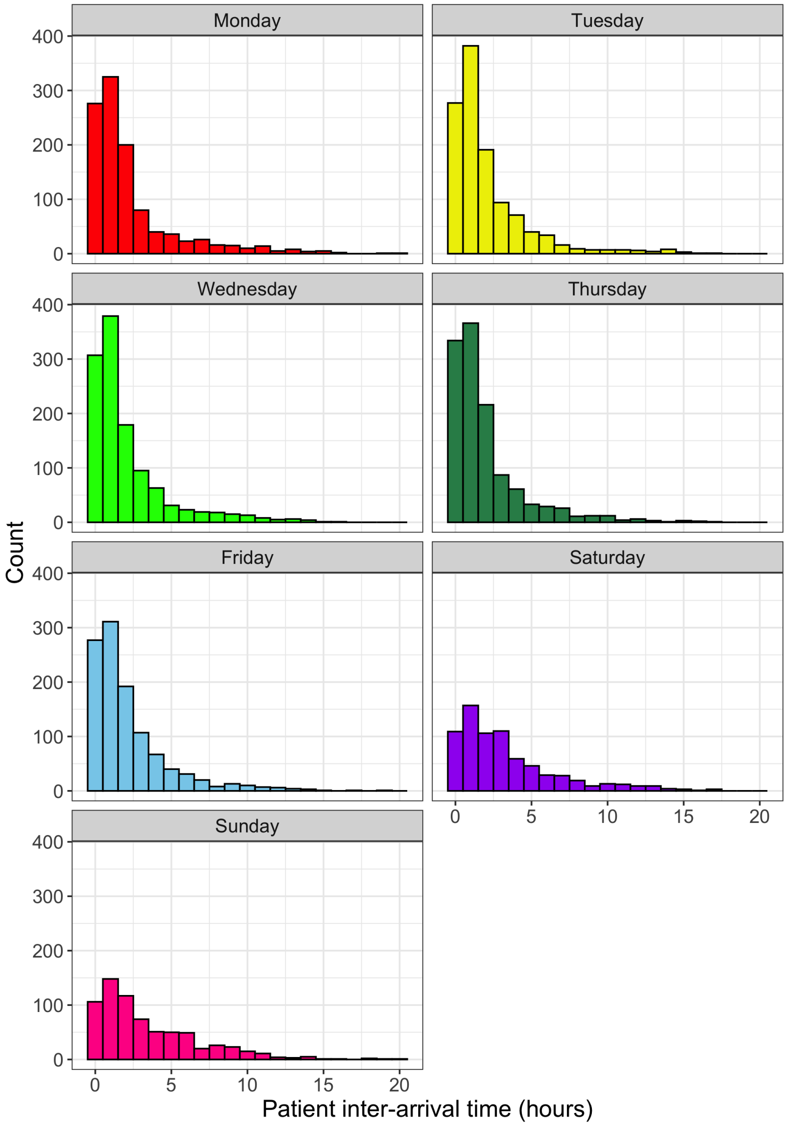


Figure 5.5.5: Histogram of patient inter-arrival times for each day of the week from the RAH ICU data set.

5.5.2 Patient length of stay

The patient length of stay in ICUs is often modelled using an exponential distribution. Recall that between November 2014 and October 2016, the average patient length of stay was 3.628 days with a variance of 26.1 days. Figure 5.5.6 plots the distribution of the patient lengths of stay, drawing attention to the drop in density for lengths of stay between 0 and 1 days.

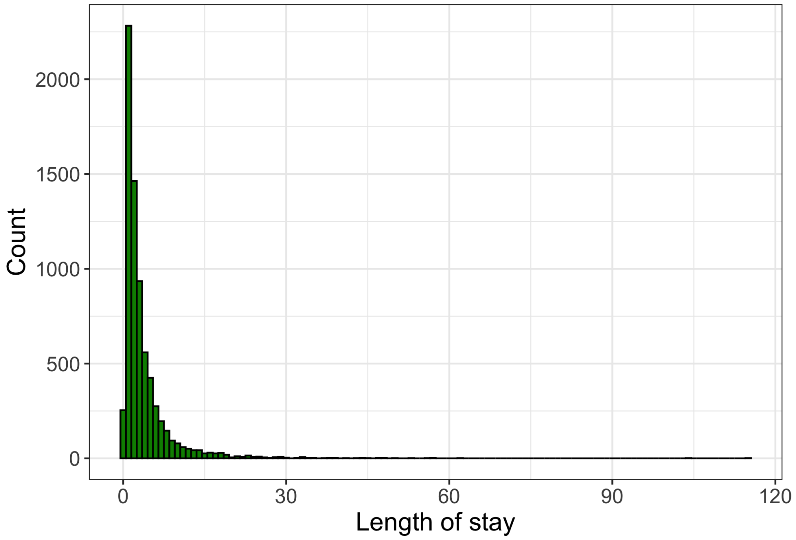


Figure 5.5.6: Histogram of patient length of stay from the RAH ICU data set.

Using maximum likelihood techniques, the probability density function of the fitted exponential distribution is

$$f(x) = 0.275 \exp(-0.275x), \quad x \geq 0,$$

and is plotted in Figure 5.5.7. If the patient lengths of stay were in fact exponentially distributed, then the points in the quantile-quantile plot in Figure 5.5.8 would fall on the solid line. However, the points deviate from the line which indicates that the patient lengths of stay are not exponentially distributed. Hence, the assumption of exponential service times is invalid for the RAH ICU.

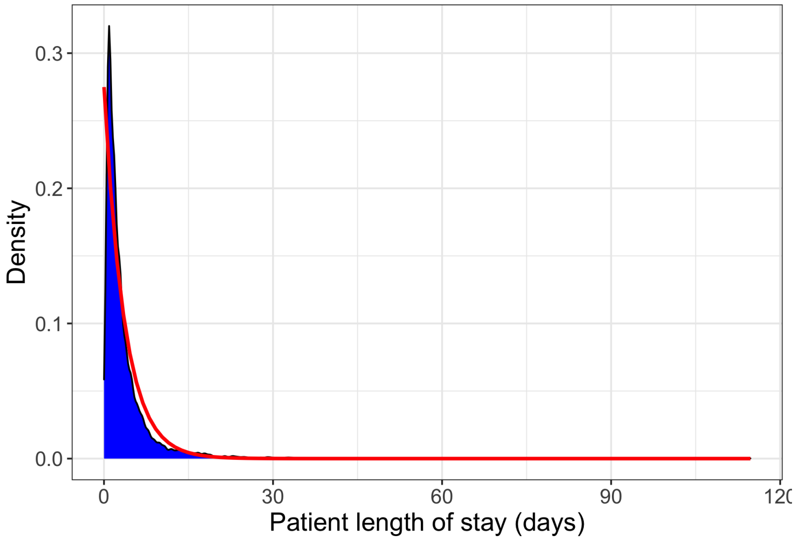


Figure 5.5.7: Observed density (blue) and the fitted exponential density (red) of the patient lengths of stay from the RAH ICU data set.

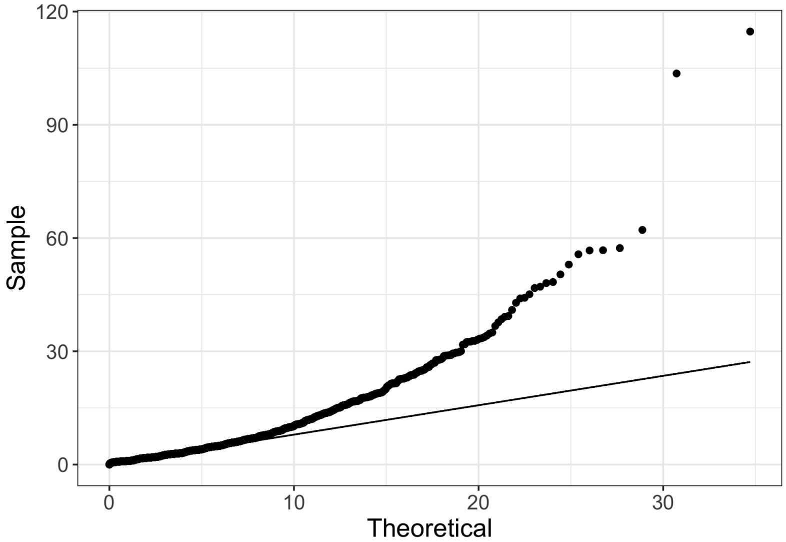


Figure 5.5.8: Quantile-quantile plot of the observed patient lengths of stay from the RAH ICU data set compared to those expected assuming an exponential distribution.

An alternative to the exponential distribution is the log-normal distribution. Using maximum likelihood estimation, the probability density function of the fitted log-normal distribution is

$$f(x) = \frac{1}{0.974x\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - 0.780)^2}{1.897}\right), \quad x \geq 0,$$

and is plotted in Figure 5.5.9.

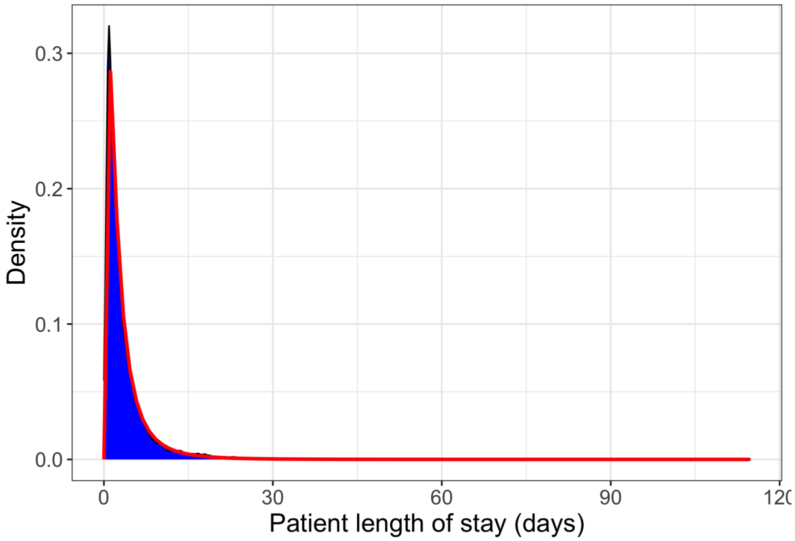


Figure 5.5.9: Observed density (blue) and the fitted log-normal density (red) of the patient lengths of stay from the RAH ICU data set.

Comparing the fit of the log-normal distribution to the fit of the exponential distribution, Figure 5.5.10 suggests that the log-normal distribution is a better fit for modelling patient lengths of stay.

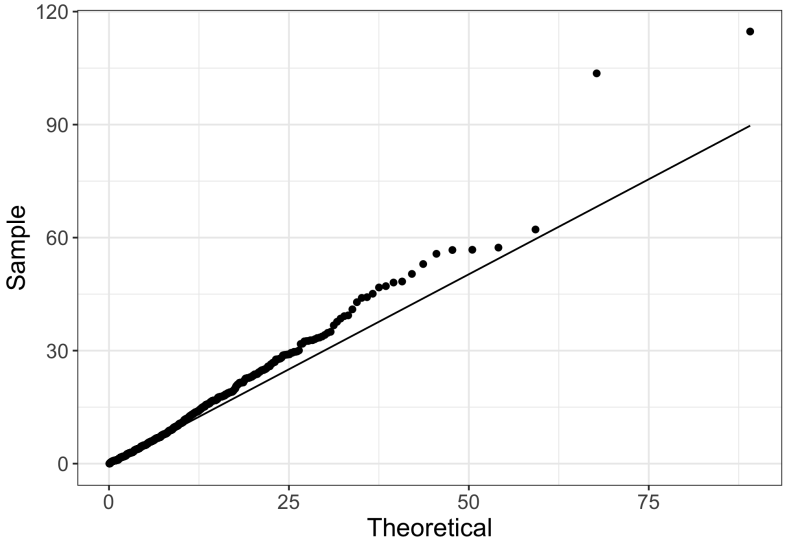


Figure 5.5.10: Quantile-quantile plot of the observed patient lengths of stay from the RAH ICU data set compared to those expected assuming a log-normal distribution.

5.5.3 Independence

A crucial assumption in all standard queueing models is that there is independence between the inter-arrival times and the distribution of service times. The assumption of independence increases the tractability of many problems and is inherent in available simulation-based methods used to analyse queueing systems.

In the context of the ICU, we require independence between the patient inter-arrival times and the distribution of patient length of stay. However, studies have suggested that a correlation structure may exist between the patient inter-arrival times and patient lengths of stay due to the moderated admissions and discharges in the ICU.

Varney *et al.* [47] used semi-experiments to show that there is evidence of a correlation structure between the patient inter-arrival times and the distribution of patient length of stay at various ICUs across hospitals in Australia and New Zealand. Within this analysis, Varney *et al.* considered the different admission and length of stay behaviour between elective and emergency patients, as well as the time-dependent behaviour of admissions and discharges. Taking this into account, Varney *et al.* showed that there was dependence between the patient inter-arrival times and the distribution of patient length of stay in the RAH ICU.

5.6 Summary

Despite the patient inter-arrival times and patient lengths of stay not following exponential distributions, $M/M/\bullet$ queueing models are still widely used to model the bed occupancy in ICUs due to their simplicity. Alternatively, simulation-based methods can be used to model the distributions of the patient inter-arrival times and patient lengths of stay [2]. However, neither approach addresses the lack of independence between the patient inter-arrival times and the patient length of stay in the ICU. Hence, the standard queueing models used to model the bed occupancy in ICUs are invalid for this research problem and alternative modelling methods are needed.

Chapter 6

Structured Markov chains

Quasi-birth-and-death (QBD) processes are Markov processes in two dimensions, the level and the phase, such that the process is skip-free in the level. That is, the QBD process does not move across several levels in a single transition. Furthermore, QBD processes provide freedom in the distribution of sojourn times for each level, as well as provide a framework to model queueing systems with a dependence structure between the arrival process and the service times. In this chapter, we first review phase-type distributions and matrix-exponential distributions which form a basis for the discussion of quasi-birth-and-death processes.

6.1 Phase-type distributions

The distributions of many natural phenomena are not symmetric or are possibly only defined for positive real numbers, thus rendering many common distributions unsuitable for such research problems. Phase-type distributions have the ability to model any given distribution with non-negative data and are commonly used in the modelling of healthcare systems and telecommunication systems due to tractability and preservation of the Markovian structure.

In this section, we define phase-type distributions, discuss the representation of phase-type distributions, and provide some examples.

6.1.1 Continuous phase-type distributions

Consider a continuous-time Markov chain $\{J(t); t \geq 0\}$ on a finite state space $\mathcal{S} = \{0, 1, 2, \ldots, m\}$, where state 0 is an absorbing state. Let the initial state probability distribution be $(\alpha_0, \boldsymbol{\alpha}) = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m)$, such that $\sum_{i=0}^{m} \alpha_i = 1$. We define the generator matrix, Q, as

$$Q = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{t} & T \end{bmatrix},$$

where T is an $m \times m$ matrix containing the transition rates between the transient states $\{1, 2, \ldots, m\}$ and $\mathbf{t} = -T\mathbf{e}$ is a $m \times 1$ vector containing the rates of absorption into the absorbing state. Given that Q is a generator matrix of a Markov process,

$$T_{i,j} \geq 0, \quad 1 \leq i \neq j \leq m,$$

$$T_{i,i} < 0 \text{ with } T_{i,i} = -t_i - \sum_{\substack{j=1 \\ j \neq i}}^{m} T_{i,j}, \quad 1 \leq i \neq j \leq m,$$

$$t_i \geq 0, \quad 1 \leq i \leq m,$$

and

$$T\mathbf{e} + \mathbf{t} = \mathbf{0},$$

where $\mathbf{e} = (1, \ldots, 1)^T$.

The distribution of time τ until the process reaches the absorbing state, state 0, is said to be phase-type distributed. The phase-type distribution is said to have representation $(\boldsymbol{\alpha}, T)$ of order m, where T is the PH-generator, $\boldsymbol{\alpha}$ is the initial state probability distribution, and α_0 is known as the point mass at zero. That is,

$$\tau \sim PH(\boldsymbol{\alpha}, T).$$

Since each non-absorbing state is transient, absorption into state 0 in finite time will occur with probability one. Phase-type distributions also have the property that each state has a positive probability of being visited before the process is absorbed into state 0.

The cumulative distribution function for a phase-type distribution with representation $(\boldsymbol{\alpha}, T)$ is defined as

$$F(t) = \begin{cases} \alpha_0, & t = 0, \\ 1 - \boldsymbol{\alpha} \exp(Tt)\mathbf{e}, & t > 0, \end{cases}$$

where the matrix exponential is defined as

$$\exp(A) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n. \tag{6.1.1}$$

For a proof of Equation 6.1.1, see Latouche and Ramaswami [34]. Differentiating $F(t)$ with respect to $t > 0$ gives the probability density function,

$$f(t) = \boldsymbol{\alpha} \exp(Tt)\mathbf{t}, \quad t > 0.$$

The Laplace-Stieltjes transform for a phase-type distribution with representation $(\boldsymbol{\alpha}, T)$ is given by

$$\phi(s) = \alpha_0 + \boldsymbol{\alpha} \left(sI - T\right)^{-1} \mathbf{t}, \quad Re(s) \geq 0.$$

Differentiating the Laplace-Stieltjes transform k times with respect to s and setting $s = 0$ gives the k^{th} non-central moment

$$m_k = (-1)^k k! \boldsymbol{\alpha} T^{-k} \mathbf{e}, \quad k \geq 0.$$

6.1.2 Representation and minimal order

Phase-type distributions have a non-uniqueness of representation, such that the density of a phase-type distribution may have two distinct representations, $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\beta}, U)$. For example, the phase-type distribution with density

$$f(t) = e^{-t}, \quad t > 0,$$

has representations $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\beta}, U)$ given by

$$\boldsymbol{\alpha} = 1,$$

$$T = -1.$$

and

$$\boldsymbol{\beta} = (1, 0),$$

$$U = \begin{pmatrix} -2 & 1 \\ 0 & -1 \end{pmatrix}.$$

For phase-type distributions, there is always a representation of minimal order which is called the *minimal representation*. That is, a representation $(\boldsymbol{\alpha}, T)$ of a phase-type distribution is of minimal order if no other representation $(\boldsymbol{\beta}, U)$ exists with dim(U) $<$ dim(T). However, representations are not necessarily unique. In the example above, $(\boldsymbol{\alpha}, T)$ is the minimal representation for the given phase-type distribution.

6.1.3 Examples

Exponential distribution

One of the most commonly used distributions in applied probability is the exponential distribution. In the context of a phase-type distribution, the exponential distribution with density function

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0,$$

has a phase-type representation

$$\boldsymbol{\alpha} = 1,$$

$$T = -\lambda.$$

Here, we can interpret the exponential distribution as the distribution of time until some Markov process changes state, as illustrated in Figure 6.1.1.

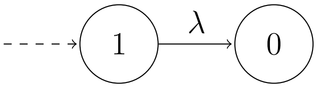


Figure 6.1.1: Phase-type representation of the exponential distribution.

Erlang distribution

The exponential distribution forms a building block to construct more complex distributions, such as the Erlang distribution which is the sum of $m \geq 1$ independent exponential random variables with parameter λ. The Erlang distribution of order $m \geq 1$ with density function

$$f(t) = \frac{\lambda^m t^{m-1} e^{-\lambda t}}{(m-1)!}, \quad t > 0,$$

has a phase-type representation

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0),$$

$$T = \begin{pmatrix} -\lambda & \lambda & 0 & \ldots & 0 & 0 \\ 0 & -\lambda & \lambda & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda & \lambda \\ 0 & 0 & 0 & \ldots & 0 & -\lambda \end{pmatrix}.$$

In this case, the time until absorption into state 0 is the sum of the exponential sojourn times in each of the states $1, 2, \ldots, m$, as illustrated in Figure 6.1.2.

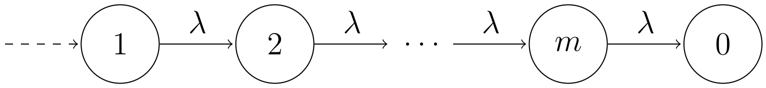


Figure 6.1.2: Phase-type representation of the Erlang distribution.

In the context of a queueing system, we can interpret this as m successive stages of service, each taking an independent, identically, and exponentially distributed amount of time.

Hyper-exponential distribution

Another complex distribution based on the exponential distribution is the hyper-exponential distribution which is defined as the convex mixture of $m \geq 1$ independent exponential distributions. The hyper-exponential distribution with density function

$$f(t) = \sum_{i=1}^{m} \alpha_i \lambda_i e^{-\lambda_i t}, \quad t > 0,$$

has a phase-type representation

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m),$$

where for $i = 1, 2, \dots, m$,

$$\alpha_i > 0,$$

$$\sum_{i=1}^{m} \alpha_i = 1, \text{ and}$$

$$T = \begin{pmatrix} -\lambda_1 & 0 & \dots & 0 \\ 0 & -\lambda_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -\lambda_m \end{pmatrix}.$$

Here, the time until absorption is the convex mixture of $m \geq 1$ independent exponential distributions where absorption into state 0 can be achieved from any state, as illustrated in Figure 6.1.3. Hence, in the context of a queueing system this may represent m different types of services that run in parallel.

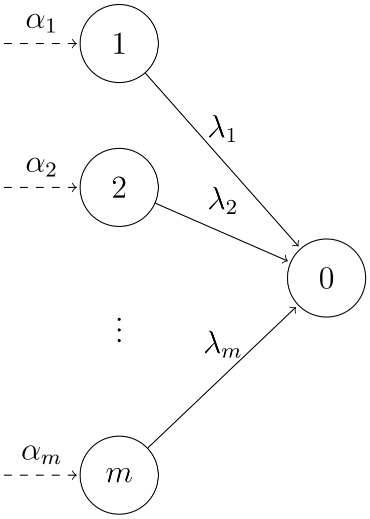


Figure 6.1.3: Phase-type representation of the hyper-exponential distribution.

Coxian distribution

The Coxian distribution of order $m \geq 1$ is a generalisation of the Erlang distribution which has a phase-type representation

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m),$$

where for $i = 1, 2, \ldots, m$, $\alpha_i > 0$ and $\sum_{i=1}^{m} \alpha_i = 1$, and for $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$,

$$T = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \ldots & 0 \\ 0 & 0 & -\lambda_3 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_m \end{pmatrix}.$$

Building upon the interpretation for the Erlang distribution, we can interpret the Coxian distribution as $m \geq 1$ successive stages of service, each taking an independent, exponentially distributed amount of time but absorption into state 0 can only achieved from state m, as illustrated in Figure 6.1.4.

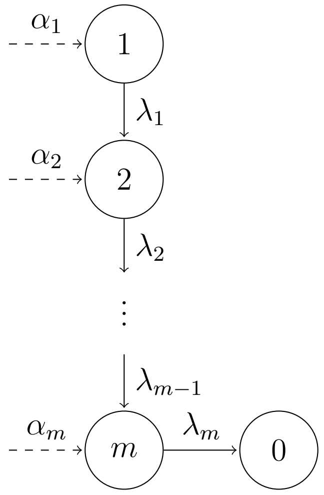


Figure 6.1.4: Phase-type representation of the Coxian distribution.

6.1.4 Parameter estimation

Various parameter estimation methods for phase-type distributions have been developed, many of which use the Expectation-Maximisation (EM) algorithm [1, 15, 42] or Markov-Chain Monte-Carlo methods [7].

The parameter estimation method for phase-type distributions developed by Asmussen *et al.* [1] used the connection between continuous-time Markov chains and phase-type distributions to calculate the maximum likelihood estimates.

Let $y \geq 0$ represent the observation of the time to absorption into state 0. Asmussen *et al.* viewed this as an incomplete observation of a Markov process, $\{X(t); t \geq 0\}$, since we only see when the process hits state 0 and do not observe how the process became to be absorbed. That is, we do not observe the starting state of the process, which states the process visited before absorption, or how long the process spent in each state.

Assume that over a time period $[0, T]$, we continuously observed K jumps of the process $\{X(t); t \geq 0\}$ until $X(t)$ was absorbed into state 0. Let S_k denote the state of the process at the k^{th} jump and let T_k denote the sojourn time before the k^{th} jump. Then the complete data in this case is defined as $\mathbf{X} = \{\mathbf{S}, \mathbf{T}\}$, where

- $\mathbf{S} = (S_0, S_1, \ldots, S_{K-1})$, where $S_K = 0$, and
- $\mathbf{T} = (T_0, T_1, \ldots, T_{K-1})$, where $T_K = \infty$.

Now suppose we have N independent realisations of the process, $X_1(t), \ldots, X_N(t)$, and let $\mathbf{S}^{(n)}$ denote the set of visited states and $\mathbf{T}^{(n)}$ denote the set of sojourn times for the n^{th} process. In this case, the observed data is now defined as $\mathbf{y} = (y_1, \ldots, y_N)$ and the complete data is now defined as $\mathbf{X} = \{\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(N)}, \mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(N)}\}$.

Assuming a phase-type distribution of order p, the density of the complete data, $\mathbf{x}$, under representation $PH(\boldsymbol{\alpha}, T)$ is of the form

$$f(\mathbf{x}|\boldsymbol{\alpha}, T) = \prod_{i=1}^{p} \alpha_i^{M_i} \prod_{i=1}^{p} e^{T_{i,i} Z_i} \prod_{i=1}^{p} \prod_{\substack{j=0 \\ j \neq i}}^{p} T_{i,j}^{N_{i,j}},$$

where the sufficient statistics are defined as follows:

- M_i is the total number of processes starting in state $i \in \{1, 2, \ldots, p\}$, such that
$$M_i = \sum_{n=1}^{N} \mathbf{1}_{\{S_0^{(n)} = i\}},$$
- Z_i is the total time spent in state $i \in \{1, 2, \ldots, p\}$, such that
$$Z_i = \sum_{n=1}^{N} \sum_{m=0}^{K^{(n)}-1} \mathbf{1}_{\{S_m^{(n)} = i\}} T_m^{(n)},$$
- and $N_{i,j}$ is the total number of jumps from state i to state j, for $i = 1, 2, \ldots, p$, $j = 0, 1, \ldots, p$, and $i \neq j$, such that
$$N_{i,j} = \sum_{n=1}^{N} \sum_{m=0}^{K^{(n)}-1} \mathbf{1}_{\{S_m^{(n)} = i,\ S_{m+1}^{(n)} = j\}}.$$

It then follows that the maximum likelihood estimates $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\alpha}}, \widehat{T}, \widehat{\mathbf{t}}\right)$ are defined as

$$\widehat{\alpha}_i = \frac{M_i}{N},$$

$$\widehat{T}_{i,j} = \frac{N_{i,j}}{Z_i},$$

$$\widehat{t}_i = \frac{N_{i,0}}{Z_i}, \text{ and}$$

$$\widehat{T}_{i,i} = -\left(\widehat{t}_i + \sum_{i=1}^{p} \sum_{\substack{j=0 \\ j \neq i}}^{p} \widehat{T}_{i,j}\right),$$

for $i, j = 1, 2, \ldots, p$, where N is the number of observed Markov processes.

6.2 Matrix exponential distributions

Matrix exponential distributions are an algebraic generalisation of phase-type distributions and were first introduced by Cox [12, 13] as distributions with a rational Laplace-Stieltjes transform. It was later shown that such distributions can be defined by vectors $\boldsymbol{\alpha}$ and $\mathbf{t}$ and a square matrix T, similar to phase-type distributions. However, the representations do not necessarily have a probabilistic interpretation [6, 36].

6.2.1 Distribution and moments

A non-negative random variable X has a matrix exponential distribution if there exists a $1 \times m$ vector $\boldsymbol{\alpha}$, a value $0 \leq \alpha_0 \leq 1$ which represents the point mass at state 0, an $m \times m$ matrix T, and an $m \times 1$ vector $\mathbf{t}$, such that

$$F_{\boldsymbol{\alpha},T,\mathbf{t}}(t) = \begin{cases} \alpha_0, & t = 0, \\ 1 + \boldsymbol{\alpha} \exp(Tt) T^{-1} \mathbf{t}, & t > 0, \end{cases}$$

is a valid cumulative distribution function; that is, it is right continuous, non-decreasing,

$$\lim_{t \to -\infty} F_{\boldsymbol{\alpha},T,\mathbf{t}}(t) = 0,$$

and,

$$\lim_{t \to \infty} F_{\boldsymbol{\alpha},T,\mathbf{t}}(t) = 1.$$

In that case, the distribution of X is said to have a matrix exponential representation $(\boldsymbol{\alpha}, T, \mathbf{t})$. That is, $X \sim ME(\boldsymbol{\alpha}, T, \mathbf{t})$.

The corresponding probability density function is defined as

$$f_{\boldsymbol{\alpha},T,\mathbf{t}}(t) = \boldsymbol{\alpha} \exp(Tt) \mathbf{t} \geq 0, \quad t > 0,$$

and the Laplace-Stieltjes transform is given by

$$\phi(s) = \alpha_0 + \boldsymbol{\alpha} \left(sI - T \right)^{-1} \mathbf{t}, \quad Re(s) \geq 0.$$

Differentiating the Laplace-Stieltjes transform k times with respect to s and setting $s = 0$ gives the k^{th} non-central moment

$$m_k = (-1)^{k+1} k! \boldsymbol{\alpha} T^{-(k+1)} \mathbf{t}, \quad k \geq 0.$$

6.2.2 Representation and minimal order

Matrix exponential distributions also have a non-uniqueness of representation, such that the density of a matrix exponential distribution may have two distinct representations, $(\boldsymbol{\alpha}, T, \mathbf{t})$ and $(\boldsymbol{\beta}, U, \mathbf{u})$. According to He *et al.* [25], the matrix exponential representations $(\boldsymbol{\alpha}, T, \mathbf{t})$ and $(\boldsymbol{\beta}, U, \mathbf{u})$ represent the same probability distribution if and only if $\boldsymbol{\alpha} T^k \mathbf{t} = \boldsymbol{\beta} U^k \mathbf{u}$ for $-\infty < k < \infty$.

The Laplace-Stieltjes transform of a matrix exponential distribution can also be expressed as a rational function of the form

$$\phi(s) = \frac{a_p s^{p-1} + a_{p-1} s^{p-2} + \ldots + a_1}{s^p + b_p s^{p-1} + b_{p-1} s^{p-2} + \ldots + b_1} + a_0,$$

where $p \geq 1$, $0 \geq a_0 \geq 1$, and $a_1, \ldots, a_p, b_1, \ldots, b_p$ are all real. If this is the case, then the matrix exponential distribution has a representation $(\boldsymbol{\alpha}, T, \mathbf{t})$ where,

$$\boldsymbol{\alpha} = (a_1, a_2, a_3, \ldots, a_{p-1}, a_p),$$

$$T = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -b_1 & -b_2 & -b_3 & \ldots & -b_{p-1} & -b_p \end{bmatrix},$$

and

$$\mathbf{t} = (0, 0, 0, \ldots, 0, 1)^T.$$

If the polynomials $a_p s^{p-1} + a_{p-1} s^{p-2} + \ldots + a_1$ and $s^p + b_p s^{p-1} + b_{p-1} s^{p-2} + \ldots + b_1$, for $Re(s) \geq 0$, have no common factors then the representation defined above is a representation of minimal order and is called the *minimal representation* [4].

6.2.3 Examples

Phase-type distribution

A simple example of a matrix-exponential distribution is any phase-type distribution, where $\boldsymbol{\alpha}$ is the initial state probability vector, T is the infinitesimal generator matrix, and $\mathbf{t} = -T\mathbf{e}$.

Matrix-exponential distribution

Consider a distribution with density,

$$f(t) = \frac{15}{7 + 15\delta} e^{-t} \left((2e^{-2x} - 1)^2 + \delta \right), \quad t > 0.$$

O'Cinneide [40] showed that this is an example of a phase-type distribution when $\delta > 0$, where the order of such distribution increases as δ tends to zero. However, this distribution is not phase-type since when $\delta = 0$, $f(0.5 \log(2)) = 0$.

By considering the Laplace-Stieltjes transform of $f(t)$,

$$\hat{f}(s) = \frac{15(1 + \delta)s^2 + 120\delta s + 225\delta + 105}{(7 + 15\delta)s^3 + (63135\delta)s^2 + (161 + 345\delta)s + 225\delta + 105},$$

we see that $f(t)$ represents a matrix-exponential distribution of (minimal) order 3 for all $\delta \geq 0$, thus presenting itself as an advantage compared to a phase-type distribution.

6.2.4 Parameter estimation

The use of matrix exponential distributions in the literature is less widespread than that of phase-type distributions. As a result, limited research has been done on the fitting of matrix exponential distributions to data. Fackrell [18] developed an estimation method using semi-infinite programming which calculates the maximum likelihood estimates of the vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$. The reader is referred to the paper for further discussion of the parameter estimation method.

6.3 Quasi-birth-and-death processes

A quasi-birth-and-death (QBD) process is a continuous-time Markov process $\{X(t), J(t); t \geq 0\}$ with state space $\mathcal{S} = \{(\ell, j); \ell \geq 0, j = 1, 2, \ldots, m\}$, where ℓ denotes the level of the QBD process and j denotes the phase of the QBD process. The state space $\mathcal{S}$ is also represented by $\ell^*(0) \cup \ell^*(1) \cup \ell^*(2) \cup \ldots$, where $\ell^*(\ell) = \{(\ell, j) : \ell \geq 0, j = 1, 2, \ldots, m\}$ defines the phases associated with level $\ell \geq 0$. The infinitesimal generator matrix for the QBD process is of the form

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \ldots \\ A_- & A_0 & A_+ & 0 & \ldots \\ 0 & A_- & A_0 & A_+ & \ldots \\ 0 & 0 & A_- & A_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where the rates of moving down one level, staying in the same level, or moving up one level are given by the matrices A_-, A_0, and A_+ respectively. Looking at the structure of the infinitesimal generator matrix for the QBD process, we see that transitions between levels are restricted to the adjacent levels, whereas the phase transitions are unrestricted.

A simple example of a QBD process is a single-server queueing system where the rates vary over time according to a phase or an environment process [34]. In this case, $X(t)$ represents the number of customers in the system at time $t \geq 0$ and $J(t)$ represents the phase or environment at time $t \geq 0$, where the number of phases or possible environments is held constant at $m \geq 1$.

Suppose that at time $s \geq 0$ the phase of the QBD process is such that $J(s) = i$, for $1 \leq i \leq m$, and at time $t > s > 0$ the phase or environment changes such that $J(t) = j$, for $1 \leq j \leq m$. Then the arrival rate is defined as $\lambda_{i,j}$ and the departure rate is defined as $\mu_{i,j}$, assuming that the server is occupied. A change in phase or environment can also occur without a change in the number of customers, which occurs at a rate of $s_{i,j}$. Table 6.3.1 summarises the possible transitions and the corresponding rates.

From	To	Rate	Conditions
(ℓ, i)	$(\ell - 1, j)$	$\mu_{i,j}$	$\ell \geq 1$
$(0, i)$	$(1, j)$	$s^*_{i,j}$	$i \neq j$
(ℓ, i)	(ℓ, j)	$s_{i,j}$	$\ell \geq 0$ and $i \neq j$
(ℓ, i)	$(\ell + 1, j)$	$\lambda_{i,j}$	$\ell \geq 0$

Table 6.3.1: Possible transitions and rates for a single-server queueing system where the rates vary over time according to a phase or environment process.

To illustrate the transitions between levels and phases, a subset of the state transition diagram is shown in Figure 6.3.1, and Equation (6.3.1) displays the infinitesimal generator matrix for this type of QBD process, where the diagonal entries are the negative of the relevant row sums.

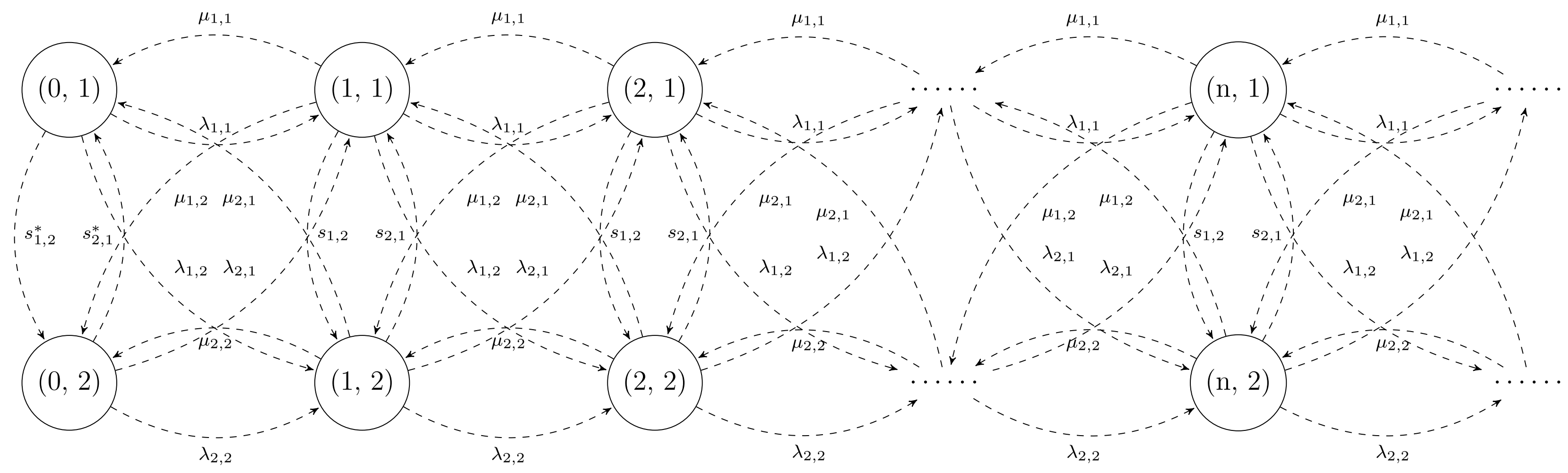


Figure 6.3.1: Subset of the state transition diagram for the single-server queueing system with two stages of service.

$$
Q = \begin{array}{r|cc|cc|cc|c}
 & (0,\,1) & (0,\,2) & (1,\,1) & (1,\,2) & (2,\,1) & (2,\,2) & \cdots \\
\hline
(0,\,1) & s^*_{1,1} & s^*_{1,2} & \lambda_{1,1} & \lambda_{1,2} & 0 & 0 & \cdots \\
(0,\,2) & s^*_{2,1} & s^*_{2,2} & \lambda_{2,1} & \lambda_{2,2} & 0 & 0 & \cdots \\
\hline
(1,\,1) & \mu_{1,1} & \mu_{1,2} & s_{1,1} & s_{1,2} & \lambda_{1,1} & \lambda_{1,2} & \cdots \\
(1,\,2) & \mu_{2,1} & \mu_{2,2} & s_{2,1} & s_{2,2} & \lambda_{2,1} & \lambda_{2,2} & \cdots \\
\hline
(2,\,1) & 0 & 0 & \mu_{1,1} & \mu_{1,2} & s_{1,1} & s_{1,2} & \cdots \\
(2,\,2) & 0 & 0 & \mu_{2,1} & \mu_{2,2} & s_{2,1} & s_{2,2} & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array} \tag{6.3.1}
$$

6.4 Level-dependent quasi-birth-and-death process

A level-dependent QBD process is also a continuous-time Markov process $\{X(t), J(t); t \geq 0\}$ with state space $\mathcal{S} = \{(\ell, j); \ell \geq 0, j = 1, 2, \ldots, J_\ell\}$, where ℓ denotes the level of the QBD process, j denotes the phase of the QBD process, and J_ℓ is the number of phases at the ℓ^{th} level. However, the transition rates and the number of phases at each level for a level-dependent QBD process depends on which level the QBD process is in. For example, the infinitesimal generator matrix for the level-dependent QBD process is of the form

$$
Q = \begin{bmatrix}
A_0^{(0)} & A_+^{(0)} & 0 & 0 & \cdots \\
A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \cdots \\
0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \cdots \\
0 & 0 & A_-^{(3)} & A_0^{(3)} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \tag{6.4.1}
$$

where the rates of moving down from level $\ell \geq 1$, staying in level $\ell \geq 0$, or moving up from level $\ell \geq 0$ are given by the matrices $A_-^{(\ell)}$, $A_0^{(\ell)}$, and $A_+^{(\ell)}$ respectively, and the phases associated with level ℓ are $\ell^*(\ell) = \{(\ell, j) : \ell \geq 0, j = 1, 2, \ldots, J_\ell\}$.

For each $\ell \geq 0$, the matrices $A_0^{(\ell)}$ are square matrices of order J_ℓ with strictly negative diagonal entries and non-negative off-diagonal entries. The matrices $A_-^{(\ell)}$ and $A_+^{(\ell)}$ are non-negative matrices with dimensions $J_\ell \times J_{\ell-1}$ and $J_\ell \times J_{\ell+1}$ respectively. Note that the row sums of the matrix $A_0^{(0)} + A_+^{(0)}$ are zero, as are the row sums for the matrices $A_-^{(\ell)} + A_0^{(\ell)} + A_+^{(\ell)}$, for $\ell \geq 1$.

The behaviour of a queueing process can be split into two categories: stationary and transient. Stationary behaviour relates to the long term behaviour of the queueing process, whereas the transient behaviour relates to short term behaviour of the queueing process.

6.4.1 Stationary distribution

If there exists a limiting distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$, then it must satisfy the system of equations $\boldsymbol{\pi} Q = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$, where $\boldsymbol{\pi}$ is a row vector and $\mathbf{e} = (1, \ldots, 1)^T$ is a column vector. The vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots$ satisfy the relation

$$\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_{\ell-1} R_{\ell-1}, \quad \ell \geq 0,$$

where the matrices R_ℓ, for $\ell \geq 0$, record the expected time spent in the states in level $\ell + 1$ before the first return to level $\ell \geq 0$ which is measured in units of sojourn time of level $\ell \geq 0$. The matrices R_ℓ are the minimal, non-negative solutions to the set of equations

$$A_+^{(\ell-1)} + R_{\ell-1} A_0^{(\ell)} + R_{\ell-1} R_\ell A_-^{(\ell+1)} = 0, \quad \ell \geq 0,$$

and $\boldsymbol{\pi}_0$ is the solution to

$$\boldsymbol{\pi}_0 (A_0^{(0)} + R_0 A_-^{(1)}) = \mathbf{0}.$$

The matrices R_ℓ, for $\ell \geq 0$ are calculated using an algorithm developed by Bright and Taylor [9].

For a finite level-dependent QBD process with C levels, we set $A_+^{(C)} = 0$ and $A_+^{(\ell)} = 0$, $A_0^{(\ell)} = 0$, and $A_-^{(\ell)} = 0$ for $\ell \geq C+1$ [21]. The vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_C$ now satisfy the relation

$$\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_{\ell-1} R_{\ell-1}, \quad 0 \leq \ell \leq C,$$

where the matrices R_ℓ, for $0 \leq \ell \leq C$, are the minimal, non-negative solutions to the set of equations

$$\begin{aligned} A_+^{(\ell-1)} + R_{\ell-1} A_0^{(\ell)} + R_{\ell-1} R_\ell A_-^{(\ell+1)} &= 0, \quad 0 \leq \ell \leq C-1, \\ A_+^{(C-1)} + R_{C-1} A_0^{(C)} &= 0, \end{aligned}$$

and $\boldsymbol{\pi}_0$ is the solution to

$$\boldsymbol{\pi}_0 (A_0^{(0)} + R_0 A_-^{(1)}) = \mathbf{0}.$$

In this case, the matrices R_ℓ, for $0 \leq \ell \leq C$, are calculated using an algorithm developed by Latouche, Jacobs, and Gaver [21].

6.4.2 Sojourn time

In the context of a level-dependent QBD process, we can think of the distribution of time spent in level $\ell \geq 0$ as the distribution of time τ_ℓ until the level-dependent QBD process leaves level $\ell \geq 0$. This type of distribution is called a phase-type distribution, such that

$$\tau_\ell \sim PH\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}\right),$$

where $\boldsymbol{\alpha}_\ell$ represents the distribution of phase upon first entry into level $\ell \geq 0$.

Let $S(t)$ represent the state of the level-dependent QBD process at time $t \geq 0$, such that

$$S(t) \in \{(\ell, j); \ell \geq 0, j = 1, 2, \ldots, J_\ell\},$$

For $\ell \geq 1$, we define the probability of being in phase $i \in J_\ell$ upon first entry into level $\ell \geq 0$ as

$$\begin{aligned}
[\boldsymbol{\alpha}_\ell]_i &= \lim_{\Delta t \to 0} P(S(t+\Delta t) = (\ell, i) | X(t) \neq \ell, X(t+\Delta t) = \ell) \\
&= \lim_{\Delta t \to 0} \frac{P(S(t+\Delta t) = (\ell, i), X(t+\Delta t) = \ell | X(t) \neq \ell)}{P(X(t+\Delta t) = \ell | X(t) \neq \ell)} \\
&= \lim_{\Delta t \to 0} \frac{\dfrac{P(S(t+\Delta t) = (\ell, i) | X(t) \neq \ell)}{\Delta t}}{\dfrac{P(X(t+\Delta t) = \ell | X(t) \neq \ell)}{\Delta t}} \\
&= \lim_{\Delta t \to 0} \frac{\begin{array}{c}\displaystyle\sum_{k \in J_{\ell-1}} \frac{P(S(t+\Delta t) = (\ell, i) | S(t) = (\ell-1, k))}{\Delta t} P(S(t) = (\ell-1, k) | X(t) \neq \ell) \; + \\ \displaystyle\sum_{k \in J_{\ell+1}} \frac{P(S(t+\Delta t) = (\ell, i) | S(t) = (\ell+1, k))}{\Delta t} P(S(t) = (\ell+1, k) | X(t) \neq \ell)\end{array}}{\begin{array}{c}\displaystyle\sum_{j \in J_\ell} \sum_{k \in J_{\ell-1}} \frac{P(S(t+\Delta t) = (\ell, j) | S(t) = (\ell-1, k))}{\Delta t} P(S(t) = (\ell-1, k) | X(t) \neq \ell) \; + \\ \displaystyle\sum_{j \in J_\ell} \sum_{k \in J_{\ell+1}} \frac{P(S(t+\Delta t) = (\ell, j) | S(t) = (\ell+1, k))}{\Delta t} P(S(t) = (\ell+1, k) | X(t) \neq \ell)\end{array}} \\
&= \frac{\displaystyle\sum_{k \in J_{\ell-1}} \left[A_+^{(\ell-1)}\right]_{k,i} \pi_{\ell-1,k} + \sum_{k \in J_{\ell+1}} \left[A_-^{(\ell+1)}\right]_{k,i} \pi_{\ell+1,k}}{\displaystyle\sum_{j \in J_\ell} \left(\sum_{k \in J_{\ell-1}} \left[A_+^{(\ell-1)}\right]_{k,j} \pi_{\ell-1,k} + \sum_{k \in J_{\ell+1}} \left[A_-^{(\ell+1)}\right]_{k,j} \pi_{\ell+1,k} \right)} \\
&= \frac{\left[\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)}\right]_i}{\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} \mathbf{e} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)} \mathbf{e}}.
\end{aligned}$$

Similarly, for $\ell = 0$, we define the probability of being in phase $i \in J_\ell$ upon first entry into level 0 as

$$\begin{aligned}
[\boldsymbol{\alpha}_0]_i &= \lim_{\Delta t \to 0} P(S(t+\Delta t) = (0,i) | X(t) \neq 0, X(t+\Delta t) = 0) \\
&= \lim_{\Delta t \to 0} \frac{P(S(t+\Delta t) = (0,i), X(t+\Delta t) = 0 | X(t) \neq 0)}{P(X(t+\Delta t) = 0 | X(t) \neq 0)} \\
&= \lim_{\Delta t \to 0} \frac{\dfrac{P(S(t+\Delta t) = (0,i) | X(t) \neq 0)}{\Delta t}}{\dfrac{P(X(t+\Delta t) = 0 | X(t) \neq 0)}{\Delta t}} \\
&= \lim_{\Delta t \to 0} \frac{\displaystyle\sum_{k \in J_1} \frac{P(S(t+\Delta t) = (0,i) | S(t) = (1,k))}{\Delta t} P(S(t) = (1,k) | X(t) \neq 0)}{\displaystyle\sum_{j \in J_0} \sum_{k \in J_1} \frac{P(S(t+\Delta t) = (0,j) | S(t) = (1,k))}{\Delta t} P(S(t) = (1,k) | X(t) \neq 0)} \\
&= \frac{\displaystyle\sum_{k \in J_1} \left[A_-^{(1)}\right]_{k,i} \pi_{1,k}}{\displaystyle\sum_{j \in J_0} \sum_{k \in J_1} \left[A_-^{(1)}\right]_{k,j} \pi_{1,k}} \\
&= \frac{\left[\boldsymbol{\pi}_1 A_-^{(1)}\right]_i}{\boldsymbol{\pi}_1 A_-^{(1)} \mathbf{e}}.
\end{aligned}$$

Therefore, the distribution of phase upon first entry to level $\ell \geq 0$ is defined as

$$\boldsymbol{\alpha}_\ell = \begin{cases} \dfrac{\boldsymbol{\pi}_1 A_-^{(1)}}{\boldsymbol{\pi}_1 A_-^{(1)} \mathbf{e}}, & \text{if } \ell = 0 \\[2ex] \dfrac{\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)}}{\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} \mathbf{e} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)} \mathbf{e}}, & \text{if } \ell \geq 1. \end{cases}$$

6.4.3 Conditional sojourn time

In some cases, the behaviour of the level-dependent QBD process before moving to the level above in taboo of the levels below (interval of time where the QBD process does not move below the current level) may be different to the behaviour of the level-dependent QBD process before moving to the level below in taboo of the levels above. Hence, sojourn times conditioned on the transition may provide more information about the transient behaviour of the level-dependent QBD process.

Here, we introduce the conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell + 1$. The conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell - 1$ follows a similar discussion.

Let τ_ℓ^+ be the time at which the level-dependent QBD process leaves level $\ell \geq 1$ and transitions to level $\ell + 1$ in taboo of levels $\ell - 1, \ell - 2, \ldots, 1, 0$. Note that if the level-dependent QBD process moves from level $\ell \geq 1$ to level $\ell - 1$ first, then $\tau_\ell^+ = \infty$.

The distribution of τ_ℓ^+ can be described by a phase-type distribution with infinitesimal generator matrix,

$$Q = \begin{bmatrix} 0 & \mathbf{0} \\ A_+^{(\ell)}\mathbf{e} & A_0^{(\ell)} \end{bmatrix},$$

where $A_0^{(\ell)}$ is the $m \times m$ matrix containing the transition rates between the phases in level $\ell \geq 1$ and $A_+^{(\ell)}\mathbf{e}$ is the $m \times 1$ vector containing the rates of moving up to level $\ell + 1$ from each phase in level $\ell \geq 1$. Given that the row sums of this generator matrix are non-positive and are not always zero, this distribution is called a *dishonest phase-type distribution.*

The dishonest phase-type distribution has presentation $PH_d\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}\right)$, where $\boldsymbol{\alpha}_\ell$ is the distribution of phase upon first entry into level $\ell \geq 1$, $A_0^{(\ell)}$ contains the transitions rates for phases $1, 2, \ldots, m$, and $A_+^{(\ell)}\mathbf{e}$ contains the rates of absorption into level $\ell + 1$ from level $\ell \geq 1$. Hence, $\tau_\ell^+ \sim PH_d\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}\right)$.

We define the cumulative distribution function of $PH_d\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}\right)$ as

$$\begin{aligned} F_{\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}}(t) &= P(\tau_\ell^+ \leq t | X(0) = \ell), \\ &= \boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} \left[I - e^{A_0^{(\ell)}t}\right] A_+^{(\ell)}\mathbf{e}, \quad t \geq 0 \end{aligned}$$

Differentiating with respect to $t \geq 0$ gives the probability density function,

$$f_{\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}}(t) = \boldsymbol{\alpha} e^{A_0^{(\ell)}t} A_+^{(\ell)}\mathbf{e}, \quad t \geq 0.$$

By the construction of this distribution, there is a chance that the level-dependent QBD process may move down to level $\ell - 1$ from level $\ell \geq 1$ instead of moving to level $\ell + 1$. In this case, we say that the level-dependent QBD process is absorbed in level $\ell - 1$ and $\tau_\ell^+ = \infty$. Hence,

$$\lim_{t \to \infty} F_{\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}}(t) = \boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} A_+^{(\ell)}\mathbf{e} < 1.$$

Now consider the conditional distribution of the time taken for the level-dependent QBD process to leave level $\ell \geq 1$, τ_ℓ, given that the transition was to level $\ell + 1$. Using the cumulative distribution function of the joint distribution, we have that for $t \geq 0$ and $T = \{0, 1, \ldots, \ell\}$,

$$\begin{aligned}
P(\tau_\ell \leq t | X(\tau_\ell) = \ell+1, X(\tau_\ell) \notin T, X(0) = \ell) &= \frac{P(\tau_\ell \leq t, X(\tau_\ell) = \ell+1 | X(0) = \ell)}{P(X(\tau_\ell) = \ell+1 | X(0) = \ell)} \\
&= \frac{F_{\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}}(t)}{F_{\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, A_+^{(\ell)}\mathbf{e}}(\infty)} \\
&= \boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} \left[I - e^{A_0^{(\ell)}t}\right] \frac{A_+^{(\ell)}\mathbf{e}}{\boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} A_+^{(\ell)}\mathbf{e}} \\
&= 1 - \frac{\boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} e^{A_0^{(\ell)}t} A_+^{(\ell)}\mathbf{e}}{\boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} A_+^{(\ell)}\mathbf{e}}.
\end{aligned}$$

Given that this is a non-negative cumulative distribution function, we find that the conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell + 1$, is defined as

$$\tau_\ell \sim ME\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, \frac{A_+^{(\ell)}\mathbf{e}}{\boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} A_+^{(\ell)}\mathbf{e}}\right).$$

By a similar method, the conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell - 1$, is defined as

$$\tau_\ell \sim ME\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}, \frac{A_-^{(\ell)}\mathbf{e}}{\boldsymbol{\alpha}_\ell \left(-A_0^{(\ell)}\right)^{-1} A_-^{(\ell)}\mathbf{e}}\right).$$

6.4.4 Transition probabilities

We do not observe the phase process when modelling queueing systems using QBD processes. Therefore, the transition probabilities we define here describe the QBD process moving between levels. That is, given that the QBD process is leaving level $\ell \geq 1$, we are interested in calculating the probability that the QBD process moves to level $\ell + 1$, as well as the probability that the QBD process moves to level $\ell - 1$. Note that the transition probability of moving from level 0 to level 1 is 1 by definition.

We begin by defining η_+ to be the first time that the level-dependent QBD process is in a level above $\ell \geq 0$, such that

$$\eta_+ = \inf\{t > 0 : X(t) > k\},$$

and η_- to be the first time that the level-dependent QBD process is in a level below $\ell \geq 1$, such that

$$\eta_- = \inf\{t > 0 : X(t) < k\}.$$

Let p_ℓ^+ denote the probability of moving up from level $\ell \geq 0$ to level $\ell + 1$, given that the QBD process is moving out of level $\ell \geq 0$. That is,

$$p_\ell^+ = \lim_{\Delta t \to 0} P(\eta_+ < \eta_- | \min(\eta_+, \eta_-) < \Delta t, X(0) = \ell).$$

Using the definition of conditional probability,

$$p_\ell^+ = \lim_{\Delta t \to 0} \frac{P(\eta_+ < \eta_- \cap \min(\eta_+, \eta_-) < \Delta t | X(0) = \ell)}{P(\min(\eta_+, \eta_-) < \Delta t | X(0) = \ell)}.$$

By the law of total probability, we sum over the possible phases $i \in J_\ell$ that the level-dependent QBD process could be moving up from, such that

$$p_\ell^+ = \lim_{\Delta t \to 0} \frac{\displaystyle\sum_{i \in J_\ell} P(\eta_+ < \eta_- \cap \min(\eta_+, \eta_-) < \Delta t | X(0) = \ell, J(0) = i) P(J(0) = i | X(0) = \ell)}{\displaystyle\sum_{i \in J_\ell} P(\min(\eta_+, \eta_-) < \Delta t | X(0) = \ell, J(0) = i) P(J(0) = i | X(0) = \ell)}.$$

Therefore,

$$p_\ell^+ = \lim_{\Delta t \to 0} \frac{\begin{array}{l}\displaystyle\sum_{i \in J_\ell} P(\eta_+ < \eta_- | \min(\eta_+, \eta_-) < \Delta t, X(0) = \ell, J(0) = i) \\ \qquad\qquad P(\min(\eta_+, \eta_-) < \Delta t | X(0) = \ell, J(0) = i) P(J(0) = i | X(0) = \ell)\end{array}}{\displaystyle\sum_{i \in J_\ell} P(\min(\eta_+, \eta_-) < \Delta t | X(0) = \ell, J(0) = i) P(J(0) = i | X(0) = \ell)}.$$

Assuming stationarity and focusing only on when the level-dependent QBD process changes level, we have that

$$p_\ell^+ = \lim_{\Delta t \to 0} \frac{\displaystyle\sum_{i \in J_\ell} \frac{\mathbf{e}_i A_+^{(\ell)} \mathbf{e}}{\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}} \left(\frac{\left(\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}\right) \Delta t + o(\Delta t)}{\Delta t} \right) [\tilde{\boldsymbol{\pi}}_\ell]_i}{\displaystyle\sum_{i \in J_\ell} \left(\frac{\left(\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}\right) \Delta t + o(\Delta t)}{\Delta t} \right) [\tilde{\boldsymbol{\pi}}_\ell]_i},$$

where $\tilde{\boldsymbol{\pi}}_\ell$ is the normalised stationary probability vector for level $\ell \geq 0$. That is,

$$\tilde{\boldsymbol{\pi}}_\ell = \frac{\boldsymbol{\pi}_\ell}{\boldsymbol{\pi}_\ell \mathbf{e}}, \quad \ell \geq 0.$$

To justify the interchange the summation and limit, we use the Monotone Convergence Theorem since all elements of this expression are non-negative and bounded by definition and both the numerator and denominator are monotonically increasing as Δt tends to zero.

Therefore,

$$\begin{aligned} p_\ell^+ &= \frac{\displaystyle\sum_{i \in J_\ell} \frac{\mathbf{e}_i A_+^{(\ell)} \mathbf{e}}{\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}} \lim_{\Delta t \to 0} \left(\frac{\left(\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}\right) \Delta t + o(\Delta t)}{\Delta t} \right) [\tilde{\boldsymbol{\pi}}_\ell]_i}{\displaystyle\sum_{i \in J_\ell} \lim_{\Delta t \to 0} \left(\frac{\left(\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}\right) \Delta t + o(\Delta t)}{\Delta t} \right) [\tilde{\boldsymbol{\pi}}_\ell]_i} \\ &= \frac{\displaystyle\sum_{i \in J_\ell} [\tilde{\boldsymbol{\pi}}_\ell]_i \mathbf{e}_i A_+^{(\ell)} \mathbf{e}}{\displaystyle\sum_{i \in J_\ell} [\tilde{\boldsymbol{\pi}}_\ell]_i \left(\mathbf{e}_i A_+^{(\ell)} \mathbf{e} + \mathbf{e}_i A_-^{(\ell)} \mathbf{e}\right)} \\ &= \frac{\tilde{\boldsymbol{\pi}}_\ell A_+^{(\ell)} \mathbf{e}}{\tilde{\boldsymbol{\pi}}_\ell A_+^{(\ell)} \mathbf{e} + \tilde{\boldsymbol{\pi}}_\ell A_-^{(\ell)} \mathbf{e}}. \end{aligned}$$

By a similar method, the transition probability of moving from level $\ell \geq 1$ to level $\ell - 1$ is defined as,

$$p_\ell^- = \frac{\tilde{\boldsymbol{\pi}}_\ell A_-^{(\ell)} \mathbf{e}}{\tilde{\boldsymbol{\pi}}_\ell A_+^{(\ell)} \mathbf{e} + \tilde{\boldsymbol{\pi}}_\ell A_-^{(\ell)} \mathbf{e}}.$$

6.5 Level-independent quasi-birth-and-death process

Now let's focus on QBD processes where the transition rates are independent of the level, except for level 0. A QBD process with an infinitesimal generator matrix of the form (6.4.1) is called a level-independent QBD process with a boundary at level 0 if

- $J_i = m$, for all $i \geq 1$,
- $A_-^{(i)} = A_-$, for all $i \geq 1$,
- $A_0^{(i)} = A_0$, for all $i \geq 0$, and
- $A_+^{(i)} = A_+$, for all $i \geq 1$.

So the infinitesimal generator matrix for a level-independent quasi-birth-and-death process is of the form

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \dots \\ A_- & A_0 & A_+ & 0 & \dots \\ 0 & A_- & A_0 & A_+ & \dots \\ 0 & 0 & A_- & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where the rates of staying in level zero are given by the matrix B_0 and the rates of moving down one level, staying in the same level, or moving up one level are given by the matrices A_-, A_0, and A_+ respectively.

The stationary and transient behaviour of a level-independent QBD process is defined similar to that of a level-dependent QBD process. Hence, we only draw attention to the changes in distributions.

6.5.1 Stationary distribution

If there exists a limiting distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$, then it must satisfy the system of equations $\boldsymbol{\pi}Q = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$. The vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots$ satisfy the relation

$$\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_{\ell-1}R, \quad \ell \geq 0,$$

where R is the minimal, non-negative solution to the set of equations

$$A_+ + RA_0 + R^2A_- = 0, \quad \ell \geq 0,$$

and $\boldsymbol{\pi}_0$ is uniquely determined by solving the boundary condition,

$$\boldsymbol{\pi}_0(B_0 + RA_-) = \mathbf{0},$$

and the normalising condition,

$$\boldsymbol{\pi}_0(I - R)^{-1}\mathbf{e} = 1,$$

where I is the identity matrix of appropriate dimension and $\mathbf{e} = (1, \ldots, 1)^T$.

6.5.2 Sojourn time

In the context of a level-independent QBD process, we focus on the distribution of time spent in level 0 and the distribution of time spent in levels 1 and above.

For the case of $\ell = 0$, we can think of the distribution of time spent in level 0 as the distribution of time τ_0 until the level-independent QBD process moves up to level 1. Hence, $\tau_0 \sim PH\left(\boldsymbol{\alpha}_0, B_0\right)$, where $\boldsymbol{\alpha}_0$ is the initial phase probability distribution for level 0, such that

$$\boldsymbol{\alpha}_0 = \frac{\boldsymbol{\pi}_1 A_-}{\boldsymbol{\pi}_1 A_- \mathbf{e}}.$$

For the case of $\ell \geq 1$, we can think of the distribution of time spent in level $\ell \geq 1$ as the distribution of time τ_ℓ until the level-independent QBD process leaves level $\ell \geq 1$. Hence, $\tau_\ell \sim PH\left(\boldsymbol{\alpha}_\ell, A_0^{(\ell)}\right)$, for level $\ell \geq 1$, where $\boldsymbol{\alpha}_\ell$ is the initial phase probability distribution for level $\ell \geq 1$, such that

$$\boldsymbol{\alpha}_\ell = \frac{\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)}}{\boldsymbol{\pi}_{\ell-1} A_+^{(\ell-1)} \mathbf{e} + \boldsymbol{\pi}_{\ell+1} A_-^{(\ell+1)} \mathbf{e}}.$$

6.5.3 Conditional sojourn time

For the level-independent QBD process, the behaviour of the QBD process before moving to the level above in taboo of the levels below may be different to the behaviour of the level-dependent QBD process before moving to the level below in taboo of the levels above.

In this section, we introduce the conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell + 1$. The conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell - 1$ follows a similar discussion.

Let τ_ℓ, for $\ell \geq 1$ be the time at which the QBD process leaves level $\ell \geq 1$ and the QBD process transitions to level $\ell + 1$ in taboo of levels $\ell - 1, \ell - 2, \ldots, 1, 0$. Then the conditional distribution of the sojourn time in level $\ell \geq 1$, given that the first transition from level $\ell \geq 1$ is to level $\ell + 1$, is defined as

$$\tau_\ell \sim ME\left(\boldsymbol{\alpha}_\ell, A_0, \frac{A_+ \mathbf{e}}{\boldsymbol{\alpha}_\ell \left(-A_0\right)^{-1} A_+ \mathbf{e}}\right).$$

6.6 Summary

Quasi-birth-and-death processes provide freedom in the distribution of inter-arrival times and service times and have the potential to explain more of the behaviour of a queueing system than explained using a general birth-and-death process. Depending on the identifiability of the model and data availability, we can fit queueing models to queueing systems by exploiting the design of the queueing system and the observed data. However, limited research has been done on the statistical modelling of queueing systems using quasi-birth-and-death processes. In the next chapter, we develop a statistical modelling method to fit arbitrary quasi-birth-and-death processes to queueing system data.

Chapter 7

Fitting Quasi-Birth-and-Death processes to queueing system data

Quasi-birth-and-death processes provide a natural framework to model queueing systems with a dependence structure between the arrival process and the distribution of service times, such as an intensive care unit which may have a dependence structure between the patient admission process and the distribution of patient length of stay in an ICU [47].

The modelling approach for QBD processes with independence between the arrival process and the distribution of service times remains a relatively straight-forward task, such that the arrival process is modelled independently to the distribution of service times. Whereas, the modelling approach becomes less clear in situations where the phases of the arrival process and the distribution of service times interact, thus leading to dependence between the arrival process and the distribution of service times. Therefore, statistical model fitting methods of arbitrary QBD processes to queueing system data are needed in order to allow the phases of the arrival process and the distribution of service times to interact where needed.

While observing a QBD queueing process, all that is observed are the changes in level and the times at which a change in level occurs. The transitions between phases remain hidden, thus leading to incomplete data. For statistical model fitting problems where we have incomplete data, the expectation-maximisation (EM) algorithm is used to compute the maximum likelihood estimators. In this chapter, we develop an EM algorithm for the parameter estimation of arbitrary quasi-birth-and-death processes.

7.1 Expectation-maximisation algorithm

The EM algorithm is an iterative algorithm used to compute the maximum likelihood estimates of a set of parameters, $\boldsymbol{\theta}$, in situations where we have incomplete data [15].

Suppose we wish to fit a probability distribution $F(\mathbf{X}, \boldsymbol{\theta})$, with k parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ to a data set of observed data $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$. In the context of maximum likelihood estimation, we aim to find an estimate $\hat{\boldsymbol{\theta}}$ given $\mathbf{X} = \mathbf{x}$ that maximises the log-likelihood function,

$$l(\boldsymbol{\theta}|\mathbf{x}) = \log\left(L(\boldsymbol{\theta}|\mathbf{x})\right),$$

where $L(\boldsymbol{\theta}|\mathbf{x})$ is the likelihood of $\mathbf{x}$ given $\boldsymbol{\theta}$. That is, $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l(\boldsymbol{\theta}|\mathbf{x})$, where $\boldsymbol{\Theta}$ is the parameter space.

Suppose we have unobserved or hidden data, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$. So the complete data is denoted by $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ and the complete data log-likelihood is defined as

$$l_C(\boldsymbol{\theta}|\mathbf{d}) = \log\left(L_C(\boldsymbol{\theta}|\mathbf{d})\right).$$

The EM algorithm maximises the conditional expected value of the complete log-likelihood, $E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x},\widehat{\boldsymbol{\theta}}\right]$ with respect to the unobserved data, where $\widehat{\boldsymbol{\theta}}$ is the current estimate of the parameters, $\boldsymbol{\theta}$.

Each iteration of the EM algorithm consists of an expectation step and a maximisation step. In the expectation step (E-step), we calculate the expected value of the log-likelihood using the current estimates of the parameters and the observed data. That is, in the E-step we calculate the conditional expectation

$$E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(t)}\right].$$

In the maximisation step (M-step), we maximise the conditional log-likelihood and obtain new estimates of the parameters. So we compute $\theta^{(t+1)}$ by maximising $Q(\theta|\widehat{\theta}^{(t)})$, such that

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \text{argmax}_\theta \left\{E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(t)}\right]\right\}.$$

The EM algorithm iterates between the E-step and the M-step until only small improvements are being made in the log-likelihood function between successive iterations. That is,

$$l_C(\widehat{\boldsymbol{\theta}}^{(t+1)}|\mathbf{d}) - l_C(\widehat{\boldsymbol{\theta}}^{(t)}|\mathbf{d}) < \epsilon,$$

where ϵ is some tolerance.

7.1.1 Expectation-conditional maximisation algorithm

An extension of the EM algorithm is the expectation-conditional maximisation (ECM) algorithm, where the maximisation step of the EM algorithm is replaced by a sequence of conditional maximisation (CM) steps [38].

Consider a set of S vector functions of θ,

$$\{g_s(\theta); s = 1, \ldots, S\}.$$

For each iteration $t = 0, 1, \ldots$, we find the value of

$$\theta^{(t+s/S)} = \{\theta_1^{(t+1)}, \ldots, \theta_s^{(t+1)}, \theta_{s+1}^{(t)}, \ldots, \theta_S^{(t)}\}$$

that maximises the conditional log-likelihood subject to the constraint $g_s(\theta) = g_s(\theta^{(t+(s-1)/S)}$.

For example, suppose the parameter $\boldsymbol{\theta}$ is partitioned into sub-vectors $\theta = \{\theta_1, \ldots, \theta_S\}$. Then the s^{th} step of the CM step maximises the conditional log-likelihood with respect to θ_s, for $1 \leq s \leq S$, subject to all other parameters held fixed.

7.2 EM algorithm for infinite level-dependent QBD processes

In this section, we first describe the data and the estimators for a level-dependent QBD process under the assumption that the phase process is unobservable. We then define the EM algorithm for a level-dependent QBD process, under the same assumption.

7.2.1 Observed data

Assume that over a time period $[0, T]$ we continuously observe the input and output of a queueing system. The observed data consists of K observed changes in level and the times at which each change in level occurred. Let $X = (\mathbf{T}, \mathbf{L})$ denote the observed random vector, where

- $\mathbf{T} = (t_1, t_2, \ldots, t_K)$, with t_k the inter-event time between the $k - 1^{th}$ and k^{th} level changes, and
- $\mathbf{L} = (l_1, l_2, \ldots, l_{K+1})$, with l_k the level after the $k - 1^{th}$ change in level.

So the time of the k^{th} change in level is defined as $s_k = \sum_{i=1}^{k} t_i$, where $s_0 = 0$.

7.2.2 Unobserved data

Since the phase process of the QBD process is unobservable, we have incomplete data $Y = \{Y_\ell = (\mathbf{Z}_\ell, S_\ell, D_\ell, U_\ell); \ell \geq 0\}$, where

- $Z_{\ell,i}^{(k)}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in level $\ell \geq 0$, for $1 \leq i \leq J_\ell$,
- $S_{\ell,i,j}^{(k)}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in level $\ell \geq 0$, for $1 \leq i, j \leq J_\ell$, and $i \neq j$,
- $D_{\ell,i,j}^{(k)}$ is the indicator random variable for the event that the transition at time s_k is from level $\ell \geq 1$ to level $\ell - 1$ and leads to a phase transition from i to j, for $1 \leq i \leq J_\ell$ and $1 \leq j \leq J_{\ell-1}$,
- $U_{\ell,i,j}^{(k)}$ is the indicator random variable for the event that the transition at time s_k is from level $\ell \geq 0$ to level $\ell + 1$ and leads to a phase transition from i to j, for $1 \leq i \leq J_\ell$ and $1 \leq j \leq J_{\ell+1}$,

and,

$$[Z_\ell]_i = Z_{\ell,i} = \sum_{k=1}^{K} Z_{\ell,i}^{(k)} \quad \ell \geq 0, 1 \leq i \leq J_\ell,$$

$$[S_\ell]_{i,j} = S_{\ell,i,j} = \sum_{k=1}^{K} S_{\ell,i,j}^{(k)} \quad \ell \geq 0, 1 \leq i \neq j \leq J_\ell, ,$$

$$[D_\ell]_{i,j} = D_{\ell,i,j} = \sum_{k=1}^{K} D_{\ell,i,j}^{(k)} \quad \ell \geq 1, 1 \leq i \leq J_\ell, \quad 1 \leq j \leq J_{\ell-1}, \text{ and,}$$

$$[U_\ell]_{i,j} = U_{\ell,i,j} = \sum_{k=1}^{K} U_{i,j}^{(k)} \quad \ell \geq 0, 1 \leq i \leq J_\ell, \quad 1 \leq j \leq J_{\ell+1}.$$

Hence, the sufficient statistic for the complete data $\mathbf{D}$ is the collection of random variables $Y = \{Y_\ell = (\mathbf{Z}_\ell, S_\ell, D_\ell, U_\ell); \ell \geq 0\}$.

7.2.3 Likelihood

Using a similar argument to that in Asmussen *et al.* [1], the likelihood of the complete data $\mathbf{D} = (X, Y)$ (*i.e.* if the phase process were observable) with parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$, is

$$L_C(\boldsymbol{\theta}|\mathbf{d}) = \prod_{\ell=0}^{\infty} \prod_{i=1}^{J_\ell} \exp(A_{0_{i,i}}^{(\ell)} Z_{\ell,i}) \prod_{\ell=0}^{\infty} \prod_{i=1}^{J_\ell} \prod_{\substack{j=1 \\ j \neq i}}^{J_\ell} {A_{0_{i,j}}^{(\ell)}}^{S_{\ell,i,j}}$$

$$\prod_{\ell=1}^{\infty} \prod_{i=1}^{J_\ell} \prod_{j=1}^{J_{\ell-1}} {A_{-_{i,j}}^{(\ell)}}^{D_{\ell,i,j}} \prod_{\ell=0}^{\infty} \prod_{i=1}^{J_\ell} \prod_{j=1}^{J_{\ell+1}} {A_{+_{i,j}}^{(\ell)}}^{U_{\ell,i,j}},$$

and the complete data log-likelihood is given by

$$\ell_C(\boldsymbol{\theta}|\mathbf{d}) = \sum_{\ell=0}^{\infty} \sum_{i=1}^{J_\ell} A_{0_{i,i}}^{(\ell)} Z_{\ell,i} + \sum_{\ell=0}^{\infty} \sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j \neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j}$$

$$+ \sum_{\ell=1}^{\infty} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=0}^{\infty} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}.$$

Using the complete data log-likelihood, we find that the entries of the maximum likelihood estimates $\widehat{\boldsymbol{\theta}} = (\{\widehat{A_0}^{(\ell)}; \ell \geq 0\}, \{\widehat{A_-}^{(\ell)}; \ell \geq 1\}, \{\widehat{A_+}^{(\ell)}; \ell \geq 0\})$ are defined as

$$\widehat{A}_{0_{i,j}}^{(\ell)} = \frac{S_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } \ell \geq 0 \text{ and } 1 \leq i \neq j \leq J_\ell, \tag{7.2.1}$$

$$\widehat{A}_{-_{i,j}}^{(\ell)} = \frac{D_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } \ell \geq 1 \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1}, \tag{7.2.2}$$

$$\widehat{A}_{+_{i,j}}^{(\ell)} = \frac{U_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } \ell \geq 0 \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1}, \tag{7.2.3}$$

and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

For each iteration of the EM algorithm, there are two steps; the E-step and the M-step. The E-step of the algorithm involves calculating the conditional expectation of the sufficient statistics given the observed data and the current estimates of the parameters, $\boldsymbol{\theta}$. In the M-step of the algorithm, the complete-data likelihood is maximised using the values of the conditional expectations calculated in the E-step. We now detail the E-step and then the M-step of the EM algorithm.

7.2.4 E-step

Assuming we know the values of the parameters, $\boldsymbol{\theta}$, the likelihood of the observed data is defined as

$$f(\mathbf{x}) = \boldsymbol{\alpha} \exp(N_1 t_1) M_1 \times \exp(N_2 t_2) M_2 \times \ldots \times \exp(N_K t_K) M_K \exp(N_{K+1}(T - s_K))\mathbf{e}, \tag{7.2.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J_0})$ is the starting distribution, $\mathbf{e}$ is a column vector of ones,

$$N_k = A_0^{(\ell_k)},$$

and

$$M_k = \begin{cases} A_+^{(\ell_k)}, & \text{if } \ell_{k+1} = \ell_k + 1, \\ A_-^{(\ell_k)}, & \text{if } \ell_{k+1} = \ell_k - 1. \end{cases}$$

Let $\mathbf{f}_k(u)$ represent the forward likelihood of the observed data before time $s_{k-1} + u$ with $u \in [0, t_k]$ and let $\mathbf{b}_k(u)$ represent the backward likelihood of the observed data after time $s_k - u$, again with $u \in [0, t_k]$. Let $N(t)$ be the number of level changes before time t and let $J(t)$ be the index of the phase at time t. Let $\mathcal{F}_k$ denote the set of events for the time interval $[0, s_k]$ and let $\mathcal{B}_k$ denote the set of events for the time interval $[s_k, s_K]$.

The i^{th} element of $f_k(u)$ is defined as

$$[\mathbf{f}_k(u)]_i = P(\mathcal{F}_{k-1}, N(s_{k-1}+u^-) - N(s_{k-1}) = 0, J(s_{k-1}+u^-) = i).$$

The i^{th} element of $b_k(u)$ is defined as

$$[\mathbf{b}_k(u)]_i = P(\mathcal{B}_k, N(s_k^-) - N(s_k - u) = 0 | J(s_k - u) = i).$$

So the forward and backward likelihoods of the observed data are defined as

$$\mathbf{f}_k(u) = \boldsymbol{\alpha}\left(\prod_{j=1}^{k-1} \exp(N_j t_j) M_j\right)\exp(N_k u), \tag{7.2.5}$$

and,

$$\mathbf{b}_k(u) = \exp(N_k u) M_k \left(\prod_{j=k+1}^{K} \exp(N_j t_j) M_j\right)\exp(N_{K+1}(T - s_K))\mathbf{e}, \tag{7.2.6}$$

where $\mathbf{f}_k(u)$ is a row vector and $\mathbf{b}_k(u)$ is a column vector. Using the forward and backward likelihoods, the conditional expectations used in the E-step of the EM algorithm are as follows.

If the process is in level $\ell \geq 0$ during the time interval $[s_{k-1}, s_k)$:

$$E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i[\mathbf{b}_k(t_k-\tau)]_i d\tau}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } 1 \leq i \leq J_\ell, \text{ and}$$

$$E\big[S_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i({A_0^{(\ell)}}_{i,j})[\mathbf{b}_k(t_k-\tau)]_j d\tau}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } 1 \leq i \neq j \leq J_\ell.$$

Here, $\bullet$ represents the dot product.

If the process is moving from level $\ell \geq 1$ to level $\ell - 1$ at the k^{th} event:

$$E\big[D_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{[\mathbf{f}_k(t_k)]_i({A_-^{(\ell)}}_{i,j})[\mathbf{b}_{k+1}(t_{k+1})]_j}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1}.$$

If the process is moving from level $\ell \geq 0$ to level $\ell + 1$ at the k^{th} event:

$$E\big[U_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{[\mathbf{f}_k(t_k)]_i({A_+^{(\ell)}}_{i,j})[\mathbf{b}_{k+1}(t_{k+1})]_j}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1}.$$

7.2.5 M-step

In the M-step of the EM algorithm, we estimate the values of the parameters by replacing the statistics in Equations (7.2.1), (7.2.2), (7.2.3) with the conditional expectations calculated in the E-step, as follows:

$$\widehat{A_0}^{(\ell)}_{i,j} = \frac{\sum_{k=1}^{K} E\big[S^{(k)}_{\ell,i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 0 \text{ and } 1 \leq i \neq j \leq J_\ell,$$

$$\widehat{A_-}^{(\ell)}_{i,j} = \frac{\sum_{k=1}^{K} E\big[D^{(k)}_{\ell,i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 1 \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

$$\widehat{A_+}^{(\ell)}_{i,j} = \frac{\sum_{k=1}^{K} E\big[U^{(k)}_{\ell,i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 0 \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1},$$

and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

7.2.6 Avoiding numerical overflow and underflow

Recall the forward and backward likelihoods of the observed data defined in Equations (7.2.5) and (7.2.6) respectively. As the number of events increases, the elements of $\mathbf{f}_k(u)$ and $\mathbf{b}_k(u)$ become progressively smaller or larger, depending on the values of the transition rates in Q.

In order to avoid numerical overflow or underflow in the E-step of the algorithm, the forward and backward likelihood vectors, $\mathbf{f}_k(u)$ and $\mathbf{b}_k(u)$, must be scaled [45]. Since $\mathbf{f}_k(u)$ and $\mathbf{b}_k(u)$ are the product of matrices and not scalars, computing the log of $\mathbf{f}_k(u)$ and $\mathbf{b}_k(u)$ is not effective in avoiding numerical overflow or underflow. Instead, we normalise the forward and backward likelihood vectors at each event $k = 1, \ldots, K$.

Consider the forward likelihood vector, $\mathbf{f}_k(u)$ and let $\widehat{\mathbf{f}}_k(u)$ denote the scaled forward likelihood vector. We start by defining $\mathbf{f}_k^{(0)}(u) = \boldsymbol{\alpha}$. Since $\boldsymbol{\alpha}$ is a probability vector, $\widehat{\mathbf{f}}_k^{(0)}(u) = \boldsymbol{\alpha}$. For $m = 1, 2, 3, \ldots, k-1$,

$$\widehat{\mathbf{f}}_k^{(m)}(u) = \frac{\widehat{\mathbf{f}}_k^{(m-1)}(u)\exp(N_m t_m)M_m}{\widehat{\mathbf{f}}_k^{(m-1)}(u)\exp(N_m t_m)M_m\mathbf{e}}$$

and,

$$\widehat{\mathbf{f}}_k^{(k)}(u) = \frac{\mathbf{f}_k^{(k-1)}(u)\exp(N_k u)}{\mathbf{f}_k^{(k-1)}(u)\exp(N_k u)\mathbf{e}}.$$

So the scaled forward likelihood vector is

$$\widehat{\mathbf{f}}_k(u) = \widehat{\mathbf{f}}_k^{(k)}(u).$$

The scaled backward likelihood vector, $\widehat{\mathbf{b}}_k(u)$, is calculated in a similar way.

7.2.7 Reducing computation time

The EM algorithm has been widely used to estimate the parameters of phase-type distributions [1], Markovian arrival processes (MAPs) [41], and Markovian binary trees (MBTs) [24]. There have been several improvements to the EM algorithm for MAPs, particularly relating to the computational time of the algorithm. Buchholz [10], Klemm *et al.* [33], and Okamura *et al.* [41] presented improvements to the EM algorithm for BMAPs and MAPs involving the uniformisation technique.

In the E-step of the EM algorithm described in Section 7.2.4, the computation of $E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]$ and $E\big[S_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]$ involves the integral of matrix expressions, thus increasing the computation time of the algorithm. In this section, we extend the uniformisation-based EM algorithm developed by Okamura *et al.* [41] to quasi-birth-and-death processes in order to reduce the computation time of the EM algorithm.

Consider a continuous-time Markov chain with transition rate matrix, Q. Let q be a constant which is at least as large as the maximum absolute value of the diagonal elements of the transition rate matrix, Q. That is,

$$q \geq \max_i \left(|q_{i,i}|\right).$$

We can write the matrix exponential as

$$\exp(Qt) = \sum_{z=0}^{\infty} \text{Poi}(z;qt)\left(\mathbf{I} + \frac{Q}{q}\right)^z,$$

where $P = I + Q/q$ is the transition probability matrix of the associated discrete-time Markov chain, and

$$\text{Poi}(z;qt) = \frac{e^{-qt}(qt)^z}{z!}, \quad z = 0, 1, 2, \ldots.$$

By finding the right truncation point, R, such that

$$\sum_{z=0}^{R} \text{Poi}(z;qt) \geq 1 - \epsilon,$$

where ϵ is some defined tolerance, we can approximate the matrix exponential as

$$\exp(Qt) \approx \sum_{z=0}^{R} \text{Poi}(z;qt)\left(\mathbf{I} + \frac{Q}{q}\right)^z.$$

Recall the joint probability density function of the inter-event times defined in Equation (7.2.4) and the forward and backward likelihood vectors defined in Equations (7.2.5) and (7.2.6) respectively.

We define the matrix H_k as

$$H_k = \int_0^{t_k} \exp(N_k\tau)M_k\mathbf{b}_{k+1}(t_{k+1})\mathbf{f}_{k-1}(t_{k-1})M_{k-1}\exp(N_k(t_k{-}\tau))d\tau, \text{ for } k = 1, 2, \ldots, K,$$

noting that $\mathbf{f}_k(u)$ is a row vector and $\mathbf{b}_k(u)$ is a column vector. Using the matrix H_k, we can re-write the equations in the M-step as follows.

If the process is in level $\ell = N(s_{k-1}) \geq 0$:

$$E\big[Z_{\ell,i}^{(k)}|\mathbf{D}, \boldsymbol{\theta}\big] = \frac{[H_k]_{i,i}}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } \ell \geq 0 \text{ and } 1 \leq i \leq J_\ell, \text{ and}$$

$$E\big[S_{\ell,i,j}^{(k)}|\mathbf{D}, \boldsymbol{\theta}\big] = \frac{(A_0^{(\ell)}{}_{i,j})\,[H_k]_{j,i}}{\boldsymbol{\alpha} \bullet \mathbf{b}_1(t_1)}, \text{ for } \ell \geq 0 \text{ and } 1 \leq i \neq j \leq J_\ell.$$

The algorithm for computing the matrix H_k is presented below. For further discussion of this algorithm, the reader is referred to Okamura *et al.* [41].

Algorithm 1: Uniformisation-based integral of matrix exponential

Step 1: Choose the value of q, such that

$$q \geq \max_i \left(|q_{i,i}| \right).$$

Step 2: Calculate the right truncation point R_k for $k = 1, 2, \ldots, K, K+1$.

Step 3: Calculate the probability mass function for the Poisson distribution, $\text{Poi}(z; qt_k)$, for $k = 1, 2, \ldots, K, K+1$ and $z = 0, 1, 2, \ldots, R_k + 1$, where

$$q \geq \max \left(\left|\text{diag}\left(N_1\right)\right|, \left|\text{diag}\left(N_2\right)\right|, \ldots, \left|\text{diag}\left(N_{K+1}\right)\right| \right).$$

Step 4: For $k = K+1, K, \ldots, 2, 1$ and $z = 1, 2, \ldots, R_k$, calculate $\mathbf{b}_{k,z}$.

$$\mathbf{b}_{k,0} = M_k \widehat{\mathbf{b}}_{k+1}(t_{k+1}), \quad \mathbf{b}_{k,z} = \left(\mathbf{I} + \frac{N_k}{q} \right) \mathbf{b}_{k,z-1}.$$

Step 5: For $k = K+1, K, \ldots, 2, 1$ and $z = R_k - 1, R_k - 2, \ldots, 2, 1, 0$, calculate $\mathbf{c}_{k,z}$, where

$$\mathbf{c}_{k,R_k} = \text{Poi}(R_k + 1; qt_k) \widehat{\mathbf{f}}_{k-1}(t_{k-1}) M_k, \text{ and}$$

$$\mathbf{c}_{k,z} = \left(\mathbf{I} + \frac{N_k}{q} \right) \mathbf{c}_{k,z+1} + \text{Poi}(z+1; qt_k) \widehat{\mathbf{f}}_k(0).$$

Step 6: For $k = K+1, K, \ldots, 2, 1$, calculate H_k.

$$H_k = \frac{1}{q} \sum_{z=0}^{R_k} \mathbf{b}_{k,z} \mathbf{c}_{k,z}.$$

7.2.8 Initial values and termination condition

The choice of initial values and a stopping condition are significant problems with the EM algorithm [29]. In terms of QBD processes, there is no standard way to determine the initial values of the parameters in the EM algorithm. Hence, we start the EM algorithm with randomly generated initial values from a uniform distribution, $U(0, 1)$.

The stopping condition must be able to detect whether the EM algorithm has found the maximum and that no further improvements can be made to the parameter estimates and the log-likelihood of the observed data given the estimated parameters. A suitable stopping condition for the EM algorithm developed in this section is based on the relative difference between successive log-likelihood values of the observed data. That is, we terminate the EM algorithm according to the stopping condition,

$$\frac{E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right] - E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}\right]}{E_Y\left[l_C(\boldsymbol{\theta}|\mathbf{d})|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right]} < \epsilon,$$

where ϵ is some tolerance.

7.2.9 Convergence and identifiably issues

It is well known that the EM algorithm has a convergence property in that the algorithm converges to any local maximum or to a saddle point of a likelihood function [8, 15]. This does not guarantee that the EM algorithm converges to the global maximisation of the conditional likelihood function. Hence, the results of the EM algorithm may depend on the initial values of the parameters.

For quasi-birth-and-death processes, there is a non-uniqueness of representation inherited from the non-uniqueness of representation in phase-type distributions [1, 39]. That is, there may be several sets of parameters $\boldsymbol{\theta}$ resulting in stochastically equivalent level processes. Therefore, the log-likelihood may have infinitely many maxima of similar magnitude. However, this does not present itself as a problem in this work as we are more interested in the behaviour of the QBD process than the specific values of the parameters.

7.2.10 Finite case

The EM algorithm for level-dependent finite QBD processes with a capacity of C customers is similar to the EM algorithm for infinite level-dependent QBD processes, where we truncate at level C such that $A_+^{(C)} = 0$ and $A_-^{(\ell)} = 0$, $A_0^{(\ell)} = 0$, and $A_+^{(\ell)} = 0$ for $\ell \geq C + 1$.

7.3 EM algorithm for infinite level-independent QBD processes

In this section, we extend the EM algorithm to estimate the parameters for a level-independent QBD process under the assumption that the phase process is unobservable. Since much of the method will remain the same for level-independent QBD processes, we only draw attention to the differences in the methods.

7.3.1 Observed data

Similar to Section 7.2, suppose we continuously observe the input and output of a queueing system over some finite time period $[0, T]$. As before, let $X = (\mathbf{T}, \mathbf{L})$ denote the observed data.

7.3.2 Unobserved data

Since the phase process of the QBD process is unobservable, we have incomplete data $Y = (\mathbf{Z}_{\ell=0}, S_{\ell=0}, \mathbf{Z}_{\ell\geq 1}, S_{\ell\geq 1}, D, U)$ defined as follows.

Given that the phase transition behaviour for level 0 is different for that in any other level, we carefully define the sufficient statistics relating to time and the transitions within each level as,

- $Z^{(k)}_{\ell=0,i}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in level 0, for $1 \leq i \leq m$,
- $S^{(k)}_{\ell=0,i,j}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in level 0, for $1 \leq i, j \leq m$ and $i \neq j$,
- $Z^{(k)}_{\ell\geq 1,i}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in any level $\ell \geq 1$, for $1 \leq i \leq m$ and
- $S^{(k)}_{\ell\geq 1,i,j}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in any level $\ell \geq 1$, for $1 \leq i, j \leq m$, and $i \neq j$,

where,

$$[Z_\ell = 0]_i = Z_{\ell=0,i} = \sum_{k=1}^{K} Z^{(k)}_{\ell=0,i} \quad 1 \leq i \leq m,$$

$$[S_\ell = 0]_{i,j} = S_{\ell=0,i,j} = \sum_{k=1}^{K} S^{(k)}_{\ell=0,i,j} \quad 1 \leq i, j \leq m, i \neq j,$$

$$[Z_\ell \geq 1]_i = Z_{\ell\geq 1,i} = \sum_{k=1}^{K} Z^{(k)}_{\ell\geq 1,i} \quad 1 \leq i \leq m, \text{ and}$$

$$[S_\ell \geq 1]_{i,j} = S_{\ell\geq 1,i,j} = \sum_{k=1}^{K} S^{(k)}_{\ell\geq 1,i,j} \quad 1 \leq i, j \leq m, i \neq j.$$

The rates describing a change in level are independent of the level. Hence, we have that

- $D^{(k)}_{i,j}$ is the indicator random variable for the event that the transition at time s_k is from any level $\ell \geq 1$ to level $\ell - 1$ and leads to a phase transition from i to j, for $1 \leq i, j \leq m$, and
- $U^{(k)}_{i,j}$ is the indicator random variable for the event that the transition at time s_k is from any level $\ell \geq 0$ to level $\ell + 1$ and leads to a phase transition from i to j, for $1 \leq i, j \leq m$,

where,

$$[D]_{i,j} = D_{i,j} = \sum_{k=1}^{K} D^{(k)}_{i,j} \quad 1 \leq i, j \leq m, \text{ and,}$$

$$[U]_{i,j} = U_{i,j} = \sum_{k=1}^{K} U^{(k)}_{i,j} \quad 1 \leq i, j \leq m.$$

Hence, the sufficient statistic for the complete data $\mathbf{D}$ is the collection of random variables $Y = (\mathbf{Z}_{\ell=0}, S_{\ell=0}, \mathbf{Z}_{\ell\geq 1}, S_{\ell\geq 1}, D, U)$.

7.3.3 Likelihood

Similar to Section 7.2, the likelihood of the complete data $\mathbf{D} = (X, Y)$ (*i.e.* if the phase process were observable) with parameters $\boldsymbol{\theta} = (A_-, B_0, A_0, A_+)$, is

$$L_C(\boldsymbol{\theta}|\mathbf{d}) = \prod_{i=1}^{m} \exp(B_{0_{i,i}} Z_{\ell=0,i}) \prod_{i=1}^{m} \exp(A_{0_{i,i}} Z_{\ell\geq 1,i}) \prod_{i=1}^{m} \prod_{\substack{j=1 \\ j\neq i}}^{m} {B_{0_{i,j}}}^{S_{\ell=0,i,j}} \prod_{i=1}^{m} \prod_{\substack{j=1 \\ j\neq i}}^{m} {A_{0_{i,j}}}^{S_{\ell\geq 1,i,j}}$$
$$\prod_{i=1}^{m} \prod_{j=1}^{m} {A_{-_{i,j}}}^{D_{i,j}} \prod_{i=1}^{m} \prod_{j=1}^{m} {A_{+_{i,j}}}^{U_{i,j}}.$$

Then the complete data log-likelihood is defined as

$$\ell_C(\boldsymbol{\theta}|\mathbf{d}) = \sum_{i=1}^{m} B_{0_{i,i}} Z_{\ell=0,i} + \sum_{i=1}^{m} A_{0_{i,i}} Z_{\ell\geq 1,i} + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} \log\left(B_{0_{i,j}}\right) S_{\ell=0,i,j} +$$
$$\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} \log\left(A_{0_{i,j}}\right) S_{\ell\geq 1,i,j} + \sum_{i=1}^{m} \sum_{j=1}^{m} \log\left(A_{-_{i,j}}\right) D_{i,j} +$$
$$\sum_{i=1}^{m} \sum_{j=1}^{m} \log\left(A_{+_{i,j}}\right) U_{i,j}.$$

Using the complete data log-likelihood we find that the entries of the maximum likelihood estimates $\widehat{\boldsymbol{\theta}} = (\widehat{A_-}, \widehat{B_0}, \widehat{A_0}, \widehat{A_+})$, are defined as

$$\widehat{B_0}_{i,j} = \frac{S_{\ell=0,i,j}}{Z_{\ell=0,i}}, \text{ for } 1 \leq i \neq j \leq m, \tag{7.3.1}$$

$$\widehat{A_0}_{i,j} = \frac{S_{\ell\geq 1,i,j}}{Z_{\ell\geq 1,i}}, \text{ for } 1 \leq i \neq j \leq m, \tag{7.3.2}$$

$$\widehat{A_-}_{i,j} = \frac{D_{i,j}}{Z_{\ell\geq 1,i}}, \text{ for } 1 \leq i, j \leq m, \text{ and} \tag{7.3.3}$$

$$\widehat{A_+}_{i,j} = \frac{U_{i,j}}{Z_{\ell\geq 0,i}}, \text{ for } 1 \leq i, j \leq m. \tag{7.3.4}$$

We now describe the E-step and M-step of the EM algorithm for level-independent QBD processes.

7.3.4 E-step

Using the forward and backward likelihoods described in Section 7.2.4, the conditional expectations used in the E-step of the EM algorithm are as follows.

If the process is in level $\ell = 0$ during the time interval $[s_{k-1}, s_k)$:

$$E\big[Z^{(k)}_{\ell=0,i}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i[\mathbf{b}_k(t_k-\tau)]_i d\tau}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i \le m, \text{ and}$$

$$E\big[S^{(k)}_{\ell=0,i,j}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i(B_{0i,j})[\mathbf{b}_k(t_k-\tau)]_j d\tau}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i \ne j \le m.$$

Here, $\bullet$ represents the dot product.

If the process is in any level $\ell \ge 1$ during the time interval $[s_{k-1}, s_k)$:

$$E\big[Z^{(k)}_{\ell\ge1,i}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i[\mathbf{b}_k(t_k-\tau)]_i d\tau}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i \le m.$$

$$E\big[S^{(k)}_{\ell\ge1,i,j}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{\int_0^{t_k}[\mathbf{f}_k(\tau)]_i(A_{0i,j})[\mathbf{b}_k(t_k-\tau)]_j d\tau}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i \ne j \le m.$$

If the process is moving from any level $\ell \ge 1$ to level $\ell - 1$ at the k^{th} event:

$$E\big[D^{(k)}_{i,j}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{[\mathbf{f}_k(t_k)]_i(A_{-i,j})[\mathbf{b}_{k+1}(t_{k+1})]_j}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i,j \le m.$$

If the process is moving from any level $\ell \ge 0$ to level $\ell + 1$ at the k^{th} event:

$$E\big[U^{(k)}_{i,j}|\mathbf{D},\boldsymbol{\theta}\big] = \frac{[\mathbf{f}_k(t_k)]_i(A_{+i,j})[\mathbf{b}_{k+1}(t_{k+1})]_j}{\boldsymbol{\alpha}\bullet\mathbf{b}_1(t_1)}, \text{ for } 1 \le i,j \le m.$$

7.3.5 M-step

In the M-step of the EM algorithm, we update the values of the parameters by replacing the statistics in Equations (7.3.1), (7.3.2), (7.3.3), (7.3.4) with the conditional expectations calculated in the E-step, as follows:

$$\widehat{B_0}_{i,j} = \frac{\sum_{k=1}^{K} E\big[S^{(k)}_{\ell=0,i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell=0,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) = 0 \text{ and } 1 \leq i \neq j \leq m,$$

$$\widehat{A_0}_{i,j} = \frac{\sum_{k=1}^{K} E\big[S^{(k)}_{\ell\geq 1,i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell\geq 1,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 1 \text{ and } 1 \leq i \neq j \leq m,$$

$$\widehat{A_-}_{i,j} = \frac{\sum_{k=1}^{K} E\big[D^{(k)}_{i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell\geq 1,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 1 \text{ and } 1 \leq i, j \leq m, \text{ and}$$

$$\widehat{A_+}_{i,j} = \frac{\sum_{k=1}^{K} E\big[U^{(k)}_{i,j}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z^{(k)}_{\ell\geq 0,i}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } \ell = N(s_{k-1}) \geq 0 \text{ and } 1 \leq i, j \leq m.$$

7.3.6 Finite case

The EM algorithm for finite level-independent QBD processes with a capacity of C customers is similar to the EM algorithm for infinite level-independent QBD processes expect for the following.

Given that the phase transition behaviour for levels 0 and C are different for that in any other level, we carefully define the sufficient statistics relating to time and the transitions within each level as,

- $Z^{(k)}_{\ell=0,i}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in level 0, $1 \leq i \leq m$,

- $S^{(k)}_{\ell=0,i,j}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in level 0, $1 \leq i, j \leq m$, and $i \neq j$,

- $Z^{(k)}_{\ell\geq 1,i}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in any level $\ell \geq 1$, $1 \leq i \leq m$,

- $S^{(k)}_{\ell\geq 1,i,j}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in any level $\ell \geq 1$, $1 \leq i, j \leq m$, and $i \neq j$,

- $Z^{(k)}_{\ell=C,i}$ is the total amount of time spent in phase i in the interval $[s_{k-1}, s_k)$ while in level C, $1 \leq i \leq m$, and

- $S^{(k)}_{\ell=C,i,j}$ is the number of transitions from phase i to phase j in the time interval $[s_{k-1}, s_k)$, while in level C, $1 \leq i, j \leq m$, and $i \neq j$,

where,

$$[Z_{\ell=0}]_i = Z_{\ell=0,i} = \sum_{k=1}^{K} Z^{(k)}_{\ell=0,i} \quad 1 \leq i \leq m,$$

$$[S_{\ell=0}]_{i,j} = S_{\ell=0,i,j} = \sum_{k=1}^{K} S^{(k)}_{\ell=0,i,j} \quad 1 \leq i, j \leq m, i \neq j,$$

$$[Z_{\ell\geq 1}]_i = Z_{\ell\geq 1,i} = \sum_{k=1}^{K} Z^{(k)}_{\ell\geq 1,i} \quad 1 \leq i \leq m,$$

$$[S_{\ell\geq 1}]_{i,j} = S_{\ell\geq 1,i,j} = \sum_{k=1}^{K} S^{(k)}_{\ell\geq 1,i,j} \quad 1 \leq i, j \leq m, i \neq j,$$

$$[Z_{\ell=C}]_i = Z_{\ell=C,i} = \sum_{k=1}^{K} Z^{(k)}_{\ell=C,i} \quad 1 \leq i \leq m, \text{ and}$$

$$[S_{\ell=C}]_{i,j} = S_{\ell=C,i,j} = \sum_{k=1}^{K} S^{(k)}_{\ell=C,i,j} \quad 1 \leq i, j \leq m, i \neq j.$$

We then define the maximum likelihood estimates of the parameters $\{B_0, A_0, C_0, A_-, A_+\}$ as

$$\widehat{B_0}_{i,j} = \frac{S_{\ell=0,i,j}}{Z_{\ell=0,i}}, \text{ for } 1 \leq i \neq j \leq m,$$

$$\widehat{A_0}_{i,j} = \frac{S_{1\leq\ell\leq C-1,i,j}}{Z_{1\leq\ell\leq C-1,i}}, \text{ for } 1 \leq i \neq j \leq m,$$

$$\widehat{C_0}_{i,j} = \frac{S_{\ell=C,i,j}}{Z_{\ell=C,i}}, \text{ for } 1 \leq i \neq j \leq m,$$

$$\widehat{A_-}_{i,j} = \frac{D_{i,j}}{Z_{1\leq\ell\leq C,i}}, \text{ for } 1 \leq i, j \leq m, \text{ and}$$

$$\widehat{A_+}_{i,j} = \frac{U_{i,j}}{Z_{0\leq\ell\leq C-1,i}}, \text{ for } 1 \leq i, j \leq m.$$

7.4 Numerical examples

In this section, we show how the EM algorithm for both level-dependent and level-independent QBD processes performs with two numerical examples. In each numerical example, we simulate data from a QBD process with known parameters and then use the simulated data as input to our EM algorithm, implemented using R [43].

We start the EM algorithm with randomly generated initial values from a uniform distribution, $U(0, 1)$ and we stop the EM algorithm once the relative difference between two successive log-likelihood values falls below 10^{-6}. We then compare the estimated QBD process to the true QBD process.

Given that there is a non-uniqueness of representation in QBD processes, we cannot simply compare the true parameters of the QBD process, $\boldsymbol{\theta}$, to the estimated parameters of the QBD process, $\widehat{\boldsymbol{\theta}}$. Instead, we choose to compare the stationary distributions and transient behaviours of each QBD process.

Additionally, the comparison of stationary distributions is by each level and not by phase due to the non-uniqueness of representation and interchangeability of phases for each estimated QBD process. Hence, we define the level-stationary distribution as $\widetilde{\boldsymbol{\pi}} = (\widetilde{\pi}_0, \widetilde{\pi}_1, \ldots)$, where $\widetilde{\pi}_n = \boldsymbol{\pi}_n \mathbf{e}$, for all $n \geq 0$. The comparison of transient behaviours, also by each level and not by phase, is by considering the conditional sojourn time spent in each level before changing level and the transition probabilities between levels.

For each type of behaviour, the mean squared error (MSE) is calculated to demonstrate the overall accuracy. Let Y_ℓ denote the value of the true QBD process corresponding to level ℓ and $\widehat{Y}_\ell^{(i)}$ denote the value of the i^{th} estimated QBD process corresponding to level ℓ, for $\ell = L_{min}, \ldots, L_{max}$ where L_{min} and L_{max} are the minimum and maximum levels associated with the calculation, respectively. Then the mean squared error for each estimated structured QBD process, i, is calculated as

$$MSE = \frac{1}{L_{max} - L_{min} + 1} \sum_{\ell = L_{min}}^{L_{max}} \left(Y_\ell - \widehat{Y}_\ell^{(i)} \right)^2 ,$$

For example, Y_ℓ could denote the upward transition probability for level ℓ of the true QBD process and $\widehat{Y}_\ell^{(i)}$ denote the estimated upward transition probability for level ℓ of the i^{th} estimated QBD process.

Lastly, we also comment on the sensitivity of the EM algorithm by analysing the log-likelihood of the observed data given the estimated parameters for various initial starting values.

7.4.1 Level-dependent QBD process

In this example, we consider a finite level-dependent QBD process of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+ & 0 & \dots & 0 & 0 & 0 \\ \frac{1}{2}A_- & A_0^{(1)} & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_- & A_0^{(2)} & \dots & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2}A_- & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 4A_- & A_0^{(8)} & A_+ \\ 0 & 0 & 0 & \dots & 0 & \frac{9}{2}A_- & A_0^{(9)} \end{bmatrix},$$

where

$$A_+ = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}, \quad A_0^{(\ell)} = \begin{pmatrix} * & 2 \\ 1 & * \end{pmatrix}, \quad A_- = \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix},$$

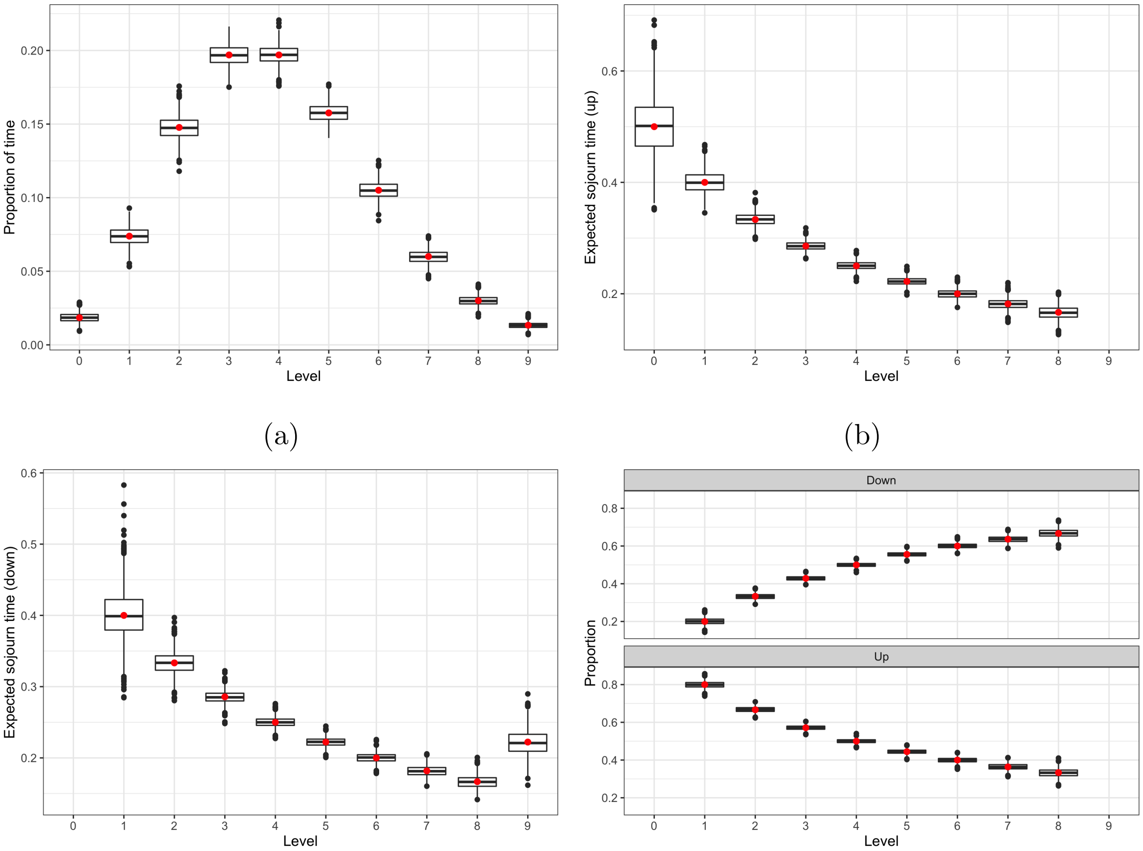
and the diagonal entries of $A_0^{(\ell)}$ for $0 \leq \ell \leq 9$ are the negative of the relevant row sums.

We simulated 1000 samples of the above level-dependent QBD process, each with 10,000 changes in level, and used the EM algorithm described in Section 7.2 to estimate the parameters of the level-dependent QBD process for each sample. We then calculated the level-stationary distribution, conditional sojourn times in levels, and the transition probabilities between levels of each estimated level-dependent QBD process and compared the values to that of the true level-dependent QBD process.

The stationary behaviour of the level-dependent QBD process is captured as illustrated in Figure 7.4.1a, which compares the estimated level-stationary distribution with the true level-stationary distribution of the true level-dependent QBD process.

In terms of transient behaviour, we first look at the conditional sojourn times. We considered both the conditional sojourn time within a level before the process moves up a level, as well as the conditional sojourn time within a level before the process moves down a level. In both situations, there is a strong similarity between the expected conditional sojourn times for each level of each estimated level-dependent QBD processes and the expected conditional sojourn times for each level of the true level-dependent QBD process, as illustrated in Figures 7.4.1b and 7.4.1c.

Secondly, we consider transitions between levels. Here, we considered the transition probability of moving up from each level and the transition probability of moving down from each level. Figure 7.4.1d illustrates the strong similarity between the transition probabilities to the level above and below of each estimated level-dependent QBD process and the transition probabilities to the level above and below of the true level-dependent QBD process. These results are also confirmed by the small MSE values summarised in Table 7.4.1.



(a)

(b)

(c)

(d)

Figure 7.4.1: Comparison of the behaviour of the 1000 estimated level-dependent QBD processes to the behaviour of the true level-dependent QBD process. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving up from level 9 are 0, the sojourn times conditioned on moving down from level 0 are 0, and the transition probabilities for levels 0 and 9 are either 0 or 1.

Type	Mean	Variance
Stationary distribution	3.121×10^{-5}	5.601×10^{-10}
Sojourn time conditioned on moving up a level	3.337×10^{-4}	1.308×10^{-7}
Sojourn time conditioned on moving down a level	2.105×10^{-4}	5.601×10^{-8}
Transition probabilities	1.880×10^{-4}	1.134×10^{-8}

Table 7.4.1: Mean and variance of the MSE for each type of behaviour for the level-dependent QBD process.

As is well known about the EM algorithm, there is a dependence on the initial values of the parameters. To assess this dependence, we simulated a single sample of the above level-dependent QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. Figure 7.4.2 plots the log-likelihood of the observed data for each of the 1000 different level-dependent QBD process estimations.

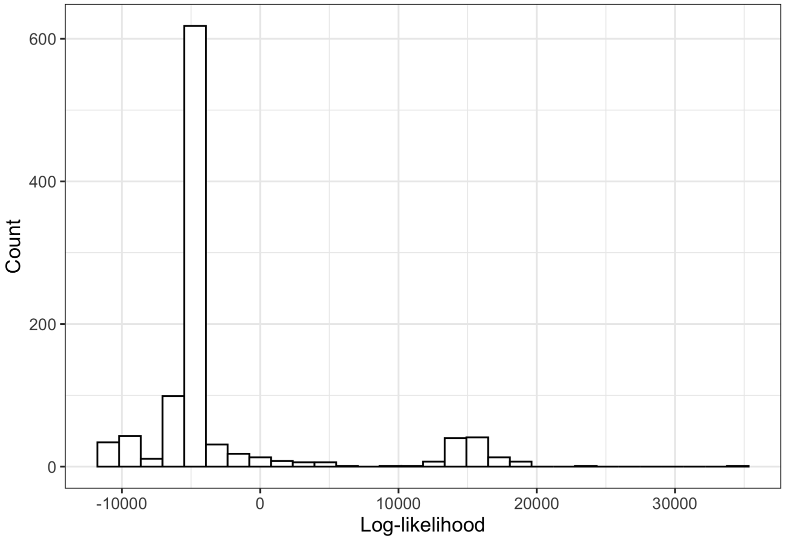
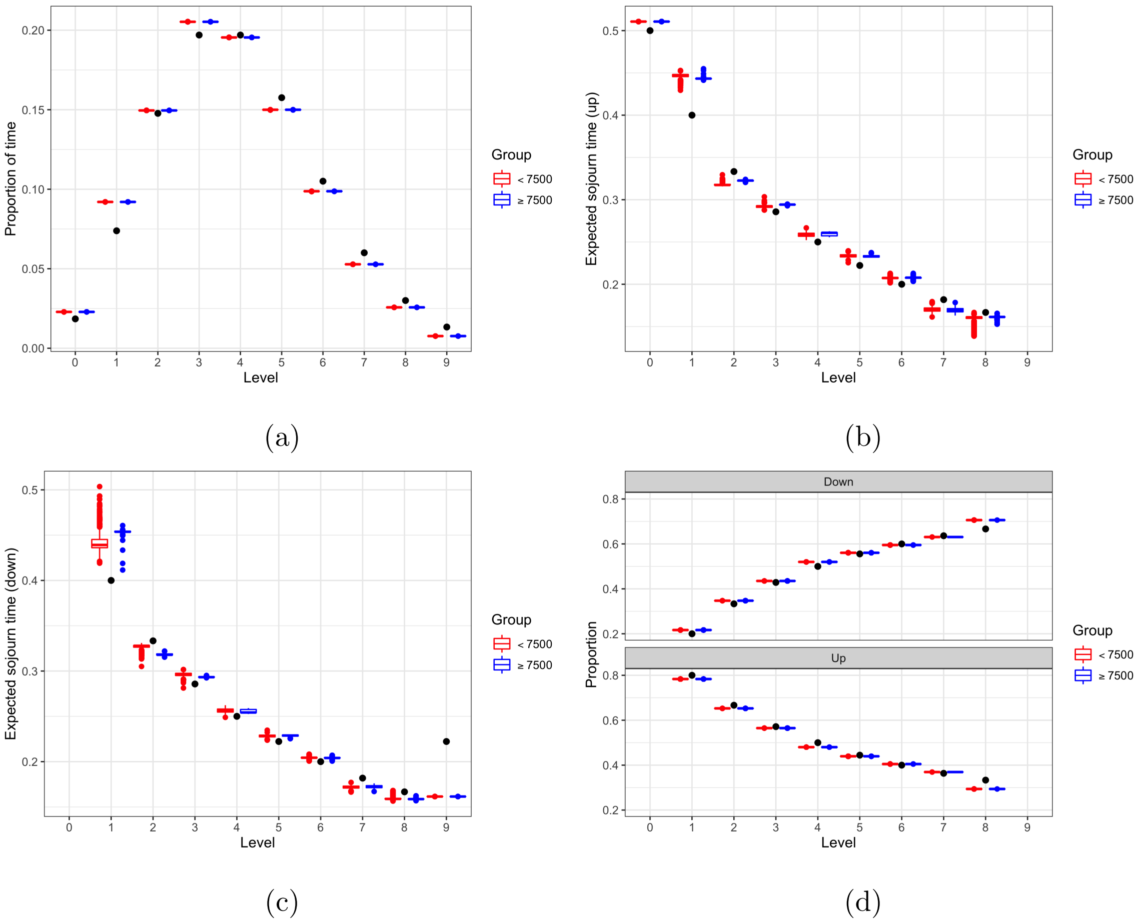


Figure 7.4.2: Log-likelihood values of the observed data for each of the fitted level-dependent QBD processes from the 1000 different starting values.

Notice the bimodal behaviour in the log-likelihood of the observed data for each of the 1000 different level-dependent QBD process estimations. To investigate the cause of the bimodal behaviour observed in Figure 7.4.2, we compared the behaviour of the estimated level-dependent QBD process with log-likelihood values less than 7500 to the estimated level-dependent QBD processes with log-likelihood values greater than or equal to 7500. Looking at Figure 7.4.3, we can see that there is little difference between the groups in terms of stationary and transient behaviour. Note that we expect to see a difference between the true and estimated behaviours as the 1000 estimations are based on a single sample of data.



(a)

(b)

(c)

(d)

Figure 7.4.3: Comparison of the behaviour of the estimated level-dependent QBD processes with log-likelihood values less than 7500 to the behaviour of the estimated level-dependent QBD process with log-likelihood values greater than or equal to 7500. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (black points). (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (black points). (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (black points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (black points). By design, the sojourn times conditioned on moving down from level 0 are 0, the sojourn times conditioned on moving up from level 9 are 0, and the transition probabilities for levels 0 and 9 are either 0 or 1.

7.4.2 Level-independent QBD process

In this example, we consider an infinite level-independent QBD process of the form

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \dots \\ A_- & A_0 & A_+ & 0 & \dots \\ 0 & A_- & A_0 & A_+ & \dots \\ 0 & 0 & A_- & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$A_+ = \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix}, \quad B_0 = \begin{pmatrix} -3 & 1 \\ 1 & -2 \end{pmatrix}, \quad A_0 = \begin{pmatrix} -4 & 0.5 \\ 0.5 & -3.5 \end{pmatrix}, \quad A_- = \begin{pmatrix} 0 & 1.5 \\ 0 & 2 \end{pmatrix}.$$

In this example, we simulated 1000 samples of the above level-independent QBD process, each with 10,000 changes in level, and used the EM algorithm described in Section 7.3 to estimate the parameters of the level-independent QBD process for each sample. We then calculated the level-stationary distribution, conditional sojourn times in levels, and the transition probabilities between levels of each estimated level-independent QBD process and compared the values to that of the true level-independent QBD process.

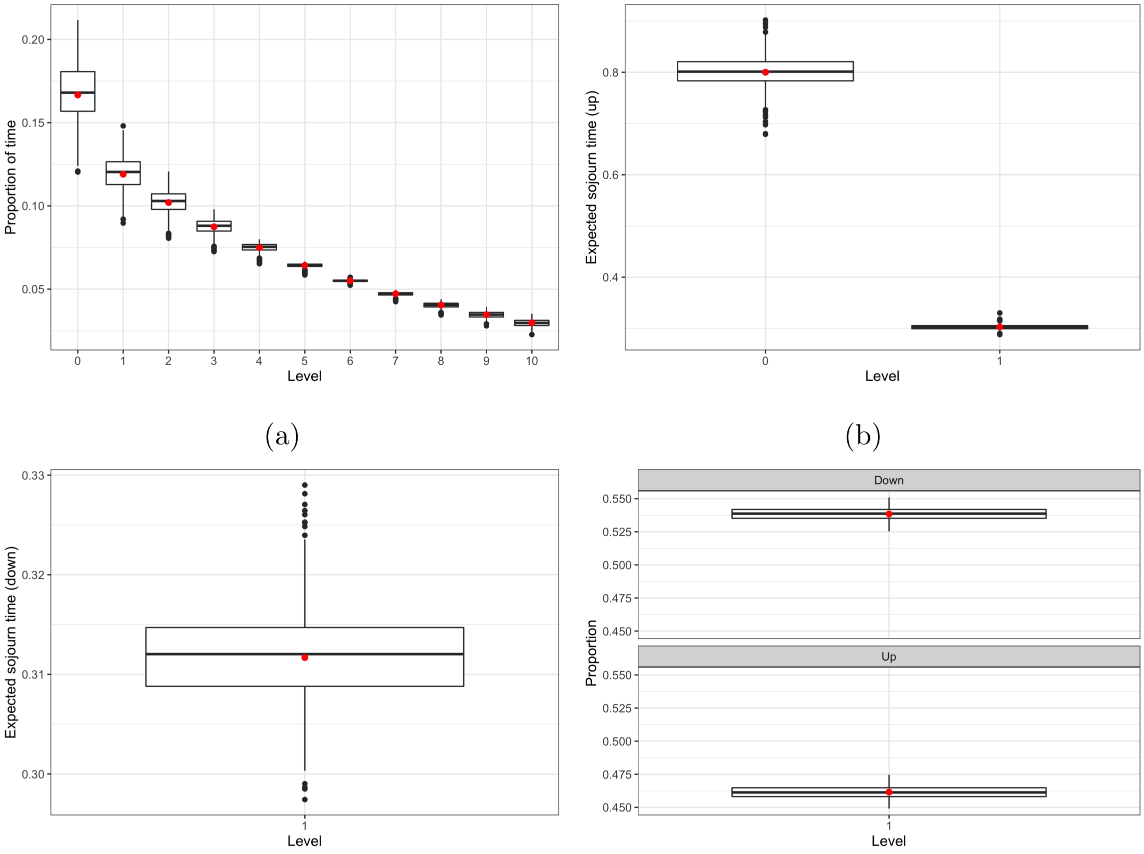
Firstly, the stationary behaviour of the level-independent QBD process is captured as illustrated in Figure 7.4.4a, which compares the estimated level-stationary distribution for the first 11 levels with the true level-stationary distribution of the true level-independent QBD process.

Secondly, we consider the both the conditional sojourn time within a level before the process moves up a level, as well as the conditional sojourn time within a level before the process moves down a level. Note that given the level-independent nature of the QBD process, behaviour above level 1 is omitted in the figures. In both cases, there is strong similarity between the expected conditional sojourn times for each level of each estimated level-independent QBD processes and the expected conditional sojourn times for each level of the true level-independent QBD process, as illustrated in Figures 7.4.4b and 7.4.4c.

Lastly, there is strong similarity between the transition probabilities to the level above and below of each estimated level-independent QBD process and the transition probabilities to the level above and below of the true level-independent QBD process, as illustrated in Figure 7.4.4d. These results are also demonstrated by the small MSE values summarised in Table 7.4.2. Similar to earlier, we omit the transition probabilities above level 1 due to the level-independent nature of the QBD process.

Type	Mean	Variance
Stationary distribution	1.411×10^{-5}	3.409×10^{-10}
Sojourn time conditioned on moving up a level	4.204×10^{-5}	2.673×10^{-9}
Sojourn time conditioned on moving down a level	3.082×10^{-3}	4.755×10^{-6}
Transition probabilities	5.928×10^{-3}	1.366×10^{-6}

Table 7.4.2: Mean and variance of the MSE for each type of behaviour for the level-independent QBD process.



(a)

(b)

(c)

(d)

Figure 7.4.4: Comparison of the behaviour of the 1000 estimated level-independent QBD processes to the behaviour of the true level-independent QBD process for the first 11 levels. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving down from level 0 are 0 and the transition probabilities for level 0 are either 0 or 1. Also note that transient behaviour above level 1 is omitted due to the level-independent nature of the QBD process.

Similar to before, we assessed the dependence on the initial values of the parameters by simulating a single sample of the above level-independent QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. As mentioned previously, this EM algorithm tends to converge to local maxima rather than the global maximum as illustrated in Figure 7.4.5, which plots the log-likelihood of the observed data for each of the 1000 different level-independent QBD process estimations.

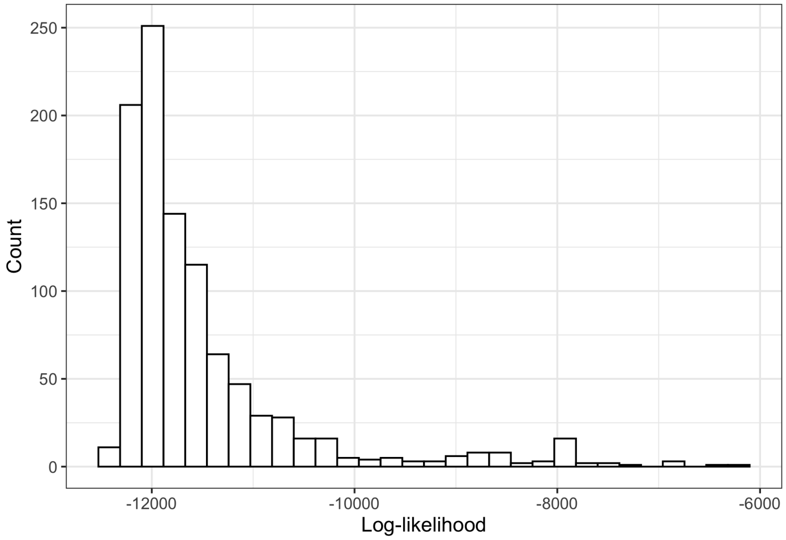


Figure 7.4.5: Log-likelihood values of the observed data for each of the fitted level-independent QBD processes from the 1000 different starting values.

7.5 Summary

Despite there being a non-uniqueness of representation of QBD processes, we have developed an EM algorithm for the model fitting of quasi-birth-and-death processes and demonstrated the accuracy of our method by means of two numerical examples. However, caution is needed when using level-dependent QBD processes to make predictions about the future state of a queueing system as there could be a risk of over-parameterisation. In the next chapter, we define a new class of QBD processes which remain level-dependent in nature but with fewer parameters than that of a level-dependent QBD process.

Chapter 8

Structured quasi-birth-and-death processes

Inherent with fitting QBD processes to queueing system data is the risk of fitting an over-parameterised model and hence being unable to use the fitted QBD process to predict behaviours of the queueing system. This is particularly true when fitting level-dependent QBD processes to queueing system data. While level-independent QBD processes have fewer parameters than a level-dependent QBD process, some aspects of the queueing process may not be explained by a level-independent QBD process. Alternatively, information about the behaviour of the transition rates of a fitted level-dependent QBD process can be used to reduce the number of parameters in the model, thus forming a structured QBD process which reflects the behaviour and properties of the queueing process.

In this chapter, we first motivate the development of structured QBD processes and then formally define various forms of structured QBD processes. Finally, we extend the statistical model fitting method developed in Chapter 7 to structured QBD processes.

8.1 Motivation

Consider a queueing system where we know that customers arrive according to a Poisson process with parameter λ, there are c servers, and the service time of each customer follows a phase-type distribution, $PH(\alpha, T)$ with b phases. This type of queueing system is an example of an $M/PH/c$ queue, where level $\ell \geq 1$ in the $M/PH/c$ queueing model has ℓbc phases. However, the dimensionality of such a model increases dramatically as the number of servers and phases increases, thus limiting the practicality of using $M/PH/\bullet$ queueing models to model real-life systems.

Now suppose we simply observe a queueing system, such that all we know is when customers arrive and leave and have no or limited information about the design of the queueing system. In this case, directly modelling the unknown structures of the queueing system using a classical queueing approach is impossible. Instead, we look towards a statistical approach to indirectly model the unknown structures using quasi-birth-and-death processes as these models have the flexibility to model the non-exponential distributions of sojourn times for each level and any dependence structure between the arrival process and the service times.

Recall that the infinitesimal generator matrix for the level-dependent QBD process is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

In terms of queueing system behaviour, $A_-^{(\ell)}$, $A_0^{(\ell)}$, and $A_+^{(\ell)}$ each play a role in modelling the arrival and departure process of a queueing system. However, $A_+^{(\ell)}$ predominately describes the arrival rate when there are $\ell \geq 0$ customers in the queueing system, whereas $A_-^{(\ell)}$ predominately describes the departure rate when there are $\ell \geq 1$ customers in the queueing system.

In some queueing systems, the departure rate depends on the number of occupied servers or the number of customers waiting in the queue, whereas the arrival rate may be independent of the number of customers in the queueing system. In terms of a birth-and-death process, such a queueing system may be represented as

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \dots \\ 0 & 2\mu & -(2\mu+\lambda) & \lambda & \dots \\ 0 & 0 & 3\mu & -(3\mu+\lambda) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

However, the departure rate need not be solely dependent on the number of customers in the queueing system. For example, the departure rate may also be effected by an external source, such that

$$A_-^{(\ell)} = a + \ell\mu, \quad \ell \geq 1,$$

where a represents the departure rate due to an external source and $\ell\mu$ represents the departure rate due to direct service of customers in the queueing system.

In the context of a birth-and-death process, such models are known as linear birth-and-death processes [28]. In these types of birth-and-death processes the arrival rate may also change at a linear rate, such that

$$A_+^{(\ell)} = b + \ell\lambda, \quad \ell \geq 0.$$

Such descriptions of $A_{-}^{(\ell)}$ and $A_{+}^{(\ell)}$ are well studied, thus providing various ways to model queueing systems by incorporating structure into the birth-and-death process. However, such descriptions do not naturally translate to QBD processes.

In this chapter, we look to use QBD processes to model queueing systems in such a way that we avoid over-parameterisation while maintaining the level-dependent nature of a process where needed. However, we assume that we have no or limited information about the structure of the queueing system including the definition of a server, the service process, and the service time distribution. As a result, a meaningful phase space of a QBD process becomes difficult to pre-determine.

Instead, we consider a statistical approach to modelling queueing systems using QBD processes that indirectly model the unknown structures of a queueing system. This approach will also allow for dependence between the arrival process and service time distribution, something which is not attainable with standard queueing models such as the $M/PH/\bullet$ queueing model.

For example, consider the application of modelling the bed occupancy of the Royal Adelaide Hospital intensive care unit. In the context of a queueing process, the definition of a server in an ICU is not clear as it could be classified as a bed, staff, or medical equipment. In addition to this, studies have found evidence of a dependence between the patient admission process and the distribution of length of stay [47]. Hence, the phase structure of any QBD process fitted to such data is complex.

Recall that the analysis of the RAH ICU data set in Section 5.1 revealed that the minimum observed bed occupancy was 20 and the maximum observed bed occupancy was 45, as illustrated in Figure 8.1.1. Hence, any QBD process fitted to the RAH ICU data set will be finite where levels 0 and 25 correspond to bed occupancies of 20 and 45 respectively.

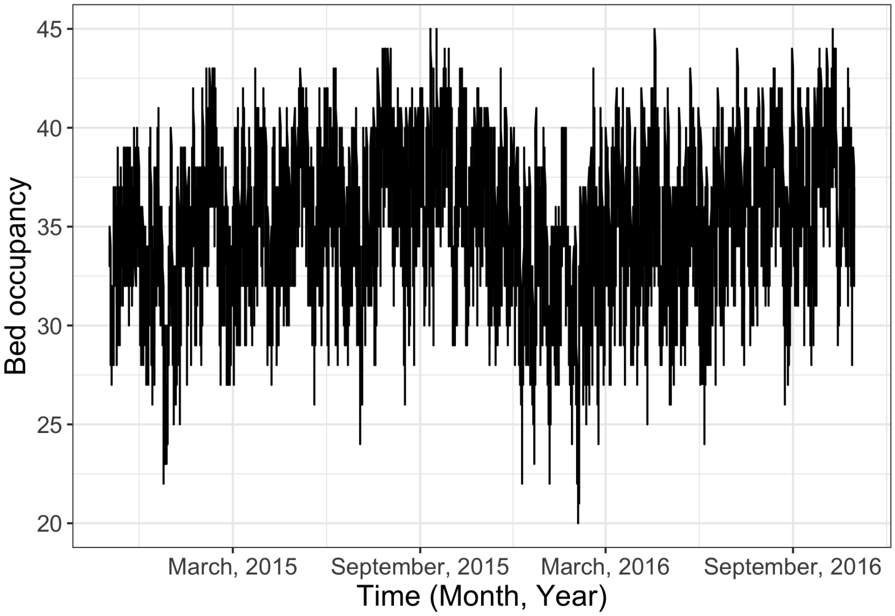


Figure 8.1.1: Time series plot of the bed occupancy in the RAH ICU.

Using the EM algorithm developed in Chapter 7, we consider a level-dependent QBD process with one phase for the RAH ICU, where the maximum likelihood estimates,

$$\widehat{\boldsymbol{\theta}} = (\{\{\widehat{A_-}^{(\ell)}; 1 \le \ell \le 25\}, \widehat{A_0}^{(\ell)}; 0 \le \ell \le 25\}, \{\widehat{A_+}^{(\ell)}; 0 \le \ell \le 24\}$$

plotted in Figure 8.1.2 show that the fitted level-dependent QBD process with one phase for the RAH ICU had a decreasing admission rate and a relatively constant discharge rate as the bed occupancy increases.

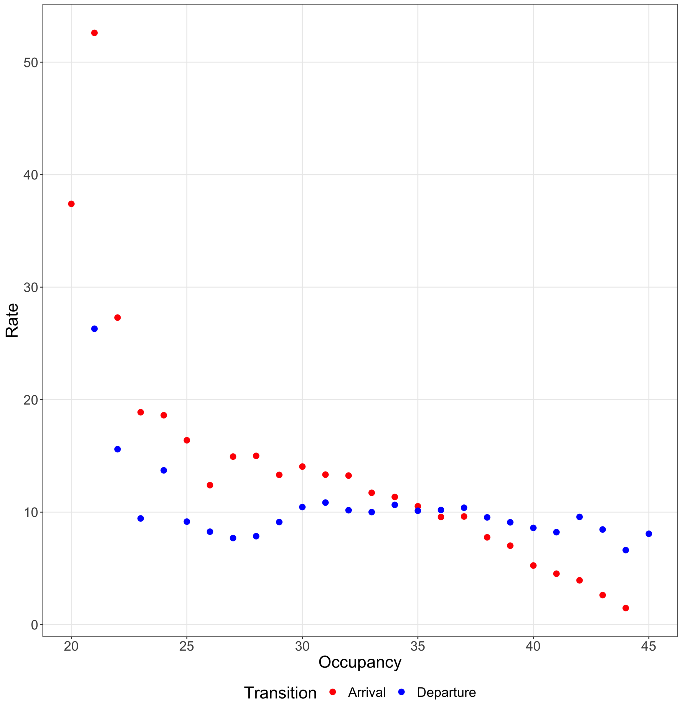


Figure 8.1.2: Maximum likelihood estimates for the level-dependent QBD process with one phase fitted to the RAH ICU data set.

While we cannot determine exactly what the servers are in this intensive care unit, we can indirectly model the unknown structures of the RAH ICU by considering the behaviour of the transition rates of the fitted level-dependent QBD process with one phase.

Using weighted least-squares regression, where the weights are the normalised expected number of times the RAH ICU is at each bed occupancy, we fitted linear regression models to the estimated arrival and departure rates of a level-dependent QBD process with one phase.

A linear term is required to explain the behaviour of the arrival rates, such that

$$\widehat{f}(\ell) = \begin{cases} 20.051 - 0.724\ell, & \text{for } 0 \leq \ell \leq 24, \\ 0, & \text{for } \ell = 25, \end{cases}$$

whereas, only a constant term is needed to explain the behaviour of the departure rates, such that

$$\widehat{h}(\ell) = \begin{cases} 8.914, & \text{for } 1 \leq \ell \leq 25, \\ 0, & \text{for } \ell = 0, \end{cases}$$

as illustrated in Figure 8.1.3.

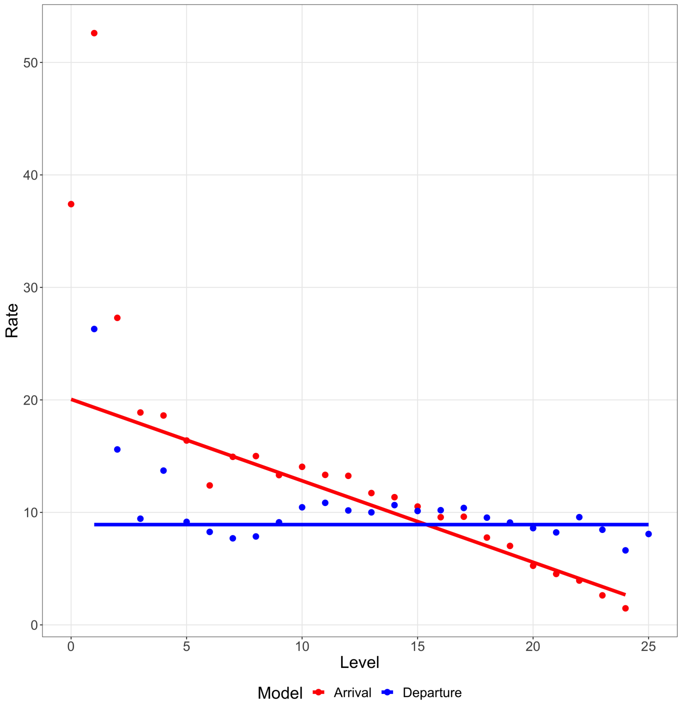


Figure 8.1.3: Fitted weighted least-squares regression models plotted against the estimated arrival and departure rates for a level-dependent QBD process with one phase for the RAH ICU.

Therefore, a possible birth-and-death process which incorporates the structural behaviour of the admission and discharge rates of the RAH ICU could have transition rates defined as

$$\widehat{A_-}^{(\ell)} = 8.914, \qquad \text{for } 1 \leq \ell \leq 25,$$

$$\widehat{A_+}^{(\ell)} = 20.051 - 0.724\ell, \qquad \text{for } 0 \leq \ell \leq 24, \text{ and}$$

$$\widehat{A_0}^{(\ell)} = \begin{cases} -20.051, & \text{for } \ell = 0, \\ -\left(8.914 + (20.051 - 0.724\ell)\right), & \text{for } 1 \leq \ell \leq 24, \\ -8.914, & \text{for } \ell = 25. \end{cases}$$

Figure 8.1.4 shows that the expected proportion of time spent at each bed occupancy differs slightly to the observed proportion of time spent at each bed occupancy, which is due to a reduction in the degree of model saturation. In addition to this, we decrease the diversity of behaviour and thus reduce the dispersion of the stationary distribution by smoothing the rates across levels.

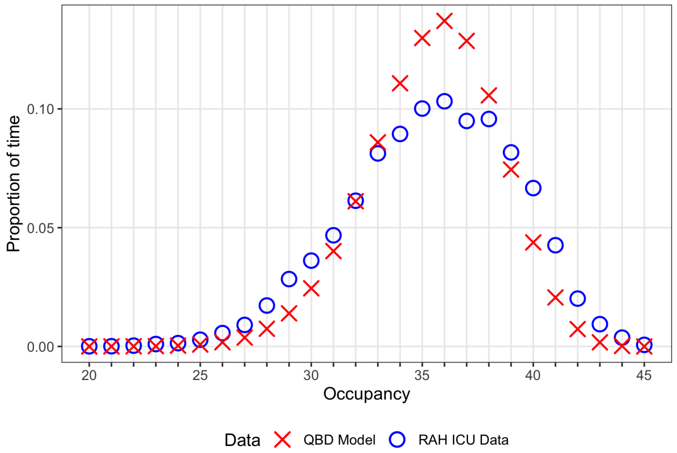


Figure 8.1.4: Observed proportion of time spent in each level (blue) versus the expected proportion of time spent in each level assuming a structured level-dependent QBD process with one phase (red).

Let's now increase the model flexibility by introducing a phase process but remain within the framework of scaled transition rates in order to model the level-dependent nature of the RAH ICU. That is, we now consider a QBD process with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$\left[A_{-}^{(\ell)}\right]_{i,j} = h_\ell \left[A_{-}\right]_{i,j}, \text{ for } \ell \geq 1,$$

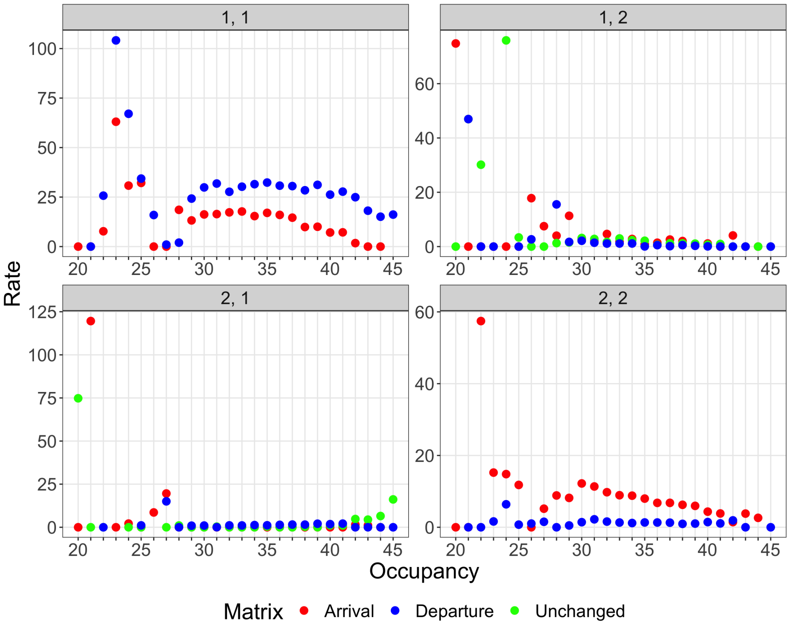
$$\left[A_{0}^{(\ell)}\right]_{i,j} = g_\ell \left[A_{0}\right]_{i,j}, \text{ for } \ell \geq 0, i \neq j,$$

$$\left[A_{+}^{(\ell)}\right]_{i,j} = f_\ell \left[A_{+}\right]_{i,j}, \text{ for } \ell \geq 0,$$

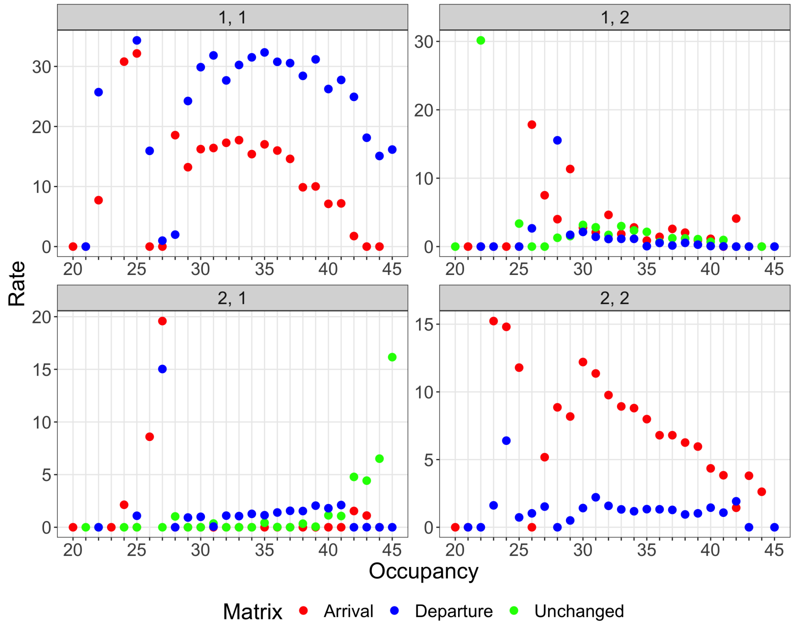
the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums, and f_ℓ, g_ℓ, and h_ℓ are non-negative scale functions of the level to ensure that the non-diagonal entries of the infinitesimal generator matrix remain non-negative.

Conceptually, this type of model replicates the notion of the linearly increasing departure rate of an $M/M/\infty$ queue (i.e. $h_\ell = \ell$ for $\ell \geq 1$ and $f_\ell = g_\ell = 1$ for $\ell \geq 0$) but it does not reflect the exponentially growing phase space of an $M/PH/\infty$ queueing model. In doing so, we introduce additional model flexibility for the purposes of statistically fitting QBD processes to data, while also limiting the dimension of the phase space and the overall number of parameters of the model.

Figure 8.1.5 plots the maximum likelihood estimates of the fitted level-dependent QBD process with two phases for the RAH ICU. Given that the dominating rate of the (1, 1) phase transition is the departure rate, we assume that phase 1 represents the discharge phase of the RAH ICU. Similarly, since the dominating rate of the (2, 2) phase transition is the arrival rate, we assume that phase 2 represents the admission phase of the RAH ICU.



(a)



(b)

Figure 8.1.5: Maximum likelihood estimates for the level-dependent QBD process with two phases fitted to the RAH ICU data set. Note that (b) zooms in on the rates presented in (a).

We can now reduce the number of parameters in the level-dependent QBD process by considering a likelihood of the form

$$L_C(\boldsymbol{\theta}|\mathbf{d}) = \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell} \exp(A_{0_{i,i}}^{(\ell)} Z_{\ell,i}) \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{\substack{j=1\\ j\neq i}}^{J_\ell} g_\ell A_{0_{i,j}}{}^{S_{\ell,i,j}}$$
$$\prod_{\ell=1}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell-1}} h_\ell A_{-_{i,j}}{}^{D_{\ell,i,j}} \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell+1}} f_\ell A_{+_{i,j}}{}^{U_{\ell,i,j}},$$

where g_ℓ, h_ℓ, and f_ℓ are non-negative polynomial forms of the level.

The degree of each polynomial form depends on the behaviour observed in the data. For example, the admission rates may decrease at a linear rate and the rates associated with discharges or no change in the number of patients in the ICU may remain relatively constant, thus suggesting that f_ℓ for $0 \leq \ell \leq 24$ is of degree 1, $h_\ell = 1$ for $1 \leq \ell \leq 25$, and $g_\ell = 1$ for $0 \leq \ell \leq 25$.

However, the behaviour of the arrival, departure, and within level rates are dependent on both the level and phase. For example, the (1, 1) and (2, 2) phase transitions appear to have a decreasing admission rate and a relatively constant discharge rate as the bed occupancy increases, whereas both the admission and discharge rates appear to be relatively constant for the (1, 2) and (2, 1) phase transitions.

Therefore, more structure may need to be incorporated into the QBD process, such that we consider a likelihood of the form

$$L_C(\boldsymbol{\theta}|\mathbf{d}) = \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell} \exp(A_{0_{i,i}}^{(\ell)} Z_{\ell,i}) \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{\substack{j=1\\ j\neq i}}^{J_\ell} g_{\ell,i,j} A_{0_{i,j}}{}^{S_{\ell,i,j}} \prod_{\ell=1}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell-1}} h_{\ell,i,j} A_{-_{i,j}}{}^{D_{\ell,i,j}} \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell+1}} f_{\ell,i,j} A_{+_{i,j}}{}^{U_{\ell,i,j}},$$

where $g_{\ell,i,j}$, $h_{\ell,i,j}$, and $f_{\ell,i,j}$ are polynomial forms of the level and of the phase transition process.

These types of QBD processes provide the motivation for the development of a new type of QBD process called the structured QBD process. Therefore, in the following sections we formally define various types of structured QBD processes and develop methodology to fit these types of structured QBD processes to queueing system data using the EM and ECM algorithms.

8.2 Structured QBD processes

A structured quasi-birth-and-death process is a continuous-time Markov process $\{X(t), J(t); t \geq 0\}$ with state space $\mathcal{S} = \{(\ell, j); i \geq 0, j = 1, 2, \ldots, J_\ell\}$, where ℓ denotes the level of the process, j denotes the phase of the process, and J_ℓ is the number of phases at the ℓ^{th} level. The infinitesimal generator matrix for the structured QBD process is the same as the infinitesimal generator matrix for the level-dependent QBD process,

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{8.2.1}$$

but with constraints defining the relationships between levels. For example, a QBD process whose infinitesimal generator matrix is of the form (8.2.1) is called a level-independent QBD process with a boundary at level 0 if

- $J_\ell = m$, for all $\ell \geq 1$,
- $A_-^{(\ell)} = A_-$, for all $\ell \geq 2$,
- $A_0^{(\ell)} = A_0$, for all $\ell \geq 1$, and
- $A_+^{(\ell)} = A_+$, for all $\ell \geq 1$.

So the infinitesimal generator matrix for a level-independent quasi-birth-and-death process with a boundary at level 0 is of the form

$$Q = \begin{bmatrix} B_0 & B_+ & 0 & 0 & \dots \\ B_- & A_0 & A_+ & 0 & \dots \\ 0 & A_- & A_0 & A_+ & \dots \\ 0 & 0 & A_- & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

If we set

- $J_i = m$, for all $1 \leq \ell \leq C - 1$,
- $A_-^{(\ell)} = A_-$, for all $2 \leq \ell \leq C - 1$,
- $A_0^{(\ell)} = A_0$, for all $1 \leq \ell \leq C - 1$,
- $A_+^{(\ell)} = A_+$, for all $1 \leq \ell \leq C - 2$, and
- $A_+^{(C)} = 0$,

we obtain the infinitesimal generator matrix for the finite level-independent QBD process with boundaries at levels 0 and C, which is of the form

$$Q = \begin{bmatrix} B_0 & B_+ & 0 & \dots & 0 & 0 & 0 \\ B_- & A_0 & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_- & A_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0 & A_+ & 0 \\ 0 & 0 & 0 & \dots & A_- & A_0 & C_+ \\ 0 & 0 & 0 & \dots & 0 & C_- & C_0 \end{bmatrix}.$$

Further to this, the relationships between the block matrices in the infinitesimal generator matrix presented in Equation (8.2.1) may resemble a functional form of the level.

To capture such relationships, we consider the infinitesimal generator matrix

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$\left[A_-^{(\ell)}\right]_{i,j} = h_\ell \left[A_-\right]_{i,j}, \text{ for } \ell \geq 1,$$

$$\left[A_0^{(\ell)}\right]_{i,j} = g_\ell \left[A_0\right]_{i,j}, \text{ for } \ell \geq 0, i \neq j,$$

$$\left[A_+^{(\ell)}\right]_{i,j} = f_\ell \left[A_+\right]_{i,j}, \text{ for } \ell \geq 0,$$

the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums, and f_ℓ, g_ℓ, and h_ℓ are non-negative scale functions of the level. For example, h_ℓ could be a linear function of the level, such that

$$h_\ell = \ell, \quad \ell \geq 1.$$

This can also be extended to situations where the relationships between the block matrices in the infinitesimal generator matrix presented in Equation (8.2.1) are explained by functional forms of the level and the phase transition process, such that the infinitesimal generator matrix is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$A_-^{(\ell)} = H_\ell \circ A_-, \text{ for } \ell \geq 1,$$

$$A_0^{(\ell)} = G_\ell \circ A_0, \text{ for } \ell \geq 0,$$

$$A_+^{(\ell)} = F_\ell \circ A_+, \text{ for } \ell \geq 0,$$

and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

The matrices F_ℓ, G_ℓ, and H_ℓ defined as

$$H_\ell = \begin{bmatrix} h_{\ell,1,1} & h_{\ell,1,2} & \dots & h_{\ell,1,J_{\ell-1}} \\ h_{\ell,2,1} & h_{\ell,1,2} & \dots & h_{\ell,1,J_{\ell-1}} \\ \vdots & \vdots & \dots & \vdots \\ h_{\ell,J_\ell,1} & h_{\ell,J_\ell,2} & \dots & h_{\ell,J_\ell,J_{\ell-1}} \end{bmatrix},$$

$$G_\ell = \begin{bmatrix} g_{\ell,1,1} & g_{\ell,1,2} & \cdots & g_{\ell,1,J_\ell} \\ g_{\ell,2,1} & g_{\ell,1,2} & \cdots & g_{\ell,1,J_\ell} \\ \vdots & \vdots & \dots & \vdots \\ g_{\ell,J_\ell,1} & g_{\ell,J_\ell,2} & \cdots & g_{\ell,J_\ell,J_\ell} \end{bmatrix},$$

and,

$$F_\ell = \begin{bmatrix} f_{\ell,1,1} & f_{\ell,1,2} & \cdots & f_{\ell,1,J_{\ell+1}} \\ f_{\ell,2,1} & f_{\ell,1,2} & \cdots & f_{\ell,1,J_{\ell+1}} \\ \vdots & \vdots & \dots & \vdots \\ f_{\ell,J_\ell,1} & f_{\ell,J_\ell,2} & \cdots & f_{\ell,J_\ell,J_{\ell+1}} \end{bmatrix},$$

are non-negative to ensure that the non-diagonal entries of the infinitesimal generator matrix remain non-negative. For example, $f_{\ell,i,j}$, for $1 \leq i \leq J_\ell$, $1 \leq j \leq J_{\ell+1}$, is some non-negative scalar function of level ℓ and the phase transition (i, j). While this increases the number of parameters in the structured QBD process, such a model may further uncover hidden features of a queueing system.

Note that $B \circ A$ indicates the Hadamard product of the matrices B and A. For example, the Hadamard product of 3×3 matrices B and A is

$$\begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} \circ \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} = \begin{bmatrix} b_{1,1}a_{1,1} & b_{1,2}a_{1,2} & b_{1,3}a_{1,3} \\ b_{2,1}a_{2,1} & b_{2,2}a_{2,2} & b_{2,3}a_{2,3} \\ b_{3,1}a_{3,1} & b_{3,2}a_{3,2} & b_{3,3}a_{3,3} \end{bmatrix}.$$

For each structured QBD process, we can identify a set of boundary levels and levels where the block matrices in the infinitesimal generator matrix presented in Equation (8.2.1) are related. We can also enforce the row sums of certain block matrices to be the same, if required. By this construction, we can consider QBD processes with specific boundary behaviours and related behaviour for levels between the boundaries.

Let L_-, U_-, L_0, U_0, L_+, and U_+ define the lower and upper boundaries for the block matrices in the generator matrix for a structured QBD process, such that

- the block matrices describing a decrease in level between the levels L_- and U_- are related,

- the block matrices describing no change in level between the levels L_0 and U_0 are related, and

- the block matrices describing an increase in level between the levels L_+ and U_+ are related.

For example, consider the generator matrix for an infinite level-independent QBD process with no unique lower boundary behaviour,

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \dots \\ A_- & A_0 & A_+ & 0 & \dots \\ 0 & A_- & A_0 & A_+ & \dots \\ 0 & 0 & A_- & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

In this case, $L_- = 1$, $L_0 = 1$, $L_+ = 0$, and $U_- = U_0 = U_+ = \infty$.

In the case of a finite level-independent QBD process with boundaries at level 0 and level C and generator matrix,

$$Q = \begin{bmatrix} B_0 & B_+ & 0 & \dots & 0 & 0 & 0 \\ B_- & A_0 & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_- & A_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0 & A_+ & 0 \\ 0 & 0 & 0 & \dots & A_- & A_0 & C_+ \\ 0 & 0 & 0 & \dots & 0 & C_- & C_0 \end{bmatrix}$$

we set $L_- = 2$, $L_0 = 1$, $L_+ = 1$, $U_- = C - 1$, $U_0 = C - 1$, and $U_+ = C - 2$.

Note that the dimensions of the matrices between the lower and upper boundaries in the generator matrix must be the same in order to define relationships of the matrices between the lower and upper boundaries.

8.2.1 Stationary and transient behaviour

The stationary and transient behaviour of a structured QBD process is defined similar to that of a level-dependent QBD process, since the structured QBD process can be level-dependent in nature.

8.2.2 Observed and unobserved data

For any QBD process, including a structured QBD process, we only observe the changes in level and the times at which a change in level occurs. The transitions between phases remain hidden, thus leading to incomplete data. In the following sections, we extend the EM algorithm developed in Chapter 7 to various types of structured QBD processes.

8.3 Structured QBD process with level-dependent scales

Firstly, we consider a structured QBD process with infinitesimal generator matrix

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A_{-_{i,j}}^{(\ell)} = h_\ell A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell-1},$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A^{(\ell)}_{0_{i,j}} = g_\ell A_{0_{i,j}}, \quad 1 \leq i \neq j \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A^{(\ell)}_{+_{i,j}} = f_\ell A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1},$$

the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums, and f_ℓ, g_ℓ, and h_ℓ are non-negative, level-dependent scales.

This form of structured QBD process has a level-dependent nature inherited from the scales but contains fewer parameters than that of a level-dependent QBD process, and hence could be a suitable alternative when predicting behaviours of a queueing system.

In this section, we first define the EM algorithm for a structured QBD process with level-dependent scales under the assumption that the phase process is unobservable. As this estimation method is an extension to the EM algorithm for level-dependent QBD processes developed in Chapter 7, we only describe the parts of the method that are different. We then focus on a special type of structured QBD process and define the associated EM algorithm in a similar way to before.

8.3.1 Likelihood

From Chapter 7, the likelihood of the complete data $\mathbf{D} = (X, Y)$ (*i.e.* if the phase process were observable) with parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$, is

$$L_C(\boldsymbol{\theta}|\mathbf{d}) = \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell} \exp(A_{0_{i,i}}^{(\ell)} Z_{\ell,i}) \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{\substack{j=1\\ j\neq i}}^{J_\ell} {A_{0_{i,j}}^{(\ell)}}^{S_{\ell,i,j}} \prod_{\ell=1}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell-1}} {A_{-_{i,j}}^{(\ell)}}^{D_{\ell,i,j}} \prod_{\ell=0}^{\infty}\prod_{i=1}^{J_\ell}\prod_{j=1}^{J_{\ell+1}} {A_{+_{i,j}}^{(\ell)}}^{U_{\ell,i,j}},$$

and the complete data log-likelihood is given by

$$\ell_C(\boldsymbol{\theta}|\mathbf{d}) = \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell} A_{0_{i,i}}^{(\ell)} Z_{\ell,i} + \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} \tag{8.3.1}$$

$$+ \sum_{\ell=1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}. \tag{8.3.2}$$

Expanding the complete data log-likelihood to show the relationships between the block matrices, we have that

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}|\mathbf{d}) = {} & \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell} A_{0_{i,i}}^{(\ell)} Z_{\ell,i} + \sum_{\ell=0}^{L_0-1}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} \\
& + \sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(g_\ell A_{0i,j}\right) S_{\ell,i,j} + \sum_{\ell=U_0+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} \\
& + \sum_{\ell=1}^{L_- -1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(h_\ell A_{-_{i,j}}\right) D_{\ell,i,j} \\
& + \sum_{\ell=U_-+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=0}^{L_+-1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j} \\
& + \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(f_\ell A_{+_{i,j}}\right) U_{\ell,i,j} + \sum_{\ell=U_++1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}.
\end{aligned}$$

Theorem 8.3.1.

Using the constrained log-likelihood, the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ *for a structured QBD process with level-dependent scales are as follows:*

$$\widehat{A}^{(\ell)}_{-_{i,j}} = \frac{D_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 1 \leq \ell \leq L_- - 1, U_- + 1 \leq \ell < \infty, \textit{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

$$\widehat{A}_{-_{i,j}} = \frac{\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\sum_{\ell=L_-}^{U_-} h_\ell Z_{\ell,i}}, \textit{ for } 1 \leq i \leq J_{L_-} \textit{ and } 1 \leq j \leq J_{L_- - 1},$$

$$\widehat{h_\ell} = \frac{\sum_{i=1}^{J_\ell} \sum_{j=1}^{J_{\ell-1}} D_{\ell,i,j}}{\sum_{i=1}^{J_\ell} \sum_{j=1}^{J_{\ell-1}} A_{-_{i,j}} Z_{\ell,i}}, \textit{ for } L_- \leq \ell \leq U_-,$$

$$\widehat{A}^{(\ell)}_{0_{i,j}} = \frac{S_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 0 \leq \ell \leq L_0 - 1, U_0 + 1 \leq \ell < \infty, \textit{ and } 1 \leq i \neq j \leq J_\ell,$$

$$\widehat{A}_{0_{i,j}} = \frac{\sum_{\ell=L_0}^{U_0} S_{\ell,i,j}}{\sum_{\ell=L_0}^{U_0} g_\ell Z_{\ell,i}}, \textit{ for } 1 \leq i \neq j \leq J_{L_0},$$

$$\widehat{g_\ell} = \frac{\sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j \neq i}}^{J_\ell} S_{\ell,i,j}}{\sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j \neq i}}^{J_\ell} A_{0_{i,j}} Z_{\ell,i}}, \textit{ for } L_0 \leq \ell \leq U_0,$$

$$\widehat{A}^{(\ell)}_{+_{i,j}} = \frac{U_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 0 \leq \ell \leq L_+ - 1, U_+ + 1 \leq \ell < \infty, \textit{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1},$$

$$\widehat{A}_{+_{i,j}} = \frac{\sum_{\ell=L_+}^{U_+} U_{\ell,i,j}}{\sum_{\ell=L_+}^{U_+} f_\ell Z_{\ell,i}}, \textit{ for } 1 \leq i \leq J_{L_+} \textit{ and } 1 \leq j \leq J_{L_+ + 1},$$

$$\widehat{f}_\ell = \frac{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} U_{\ell,i,j}}{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} A_{+_{i,j}} Z_{\ell,i}}, \textit{ for } L_+ \leq \ell \leq U_+.$$

Note: the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

Proof: See Appendix A.

8.3.2 E-step and CM-step

We use the ECM algorithm to estimate the parameters of the structured QBD process with level-dependent scales. The calculations involved in the E-step follow those described in Section 7.2.4. The calculations involved in the CM-step are as follows.

In the CM-step of the ECM algorithm, we update the values of the parameters by replacing the statistics in Theorem 8.3.1 with the conditional expectations calculated in the E-step.

At the $(t+1)^{th}$ iteration, the conditional maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ given $h_\ell = h_\ell^{(t)}$, $g_\ell = g_\ell^{(t)}$, and $f_\ell = f_\ell^{(t)}$ are as follows.

For $1 \leq \ell \leq L_- - 1$, $U_- + 1 \leq \ell < \infty$, $1 \leq i \leq J_\ell$, and $1 \leq j \leq J_{\ell-1}$:

$$\widehat{A_{-}}_{i,j}^{(\ell)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \le i \le J_\ell$ and $1 \le j \le J_{L_-}$:

$$\widehat{A_{-}}_{i,j} = \frac{\displaystyle\sum_{\ell=L_-}^{U_-} \sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{\ell=L_-}^{U_-} \widehat{h}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $0 \le \ell \le L_0 - 1$, $U_0 + 1 \le \ell < \infty$, and $1 \le i \ne j \le J_\ell$:

$$\widehat{A_{0}}_{i,j}^{(\ell)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \le i \ne j \le J_{L_0}$:

$$\widehat{A_{0}}_{i,j} = \frac{\displaystyle\sum_{\ell=L_0}^{U_0} \sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{\ell=L_0}^{U_0} \widehat{g}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $0 \le \ell \le L_+ - 1$, $U_+ + 1 \le \ell < \infty$, $1 \le i \le J_\ell$, and $1 \le j \le J_{\ell+1}$:

$$\widehat{A_{+}}_{i,j}^{(\ell)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \le i \le J_\ell$ and $1 \le j \le J_{L_+}$:

$$\widehat{A_{+}}_{i,j}^{=} \frac{\displaystyle\sum_{\ell=L_+}^{U_+} \sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{\ell=L_+}^{U_+} \widehat{f}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

Then, the conditional maximum likelihood estimates for $\{h_\ell; L_- \le \ell \le U_-\}$, $\{g_\ell; L_0 \le \ell \le U_0\}$, and $\{f_\ell; L_+ \le \ell \le U_+\}$ given $\widehat{A_-}_{i,j} = \widehat{A_-}_{i,j}^{(t+1)}$, $\widehat{A_0}_{i,j} = \widehat{A_0}_{i,j}^{(t+1)}$, and $\widehat{A_+}_{i,j} = \widehat{A_+}_{i,j}^{(t+1)}$ are

$$\widehat{h}_\ell = \frac{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\widehat{A_-}_{i,j}^{(t+1)}\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } L_- \le \ell \le U_-,$$

$$\widehat{g}_\ell = \frac{\displaystyle\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\widehat{A_0}_{i,j}^{(t+1)}\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } L_0 \le \ell \le U_0, \text{ and}$$

$$\widehat{f}_\ell = \frac{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\widehat{A_+}_{i,j}^{(t+1)}\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}, \text{ for } L_+ \le \ell \le U_+.$$

8.3.3 Special case

Now let's consider the special case of a finite level-independent QBD process with boundaries at level 0 and level C. To obtain the maximum likelihood estimators for the parameters in the infinitesimal generator matrix

$$\begin{bmatrix} B_0 & B_+ & 0 & \dots & 0 & 0 & 0 \\ B_- & A_0 & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_- & A_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0 & A_+ & 0 \\ 0 & 0 & 0 & \dots & A_- & A_0 & C_+ \\ 0 & 0 & 0 & \dots & 0 & C_- & C_0 \end{bmatrix},$$

we set $L_- = 2$, $L_0 = 1$, $L_+ = 1$, $U_- = C-1$, $U_0 = C-1$, $U_+ = C-2$, $h_\ell 1$ for all $1 \le \ell \le C$, and $g_\ell = f_\ell = 1$ for all $0 \le \ell \le C$.

Note that the rows sums of the block matrices B_- and A_- need to be the same. Hence, we must include the following Lagrangian term:

$$\sum_{i=1}^{J} \lambda_i^{(L)} \left(\sum_{j=1}^{J_{L_- -2}} A_{-_{i,j}}^{(L_- -1)} - \sum_{j=1}^{J_{L_- -1}} A_{-_{i,j}}^{(L_-)} \right).$$

Similar issues occur for the rows sums of the block matrices A_+ and C_+. Hence, we must include the following additional Lagrangian term:

$$\sum_{i=1}^{J} \lambda_i^{(U)} \left(\sum_{j=1}^{J_{U_+ +1}} A_{+_{i,j}}^{(U_+)} - \sum_{j=1}^{J_{U_+ +2}} A_{+_{i,j}}^{(U_+ +1)} \right).$$

Corollary 8.3.1.1. *The maximum likelihood estimates of the parameters*

$$\boldsymbol{\theta} = (B_-, B_0, B_+, A_-, A_0, A_+, C_-, C_0, C_+)$$

are identical to those presented in Theorem 8.3.1, except for the following:

$$\widehat{B}_{-_{i,j}} = \widehat{A}_{-_{i,j}}^{(1)} = \frac{D_{1,i,j}}{Z_{1,i} - \lambda_i^{(L)}}, \textit{ for } 1 \le i \le J_1, 1 \le j \le J_0,$$

and,

$$\widehat{A}_{-i,j} = \widehat{A}^{(2)}_{-i,j} = \frac{\sum_{\ell=2}^{C-1} D_{\ell,i,j}}{\sum_{\ell=2}^{C-1} Z_{\ell,i} + \lambda_i^{(L)}}, \textit{ for } 1 \leq i \leq J_2, 1 \leq j \leq J_1,$$

where,

$$\lambda_i^{(L)} = \frac{Z_{1,i} \sum_{j=1}^{J_1} \sum_{\ell=2}^{C-1} D_{\ell,i,j} - \left(\sum_{\ell=2}^{C-1} Z_{\ell,i} \right) \sum_{j=1}^{J_0} D_{1,i,j}}{\sum_{j=1}^{J_0} D_{1,i,j} + \sum_{j=1}^{J_1} \sum_{\ell=2}^{C-1} D_{\ell,i,j}},$$

$$\widehat{A}_{+i,j} = \widehat{A}^{(1)}_{+i,j} = \frac{\sum_{\ell=1}^{C-2} U_{\ell,i,j}}{\sum_{\ell=1}^{C-2} Z_{\ell,i} - \lambda_i^{(U)}}, \textit{ for } 1 \leq i \leq J_1, 1 \leq j \leq J_2,$$

and,

$$\widehat{C}_{+i,j} = \widehat{A}^{(C-1)}_{+i,j} = \frac{U_{C-1,i,j}}{Z_{C-1,i} + \lambda_i^{(U)}}, \textit{ for } 1 \leq i \leq J_{C-1}, 1 \leq j \leq J_C,$$

where,

$$\lambda_i^{(U)} = \frac{\left(\sum_{\ell=1}^{C-2} Z_{\ell,i} \right) \sum_{j=1}^{J_C} U_{C-1,i,j} - Z_{C-1,i} \sum_{j=1}^{J_2} \sum_{\ell=1}^{C-2} U_{\ell,i,j}}{\sum_{j=1}^{J_2} \sum_{\ell=1}^{C-2} U_{\ell,i,j} + \sum_{j=1}^{J_C} U_{C-1,i,j}}, \textit{ for } 1 \leq i \leq J_\ell,$$

and $\widehat{B}_0 = \widehat{A}_0^{(0)}$, $\widehat{A}_0 = \widehat{A}_0^{(1)}$, $\widehat{C}_0 = \widehat{A}_0^{(C)}$, $\widehat{C}_- = \widehat{A}_-^{(C)}$, $\widehat{B}_+ = \widehat{A}_+^{(0)}$.

Note: to obtain the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (B_0, B_-, B_+, A_0, A_-, A_+)$ *for an infinite level-independent QBD process with a boundary at level 0, set* $C = \infty$.

Proof: See Appendix B.

E-step and M-step

In this case, we use the standard EM algorithm to estimate the parameters of the finite level-independent QBD process with boundaries at level 0 and level C. The calculations involved in the E-step follow those described in Section 7.2.4. The calculations involved in the M-step are as follows.

In the M-step, we update the values of the parameters by replacing the statistics in Corollary 8.3.1.1 with the conditional expectations calculated in the E-step.

At the $(t+1)^{th}$ iteration, the conditional maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ are as follows.

For $1 \leq i \leq J_1$ and $1 \leq j \leq J_0$:

$$\widehat{A}_{-i,j}^{(1)} = \frac{\sum_{k=1}^{K} E\big[D_{1,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{1,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] - \lambda_i^{(L)}}.$$

For $1 \leq i \leq J_2$ and $1 \leq j \leq J_1$:

$$\widehat{A}_{-i,j} = \frac{\sum_{\ell=2}^{C-1}\sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] + \lambda_i^{(L)}}.$$

For $1 \leq i \leq J_C$ and $1 \leq j \leq J_{C-1}$:

$$\widehat{A}_{-i,j}^{(C)} = \frac{\sum_{k=1}^{K} E\big[D_{C,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{C,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] - \lambda_i^{(L)}}.$$

For $1 \leq i \neq j \leq J_0$:

$$\widehat{A}_{0i,j}^{(0)} = \frac{\sum_{k=1}^{K} E\big[S_{0,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{0,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}.$$

For $1 \le i \ne j \le J_1$:

$$\widehat{A_0}_{i,j} = \frac{\displaystyle\sum_{\ell=1}^{C-1}\sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}.$$

For $1 \le i \ne j \le J_C$:

$$\widehat{A_0}_{i,j}^{(C)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[S_{C,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{C,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}.$$

For $1 \le i \le J_0$, and $1 \le j \le J_1$:

$$\widehat{A_+}_{i,j}^{(0)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[U_{0,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{0,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] + \lambda_i^{(U)}}.$$

For $1 \le i \le J_1$ and $1 \le j \le J_2$:

$$\widehat{A_+}_{i,j} = \frac{\displaystyle\sum_{\ell=1}^{C-2}\sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{\ell=L_+}^{U_+}\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] - \lambda_i^{(U)}}.$$

For $1 \le i \le J_{C-1}$, and $1 \le j \le J_C$:

$$\widehat{A_+}_{i,j}^{(C-1)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[U_{C-1,i,j}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{C-1,i}^{(k)}|\mathbf{D},\boldsymbol{\theta}\big] + \lambda_i^{(U)}}.$$

8.3.4 Discussion

The structured QBD process with level-dependent scales remains level-dependent in nature but with fewer parameters than the standard level-dependent QBD process. This type of structured QBD process also allows us to define structural relationships between the levels to define special types of QBD processes, such as the finite level-independent QBD process. In the following sections, we structurally reduce the number of parameters again by considering functional forms of the level.

8.4 Structured QBD processes with linear functional forms

The arrival and departure rates of queueing systems sometimes increase or decrease at a linear rate, which resembles the nature of a M/M/c or a M/M/∞ queue. Using this as motivation, we now consider a structured QBD process with infinitesimal generator matrix

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \le \ell \le U_-$,

$$A_{-_{i,j}}^{(\ell)} = (1 + \beta_1^h(\ell - 1))A_{-_{i,j}}, \quad 1 \le i \le J_\ell,\ 1 \le j \le J_{\ell-1},$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A^{(\ell)}_{0_{i,j}} = (1 + \beta^g_1 \ell) A_{0_{i,j}}, \quad 1 \leq i \neq j, \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A^{(\ell)}_{+_{i,j}} = (1 + \beta^f_1 \ell) A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell, \ 1 \leq j \leq J_{\ell+1},$$

and the diagonal entries of $A^{(\ell)}_0$ for $\ell \geq 0$ are the negative of the relevant row sums.

Given that the non-diagonal entries of the infinitesimal generator matrix must be non-negative, the following constraints must be satisfied:

$$1 + \beta^h_1(\ell - 1) \geq 0 \implies \beta^h_1 \geq -\frac{1}{\ell - 1}, \text{ for all } \ell \in [L_-, U_-],$$

$$1 + \beta^g_1 \ell \geq 0 \implies \beta^g_1 \geq -\frac{1}{\ell}, \text{ for all } \ell \in [L_0, U_0], \text{ and}$$

$$1 + \beta^f_1 \ell \geq 0 \implies \beta^f_1 \geq -\frac{1}{\ell}, \text{ for all } \ell \in [L_+, U_+].$$

In this section, we define the EM algorithm for a structured QBD process with linear functional forms under the assumption that the phase process is unobservable. Note that the method described in this section will be similar to that presented in Section 8.3. Hence, we only describe the parts of the method which are different.

8.4.1 Likelihood

Similar to Section 8.3, we expand the complete data log-likelihood in Equation (8.3.2) to show the relationships between the block matrices, such that

$$
\begin{aligned}
\ell_C(\boldsymbol{\theta}|\mathbf{d}) = {}& \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell} A^{(\ell)}_{0_{i,i}} Z_{\ell,i} + \sum_{\ell=0}^{L_0-1}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A^{(\ell)}_{0_{i,j}}\right) S_{\ell,i,j} \\
&+ \sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left((1+\beta_1^g\ell)A_{0_{i,j}}\right) S_{\ell,i,j} + \sum_{\ell=U_0+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A^{(\ell)}_{0_{i,j}}\right) S_{\ell,i,j} \\
&+ \sum_{\ell=1}^{L_- -1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A^{(\ell)}_{-_{i,j}}\right) D_{\ell,i,j} + \sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left((1+\beta_1^h(\ell-1))A_{-_{i,j}}\right) D_{\ell,i,j} \\
&+ \sum_{\ell=U_-+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A^{(\ell)}_{-_{i,j}}\right) D_{\ell,i,j} + \sum_{\ell=0}^{L_+-1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A^{(\ell)}_{+_{i,j}}\right) U_{\ell,i,j} \\
&+ \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left((1+\beta_1^f\ell)A_{+_{i,j}}\right) U_{\ell,i,j} + \sum_{\ell=U_++1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A^{(\ell)}_{+_{i,j}}\right) U_{\ell,i,j}.
\end{aligned}
$$

Theorem 8.4.1. *Using the complete data log-likelihood, the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ for a structured QBD process with linear functional forms of the level are as follows:*

$$\widehat{A}^{(\ell)}_{-_{i,j}} = \frac{D_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 1 \leq \ell \leq L_- - 1, \ell \geq U_- + 1, \textit{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

$$\widehat{A}_{-_{i,j}} = \frac{\sum\limits_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\sum\limits_{\ell=L_-}^{U_-} (1+\beta_1^h(\ell-1))Z_{\ell,i}}, \textit{ for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_- -1},$$

$$\widehat{A}^{(\ell)}_{0_{i,j}} = \frac{S_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 0 \leq \ell \leq L_0 - 1, \ell \geq U_0 + 1, \textit{ and } 1 \leq i \neq j \leq J_\ell,$$

$$\widehat{A}_{0_{i,j}} = \frac{\sum\limits_{\ell=L_0}^{U_0} S_{\ell,i,j}}{\sum\limits_{\ell=L_0}^{U_0} (1+\beta_1^g\ell)Z_{\ell,i}}, \textit{ for } 1 \leq i \neq j \leq J_{L_0},$$

$$\widehat{A}^{(\ell)}_{+_{i,j}} = \frac{U_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } 0 \leq \ell \leq L_+ - 1, \ell \geq U_+ + 1, \textit{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1},$$

$$\widehat{A}_{+i,j} = \frac{\sum_{\ell=L_+}^{U_+} U_{\ell,i,j}}{\sum_{\ell=L_+}^{U_+} (1+\beta_1^f \ell) Z_{\ell,i}}, \textit{ for } 1 \leq i \leq J_{L_+}, 1 \leq j \leq J_{L_+ + 1}.$$

Note: the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

Proof: Similar to the proof of Theorem 8.3.1.

Taking the partial derivative of the complete data log-likelihood with respect to β_1^h, β_1^g, and β_1^f, we obtain the following expressions:

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^h} = -\sum_{\ell=L_-}^{U_-} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell - 1} (\ell-1) A_{-i,j} Z_{\ell,i} + \sum_{\ell=L_-}^{U_-} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell - 1} \frac{\ell D_{\ell,i,j}}{1+\beta_1^h(\ell-1)}, \tag{8.4.1}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^g} = -\sum_{\ell=L_0}^{U_0} \sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j\neq i}}^{J_\ell} \ell A_{0i,j} Z_{\ell,i} + \sum_{\ell=L_0}^{U_0} \sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j\neq i}}^{J_\ell} \frac{\ell S_{\ell,i,j}}{1+\beta_1^g \ell}, \text{ and} \tag{8.4.2}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^f} = -\sum_{\ell=L_+}^{U_+} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell} \ell A_{+i,j} Z_{\ell,i} + \sum_{\ell=L_+}^{U_+} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell} \frac{\ell U_{\ell,i,j}}{1+\beta_1^f \ell}. \tag{8.4.3}$$

Explicit expressions for the maximum likelihood estimates of β_1^h, β_1^g, and β_1^f cannot be found. As a result, we numerically solve for β_1^h, β_1^g, and β_1^f using the bisection method.

8.4.2 Bisection Method

The bisection method is a root-finding algorithm that approximates the root of a continuous function, $f(x)$, that changes sign at least once for $x \in [a_0, b_0]$. At each iteration, the algorithm halves the interval and selects the sub-interval which contains the root, thus refining the approximation of the root of $f(x)$. The algorithm of the bisection method is presented in Algorithm 2.

Algorithm 2: Bisection method

```
Initialise
    • $a_0$ and $b_0$, such that $f(a_0) \times f(b_0) < 0$
    • $m_0 = \frac{a_0+b_0}{2}$
while $|f(m_n)| \geq \epsilon$ do
    if $f(m_n) = 0$ then
        Stop, as we have found the solution.
    else
        Calculate the midpoint of the interval:
            $m_n = \frac{a_n + b_n}{2}$
        if $f(a_n) \times f(m_n) < 0$ then
            $a_{n+1} = a_n$
            $b_{n+1} = m_n$
        else
            $a_{n+1} = m_n$
            $b_{n+1} = b_n$
```

8.4.3 Feasible regions

Consider the partial derivative of the complete data log-likelihood with respect to β_1^f,

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^f} = -\sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell} \ell A_{+i,j} Z_{\ell,i} + \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell} \frac{\ell U_{\ell,i,j}}{1+\beta_1^f \ell}.$$

Notice that this function is undefined when $1 + \beta_1^f \ell = 0$, for $L_+ \leq \ell \leq U_+$. The vertical asymptotes of this function are then defined as

$$-\frac{1}{\ell}, \quad \text{for } L_+ \leq \ell \leq U_+.$$

Recall that the non-diagonal entries of the infinitesimal generator matrix must be non-negative, and so we have the constraint

$$\beta_1^f \geq -\frac{1}{\ell}, \qquad \text{for all } \ell \in [L_+, U_+].$$

Therefore, we use the bisection method to numerically solve for β_1^f by searching for a solution within the regions defined not only by the vertical asymptotes but also those that satisfy the constraint

$$\beta_1^f \geq -\frac{1}{\ell}, \qquad \text{for all } \ell \in [L_+, U_+].$$

That is, we use the bisection method to numerically solve for β_1^f by searching for a solution within the region $\left(-\frac{1}{U_+}, \infty\right)$. Note that we set the solution to be $-\frac{1}{U_+} + \delta$ for small δ if the bisection method is searching for a solution too close to $-\frac{1}{U_+}$, so as to avoid undefined values.

Similarly, we numerically solve for β_1^g by searching for a solution within the region $\left(-\frac{1}{U_0}, \infty\right)$ and we numerically solve for β_1^h by searching for a solution within the region $\left(-\frac{1}{U_- - 1}, \infty\right)$.

8.4.4 E-step and CM-step

Similar to before, we use the ECM algorithm to estimate the parameters of the structured QBD process with linear functional forms. The calculations involved in the E-step follow those described in Section 7.2.4. The calculations involved in the CM-step are as follows.

In the CM-step of the ECM algorithm, we update the values of the parameters by replacing the statistics in Theorem 8.4.1 with the conditional expectations calculated in the E-step.

At the $(t+1)^{th}$ iteration, the conditional maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ given $h_\ell = h_\ell^{(t)}$, $g_\ell = g_\ell^{(t)}$, and $f_\ell = f_\ell^{(t)}$ are as follows.

For $1 \leq \ell \leq L_- - 1$, $U_- + 1 \leq \ell < \infty$, $1 \leq i \leq J_\ell$, and $1 \leq j \leq J_{\ell-1}$:

$$\widehat{A_-}_{i,j}^{(\ell)} = \frac{\sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \leq i \leq J_\ell$ and $1 \leq j \leq J_{L_-}$:

$$\widehat{A_-}_{i,j} = \frac{\sum_{\ell=L_-}^{U_-} \sum_{k=1}^{K} E\big[D_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\sum_{\ell=L_-}^{U_-} \widehat{h}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $0 \leq \ell \leq L_0 - 1$, $U_0 + 1 \leq \ell < \infty$, and $1 \leq i \neq j \leq J_\ell$:

$$\widehat{A_0}_{i,j}^{(\ell)} = \frac{\sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \leq i \neq j \leq J_{L_0}$:

$$\widehat{A_0}_{i,j} = \frac{\sum_{\ell=L_0}^{U_0} \sum_{k=1}^{K} E\big[S_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\sum_{\ell=L_0}^{U_0} \widehat{g}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $0 \leq \ell \leq L_+ - 1$, $U_+ + 1 \leq \ell < \infty$, $1 \leq i \leq J_\ell$, and $1 \leq j \leq J_{\ell+1}$:

$$\widehat{A_{+}}_{i,j}^{(\ell)} = \frac{\displaystyle\sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

For $1 \leq i \leq J_\ell$ and $1 \leq j \leq J_{L_+}$:

$$\widehat{A_{+}}_{i,j}^{=} \frac{\displaystyle\sum_{\ell=L_+}^{U_+} \sum_{k=1}^{K} E\big[U_{\ell,i,j}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}{\displaystyle\sum_{\ell=L_+}^{U_+} \widehat{f}_\ell^{(t)} \sum_{k=1}^{K} E\big[Z_{\ell,i}^{(k)} | \mathbf{D}, \boldsymbol{\theta}\big]}.$$

Then, the conditional maximum likelihood estimates for $\{h_\ell; L_- \leq \ell \leq U_-\}$, $\{g_\ell; L_0 \leq \ell \leq U_0\}$, and $\{f_\ell; L_+ \leq \ell \leq U_+\}$ given $\widehat{A_{-}}_{i,j} = \widehat{A_{-}}_{i,j}^{(t+1)}$, $\widehat{A_{0}}_{i,j} = \widehat{A_{0}}_{i,j}^{(t+1)}$, and $\widehat{A_{+}}_{i,j} = \widehat{A_{+}}_{i,j}^{(t+1)}$ are found by solving Equations (8.4.1), (8.4.2), and (8.4.3) using the bisection method, respectively.

8.4.5 Discussion

The structured QBD process with linear functional forms has more model flexibility than standard queueing models of the form $M/M/c$ or $M/M/\infty$, as well as fewer parameters that that of a level-dependent QBD process whilst remaining level-dependent in nature. In the next section, we consider quadratic functional forms of the level which describes unique behaviour in queueing systems.

8.5 Structured QBD processes with quadratic functional forms

The behaviour of certain queueing systems may be unexpected, such that the arrival and departure rates may be convex or concave in nature. For example, such behaviour may reflect demand for a service where the server's productivity is low when the system is near empty and close to capacity (too much stress). Therefore, in this section we consider a structured QBD process with infinitesimal generator matrix

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A_{-_{i,j}}^{(\ell)} = (1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2)A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell-1},$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A_{0_{i,j}}^{(\ell)} = (1 + \beta_1^g \ell + \beta_2^g \ell^2)A_{0_{i,j}}, \quad 1 \leq i \neq j, \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A_{+_{i,j}}^{(\ell)} = (1 + \beta_1^f \ell + \beta_2^f \ell^2)A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1},$$

and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

Given that the non-diagonal entries of the infinitesimal generator matrix must be non-negative, the following constraints must be satisfied:

$$1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2 \geq 0 \text{ for all } \ell \in [L_-, U_-],$$

$$1 + \beta_1^g \ell + \beta_2^g \ell^2 \geq 0, \text{ for all } \ell \in [L_0, U_0], \text{ and}$$

$$1 + \beta_1^f \ell + \beta_2^f \ell^2 \geq 0 \text{ for all } \ell \in [L_+, U_+].$$

For example, suppose the quadratic associated with the block matrices describing no change in level is convex. If the vertex of the quadratic lies within the domain $[L_0, U_0]$ then we must have that the y-coordinate of the vertex is non-negative, as shown in Figure 8.5.1.

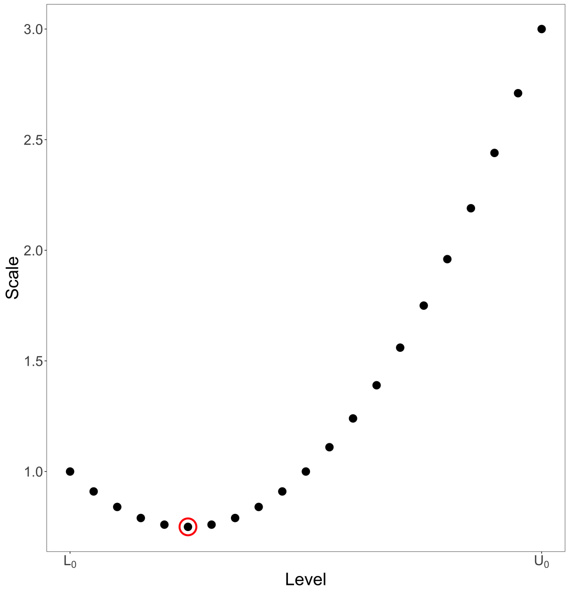


Figure 8.5.1: Illustration of the non-negative requirement of a convex quadratic when the vertex of the quadratic lies within the domain $[L_0, U_0]$. Note that the red circle indicates the vertex of the quadratic.

If the vertex does not lie within the domain $[L_0, U_0]$ then we must have that the minimum value of the quadratic within the domain $[L_0, U_0]$ is non-negative, as demonstrated in Figure 8.5.2.

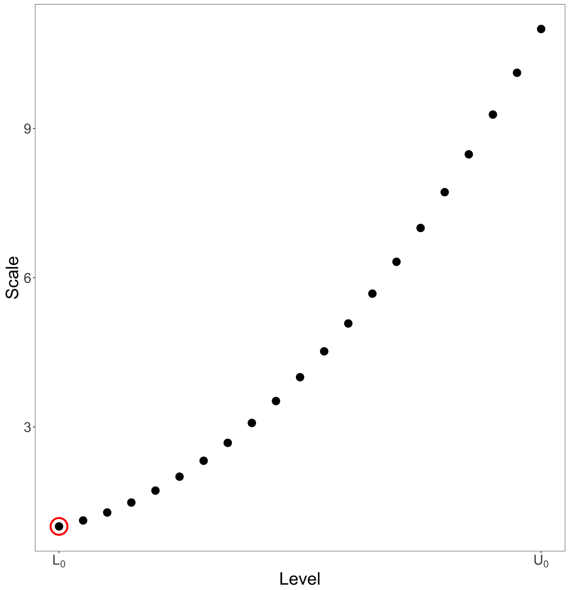


Figure 8.5.2: Illustration of the non-negative requirement of a convex quadratic when the vertex of the quadratic does not lie within the domain $[L_0, U_0]$. Note that the red circle indicates the minimum values within the domain.

Suppose now that the quadratic associated with the block matrices describing no change in level is concave. In this case, we must ensure that the minimum value of the quadratic within the domain $[L_0, U_0]$ is non-negative, as illustrated in Figure 8.5.3.

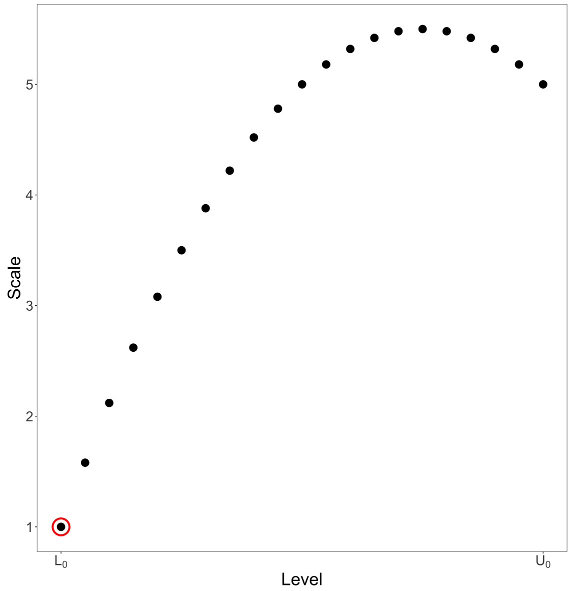


Figure 8.5.3: Illustration of the non-negative requirement of a concave quadratic. Note that the red circle indicates the minimum values within the domain.

The requirement for such detailed constraints is motivated by the potential to use fitted structured QBD processes for prediction modelling. For example, suppose we fit a structured QBD process to observed data associated with a queueing system with a capacity of 10 customers. In this case, we consider an infinitesimal generator matrix defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_-^{(9)} & A_0^{(9)} & A_+^{(9)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(10)} & A_0^{(10)} \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels 1 and 10 are related, such that for $1 \leq \ell \leq 10$,

$$A^{(\ell)}_{-_{i,j}} = (1 + \beta^h_1(\ell - 1) + \beta^h_2(\ell - 1)^2)A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell-1},$$

the block matrices describing no change in level between the levels 0 and 10 are related, such that for $0 \leq \ell \leq 10$,

$$A^{(\ell)}_{0_{i,j}} = A_{0_{i,j}}, \quad 1 \leq i \neq j, \leq J_\ell,$$

the block matrices describing an increase in level between the levels 0 and 9 are related, such that for $0 \leq \ell \leq 9$,

$$A^{(\ell)}_{+_{i,j}} = A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1},$$

and the diagonal entries of $A^{(\ell)}_0$ for $0 \leq \ell \leq 10$ are the negative of the relevant row sums.

Now suppose we wanted to reduce the capacity of the queueing system to 8 customers in such a way that the smaller queueing system will have similar arrival and service processes to those of the original queueing system. Consider now the infinitesimal generator matrix of the smaller queueing system,

$$\tilde{Q} = \begin{bmatrix} \tilde{A}^{(0)}_0 & \tilde{A}^{(0)}_+ & 0 & \dots & 0 & 0 & 0 \\ \tilde{A}^{(1)}_- & \tilde{A}^{(1)}_0 & \tilde{A}_+ & \dots & 0 & 0 & 0 \\ 0 & \tilde{A}^{(2)}_- & \tilde{A}^{(2)}_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \tilde{A}^{(7)}_- & \tilde{A}^{(7)}_0 & \tilde{A}^{(7)}_+ \\ 0 & 0 & 0 & \dots & 0 & \tilde{A}^{(8)}_- & \tilde{A}^{(8)}_0 \end{bmatrix}.$$

Since the scales associated with the $A_0^{(\ell)}$ and $A_+^{(\ell)}$ block matrices are constant, we set

$$\tilde{A}_0^{(\ell^*)} = A_0^{(\ell)}, \quad 0 \leq \ell^* \leq 8,$$

and

$$\tilde{A}_+^{(\ell^*)} = A_0^{(\ell)}, \quad 0 \leq \ell^* \leq 7.$$

Note that the diagonal entries of $\tilde{A}_0^{(\ell)}$ for $0 \leq \ell \leq 8$ are the negative of the relevant row sums.

In the case of the $\tilde{A}_-^{(\ell)}$ block matrices, we transform the domain from $[a, b]$ to $[c, d]$ by using the transformation,

$$\ell^* = \frac{(b-a)(\ell-c)}{(d-c)} + a,$$

That is,

$$\tilde{A}_-^{(\ell^*)} = (1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2)A_-, \quad 1 \leq \ell^* \leq 8,$$

such that

$$\ell = \frac{(10-1)(\ell^*-1)}{(8-1)} + 1, \quad \ell^* \in [1, 8].$$

Even though the level-dependent scales associated with the fitted structured QBD process are non-negative, the vertex of the quadratic associated with the block matrices describing a decrease in level may have a negative y-coordinate. Hence, the transformation in domain may introduce negative scales and thus lead to an invalid structured QBD process. Therefore, we must ensure that the quadratic is non-negative for all values between the minimum and maximum levels.

Similar to previous sections, we now define the EM algorithm for a structured QBD process with quadratic functional forms under the assumption that the phase process is unobservable. Note that the method described in this section will be similar to that presented in Section 8.4. Hence, we only describe parts of the method which are different.

8.5.1 Likelihood

Similar to Section 8.4, we expand the complete data log-likelihood in Equation (8.3.2) to show the relationships between the block matrices, such that

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}|\mathbf{d}) &= \sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell} A_{0_{i,i}}^{(\ell)} Z_{\ell,i} + \sum_{\ell=0}^{L_0-1}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} \\
&+ \sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left((1+\beta_1^g\ell+\beta_2^g\ell^2)A_{0_{i,j}}\right) S_{\ell,i,j} + \sum_{\ell=U_0+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} \\
&+ \sum_{\ell=1}^{L_--1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left((1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2)A_{-_{i,j}}\right) D_{\ell,i,j} \\
&+ \sum_{\ell=U_-+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=0}^{L_+-1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j} \\
&+ \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left((1+\beta_1^f\ell+\beta_2^f\ell^2)A_{+_{i,j}}\right) U_{\ell,i,j} + \sum_{\ell=U_++1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}.
\end{aligned}$$

Theorem 8.5.1. *Using the complete data log-likelihood, the maximum likelihood estimates of the parameters* $\boldsymbol{\theta} = (\{A_-^{(\ell)}; \ell \geq 1\}, \{A_0^{(\ell)}; \ell \geq 0\}, \{A_+^{(\ell)}; \ell \geq 0\})$ *for a structured QBD process with quadratic functional forms of the level are as follows:*

$$\widehat{A}_{-_{i,j}}^{(\ell)} = \frac{D_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } 1 \leq \ell \leq L_- - 1, \ell \geq U_- + 1, \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

$$\widehat{A}_{-_{i,j}} = \frac{\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\sum_{\ell=L_-}^{U_-} (1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2)Z_{\ell,i}}, \text{ for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_--1},$$

$$\widehat{A}_{0_{i,j}}^{(\ell)} = \frac{S_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } 0 \leq \ell \leq L_0 - 1, \ell \geq U_0 + 1, \text{ and } 1 \leq i \neq j \leq J_\ell,$$

$$\widehat{A}_{0_{i,j}} = \frac{\sum_{\ell=L_0}^{U_0} S_{\ell,i,j}}{\sum_{\ell=L_0}^{U_0} (1+\beta_1^g\ell+\beta_2^g\ell^2)Z_{\ell,i}}, \text{ for } 1 \leq i \neq j \leq J_{L_0},$$

$$\widehat{A}_{+_{i,j}}^{(\ell)} = \frac{U_{\ell,i,j}}{Z_{\ell,i}}, \text{ for } 0 \leq \ell \leq L_+ - 1, \ell \geq U_+ + 1, \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1},$$

$$\widehat{A}_{+i,j} = \frac{\sum\limits_{\ell=L_+}^{U_+} U_{\ell,i,j}}{\sum\limits_{\ell=L_+}^{U_+} (1+\beta_1^f\ell+\beta_2^f\ell^2)Z_{\ell,i}}, \; \textit{for } 1 \le i \le J_{L_+}, 1 \le j \le J_{L_++1},$$

Note: the diagonal entries of $A_0^{(\ell)}$ for $\ell \ge 0$ are the negative of the relevant row sums.

Proof: Similar to the proof of Theorem 8.3.1.

Taking the partial derivative of the complete data log-likelihood with respect to β_1^h, β_2^h, β_1^g, β_2^g, β_1^f, and β_2^f, we obtain the following expressions:

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^h} = -\sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell-1}(\ell-1)A_{-i,j}Z_{\ell,i} + \sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell-1}\frac{\ell D_{\ell,i,j}}{1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2}, \tag{8.5.1}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_2^h} = -\sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell-1}(\ell-1)^2A_{-i,j}Z_{\ell,i} + \sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell-1}\frac{\ell^2 D_{\ell,i,j}}{1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2}, \tag{8.5.2}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^g} = -\sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\ell A_{0i,j}Z_{\ell,i} + \sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\frac{\ell S_{\ell,i,j}}{1+\beta_1^g\ell+\beta_2^g\ell^2}, \tag{8.5.3}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_2^g} = -\sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\ell^2 A_{0i,j}Z_{\ell,i} + \sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\ j\neq i}}^{J_\ell}\frac{\ell^2 S_{\ell,i,j}}{1+\beta_1^g\ell+\beta_2^g\ell^2}, \tag{8.5.4}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_1^f} = -\sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\ell A_{+i,j}Z_{\ell,i} + \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\frac{\ell U_{\ell,i,j}}{1+\beta_1^f\ell+\beta_2^f\ell^2}, \tag{8.5.5}$$

and,

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_2^f} = -\sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\ell^2 A_{+i,j}Z_{\ell,i} + \sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_\ell}\frac{\ell^2 U_{\ell,i,j}}{1+\beta_1^f\ell+\beta_2^f\ell^2}. \tag{8.5.6}$$

Explicit expressions for the maximum likelihood estimates of β_1^h, β_2^h, β_1^g, β_2^g, β_1^f, and β_2^f cannot be found. As a result, we numerically solve for β_1^h, β_2^h, β_1^g, β_2^g, β_1^f, and β_2^f using the bisection method. Note that to ensure h_ℓ, g_ℓ, and f_ℓ remain non-negative, we iteratively update each value, such that

$$\beta_1^h \geq \frac{-1-\beta_2^h(\ell-1)^2}{(\ell-1)}, \quad \beta_2^h \geq \frac{-1-\beta_1^h(\ell-1)}{(\ell-1)^2}, \qquad \text{for all } \ell \in [L_-, U_-],$$

$$\beta_1^g \geq \frac{-1-\beta_2^g\ell^2}{\ell}, \quad \beta_2^g \geq \frac{-1-\beta_1^g\ell}{\ell^2}, \qquad \text{for all } \ell \in [L_0, U_0],$$

and

$$\beta_1^f \geq \frac{-1-\beta_2^f\ell^2}{\ell}, \quad \beta_2^f \geq \frac{-1-\beta_1^f\ell}{\ell^2}, \qquad \text{for all } \ell \in [L_+, U_+],$$

while also ensuring that each quadratic remains non-negative in the respective domains.

8.5.2 E-step and CM-step

Each iteration of the ECM algorithm consists of an E-step and a CM-step. The calculations involved in the E-step follow those described in Section 7.2.4. In the CM-step of the ECM algorithm, we update the values of the parameters by replacing the statistics in Theorem 8.5.1 with the conditional expectations calculated in the E-step, similar to those described in Section 8.4.

The conditional maximum likelihood estimates for $\{h_\ell; L_- \leq \ell \leq U_-\}$, $\{g_\ell; L_0 \leq \ell \leq U_0\}$, and $\{f_\ell; L_+ \leq \ell \leq U_+\}$ given $\widehat{A_{-i,j}} = \widehat{A_{-i,j}}^{(t+1)}$, $\widehat{A_{0i,j}} = \widehat{A_{0i,j}}^{(t+1)}$, and $\widehat{A_{+i,j}} = \widehat{A_{+i,j}}^{(t+1)}$ are found by solving Equations (8.5.1), (8.5.2), (8.5.3), (8.5.4), (8.5.5), and (8.5.6) using the bisection method, respectively.

8.5.3 Discussion

The structured QBD process with quadratic forms of the level describes the behaviour queueing systems with non-linear rates without having to consider over-parameterised models such as the level-dependent QBD process. Higher order forms of the level, such as cubics or quartics, have the ability to explain more of the behaviour of the queueing system compared to linear and quadratic forms. For completeness, in the next section we consider structured QBD processes with higher order polynomial forms of the level.

8.6 Structured QBD processes with polynomial functional forms

The statistical model fitting method developed in Section 8.5 can be extended to structured QBD processes with polynomial forms, such that the infinitesimal generator matrix is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A_{-_{i,j}}^{(\ell)} = (1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2+\ldots+\beta_p^h(\ell-1)^p)A_{-_{i,j}},\ 1 \leq j \leq J_{\ell-1},\ 1 \leq i \leq J_\ell,$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A^{(\ell)}_{0_{i,j}} = (1 + \beta^g_1 \ell + \beta^g_2 \ell^2 + \ldots + \beta^g_p \ell^p) A_{0_{i,j}},\ 1 \leq i \neq j, \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A^{(\ell)}_{+_{i,j}} = (1 + \beta^f_1 \ell + \beta^f_2 \ell^2 + \ldots + \beta^f_p \ell^p) A_{+_{i,j}},\ 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1}.$$

and the diagonal entries of $A^{(\ell)}_0$ for $\ell \geq 0$ are the negative of the relevant row sums.

Given that the non-diagonal entries of the infinitesimal generator matrix must be non-negative, the following constraints must be satisfied:

$$1 + \beta^h_1(\ell - 1) + \beta^h_2(\ell - 1)^2 + \ldots + \beta^h_p(\ell - 1)^p \geq 0 \text{ for all } \ell \in [L_-, U_-],$$

$$1 + \beta^g_1 \ell + \beta^g_2 \ell^2 + \ldots + \beta^g_p \ell^p \geq 0, \text{ for all } \ell \in [L_0, U_0], \text{ and}$$

$$1 + \beta^f_1 \ell + \beta^f_2 \ell^2 + \ldots + \beta^f_p \ell^p \geq 0 \text{ for all } \ell \in [L_+, U_+].$$

The method described in this section is an extension of the method presented in Section 8.5. Therefore, we only describe parts of the method which are different.

8.6.1 Likelihood

Similar to Section 8.5, the complete data log-likelihood is defined as

$$\begin{aligned}\Delta_C(\boldsymbol{\theta}|\mathbf{d}) = &\sum_{\ell=0}^{\infty}\sum_{i=1}^{J_\ell} A_{0_{i,i}}^{(\ell)} Z_{\ell,i} + \sum_{\ell=0}^{L_0-1}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j}\\
&+\sum_{\ell=L_0}^{U_0}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\j\neq i}}^{J_\ell} \log\left((1+\beta_1^g\ell+\beta_2^g\ell^2+\ldots+\beta_p^g\ell^p)A_{0i,j}\right) S_{\ell,i,j}\\
&+\sum_{\ell=U_0+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{\substack{j=1\\j\neq i}}^{J_\ell} \log\left(A_{0_{i,j}}^{(\ell)}\right) S_{\ell,i,j} + \sum_{\ell=1}^{L_--1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j}\\
&+\sum_{\ell=L_-}^{U_-}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left((1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2+\ldots+\beta_p^h(\ell-1)^p)A_{-_{i,j}}\right) D_{\ell,i,j}\\
&+\sum_{\ell=U_-+1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell-1}} \log\left(A_{-_{i,j}}^{(\ell)}\right) D_{\ell,i,j} + \sum_{\ell=0}^{L_+-1}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}\\
&+\sum_{\ell=L_+}^{U_+}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left((1+\beta_1^f\ell+\beta_2^f\ell^2+\ldots+\beta_p^f\ell^p)A_{+_{i,j}}\right) U_{\ell,i,j}\\
&+\sum_{\ell=U_++1}^{\infty}\sum_{i=1}^{J_\ell}\sum_{j=1}^{J_{\ell+1}} \log\left(A_{+_{i,j}}^{(\ell)}\right) U_{\ell,i,j}.\end{aligned}$$

Theorem 8.6.1. *Using the complete data log-likelihood, the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_-^{(\ell)};\ell\geq 1\},\{A_0^{(\ell)};\ell\geq 0\},\{A_+^{(\ell)};\ell\geq 0\})$ for a structured QBD process with polynomial functional forms of the level are as follows:*

$$\widehat{A}_{-_{i,j}}^{(\ell)} = \frac{D_{\ell,i,j}}{Z_{\ell,i}},\ \textit{for } 1\leq\ell\leq L_--1, \ell\geq U_-+1,\ \textit{and } 1\leq i\leq J_\ell, 1\leq j\leq J_{\ell-1},$$

$$\widehat{A}_{-_{i,j}} = \frac{\sum\limits_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\sum\limits_{\ell=L_-}^{U_-}(1+\beta_1^h(\ell-1)+\beta_2^h(\ell-1)^2+\ldots+\beta_p^h(\ell-1)^p)Z_{\ell,i}},\ \textit{for } 1\leq i\leq J_{L_-}, 1\leq j\leq J_{L_--1},$$

$$\widehat{A}_{0_{i,j}}^{(\ell)} = \frac{S_{\ell,i,j}}{Z_{\ell,i}},\ \textit{for } 0\leq\ell\leq L_0-1, \ell\geq U_0+1,\ \textit{and } 1\leq i\neq j\leq J_\ell,$$

$$\widehat{A}_{0_{i,j}} = \frac{\sum\limits_{\ell=L_0}^{U_0} S_{\ell,i,j}}{\sum\limits_{\ell=L_0}^{U_0} (1+\beta_1^g \ell + \beta_2^g \ell^2 + \ldots + \beta_p^g \ell^p) Z_{\ell,i}}, \ \textit{for } 1 \le i \ne j \le J_{L_0},$$

$$\widehat{A}_{+_{i,j}}^{(\ell)} = \frac{U_{\ell,i,j}}{Z_{\ell,i}}, \ \textit{for } 0 \le \ell \le L_+ - 1, \ell \ge U_+ + 1, \ \textit{and } 1 \le i \le J_\ell, 1 \le j \le J_{\ell+1},$$

$$\widehat{A}_{+_{i,j}} = \frac{\sum\limits_{\ell=L_+}^{U_+} U_{\ell,i,j}}{\sum\limits_{\ell=L_+}^{U_+} (1+\beta_1^f \ell + \beta_2^f \ell^2 + \ldots + \beta_p^f \ell^p) Z_{\ell,i}}, \ \textit{for } 1 \le i \le J_{L_+}, 1 \le j \le J_{L_+ + 1},$$

Note: the diagonal entries of $A_0^{(\ell)}$ for $\ell \ge 0$ are the negative of the relevant row sums.

Proof: Similar to the proof of Theorem 8.3.1.

Taking the partial derivative of the complete data log-likelihood with respect to β_k^f, β_k^g, and β_k^h, we obtain the following expressions:

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_k^h} = -\sum_{\ell=L_-}^{U_-} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell - 1} (\ell-1)^k A_{-_{i,j}} Z_{\ell,i} \tag{8.6.1}$$

$$+ \sum_{\ell=L_-}^{U_-} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell - 1} \frac{\ell^k D_{\ell,i,j}}{1 + \beta_1^h(\ell-1) + \beta_2^h(\ell-1)^2 + \ldots + \beta_p^h(\ell-1)^p}, \tag{8.6.2}$$

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_k^g} = -\sum_{\ell=L_0}^{U_0} \sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j \ne i}}^{J_\ell} \ell^k A_{0i,j} Z_{\ell,i} + \sum_{\ell=L_0}^{U_0} \sum_{i=1}^{J_\ell} \sum_{\substack{j=1 \\ j \ne i}}^{J_\ell} \frac{\ell^k S_{\ell,i,j}}{1 + \beta_1^g \ell + \beta_2^g \ell^2 + \ldots + \beta_p^g \ell^p}, \tag{8.6.3}$$

and,

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \beta_k^f} = -\sum_{\ell=L_+}^{U_+} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell} \ell^k A_{+i,j} Z_{\ell,i} + \sum_{\ell=L_+}^{U_+} \sum_{i=1}^{J_\ell} \sum_{j=1}^{J_\ell} \frac{\ell^k U_{\ell,i,j}}{1 + \beta_1^f \ell + \beta_2^f \ell^2 + \ldots + \beta_p^f \ell^p}. \tag{8.6.4}$$

Explicit expressions for the maximum likelihood estimates of β_k^h, β_k^g, and β_k^f cannot be found. As a result, we numerically solve for β_k^h, β_k^g, and β_k^f using the bisection method. Note that to ensure h_ℓ, g_ℓ, and f_ℓ remain non-negative we iteratively update each value, such that

$$\beta_k^h \geq \frac{-1 - \beta_1^h(\ell-1) - \ldots - \beta_{k-1}^h(\ell-1)^{k-1} - \beta_{k+1}^h(\ell-1)^{k+1} - \ldots - \beta_p^h(\ell-1)^p}{\ell^k},$$

$$\beta_k^g \geq \frac{-1 - \beta_1^g\ell - \ldots - \beta_{k-1}^g\ell^{k-1} - \beta_{k+1}^g\ell^{k+1} - \ldots - \beta_p^g\ell^p}{\ell^k}, \text{ and}$$

$$\beta_k^f \geq \frac{-1 - \beta_1^f\ell - \ldots - \beta_{k-1}^f\ell^{k-1} - \beta_{k+1}^f\ell^{k+1} - \ldots - \beta_p^f\ell^p}{\ell^k},$$

while also ensuring that each polynomial remains non-negative in the respective domains.

8.6.2 E-step and CM-step

Similar to before, the calculations involved in the E-step follow those described in Section 7.2.4. In the CM-step of the ECM algorithm, we update the values of the parameters by replacing the statistics in Theorem 8.6.1 with the conditional expectations calculated in the E-step, similar to those described in Section 8.5.

The conditional maximum likelihood estimates for $\{h_\ell; L_- \leq \ell \leq U_-\}$, $\{g_\ell; L_0 \leq \ell \leq U_0\}$, and $\{f_\ell; L_+ \leq \ell \leq U_+\}$ given $\widehat{A_{-i,j}} = \widehat{A_{-i,j}}^{(t+1)}$, $\widehat{A_{0i,j}} = \widehat{A_{0i,j}}^{(t+1)}$, and $\widehat{A_{+i,j}} = \widehat{A_{+i,j}}^{(t+1)}$ are found by solving Equations (8.6.2), (8.6.3), and (8.6.4) for all parameters using the bisection method, respectively.

8.6.3 Discussion

Higher order forms of the level are generally able to explain more of the behaviour of the queueing system compared to linear and quadratic forms. However, the interpretability of structured QBD processes with higher order polynomial forms of the level in the context of a queueing process may be more complicated than that of linear and quadratic forms. As such, the benefits of a potentially better fitting structured QBD process need to be considered alongside an interpretable structured QBD process.

8.7 Finite case

Note that the EM algorithm for a finite structured QBD process is similar to the EM algorithm for an infinite structured QBD process where we truncate at level $C < \infty$, such that $A_{+}^{(C)} = 0$ and $A_{-}^{(\ell)} = 0$, $A_{0}^{(\ell)} = 0$, and $A_{+}^{(\ell)} = 0$ for $\ell \geq C + 1$.

8.8 Extensions

Firstly, the methods presented in this chapter can be combined to form a structured QBD process with various types of functional forms. For example, we could consider a structured QBD process where

- the rates corresponding to an increase in level decrease at a linear rate,
- the rates corresponding to no change in level increase for lower levels but decrease for higher levels, and
- the rates corresponding to a decrease in level are independent of level change.

That is, we could consider a structured QBD process such that

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A_{-_{i,j}}^{(\ell)} = A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell, \; 1 \leq j \leq J_{\ell-1}$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A_{0_{i,j}}^{(\ell)} = (1 + \beta_1^g \ell + \beta_2^g \ell^2) A_{0_{i,j}}, \quad 1 \leq i \neq j, \leq J_\ell$$

and the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A_{+_{i,j}}^{(\ell)} = (1 + \beta_1^f \ell) A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell, \text{ and } 1 \leq j \leq J_{\ell+1}.$$

Secondly, this work can be extended to general and polynomial functional forms of the level and the phase transition process, such that

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \dots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A^{(\ell)}_{-_{i,j}} = h_{\ell,i,j} A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A^{(\ell)}_{0_{i,j}} = g_{\ell,i,j} A_{0_{i,j}}, \quad 1 \leq i \neq j \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A^{(\ell)}_{+_{i,j}} = f_{\ell,i,j} A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1},$$

the diagonal entries of $A^{(\ell)}_0$ for $\ell \geq 0$ are the negative of the relevant row sums, and $g_{\ell,i,i} = 1$ for $i \neq j$ and all levels $L_0 \leq \ell \leq U_0$.

For example, the parameters of a structured QBD process with polynomial functional forms of the level and the phase transition process are such that for $L_- \leq \ell \leq U_-$ and $1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell-1}$,

$$A^{(\ell)}_{-_{i,j}} = h_{\ell,i,j} A_{-_{i,j}},$$

for $L_0 \leq \ell \leq U_0$ and $1 \leq i \neq j \leq J_\ell$,

$$A^{(\ell)}_{0_{i,j}} = g_{\ell,i,j} A_{0_{i,j}},$$

and for $L_+ \leq \ell \leq U_+$ and $1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1}$,

$$A^{(\ell)}_{+_{i,j}} = f_{\ell,i,j} A_{+_{i,j}},$$

where

$$h_{\ell,i,j} = 1 + \beta_1^{h,i,j}(\ell-1) + \beta_2^{h,i,j}(\ell-1)^2 + \ldots + \beta_p^{h,i,j}(\ell-1)^p,$$

$$g_{\ell,i,j} = 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2 + \ldots + \beta_p^{g,i,j}\ell^p,$$

$$f_{\ell,i,j} = 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2 + \ldots + \beta_p^{f,i,j}\ell^p.$$

Note that the non-diagonal entries of the infinitesimal generator matrix must be non-negative, such that

$$h_{\ell,i,j} \geq 0 \text{ for all } \ell \in [L_-, U_-], 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

$$g_{\ell,i,j} \geq 0 \text{ for all } \ell \in [L_0, U_0], 1 \leq i \neq j \leq J_\ell, \text{ and}$$

$$f_{\ell,i,j} \geq 0 \text{ for all } \ell \in [L_+, U_+], 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell+1}.$$

Lastly, this work can be extended to other functional forms of the level and phase transition. Note that the functional forms must be non-negative to ensure that the off-diagonal entries of the infinitesimal generator matrix remain non-negative. For example, consider a structured QBD process with infinitesimal generator matrix with exponential functional forms of the level,

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \ldots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \ldots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \ldots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the block matrices describing a decrease in level between the levels L_- and U_- are related, such that for $L_- \leq \ell \leq U_-$,

$$A^{(\ell)}_{-_{i,j}} = \exp(\beta_h(\ell - 1))A_{-_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell-1},$$

the block matrices describing no change in level between the levels L_0 and U_0 are related, such that for $L_0 \leq \ell \leq U_0$,

$$A^{(\ell)}_{0_{i,j}} = \exp(\beta_g \ell))A_{0_{i,j}}, \quad 1 \leq i \neq j \leq J_\ell,$$

the block matrices describing an increase in level between the levels L_+ and U_+ are related, such that for $L_+ \leq \ell \leq U_+$,

$$A^{(\ell)}_{+_{i,j}} = \exp(\beta_f \ell))A_{+_{i,j}}, \quad 1 \leq i \leq J_\ell,\ 1 \leq j \leq J_{\ell+1},$$

and the diagonal entries of $A^{(\ell)}_0$ for $\ell \geq 0$ are the negative of the relevant row sums.

8.9 Model fitting and selection

Each of the structured QBD processes presented in this chapter provide the opportunity to capture the stationary and transient behaviour of a QBD process without over-fitting the model. However, the choice of a structured QBD process may not always be clear when modelling a queueing process. Therefore, we describe a method of choosing which structured QBD processes to fit based on the information available, which is presented below.

1. If **there is no or limited information** on whether the arrival or service processes exhibit any form of polynomial characteristics, fit a structured QBD process to the data where the scales are both level and phase dependent.

   (a) If the fitted values of $f_{\ell,i,j}$, $g_{\ell,i,j}$, and $h_{\ell,i,j}$ appear to be **independent** of phase, then fit a structured QBD process to the data where the scales are level-dependent.

i. Assess the forms of f_ℓ, g_ℓ, and h_ℓ for all possible levels ℓ respectively.

ii. Identify which, if any, of the functional forms can be represented by a polynomial form. Statistical modelling such as weighted least squares regression and likelihood ratio tests can be used to identify suitable polynomial forms.

iii. Decide on which, if any, polynomial forms are to be used in the structured QBD process.

(b) If the fitted values of $f_{\ell,i,j}$, $g_{\ell,i,j}$, and $h_{\ell,i,j}$ appear to be **dependent** on phase, then assess the forms of $f_{\ell,i,j}$, $g_{\ell,i,j}$, and $h_{\ell,i,j}$ for all possible levels ℓ and phase transitions (i, j) respectively.

i. Identify which, if any, of the functional forms can be represented by a polynomial form. Statistical modelling such as weighted least squares regression and likelihood ratio tests can be used to identify suitable polynomial forms.

ii. Decide on which, if any, polynomial forms are to be used in the structured QBD process.

2. If **there is relevant information** on whether the arrival or service processes exhibit any form of polynomial characteristics, identify which of the functional forms can be represented by a polynomial form.

(a) Decide on which polynomial forms are to be used in the structured QBD process.

8.9.1 Likelihood ratio test

The likelihood ratio test (LRT) is used to compare the goodness of fit of two statistical models, where the null and alternative hypotheses are defined as

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \boldsymbol{\Theta}_0, \\ H_a &: \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is the parameter vector and $\boldsymbol{\Theta}_0$ is a subset of the parameter space $\boldsymbol{\Theta}$. The likelihood ratio test statistic is defined as

$$\lambda = -2 \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \log(L(\boldsymbol{\theta}; \mathbf{Y})) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log(L(\boldsymbol{\theta}; \mathbf{Y})) \right),$$

where $\mathbf{Y}$ denotes the data and $L(\boldsymbol{\theta}; \mathbf{Y})$ denotes the likelihood function. Under the null hypothesis, this test statistic has an approximate chi-square distribution with p degrees of freedom.

The LRT is also used to compare nested models $\mathcal{M}_1$ and $\mathcal{M}_2$, such that $\mathcal{M}_1$ is the restricted model or simpler model and $\mathcal{M}_2$ is the less restricted model or the model with more parameters. In this case, the likelihood ratio test statistic is defined as

$$\lambda = -2 \left(\log(L(\widehat{\boldsymbol{\theta}_1}; \mathbf{Y})) - \log(L(\widehat{\boldsymbol{\theta}_2}; \mathbf{Y})) \right),$$

where $\widehat{\boldsymbol{\theta}_1} = (\widehat{\theta}_{11}, \ldots, \widehat{\theta}_{1p_1})$ denotes the maximum likelihood estimate of the parameters for the restricted model and $\widehat{\boldsymbol{\theta}_2} = (\widehat{\theta}_{21}, \ldots, \widehat{\theta}_{2p_2})$ denotes the maximum likelihood estimate of the parameters for the less restricted model.

Under the null hypothesis, this test statistic has an approximate chi-square distribution with $p_2 - p_1$ degrees of freedom.

For example, consider a two-phase structured QBD process with infinitesimal generator matrix

$$Q = \begin{bmatrix} A_0^{(0)} & f_0 A_+ & 0 & 0 & 0 \\ A_- & A_0^{(1)} & f_1 A_+ & 0 & 0 \\ 0 & A_- & A_0^{(2)} & f_2 A_+ & 0 \\ 0 & 0 & A_- & A_0^{(3)} & f_3 A_+ \\ 0 & 0 & 0 & A_- & A_0^{(4)} \end{bmatrix}.$$

Suppose we were considering two structured QBD processes; one with linear forms for the scales associated with an increase in level and the other with quadratic forms for the scales associated with an increase in level. The likelihood ratio test statistic is

$$\lambda = -2\left(\log(L(\widehat{\boldsymbol{\theta}_1}; \mathbf{Y})) - \log(L(\widehat{\boldsymbol{\theta}_2}; \mathbf{Y}))\right),$$

where $\widehat{\boldsymbol{\theta}}_1 = \left(\widehat{A_0}, \widehat{A_-}, \widehat{A_+}, \widehat{\beta_1^f}\right)$ denotes the maximum likelihood estimates of the parameters for the structured QBD process with linear forms for the scales associated with an increase in level and $\widehat{\boldsymbol{\theta}}_2 = \left(\widehat{A_0}, \widehat{A_-}, \widehat{A_+}, \widehat{\beta_1^f}, \widehat{\beta_2^f}\right)$ denotes the maximum likelihood estimates of the parameters for the structured QBD process with quadratic forms for the scales associated with an increase in level.

Under the null hypothesis,

$$\lambda \sim \chi_1^2.$$

Therefore, we reject the null hypothesis if λ is greater than the $100(1-\alpha)$ percentile of χ_1^2, where α is a chosen significance level.

8.10 Numerical examples

In this section, we show how the EM algorithm for structured QBD processes performs with three numerical examples. In each numerical example, we simulate data from a structured QBD process with known parameters and then use the simulated data as input to the appropriate EM algorithm, implemented using R [43].

We start the EM algorithm with randomly generated initial values from a uniform distribution, $U(0,1)$ and we stop the EM algorithm once the relative difference between two successive log-likelihoods falls below 10^{-6}. We then compare the estimated structured QBD process to the true structured QBD process by considering the stationary distribution, conditional sojourn time before the level change, and transition probabilities for each level of each structured QBD process. Note that the comparison of stationary distributions and transient behaviour is by each level and not by phase due to the non-uniqueness of representation of structured QBD processes and interchangeability of phases for each estimated structured QBD process.

For each type of behaviour, the mean squared error (MSE) is calculated to demonstrate the overall accuracy. Let Y_ℓ denote the value of the true structured QBD process corresponding to level ℓ and $\widehat{Y}_\ell^{(i)}$ denote the value of the i^{th} estimated structured QBD process corresponding to level ℓ, for $\ell = L_{min}, \ldots, L_{max}$ where L_{min} and L_{max} are the minimum and maximum levels associated with the calculation, respectively. Then the mean squared error for each estimated structured QBD process, i, is calculated as

$$MSE = \frac{1}{L_{max} - L_{min} + 1} \sum_{\ell = L_{min}}^{L_{max}} \left(Y_\ell - \widehat{Y}_\ell^{(i)} \right)^2 .$$

For example, Y_ℓ could denote the upward transition probability for level ℓ of the true structured QBD process and $\widehat{Y}_\ell^{(i)}$ denote the estimated upward transition probability for level $\ell = L_{min}, \dots, L_{max}$ of the i^{th} estimated structured QBD process.

8.10.1 Structured QBD process with linearly increasing scales associated with a decrease in level

In many queueing systems, the service rate often depends on the workload. For example, the productivity of a bank teller will be relatively low when few customers are in the bank but as the number of customers increases, the productivity of the bank teller increases. This type of service reflects the service behaviour described in an M/M/∞ queue.

In this example, we consider a structured QBD process with linearly increasing scales associated with a decrease in level, where the infinitesimal generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+ & 0 & 0 & \dots \\ h_1 A_- & A_0^{(1)} & A_+ & 0 & \dots \\ 0 & h_2 A_- & A_0^{(2)} & A_+ & \dots \\ 0 & 0 & h_3 A_- & A_0^{(3)} & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix},$$

where

$$A_{0i,j} = \begin{bmatrix} * & 1 \\ 1 & * \end{bmatrix}, \quad A_{+i,j} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}, \quad A_{-i,j} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

$$h_\ell = \ell, \text{ for } \ell \geq 1,$$

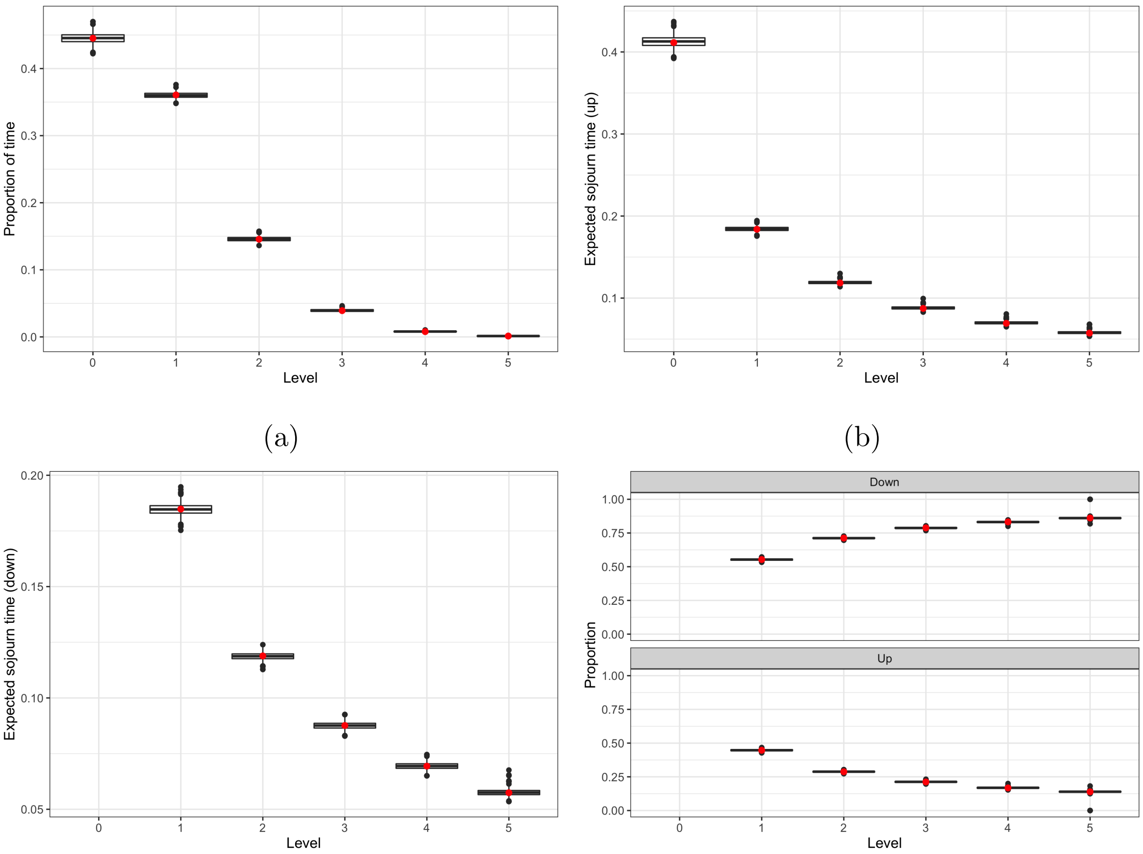
and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

We simulated 1000 samples of the above structured QBD process, each with 10,000 changes in level and used the simulated data as input into the EM algorithm described in Section 8.4. For both stationary and transient behaviour, the estimated structured QBD processes capture the behaviour observed in the true structured QBD process as demonstrated by the plots presented in Figure 8.10.1 and the small MSE values presented in Table 8.10.1.

Type	Mean	Variance
Stationary distribution	1.190×10^{-5}	2.023×10^{-10}
Sojourn time conditioned on moving up a level	9.964×10^{-6}	1.266×10^{-10}
Sojourn time conditioned on moving down a level	6.618×10^{-6}	2.246×10^{-11}
Transition probabilities	1.630×10^{-3}	1.581×10^{-7}

Table 8.10.1: Mean and variance of the MSE for each type of behaviour for the structured QBD process with linearly increasing scales associated with a decrease in level.

Similar to Chapter 7, we assessed the dependence on the initial values of the parameters by simulating a single sample of the above QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. Little difference was found between the stationary and transient behaviour of each estimated QBD process despite seeing variation in the complete data log-likelihood for each of the 1000 different structured QBD process estimations, as illustrated in Figure 8.10.2. Refer to Appendix C.1 for a summary of the analysis.



(a)

(b)

(c)

(d)

Figure 8.10.1: Comparison of the behaviour of the 1000 estimated structured QBD processes to the behaviour of the true structured QBD process with linearly increasing scales associated with a decrease in level, for the first six levels. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the estimated expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the estimated expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving down from level 0 are 0 and the transition probabilities for level 0 are omitted as they are either 0 or 1.

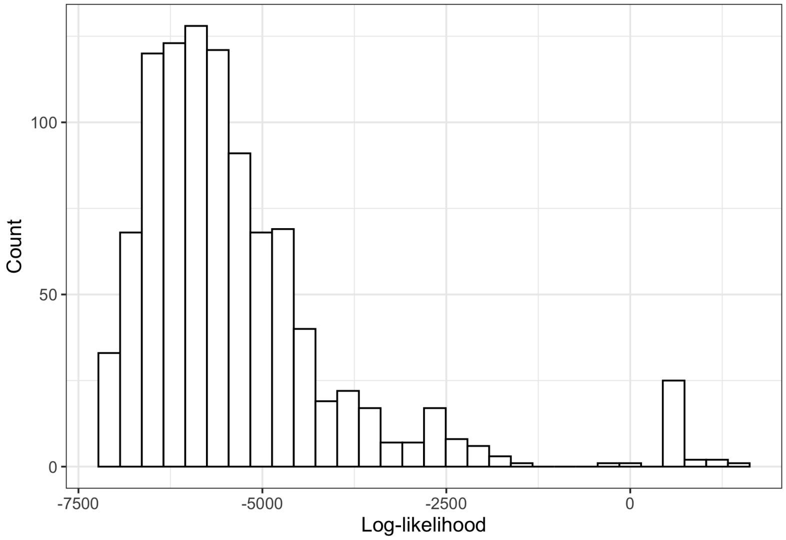


Figure 8.10.2: Log-likelihoods of the observed data for each of the fitted structured QBD process with linearly increasing scales associated with a decrease in level from the 1000 different starting values.

8.10.2 Structured QBD process with linearly decreasing scales associated with an increase in level

In other queueing systems, the rate of arrivals may depend on the amount of customers presently in the system. For example, customers may avoid the bank if they see that the bank is busy. Alternatively, they may be a finite number of customers attending a service facility, and so the rate of arrivals may decrease as service continues.

In this example, we consider a structured QBD process with linearly decreasing scales associated with an increase in level, where the infinitesimal generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & f_0 A_+ & 0 & 0 & \dots \\ A_- & A_0^{(1)} & f_1 A_+ & 0 & \dots \\ 0 & A_- & A_0^{(2)} & f_2 A_+ & \dots \\ 0 & 0 & A_- & A_0^{(3)} & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix},$$

where

$$A_{0i,j} = \begin{bmatrix} * & 1 \\ 1 & * \end{bmatrix}, \quad A_{+i,j} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}, \quad A_{-i,j} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

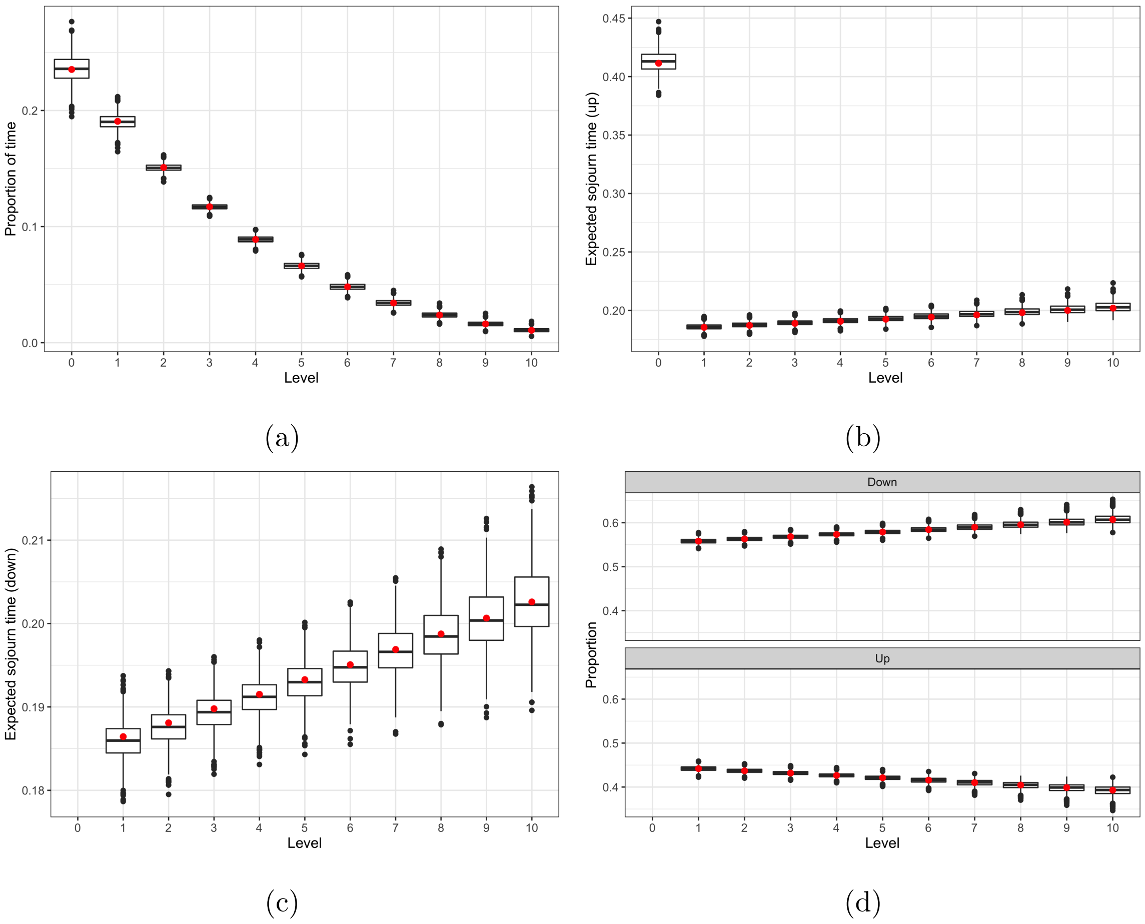
$$f_\ell = 1 - \frac{1}{50}\ell, \text{ for } \ell \geq 0,$$

and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

Similar to before, we simulated 1000 samples of the above QBD process, each with 10,000 changes in level and used the simulated data as input into the EM algorithm described in Section 8.4. Figure 8.10.3 and Table 8.10.2 compare the stationary and transient behaviour of the estimated structured QBD process to the true structured QBD process. In each case, there is similarity between the behaviours of the estimated structured QBD processes to the true structured QBD process.

Type	Mean	Variance
Stationary distribution	1.542×10^{-5}	4.059×10^{-10}
Sojourn time conditioned on moving up a level	2.608×10^{-5}	1.072×10^{-9}
Sojourn time conditioned on moving down a level	7.582×10^{-4}	2.424×10^{-8}
Transition probabilities	6.960×10^{-3}	1.752×10^{-6}

Table 8.10.2: Mean and variance of the MSE for each type of behaviour for the structured QBD process with linearly decreasing scales associated with an increase in level.



(a)

(b)

(c)

(d)

Figure 8.10.3: Comparison of the behaviour of the 1000 estimated structured QBD processes to the behaviour of the true structured QBD process with linearly decreasing scales associated with an increase in level for the first eleven levels. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the estimated expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the estimated expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving down from level 0 are 0 and the transition probabilities for level 0 are omitted as they are either 0 or 1.

Similar to before, we assessed the dependence on the initial values of the parameters by simulating a single sample of the above structured QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. Despite seeing variation in the estimated log-likelihood for each of the 1000 different structured QBD process estimations, we found little difference between the stationary and transient behaviour of each estimated QBD process, as illustrated in Figure 8.10.4. Refer to Appendix C.2 for a summary of the analysis.

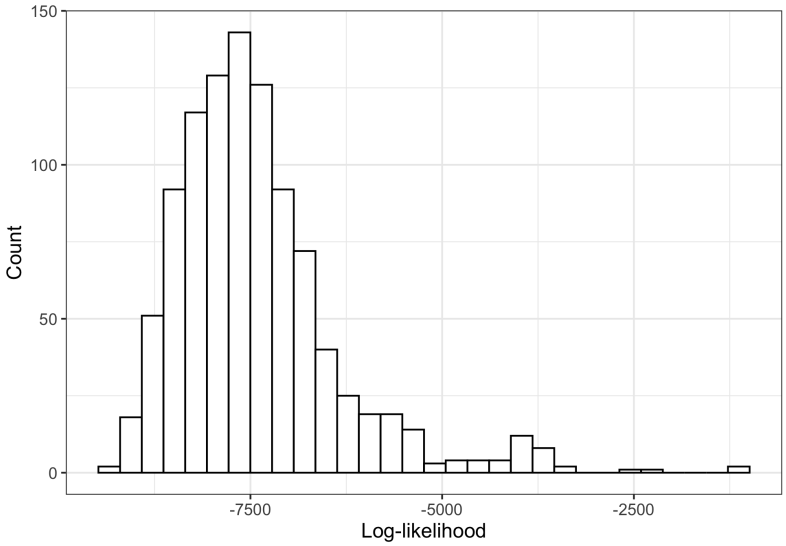


Figure 8.10.4: Log-likelihoods of the observed data for each of the fitted structured QBD process with linearly decreasing scales associated with an increase in level from the 1000 different starting values.

8.10.3 Structured QBD process with linearly decreasing scales associated with an increase in level and quadratic scales associated with a decrease in level

In this example, we consider a structured QBD process with linearly decreasing scales associated with an increase in level and quadratic scales associated with a decrease in level, where the infinitesimal generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & f_0 A_+ & 0 & 0 & 0 \\ h_1 A_- & A_0^{(1)} & f_1 A_+ & 0 & 0 \\ 0 & h_2 A_- & A_0^{(2)} & f_2 A_+ & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & h_9 A_- & A_0^{(9)} & f_9 A_+ \\ 0 & 0 & 0 & h_{10} A_- & A_0^{(10)} \end{bmatrix},$$

$$A_{0i,j} = \begin{bmatrix} * & 1 \\ 1 & * \end{bmatrix}, \quad A_{+i,j} = \begin{bmatrix} 2 & 2 \\ 1 & 2 \end{bmatrix}, \quad A_{-i,j} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},$$

$$f_\ell = 1 - \frac{1}{50}\ell, \text{ for } 0 \leq \ell \geq 9,$$

$$h_\ell = 1 + 1.8 \times (\ell - 1) - 0.1 \times (\ell - 1)^2, \text{ for } 1 \leq \ell \geq 10,$$

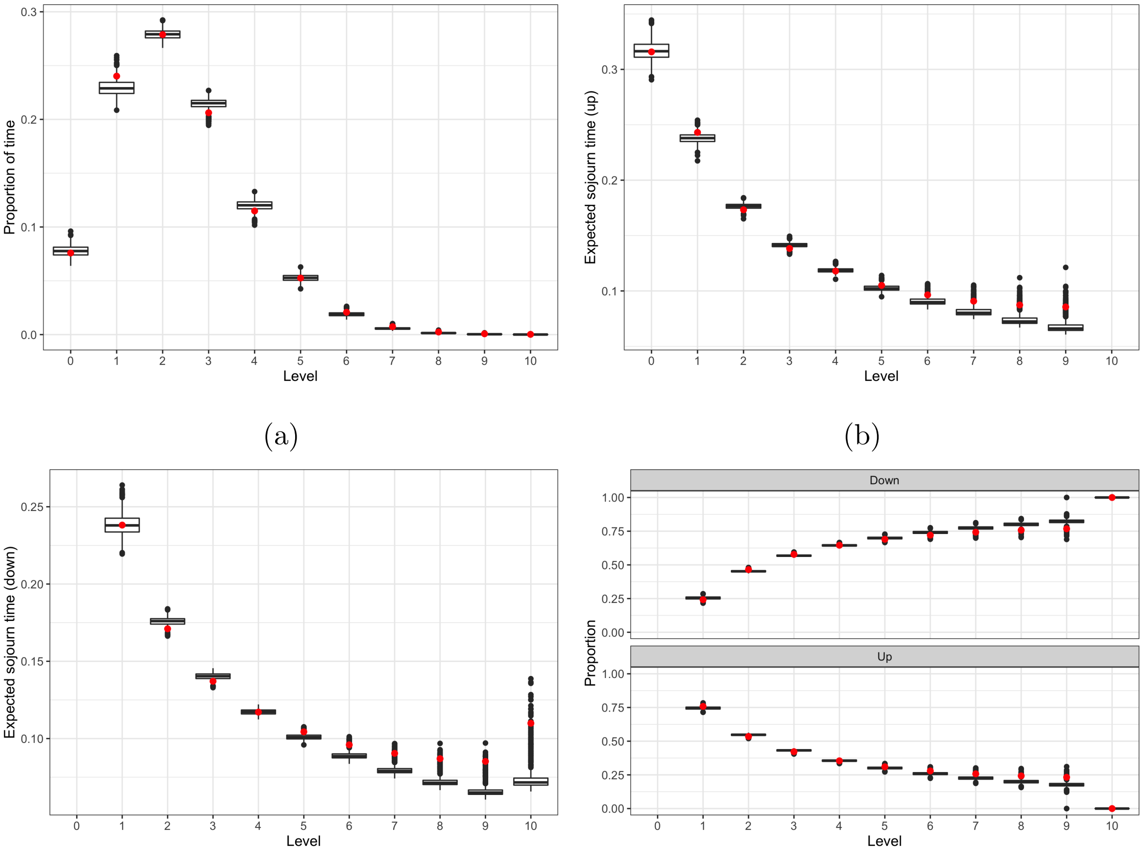
and the diagonal entries of $A_0^{(\ell)}$ for $\ell \geq 0$ are the negative of the relevant row sums.

Such a queueing model may reflect behaviours where the server's productivity is low when the system is near empty (due to limited work available) and close to capacity (due to being overworked and stressed).

Similar to before, we simulated 1000 samples of the above structured QBD process, each with 10,000 changes in level. In this example, we first use the simulated data as input into the EM algorithm described in Section 8.4. Figure 8.10.5 compares the stationary and transient behaviour of the estimated structured QBD process to the true structured QBD process. These results illustrate the need for a quadratic term, as the stationary and transient behaviour are poorly estimated.

Type	Mean	Variance
Stationary distribution	3.569×10^{-5}	6.628×10^{-10}
Sojourn time conditioned on moving up a level	7.950×10^{-5}	1.024×10^{-9}
Sojourn time conditioned on moving down a level	1.863×10^{-4}	3.134×10^{-9}
Transition probabilities	7.066×10^{-4}	1.826×10^{-7}

Table 8.10.3: Mean and variance of the MSE for each type of behaviour for the structured QBD process with linearly decreasing scales associated with an increase in level and linear scales associated with a decrease in level.



(a)

(b)

(c)

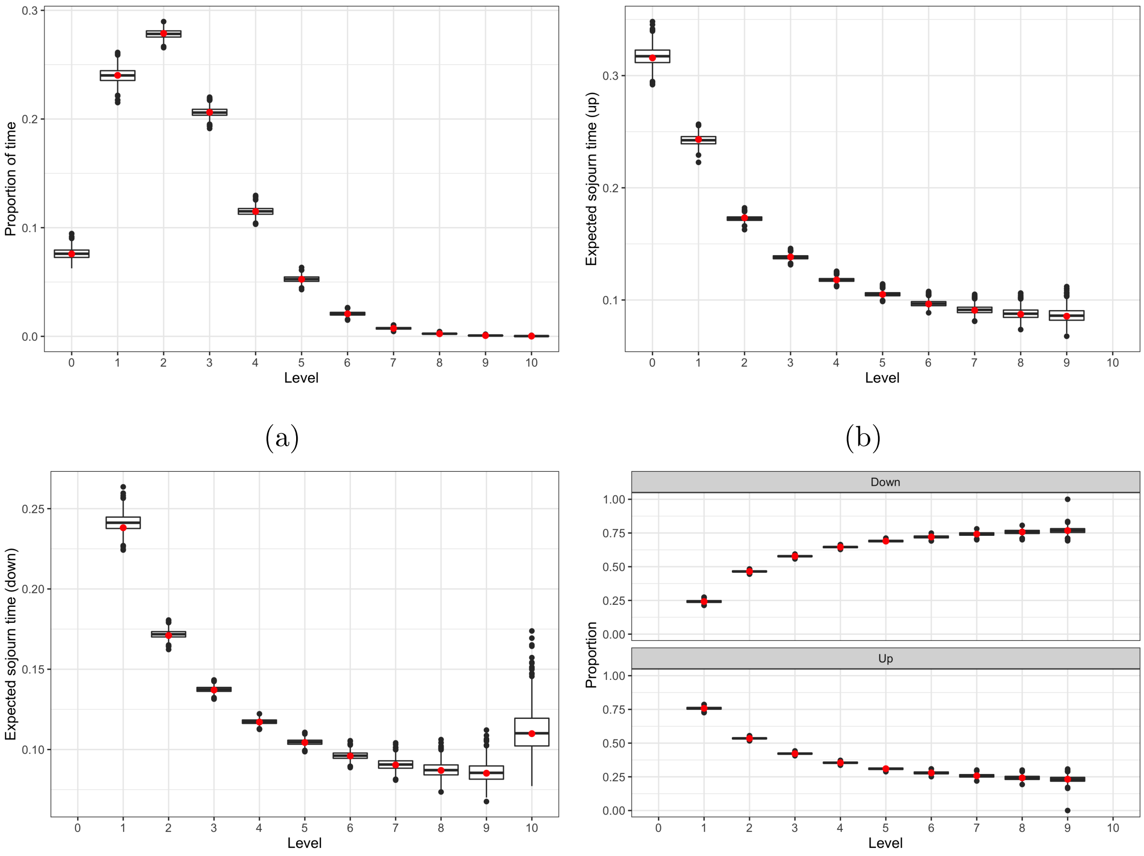
(d)

Figure 8.10.5: Comparison of the behaviour of the 1000 estimated structured QBD processes to the behaviour of the true structured QBD process with linearly increasing scales associated with an increase in level and quadratic scales associated with a decrease in level. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the estimated expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the estimated expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving up from level 10 are 0, the sojourn times conditioned on moving down from level 0 are 0, and the transition probabilities for levels 0 and 10 are omitted as they are either 0 or 1.

We now use the simulated data as input into the EM algorithm described in Section 8.5. Of particular interest is the improvement in capturing the stationary and transient behaviour of the true structured QBD process, as demonstrated in Figure 8.10.6 and Table 8.10.3.

Type	Mean	Variance
Stationary distribution	1.232×10^{-5}	1.824×10^{-10}
Sojourn time conditioned on moving up a level	1.688×10^{-5}	2.819×10^{-10}
Sojourn time conditioned on moving down a level	2.722×10^{-5}	1.507×10^{-9}
Transition probabilities	1.160×10^{-4}	4.525×10^{-8}

Table 8.10.4: Mean and variance of the MSE for each type of behaviour for the structured QBD process with linearly decreasing scales associated with an increase in level and quadratic scales associated with a decrease in level.



(a)

(b)

(c)

(d)

Figure 8.10.6: Comparison of the behaviour of the 1000 estimated structured QBD processes to the behaviour of the true structured QBD process with linearly increasing scales associated with an increase in level and quadratic scales associated with a decrease in level. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (red points). (b) Box-plots of the estimated expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (red points). (c) Box-plots of the estimated expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (red points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (red points). By design, the sojourn times conditioned on moving up from level 10 are 0, the sojourn times conditioned on moving down from level 0 are 0, and the transition probabilities for levels 0 and 10 are omitted as they are either 0 or 1.

A more statistical approach to demonstrate the improvement in capturing the stationary and transient behaviour is to use the likelihood ratio test. Denote $\mathcal{M}_1$ as the model with linearly decreasing scales associated with moving down a level and let $\mathcal{M}_2$ as the model with quadratic scales associated with moving down a level. The observed value of the likelihood ratio test statistic is

$$\begin{aligned}\lambda &= -2\left(\log(L(\widehat{\boldsymbol{\theta}}_1;\mathbf{Y})) - \log(L(\widehat{\boldsymbol{\theta}}_2;\mathbf{Y}))\right)\\ &= -2(-7159.59 + 1048.13)\\ &= 12222.92,\end{aligned}$$

where $\widehat{\boldsymbol{\theta}}_1 = \left(\widehat{A_0}, \widehat{A_-}, \widehat{A_+}, \widehat{\beta_1^f}\right)$ denotes the maximum likelihood estimate of the parameters for the structured QBD process with linearly decreasing scales associated with a decrease in level and $\widehat{\boldsymbol{\theta}}_2 = \left(\widehat{A_0}, \widehat{A_-}, \widehat{A_+}, \widehat{\beta_1^f}, \widehat{\beta_2^f}\right)$ for the structured QBD process with quadratic forms for the scales associated with a decrease in level. Note that the model with the highest observed log-likelihood of the data was used in each case.

Under the null hypothesis, $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$,

$$\lambda \sim \chi_1^2,$$

which leads to a p-value approximately equal to zero. Therefore, we reject the null hypothesis and conclude that the structured QBD process with quadratic scales associated with a decrease in level is the better suited model, as expected.

Similar to before, we assessed the dependence on the initial values of the parameters by simulating a single sample of the above structured QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. After an investigation, there was little difference between the stationary and transient behaviour of each estimated QBD process despite seeing bi-modal behaviour in the estimated log-likelihood for each of the 1000 different structured QBD process estimations, as illustrated in Figure 8.10.7. Refer to Appendix C.3 for a summary of the analysis.

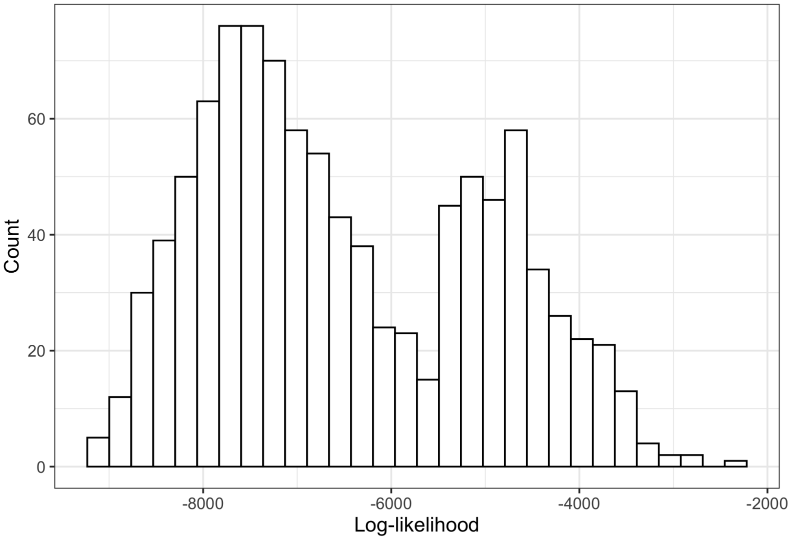


Figure 8.10.7: Log-likelihoods of the observed data for each of the fitted structured QBD process with linearly decreasing scales associated with an increase in level and quadratic scales associated with a decrease in level from the 1000 different starting values.

8.11 Summary

In this chapter, we have developed a new type of QBD process which offers a reduction in the number of parameters whilst remaining level-dependent in nature. In doing so, we are able to predict behaviours of a queueing system without using an over-parameterised queueing model.

While the development of structured QBD processes expands the range of models that can be used to model queueing systems with dependence, there remains a need to develop statistical methods to assess the fit of such queueing models to queueing system data. In the next chapter, we review goodness of fit tests in a statistical framework and then develop a goodness of fit test that assesses the suitability of any QBD process to queueing system data.

Chapter 9

Goodness of fit

Assessing the fit of a model to data is crucial before using such a model to predict behaviours of a process. For QBD processes, there are two different types of behaviour; stationary and transient. Stationary behaviour relates to the long term behaviour of the queueing system, such as the long term probability of being in each level, and transient behaviour relates to the short term behaviour, such as conditional sojourn times in each level and the transitions between levels.

In this chapter, we first review distance measures in a statistical setting and then develop a goodness of fit test of QBD processes to queueing system data, as well as diagnostic plots to be used in situations where the QBD process is not adequate. We then demonstrate how our goodness of fit test performs with several examples and complete a small simulation study to explore the statistical power and significance.

9.1 Distance measures

Distance measures greatly contribute to the development of hypothesis tests and goodness of fit tests as they quantify the distance between two statistical objects, such as two random variables, two probability distributions or an individual sample point and a probability distribution [22].

A function $d : X \times X \to \mathbb{R}^+$, where $\mathbb{R}^+ = \{x \in \mathbb{R} : x \geq 0\}$, is distance measure on the set X if for all $x, y, z \in X$,

- $d(x, y) \geq 0$,
- $d(x, y) = 0$ if and only if $x = y$,
- $d(x, y) = d(y, x)$, and
- $d(x, z) \leq d(x, y) + d(y, z)$.

9.1.1 Euclidean distance

Consider the distance between two numbers $x, y \in \mathbb{R}$. The distance between numbers x and y is given by

$$d_E(x, y) = |x - y|,$$

which satisfies the conditions of a distance measure defined above.

This distance measure is known as the Euclidean distance in one dimension, which is equivalently defined as

$$d_E(x, y) = \sqrt{(x - y)^2}.$$

In two dimensions, the Euclidean distance between points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ is defined as

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$. This is also illustrated in Figure 9.1.1.

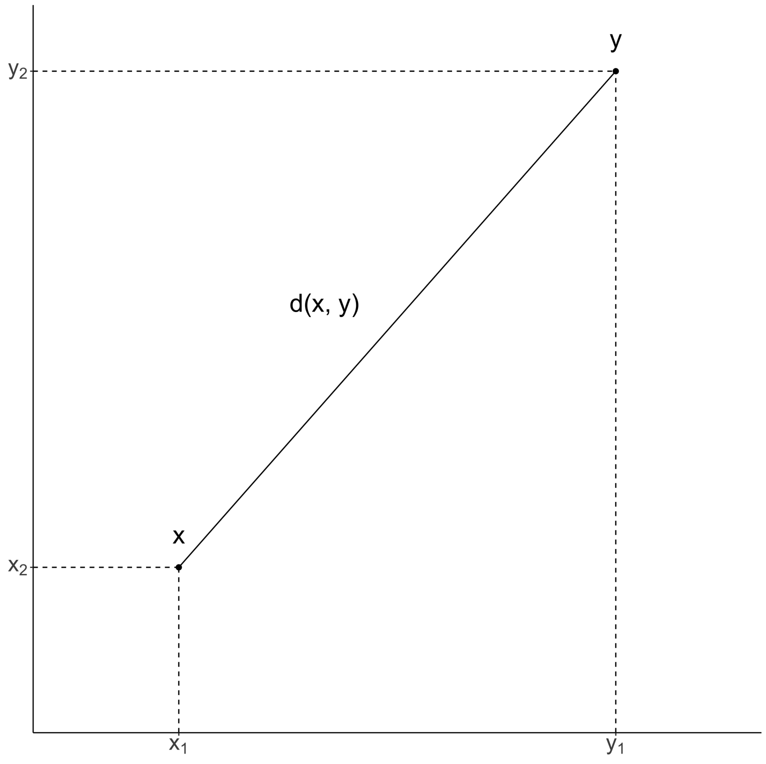


Figure 9.1.1: Example of the Euclidean distance in two dimensions.

Generalising this to higher dimensions, we have that for points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the Euclidean distance is defined as

$$\begin{aligned} d_E(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}, \\ &= \sqrt{(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})}, \end{aligned}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

Suppose we are interested in calculating the distance between the individual sample point (blue) and the distribution of data (red points) shown in Figure 9.1.2.

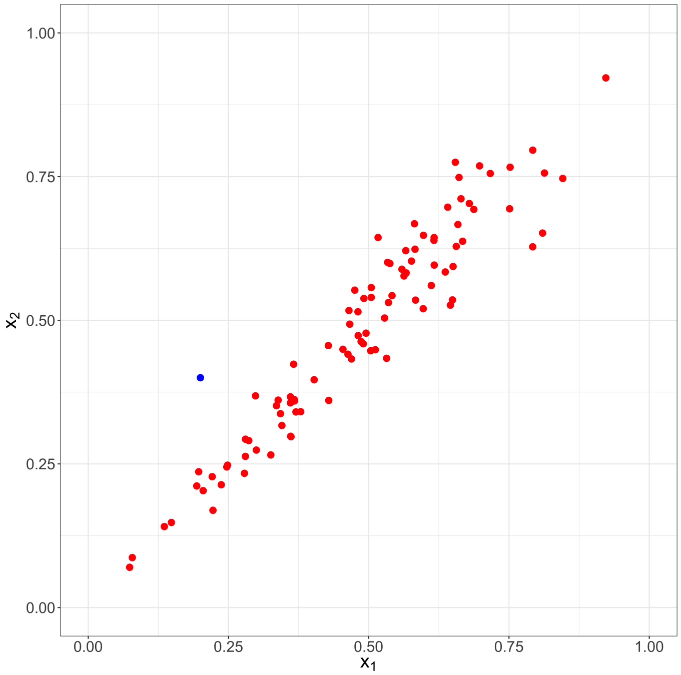
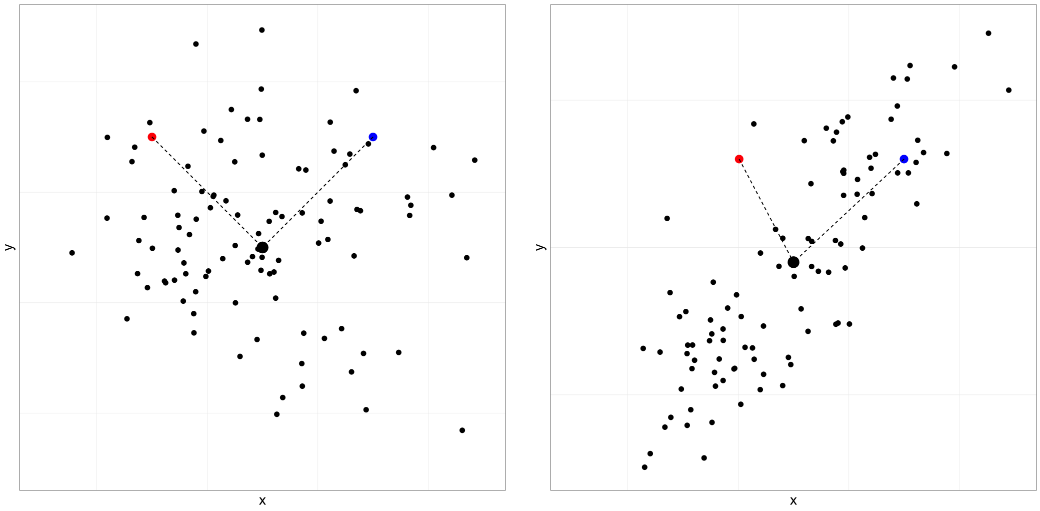


Figure 9.1.2: Example of two dimensional data.

In the Euclidean setting, the distance between the individual sample point, $\mathbf{x} \in \mathbb{R}^2$, and a distribution of data is

$$d_E(\mathbf{x}, \bar{\mathbf{x}}) = \sqrt{(\mathbf{x} - \bar{\mathbf{x}})^T(\mathbf{x} - \bar{\mathbf{x}})},$$

where $\bar{\mathbf{x}}$ is the sample mean of the data. However, this calculation does not take the positive correlation between x_1 and x_2 into account. For example, consider uncorrelated and correlated data presented in Figures 9.1.3a and 9.1.3b. The red and blue points are equally distant from the centre of the distribution in terms of Euclidean distance for both the uncorrelated and correlated data sets. However, it is clear that the blue point is more typical of the correlated data than the red point.



(a) Uncorrelated data.

(b) Correlated data.

This example illustrates that the Euclidean distance is just a measure of distance between two points and it does not consider correlation between variables or the variability present in the data. As a result, the Euclidean distance is not usually used to determine how close a point is to a distribution of data.

9.1.2 Mahalanobis distance

The Mahalanobis distance measure is used to measure the distance between an individual sample point and a probability distribution and is typically used in multivariate outlier problems [11]. The Mahalanobis distance is a scale invariant and unitless distance measure and provides a measure of distance between a point $\mathbf{x} \in \mathbb{R}^p$ generated from a p-variate probability distribution $f_{\mathbf{X}}(\cdot)$ and $\boldsymbol{\mu} = E(\mathbf{X})$.

First, we provide the formal definition of the Mahalanobis distance [11],

$$d_{\mathbf{\Sigma}}(\mathbf{X}, \boldsymbol{\mu}) = \sqrt{(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})},$$

which provides a measure of distance between a $p \times 1$ random vector $\mathbf{X}$ and it's mean vector $E[\mathbf{X}] = \boldsymbol{\mu}$, where $E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix. Note that if $\mathbf{\Sigma} = I$, such that there is no correlation between variables and each variable has unit-variance, then the Mahalanobis distance reduces to the Euclidean distance.

A special property of the Mahalanobis distance is that if $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ is an m-variate normal random vector such that $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is invertible, then the square Mahalanobis distance $d(\mathbf{X}, \boldsymbol{\mu}, \mathbf{\Sigma})^2$ is distributed as a χ^2 variable with n degrees of freedom [26]. This result is particularly useful in determining whether a point is identified as an outlier, such that points that do not fit the distribution of data will have a squared Mahalanobis distance lying above a pre-defined cut-off value of the χ^2 distribution with p degrees of freedom [20].

When the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ are unknown, the sample mean $\bar{\mathbf{X}}$ and the sample covariance matrix $\mathbf{S}$ are used as estimators of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ instead. For example, suppose that $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ are independent realisations of the random vector $\mathbf{X}$. Then the sample mean vector is calculated as

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i,$$

and the sample covariance matrix is calculated as

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T.$$

In this case, the Mahalanobis distance is defined as

$$d_{\mathbf{S}}(\mathbf{X}, \bar{\mathbf{X}}) = \sqrt{(\mathbf{X} - \bar{\mathbf{X}})^T \mathbf{S}^{+} (\mathbf{X} - \bar{\mathbf{X}})},$$

where $\mathbf{S}^{+}$ is the inverse of the sample covariance matrix calculated using the Moore-Penrose pseudo-inverse [5].

The Moore-Penrose pseudo-inverse is calculated using the singular value decomposition method, where if a matrix $A \in \mathbb{R}^{n \times n}$ has the singular value decomposition $U\Sigma V^T$, then

$$A^{+} = V\Sigma^{+}U^T.$$

Here, $\Sigma \in \mathbb{R}^{n \times n}$ is a diagonal matrix of ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$,

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix},$$

where Σ^{+} is calculated by taking the reciprocal of all non-zero elements,

$$\Sigma^{+} = \begin{bmatrix} 1/\lambda_1 & 0 & \ldots & 0 \\ 0 & 1/\lambda_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\lambda_n \end{bmatrix},$$

and $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{n \times n}$ contain the left and right eigenvectors, respectively.

If the matrix A is invertible, such that the determinant is non-zero, then the pseudo-inverse of A is equal to the usual matrix inverse of A, A^{-1}.

If A is a singular matrix with rank $q < n$, then the last $p = n - q$ eigenvalues will be zero. Therefore, the eigenvectors corresponding to non-zero eigenvalues form a basis for the range of A and the eigenvectors corresponding to the zero eigenvalues form a basis for the null space of A. In this case, $\mathbf{\Sigma}^+$ is calculated as

$$\Sigma^+ = \begin{bmatrix} 1/\lambda_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1/\lambda_q & 0 & \dots & \vdots \\ 0 & \dots & \dots & 0 & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \dots & 0 \end{bmatrix}.$$

The use of the Moore-Penrose pseudo-inverse may be problematic in situations where the covariance matrix, $\mathbf{\Sigma}$, is singular. For example, consider the distance between the point $\mathbf{x} \in \mathbb{R}^2$ and the sample mean $\bar{\mathbf{x}}$, where the sample covariance matrix $\mathbf{S}$ is singular. Suppose $\mathbf{x}$ can be written as a linear combination of the eigenvectors of $\mathbf{S}$, such that

$$\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2,$$

where $a_1, a_2 > 0$. If $\lambda_2 < 10^{-10}$, for example, then

$$\Sigma^+ = \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 0 \end{bmatrix},$$

which suggests that the Mahalanobis distance between $\mathbf{x}$ and $\bar{\mathbf{x}}$,

$$d_{\mathbf{S}}(\mathbf{x}, \bar{\mathbf{x}}) = \sqrt{(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^+ (\mathbf{x} - \bar{\mathbf{x}})},$$

is unreliable, since critical information associated with the second eigenvalue and eigenvector will be discounted.

For example, consider two variables $X \sim N(0, 1)$ and $Y = X + Z$, where $Z \sim N(0, \sigma)$ for some $\sigma > 0$. As σ gets smaller in value, the second eigenvalue of the covariance matrix decreases in magnitude, as illustrated in Figure 9.1.4.

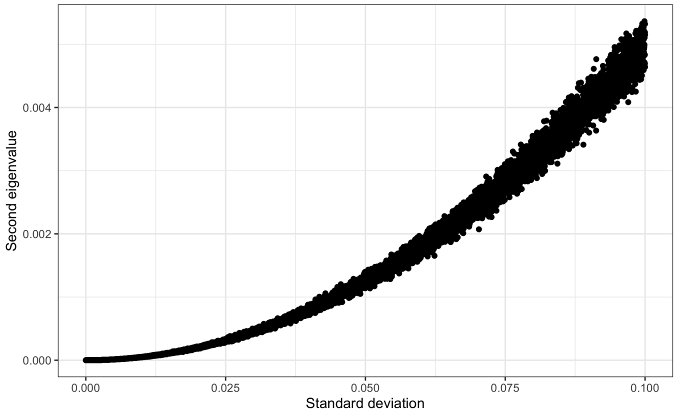
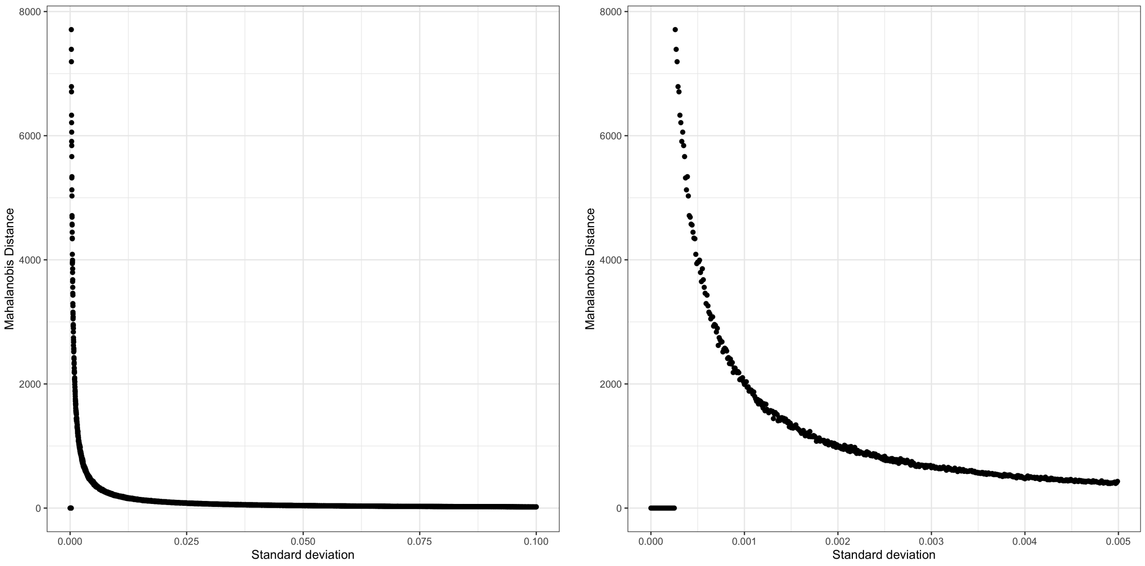


Figure 9.1.4: Change in the value of the second eigenvalue of the covariance matrix as σ decreases.

Given that Σ^+ is calculated by taking the reciprocal of all non-zero elements, the magnitude of the Mahalanobis distance increases as σ decreases. However, there is a sharp decline in the value of the Mahalanobis distance when the second eigenvalue becomes too small which incorrectly suggests that the point (1, -1) belongs to the bivariate distribution of X and Y, as illustrated in Figure 9.1.5.



(a) (b)

Figure 9.1.5: Change in the value of the Malahanobis distance between the bivariate distribution of X and Y and the three points, (0, 0), (1, -1), and (2, 1) as σ decreases. Note that plot (b) zooms in on the sharp decline in the value of the Mahalanobis distance.

Therefore, caution should be taken when using the Moore-Penrose pseudo-inverse to calculate the Mahalanobis distance, such that critical information is not lost when comparing a point $\mathbf{x} \in \mathbb{R}^n$ generated from an n-variate probability distribution $f_{\mathbf{X}}(\cdot)$ to $\boldsymbol{\mu} = E(\mathbf{X})$.

9.1.3 Distance between two probability distributions

Lastly, we also consider how well a continuous distribution fits a set of observations. That is, we are interested in calculating the distance between an empirical cumulative distribution function and a known cumulative distribution function.

For example, let $x_1, x_2, \ldots, x_n$ be observations of continuous random variables $X_1, X_2, \ldots, X_n$ with a cumulative distribution function, F_0. If F_0 is a known cumulative distribution function, the Kolmogorov-Smirnov statistic [37] measures the distance between an empirical cumulative distribution function $\hat{F}(x)$ defined as

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

and F_0, such that

$$D_n = \max_x |\hat{F}(x) - F_0(x)|.$$

9.1.4 Discussion

Distance measures provide a measure of the distance between statistical objects and are important to the development of goodness of fit tests. In the following section, we develop a goodness of fit test to assess the fit of a QBD process to queueing system data using the distance measures described above.

9.2 Goodness of fit test

To assess the fit of a QBD process to queueing system data, we must focus on whether the stationary and transient behaviour observed in the data is captured by the QBD process of interest. In particular, we focus on the stationary distribution, conditional sojourn times, and transitions between levels. Note that since observed queueing system data only consists of the changes in level and the times at which those changes in level occurred, the goodness of fit test developed in this section will compare the QBD process to the observed queueing system data by each level and not by phase.

9.2.1 Null and alternative hypotheses

Let X represent a sample trajectory of a ergodic, time-homogeneous, random process, such that it contains information on the changes in level and the times at which those changes occurred. For example, X may represent a sample trajectory of a queueing process, such that it describes the changes in the number of customers in the queueing system and the times at which those changes occurred.

The stationary and transient behaviour of X relate to the long term and short term behaviour of the process, respectively. More specifically, we are interested in four specific types of behaviour.

- O_ℓ: the long term proportion of time the process X spends in level $\ell \geq 0$.
- R_ℓ: an indicator variable which describes a transition out of level ℓ, such that

$$R_\ell = \begin{cases} 1 & \text{if the process } X \text{ moves from level } \ell \geq 0 \text{ to level } \ell + 1, \\ 0 & \text{if the process } X \text{ moves from level } \ell \geq 1 \text{ to level } \ell - 1. \end{cases}$$

- S_ℓ^-: the distribution of time spent in level $\ell \geq 1$ before the process moves down from level $\ell \geq 1$.
- S_ℓ^+: the distribution of time spent in level $\ell \geq 0$ before the process moves up from level $\ell \geq 0$.

The goodness of fit test developed in this section addresses the question of whether observed data $\tilde{X}$ can be modelled as a realisation of a QBD process $\mathcal{Q}$, where $\mathcal{Q}$ belongs to a given class of QBD processes Ω. More specifically, we are interested in whether a QBD process $\mathcal{Q} \in \Omega$ captures both the stationary and transient behaviour observed in data $\tilde{X}$. Hence, we define g as the set of expected behaviours,

$$g = \{\boldsymbol{\rho}, \{p_\ell^+; \ell \geq 0\}, \{F_\ell^-; \ell \geq 0\}, \{F_\ell^+; \ell \geq 0\}\},$$

where, conditional on some sample trajectory X of $\mathcal{Q} \in \Omega$ and assuming a maximum observed level of L,

- $\boldsymbol{\rho} = E[\mathbf{O}]$ is the expected level frequencies for $0 \leq \ell \leq L$, where $\mathbf{O} = (O_0, O_1, \ldots, O_L)$ and $\boldsymbol{\rho} = (\rho_0, \rho_1, \ldots, \rho_L)$,
- $\mathbf{p}^+ = E[R]$ is the expected proportion of times the process X moves up from each level $1 \leq \ell \leq L$, where $\mathbf{p}^+ = (p_1^+, p_2^+, \ldots, p_L^+)$,
- F_ℓ^- represents the cumulative distribution function of S_ℓ^- for $\ell \geq 1$, and
- F_ℓ^+ represents the cumulative distribution function of S_ℓ^+ for $\ell \geq 0$.

Therefore, our null and alternative hypotheses are defined as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

and

$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

respectively.

Here, the function $f(\mathcal{Q})$ returns the required stationary and transient behaviour of any given QBD process, $\mathcal{Q}$, such that

$$f(\mathcal{Q}) : \mathcal{Q} \to G_{\mathcal{Q}},$$

where $G_{\mathcal{Q}}$ represents the set of all behaviours of QBD processes $\mathcal{Q} \in \Omega$.

9.2.2 Multivariate setting

Summarising the stationary and transient behaviour into a single object, such as a vector, is problematic since the stationary distribution and transition probabilities are contained within a vector but the conditional sojourn times are summarised by continuous distributions.

Instead, we define $\mathbf{V}(\tilde{X}) = \{\mathbf{V}_{\tilde{\boldsymbol{\pi}}}, \mathbf{V}_{\mathbb{P}^+}, \mathbf{V}_{\boldsymbol{\tau}^-}, \mathbf{V}_{\boldsymbol{\tau}^+}\}$ to be an n-variate random vector which quantifies the difference between the stationary and transient behaviour of observed data $\tilde{X}$ and the expected stationary and transient behaviour $\hat{\mathbf{g}}$, where $\hat{\mathbf{g}}$ is estimated from a QBD process $\widehat{\mathcal{Q}} \in \Omega$ fitted to $\tilde{X}$.

That is, for stationary behaviour,

$$\mathbf{V}_{\tilde{\boldsymbol{\pi}}} = \mathbf{O} - \boldsymbol{\rho},$$

where $\mathbf{O} = (O_0, O_1, \ldots, O_L)$, such that O_ℓ is the observed proportion of time spent in each level $\ell \geq 0$, and $[\boldsymbol{\rho}]_\ell$ is the estimated proportion of time spent in each level $\ell \geq 0$ from $\hat{\mathbf{g}}$. Note that the last element of this vector is omitted since the stationary vector is a probability vector and hence has $L - 1$ degrees of freedom, assuming L is the maximum observed level.

For transient behaviour relating to the transition probability of moving up from level $\ell \geq 0$ to level $\ell + 1$,

$$\mathbf{V}_{\mathbb{P}^+} = \mathbf{P} - \mathbb{P}^+,$$

where $\mathbf{P} = (P_1, P_2, \ldots, P_L)$, such that P_ℓ is the observed proportion of times the process moves up from level $\ell \geq 1$ to level $\ell + 1$, and $[\mathbb{P}^+]_\ell$ contains the estimated transition probability of moving up from level $\ell \geq 1$ to level $\ell + 1$ from $\hat{\mathbf{g}}$. Note that the transition probability of moving up from level 0 to level 1 is 1 by design and hence is omitted from the goodness of fit test.

For transient behaviour relating to the sojourn time conditioned on moving down from level $\ell \geq 1$, we consider a signed Kolmogorov-Smirnov statistic defined as

$$\mathbf{V}_{\boldsymbol{\tau}^-} = \mathrm{sign}(\hat{F}_\ell^-(x^*) - F_\ell^-(x^*))|\hat{F}_\ell^-(x^*) - F_\ell^-(x^*)|,$$

where $x^* = \mathrm{argmax}_x|\hat{F}_\ell^-(x) - F_\ell^-(x)|$, $\hat{F}_\ell^-$ is the empirical distribution of the sojourn time conditioned on moving down from level $\ell \geq 1$, $F_\ell^-(x)$ is the estimated distribution of the sojourn time conditioned on moving down from level $\ell \geq 1$ from $\hat{\mathbf{g}}$, and

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Lastly, for transient behaviour relating to the sojourn time conditioned on moving up from level $\ell \geq 0$, we again consider a signed Kolmogorov-Smirnov statistic defined as

$$\mathbf{V}_{\boldsymbol{\tau}^+} = \mathrm{sign}(\hat{F}_\ell^+(x^*) - F_\ell^+(x^*))|\hat{F}_\ell^+(x^*) - F_\ell^+(x^*)|,$$

where $x^* = \mathrm{argmax}_x|\hat{F}_\ell^+(x) - F_\ell^+(x)|$, $\hat{F}_\ell^+$ is the empirical distribution of the sojourn time conditioned on moving up from level $\ell \geq 0$ and $F_\ell^+(x)$ is the estimated distribution of the sojourn time conditioned on moving up from level $\ell \geq 0$ from $\hat{\mathbf{g}}$.

9.2.3 Test statistic

Using the Mahalanobis distance, we define the test statistic as

$$d_{\boldsymbol{\Sigma}}(\mathbf{V}(\tilde{X}), \boldsymbol{\mu}) = \sqrt{(\mathbf{V}(\tilde{X}) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{V}_{\mathcal{Q}}(\tilde{X}) - \boldsymbol{\mu})},$$

where $\boldsymbol{\mu} = E[\mathbf{V}(X)]$, $\boldsymbol{\Sigma} = E[(\mathbf{V}(X) - \boldsymbol{\mu})(\mathbf{V}(X) - \boldsymbol{\mu})^T]$, and X is a trajectory of $\mathcal{Q} \in \Omega$.

Given that the differences associated with conditional sojourn times are calculated using a modified version of the Kolmogorov-Smirnov statistic, the distribution of $\mathbf{V}(X)$ is not multivariate normal (see Appendix D for an illustrative example). Hence, we cannot assume that $d_{\mathbf{\Sigma}}(\mathbf{V}(X), \boldsymbol{\mu})^2$ follows a Chi-Square distribution with p degrees of freedom. Instead, we consider empirical estimation.

9.2.4 Parametric bootstrap

We empirically estimate the distribution of $\mathbf{V}(X)$, where X is a trajectory from $\widehat{\mathcal{Q}} \in \Omega$, using the parametric bootstrap method, which involves simulating data from a model that has been fitted to observed data [17, 16]. The method to calculate the empirical cumulative distribution function, $F_D(t)$ for $t \geq 0$, is described below.

1. Fit a QBD process, $\widehat{\mathcal{Q}} \in \Omega$ to data X using maximum likelihood techniques.

2. Simulate D data sets of K level changes, X_j, for $j = 1, \ldots, D$, from the fitted QBD process, $\widehat{\mathcal{Q}}$.

3. Fit a QBD process, $\widehat{\mathcal{Q}}_j \in \Omega$, for $j = 1, ..., D$, to each simulated data set using an appropriate method from Chapter 7 or 8.

4. Calculate the distance vector $\mathbf{v}(X_j) = \{\mathbf{v}_{\tilde{\boldsymbol{\pi}}}, \mathbf{v}_{\mathbb{P}^+}, \mathbf{v}_{\boldsymbol{\tau}^-}, \mathbf{v}_{\boldsymbol{\tau}^+}\}$ which quantifies the difference between the stationary and transient behaviour of simulated data X_d to that expected under the associated fitted QBD process $\widehat{\mathcal{Q}}_j \in \Omega$.

5. Calculate the Mahalanobis distance for each distance vector $\mathbf{v}(X_j)$, for $j = 1, \ldots, D$,

   $$d_{\mathbf{S}^{|j}}(\mathbf{v}(X_j), \bar{\mathbf{x}}^{|j}) = \sqrt{(\mathbf{v}(X_j) - \bar{\mathbf{x}}^{|j})^T \mathbf{S}^{|j^+}(\mathbf{v}(X_j) - \bar{\mathbf{x}}^{|j})},$$

   where $\bar{\mathbf{x}}^{|j}$ is the sample mean defined as

   $$\bar{\mathbf{x}}^{|j} = \frac{1}{D} \sum_{\substack{i=1 \\ i \neq j}}^{D} \mathbf{v}(X_i),$$

and $\mathbf{S}^{|j}$ is the sample covariance matrix defined as

$$\mathbf{S}^{|j} = \frac{1}{D-2}\sum_{\substack{i=1\\ i\neq j}}^{D}((\mathbf{v}(X_i) - \bar{\mathbf{x}}^{|j})(d(\mathbf{v}(X_i) - \bar{\mathbf{x}}^{|j})^T.$$

6. Calculate the empirical cumulative distribution function,

$$F_D(t) = \frac{1}{D}\sum_{j=1}^{D} I(d_{\mathbf{S}^{|j}}(\mathbf{v}(X_j), \bar{\mathbf{x}}^{|j}) \leq t).$$

9.2.5 P-value

We empirically estimate the p-value as

$$\text{p-value} = \frac{1}{D}\sum_{j=1}^{D} I\left(d_{\mathbf{S}^{|j}}(\mathbf{v}(X_j), \bar{\mathbf{x}}^{|j}) > d_{\mathbf{S}}(\mathbf{v}(X), \bar{\mathbf{x}})\right),$$

where $d(\mathbf{v}(X), \bar{\mathbf{x}}, \mathbf{S})$ is the observed value of the test statistic. If the p-value is less than α, then we reject H_0 at the $\alpha\%$ level of significance. Otherwise, we retain H_0.

9.2.6 Preliminary checks and level aggregation

Firstly, we note that a QBD process can be discounted if it structurally unable to produce the same behaviour to that observed in the data. For example, if the observed queueing system data has a maximum level of 10, then the maximum level of the fitted QBD process must be at least 10. A fitted QBD process with less than the required number of levels will not fully capture the behaviour of the queueing process.

Secondly, inherent with fitting some QBD processes, such as infinite QBD processes, to observed queueing system data is the possibility of visiting levels higher than that actually observed in the data. For example, the maximum level observed in the data may be 10 but an infinite QBD process may allow transitions to levels higher than 10.

To ensure there is enough data within each level and to avoid missing data in the estimation of the null distribution, we aggregate levels such that

- the observed number of visits to each level is at least 5,
- the expected number of visits to each level is at least 5, and
- the simulated number of visits to each level is at least 5.

Note that the expected number of visits to each level is taken from the stationary distribution of the embedded discrete-time jump chain, $\tilde{X}_t$, of the fitted continuous-time QBD process defined by the probability transition matrix

$$\tilde{P} = \begin{bmatrix} 0 & \widetilde{A}_{+}^{(0)} & 0 & 0 & \dots \\ \widetilde{A}_{-}^{(1)} & 0 & \widetilde{A}_{+}^{(1)} & 0 & \dots \\ 0 & \widetilde{A}_{-}^{(2)} & 0 & \widetilde{A}_{+}^{(2)} & \dots \\ 0 & 0 & \widetilde{A}_{-}^{(3)} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where $\tilde{A}_{+}^{(\ell)} = \left(-A_{0}^{(\ell)}\right)^{-1} A_{+}^{(\ell)}$ and $\tilde{A}_{-}^{(\ell)} = \left(-A_{0}^{(\ell)}\right)^{-1} A_{-}^{(\ell)}$, for all $\ell \geq 0$ [34].

Note that the method of aggregating levels to ensure that the expected number of visits is sufficiently large is similar to the method of combining bins in the Chi-Square test to ensure each cell has an expected count of at least 5.

For example, consider a level-dependent QBD process with a capacity of 10 and suppose the expected visit counts of levels 8, 9, and 10 are 7, 3, and 1 respectively, these levels will be aggregated to form a new level ≥ 8 with an expected visit count of 11.

More specifically, suppose $\tilde{\boldsymbol{\pi}}$ contains the proportion of time spent in each level. We update $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{10})$ such that $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{\geq 8})$, where

$$\tilde{\pi}_{\geq 8} = \sum_{\ell=8}^{10} \tilde{\pi}_{\ell}.$$

In the case of transitions between levels, suppose $\mathbf{p}$ contains the upward transition probabilities for each level. We update $\mathbf{p}^{+} = (p_0^{+}, p_1^{+}, \ldots, p_{10}^{+})$ such that $\mathbf{p}^{+} = (p_0^{+}, p_1^{+}, \ldots, p_{\geq 8}^{+})$, where

$$p_{\geq 8}^{+} = \frac{\displaystyle\sum_{\ell=8}^{10} n_{\ell,\ell+1}^{E} p_{\ell}^{+}}{\displaystyle\sum_{\ell=8}^{10} n_{\ell,\ell+1}^{E} p_{\ell}^{+} + \sum_{\ell=8}^{10} n_{\ell,\ell-1}^{E} \left(1 - p_{\ell}^{+}\right)},$$

$n_{\ell,\ell+1}^{E}$ is the expected number of times the QBD process moves from level $\ell \geq 1$ to $\ell + 1$, and $n_{\ell,\ell-1}^{E}$ is the expected number of times the QBD process moves from level $\ell \geq 1$ to $\ell - 1$.

Lastly, we consider the case of sojourn times conditioned on moving down a level. Suppose $\{F_1^{-}, F_2^{-}, \ldots, F_{10}^{-}\}$ represents the set of distributions of sojourn times conditioned on moving down from each level. We update this set of distributions to achieve the set $\{F_1^{-}, F_2^{-}, \ldots, F_{\geq 8}^{-}\}$ where

$$F_{\geq 8}^{-} = \sum_{\ell=8}^{10} \omega_{\ell} F_{\ell}^{-},$$

such that

$$\omega_{\ell} = \frac{\tilde{\pi}_{\ell}^{E}}{\displaystyle\sum_{\ell=8}^{10} \tilde{\pi}_{\ell}^{E}}$$

and $\tilde{\pi}_{\ell}^{E}$ is the expected proportion of times the QBD process visits level 8, 9, and 10. The method for sojourn times conditioned on moving up a level is similar.

9.2.7 Discussion

In the goodness of fit method described above, we test all relevant behavioural components at the same time and the reasoning for this is as follows. Firstly, it provides a single test rather than individual tests for each type of behaviour and each level, thus leading to multiple hypothesis testing adjustments. Secondly, it is not easily attainable to always perform simulations that have the same number of upward and downward transitions from each $\ell \geq 1$. For example, the higher levels of an infinite level-independent QBD process may be infrequently visited compared to the lower levels and thus harder to achieve the same number of visits to each level across all simulations.

9.2.8 Discussion

We have developed a goodness of fit method using the Mahalanobis distance, the Kolmogorov-Smirnov statistic, and statistical theory to determine if observed data $\tilde{X}$ can be modelled as a realisation of a QBD process $\mathcal{Q} \in \Omega$, where Ω is a particular class of QBD processes. However, this method does not identify which behavioural component caused the test to fail in the event of a QBD process inadequately modelling the stationary and transient behaviour observed in data $\tilde{X}$. In the next section, we develop a diagnostic method to determine where, if at all, the observed stationary and transient behaviour deviates from the expected stationary and transient behaviour.

9.3 Diagnostic plots

In this section, we develop a diagnostic method which illustrates where, if at all, the stationary and transient behaviour of a fitted QBD process $\widehat{\mathcal{Q}} \in \Omega$ differs to that observed in the data, $\tilde{X}$. In particular, we compare the observed stationary distribution, upward transition probabilities, and conditional sojourn times to that of the simulated data sets, X_j for $j = 1, \ldots, D$, used in the goodness of fit test. Note that for each behavioural component, we present an example of a diagnostic plot which is based on data generated from a level-dependent QBD process with a maximum capacity of 5.

9.3.1 Stationary distribution

In this section, we describe a diagnostic method to illustrate where, if at all, the stationary behaviour of the fitted QBD process differs to that observed in the data. In particular, we use a parametric bootstrap approach as described in Algorithm 3.

Figure 9.3.1 shows an example of a diagnostic plot, where the red points illustrate the observed proportion of time spent in each level and the box-plots represent the proportion of time spent in each level observed in the simulated data sets.

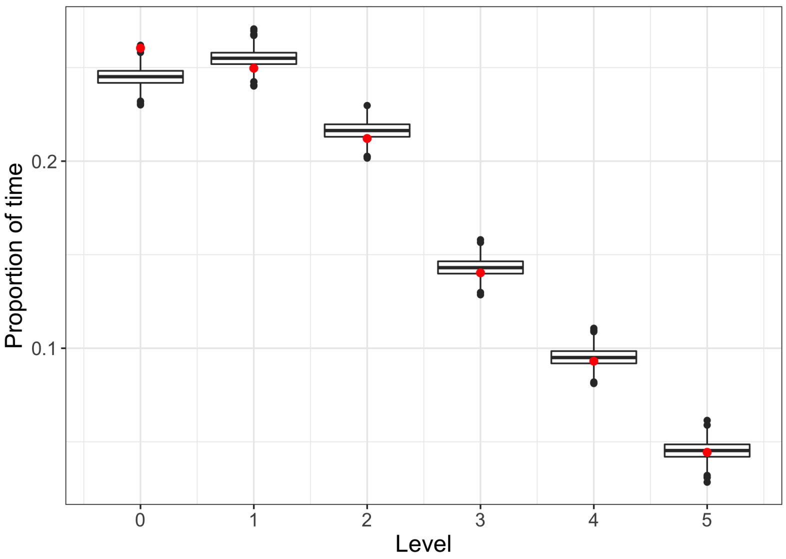


Figure 9.3.1: Example of a diagnostic plot for the stationary distribution.

Algorithm 3: Stationary distribution - Diagnostic method.

Step 1: If no level aggregation is required, calculate the observed proportion of time spent in each level, ℓ, and denote it $O_\ell(X)$. If levels need to be aggregated such that $L^* = \{\ell^*, \dots, L\}$, then

$$O_{L^*}(X) = \sum_{\ell=\ell*}^{L} O_\ell(X).$$

Note that if levels need to be aggregated at the lower boundary, then we also need to calculate

$$O_{L^*}(X) = \sum_{\ell=0}^{\ell^*} O_\ell(X),$$

where $L^* = \{0, \dots, \ell^*\}$.

Step 2: For each simulated data set, X_j, and each level ℓ, calculate the proportion of time spent in each level. We call this value the simulated summary statistic for level ℓ and denote it $O_\ell(X_j)$. If levels need to be aggregated such that $L^* = \{\ell^*, \dots, L\}$, then

$$O_{d,L^*}(X_j) = \sum_{\ell=\ell*}^{L} O_\ell(X_j).$$

Note that if levels need to be aggregated at the lower boundary, then we also need to calculate

$$O_{d,L^*}(X_j) = \sum_{\ell=0}^{\ell*} O_\ell(X_j),$$

where $L^* = \{0, \dots, \ell^*\}$.

Step 3: Use box-plots to compare the observed proportion of time spent in each level to that observed in each simulated data set.

9.3.2 Transitions between levels

In this section, we describe a diagnostic method to illustrate where, if at all, the transition behaviour of the fitted QBD process differs to that observed in the data. We again use a parametric bootstrap approach as described in Algorithm 4.

Note that if the stationary behaviour of the observed data $\tilde{X}$ is not captured by the fitted QBD process $\widehat{\mathcal{Q}} \in \Omega$, then the observed number of times each level is visited may be different to the simulated number of times each level is visited. Therefore, we truncate the data such that the observed number of times level $\ell \geq 0$ is visited is equal to the simulated number of times level $\ell \geq 0$ is visited to avoid under-sampling or over-sampling of levels.

Figure 9.3.2 shows an example of a diagnostic plot, where the red points illustrate the observed transition probabilities between levels and the box-plots represent the transition probabilities between levels observed in the simulated data sets. Note that the data for the minimum and maximum levels are omitted as the probabilities are 0 and 1 by definition.

Given that the downward transition probability from level $\ell \geq 1$ is equal to one minus the upward transition probability, we can omit the plot of downward transition probabilities from the diagnostic analysis if it is not required.

Algorithm 4: Transition probabilities - Diagnostic method.

Step 1: If no level aggregation is required, calculate the observed transition probability of moving up a level and denote it $R_\ell(X)$. If levels need to be aggregated such that $L^* = \{\ell^*, \ldots, L\}$, then

$$R_{L^*}(X) = \frac{\sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}}{\sum_{\ell=\ell*}^{L} n_{\ell,\ell-1} + \sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}}.$$

Note that if levels need to be aggregated at the lower boundary, then we also need to calculate

$$R_{L^*}(X) = \frac{\sum_{\ell=0}^{\ell*} n_{\ell,\ell+1}}{\sum_{\ell=0}^{\ell*} n_{\ell,\ell-1} + \sum_{\ell=0}^{\ell*} n_{\ell,\ell+1}},$$

where $L^* = \{0, \ldots, \ell^*\}$. Also note that the transition probability of moving down a level is calculated as $1 - R_\ell(X)$.

Step 2: For each simulated data set, X_j, and each level ℓ, calculate the transition probability of moving up a level and denote it $R_\ell(X_j)$. If levels need to be aggregated such that $L^* = \{\ell^*, \ldots, L\}$, then

$$R_{d,L^*}(X_j)(X) = \frac{\sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}}{\sum_{\ell=\ell*}^{L} n_{\ell,\ell-1} + \sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}}$$

Note that if levels need to be aggregated at the lower boundary, then we also need to calculate

$$R_{d,L^*}(X_j)(X) = \frac{\sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}}{\sum_{\ell=\ell*}^{L} n_{\ell,\ell-1} + \sum_{\ell=\ell*}^{L} n_{\ell,\ell+1}},$$

where $L^* = \{0, \ldots, \ell^*\}$. Also note that the transition probability of moving down from level $\ell \geq 1$ is calculated as $1 - R_\ell(X_j)$.

Step 3: Use box-plots to compare the observed transition probabilities between levels to that observed in each simulated data set.

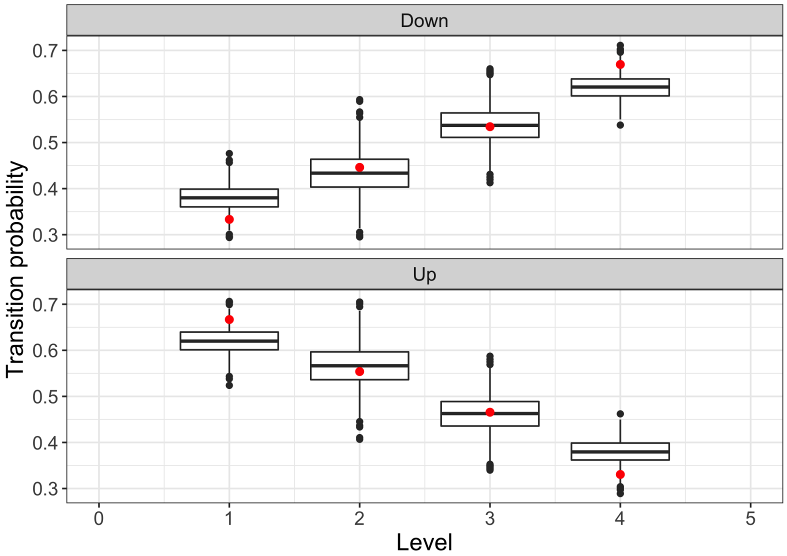


Figure 9.3.2: Example of a diagnostic plot for the transition probabilities between levels.

9.3.3 Conditional sojourn time

In this section, we develop a diagnostic method to illustrate where, if at all, the distribution of conditional sojourn time for each level of the fitted QBD process differs to that observed in the data. Similar to before, we truncate the data such that the observed number of times level $\ell \geq 0$ is visited is equal to the simulated number of times level $\ell \geq 0$ is visited to avoid under-sampling of levels.

The diagnostic method in this case is a moment based approach, where we compare the first three moments (mean, variance, and skewness) of the conditional sojourn time for each level of the fitted QBD process to that of the observed data. The parametric bootstrap approach described in Algorithm 5 focuses on the first moment. The methods for the second and third moments are similar.

Note that we only describe the goodness of fit test for the sojourn time conditioned on the process moving up a level. The diagnostic method for the sojourn time conditioned on the process moving down a level is similar.

Algorithm 5: Conditional sojourn time - Moment based diagnostic method.

Step 1: If no level aggregation is required, calculate the observed average time spent in each level ℓ before moving up a level and denote it $S_\ell(X)$. If levels need to be aggregated such that $L^* = \{\ell^*, \ldots, L\}$, then

$$S_{L^*}(X) = \sum_{\ell=\ell*}^{L} S_\ell(X).$$

Note that if levels need to be aggregated at the lower boundary, then we also need to calculate

$$S_{L^*}(X) = \sum_{\ell=0}^{\ell*} S_\ell(X),$$

where $L^* = \{0, \ldots, \ell^*\}$.

Step 2: For each simulated data set, X_j, and each level ℓ, calculate the observed average time spent in each level. We call this value the simulated summary statistic for level ℓ and denote it $S_\ell(X_j)$.

If levels need to be aggregated such that $L^* = \{\ell^*, \ldots, L\}$, then $S_{d,L^*}(X_j)$ is the average of the set of conditional sojourn times observed across levels $\ell^*, \ldots, L$. Note that if levels need to be aggregated at the lower boundary, then we also need to calculate $S_{d,L^*}(X_j)$, where $L^* = \{0, \ldots, \ell^*\}$.

Step 3: Use box-plots to compare the observed conditional sojourn times to that observed in each simulated data set.

Figure 9.3.3 shows an example of a diagnostic plot, where the red points illustrate the average conditional sojourn times for each level in the observed data and the box-plots represent the average conditional sojourn times for each level observed in each of the simulated data sets. Note that in a finite QBD process, the data associated with the average sojourn time conditioned on moving up from the maximum level is omitted as the process cannot transition to levels above what is defined in the QBD process.

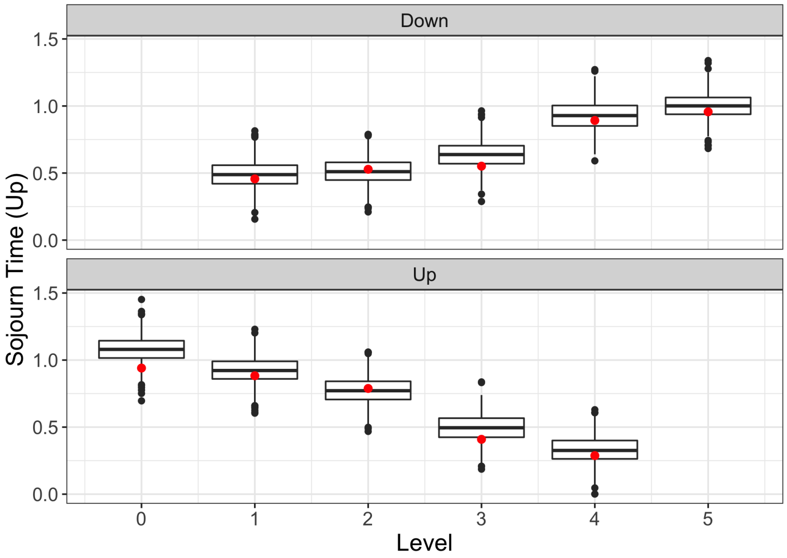


Figure 9.3.3: Example of a diagnostic plot for the conditional sojourn times.

For visualisation purposes, we also plot the empirical densities of the conditional sojourn time for each level of each simulated data set against that of the observed data. Similar to before, if levels need to be aggregated such that $L^* = \{\ell^*, \ldots, L\}$, then we calculate the empirical density of the conditional sojourn time across levels $\ell^*, \ldots, L$. Similarly for levels that need to be aggregated at the lower boundary.

Figure 9.3.4 shows an example of such a plot, where the red lines represent the empirical densities of the conditional sojourn time for each level of each simulated data set and the black line represents the empirical density of the conditional sojourn time for each level of the observed data.

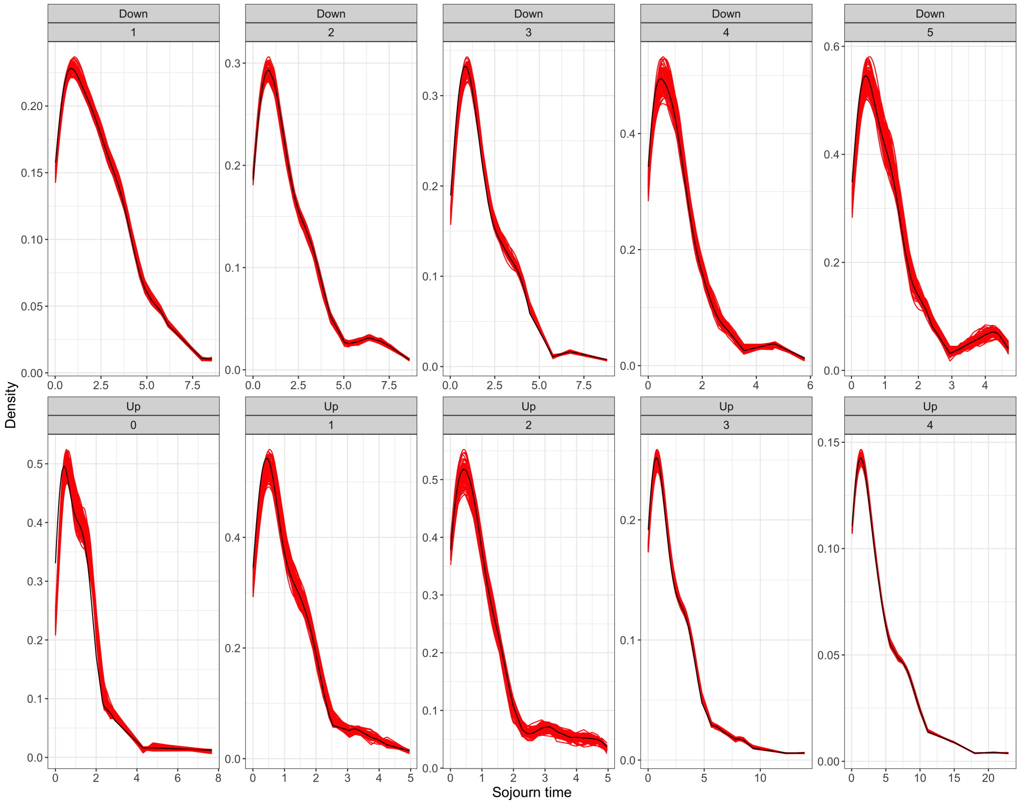


Figure 9.3.4: Example of a density diagnostic plot for the conditional sojourn times.

9.4 Numerical Examples

In the following sections, we demonstrate how our goodness of fit method performs with several numerical examples. In each example, an original data set, $\tilde{X}$, will be simulated from an example QBD process, $\mathcal{Q}$. The best fitting QBD process, $\widehat{\mathcal{Q}}$ based on the value of the log-likelihood will be chosen from several QBD process estimations. We then simulate 1000 data sets, X_j, for $j = 1, \ldots, 1000$, from $\widehat{\mathcal{Q}}$. In each example, we fit a QBD process, $\widehat{\mathcal{Q}}_j$, to each simulated data set, X_j, for $j = 1, \ldots, 1000$ and then proceed with the goodness of fit test as developed in Sections 9.2 and 9.3. This process is illustrated in Figure 9.4.1.

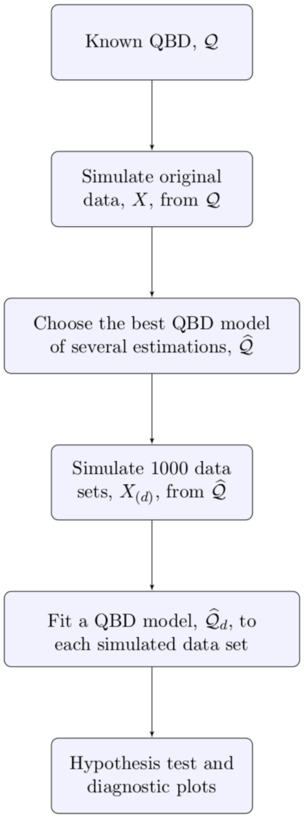

Known QBD, $\mathcal{Q}$
Simulate original data, X, from $\mathcal{Q}$
Choose the best QBD model of several estimations, $\widehat{\mathcal{Q}}$
Simulate 1000 data sets, $X_{(d)}$, from $\widehat{\mathcal{Q}}$
Fit a QBD model, $\widehat{\mathcal{Q}}_d$, to each simulated data set
Hypothesis test and diagnostic plots

Figure 9.4.1: Illustration of the goodness of fit process for the numerical examples.

9.4.1 Fitting a finite level-independent QBD process to infinite level-independent QBD process data

In this example, we consider the continuous-time equivalent of the discrete-time QBD process described by Latouche *et al.* [34]. This infinite level-independent QBD process has the following properties.

- 2 phases per level.
- Process moves from $(0,1)$ to $(0,2)$ with rate 1.
- Process moves from $(\ell,1)$ to $(\ell,2)$ with rate p, for $\ell \geq 1$.
- Process moves from $(\ell,1)$ to $(\ell-1,1)$ with rate $1-p$, for $\ell \geq 1$.
- Process moves from $(\ell,2)$ to $(\ell,1)$ with rate $2p$, for $\ell \geq 0$.
- Process moves from $(\ell,2)$ to $(\ell+1,2)$ with rate $1-2p$, for $\ell \geq 0$.

The infinitesimal generator for this infinite level-independent QBD process is defined as

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \dots \\ A_- & A_0 & A_+ & 0 & \dots \\ 0 & A_- & A_0 & A_+ & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where

$$B_0 = \begin{bmatrix} -1 & 1 \\ 2p & -1 \end{bmatrix}, \quad A_0 = \begin{bmatrix} -1 & p \\ 2p & -1 \end{bmatrix},$$

$$A_+ = \begin{bmatrix} 0 & 0 \\ 0 & 1-2p \end{bmatrix}, \quad A_- = \begin{bmatrix} 1-p & 0 \\ 0 & 0 \end{bmatrix}.$$

We use this example to show how our goodness of fit method performs when fitting a finite level-independent QBD process to infinite level-independent QBD process data. In particular, we fit a finite level-independent QBD process with two phases to observed data, $\tilde{X}$, generated from an infinite level-independent QBD process with two phases and $p = 0.3$.

Goodness of fit test

In this example, we define the null and alternative hypotheses as

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all finite level-independent QBD processes with two phases.

Using the goodness of fit test developed in Section 9.2, we obtain an empirical p-value of 0.54 which suggests that there is no evidence of dissimilarity between the observed data and that expected under a finite level-independent QBD process. This result is also demonstrated in Figure 9.4.2 which plots the observed Mahalanobis distance (red line) against the expected Mahalanobis distances under the null hypothesis (histogram).

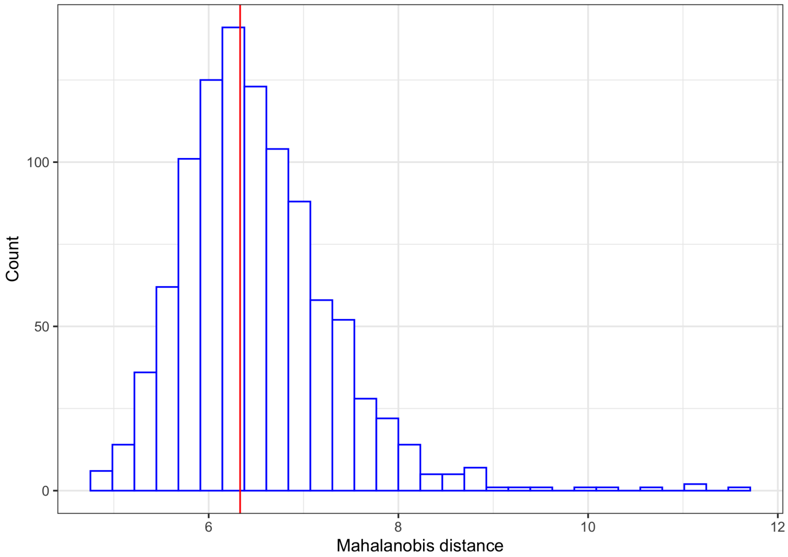


Figure 9.4.2: Comparison of the observed Mahalanobis distance (red line) against the expected Mahalanobis distances assuming a finite level-independent QBD process (histogram).

Discussion

The finite and infinite level-independent QBD processes are very similar in behaviour, with the exception of the upper boundary. Given the nature of this example, this result was expected and therefore requires no further investigation.

We now consider the reverse situation; fitting an infinite level-independent QBD process to data generated from a finite level-independent QBD process.

9.4.2 Fitting an infinite level-independent QBD process to finite level-independent QBD process data

In this example, we consider a breakdown and repair model, which is often used to model production lines [19]. This queueing system has the following properties.

- 2 phases per level.
- Poisson arrivals with rate λ.
- Exponential service times with rate μ.
- 1 server.
- If the server is operating, it breaks down exponentially with rate α.
- Once the breakdown occurs, the repair starts. The time to the end of the repair is exponential with rate β.
- At the time of a breakdown, the part that was being processed is not processed. Once repaired, service starts on the same part.

The infinitesimal generator for this finite level-independent QBD process is defined as

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 \\ A_- & A_0 & A_+ & 0 \\ 0 & A_- & A_0 & A_+ \\ 0 & 0 & A_- & C_0 \end{bmatrix},$$

where

$$B_0 = \begin{bmatrix} -\alpha-\lambda & \alpha \\ \beta & -\beta-\lambda \end{bmatrix}, \quad A_0 = \begin{bmatrix} -\alpha-\lambda-\mu & \alpha \\ \beta & -\beta-\lambda \end{bmatrix}, \quad C_0 = \begin{bmatrix} -\alpha-\mu & \alpha \\ \beta & -\beta \end{bmatrix},$$

$$A_+ = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad A_- = \begin{bmatrix} \mu & 0 \\ 0 & 0 \end{bmatrix}.$$

We use this example to show how our goodness of fit method performs when fitting an infinite level-independent QBD process to finite level-independent QBD process data. In particular, we fit an infinite level-independent QBD process with two phases to data generated from a finite level-independent QBD process with two phases, such that $\lambda = 1$, $\mu = 2$, $\alpha = 1$, and $\beta = 2$.

Goodness of fit test

In this example, we define the null and alternative hypotheses as

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all infinite level-independent QBD processes with two phases.

Using the goodness of fit test developed in Section 9.2, we obtain an empirical p-value of 0 which suggests that the behaviour of the finite level-independent QBD process is not captured by the infinite level-independent QBD process. This result is also illustrated in Figure 9.4.3 which shows how far the observed Mahalanobis distance (red line) is from the expected Mahalanobis distances under the null hypothesis (histogram). Let's now consider the diagnostic plots to see where the dissimilarities exist.

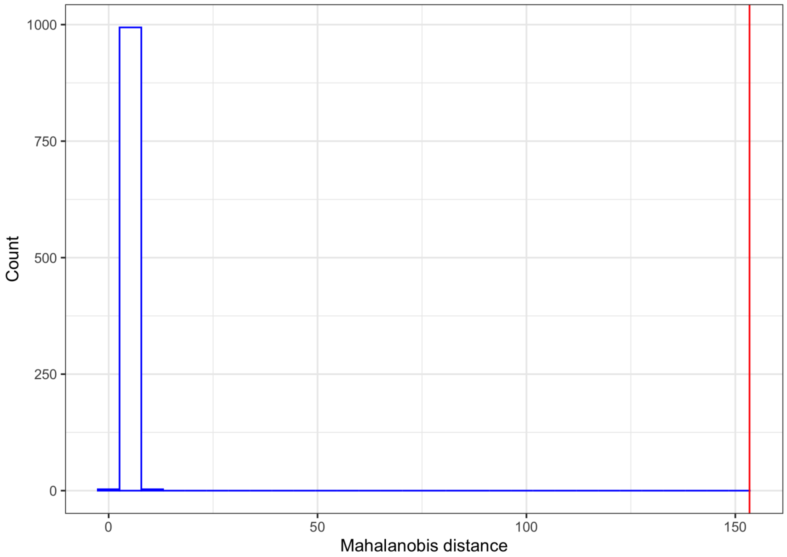


Figure 9.4.3: Comparison of the observed Mahalanobis distance (red line) against the expected Mahalanobis distances assuming an infinite level-independent QBD process (histogram).

Stationary Distribution

Figure 9.4.4 compares the observed proportion of time spent in each level to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process. By design, the infinite level-independent QBD process can go beyond the maximum level observed in the data which in this case is level 4. Therefore, the stationary distribution of the infinite level-independent QBD process does not match the stationary distribution corresponding to the observed data.

Given that the stationary behaviour is incorrectly estimated, the data used to generate the diagnostic plots for the transient behaviour will be truncated to ensure equal comparisons between observed and simulated data.

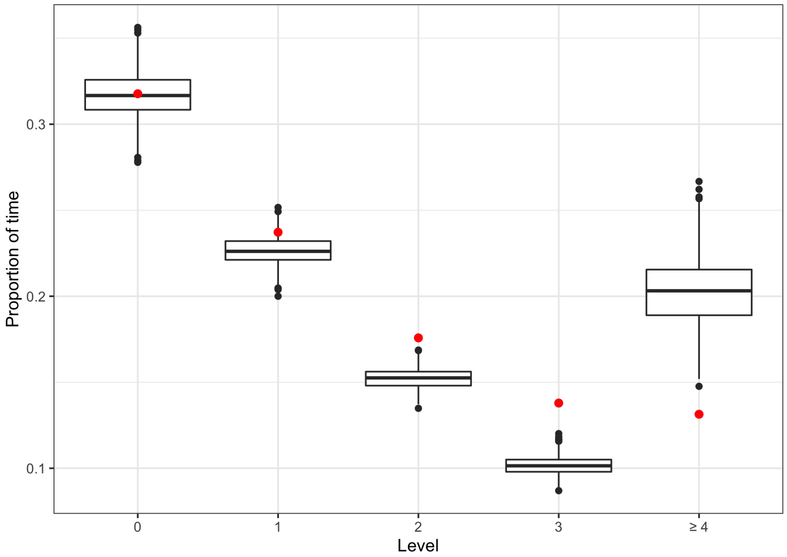


Figure 9.4.4: Comparison of the observed proportion of time spent in each level (red points) to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process (box-plots).

Upward transition Probabilities

The observed upward transition probabilities for each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process slightly underestimate the upward transition probabilities observed in the data, as illustrated in Figure 9.4.5. In addition to this, the infinite level-independent QBD process allows transitions above the maximum observed level in the data, thus leading to omitted data for levels 4 and above.

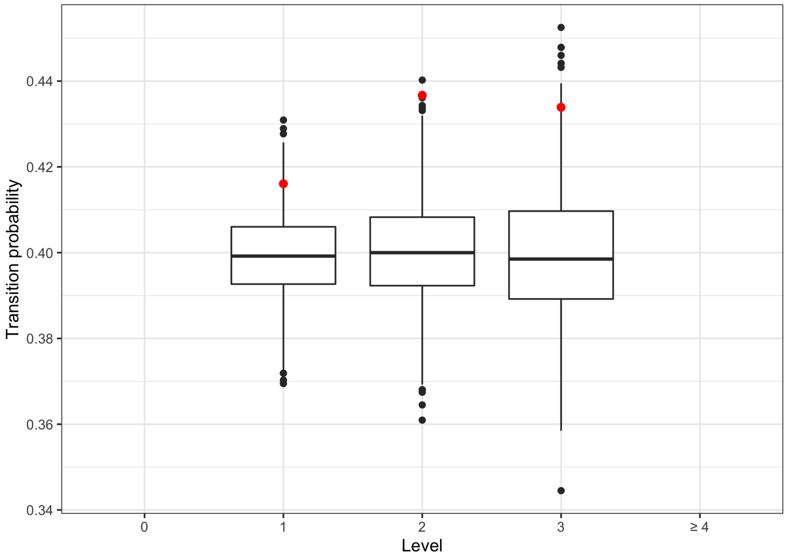
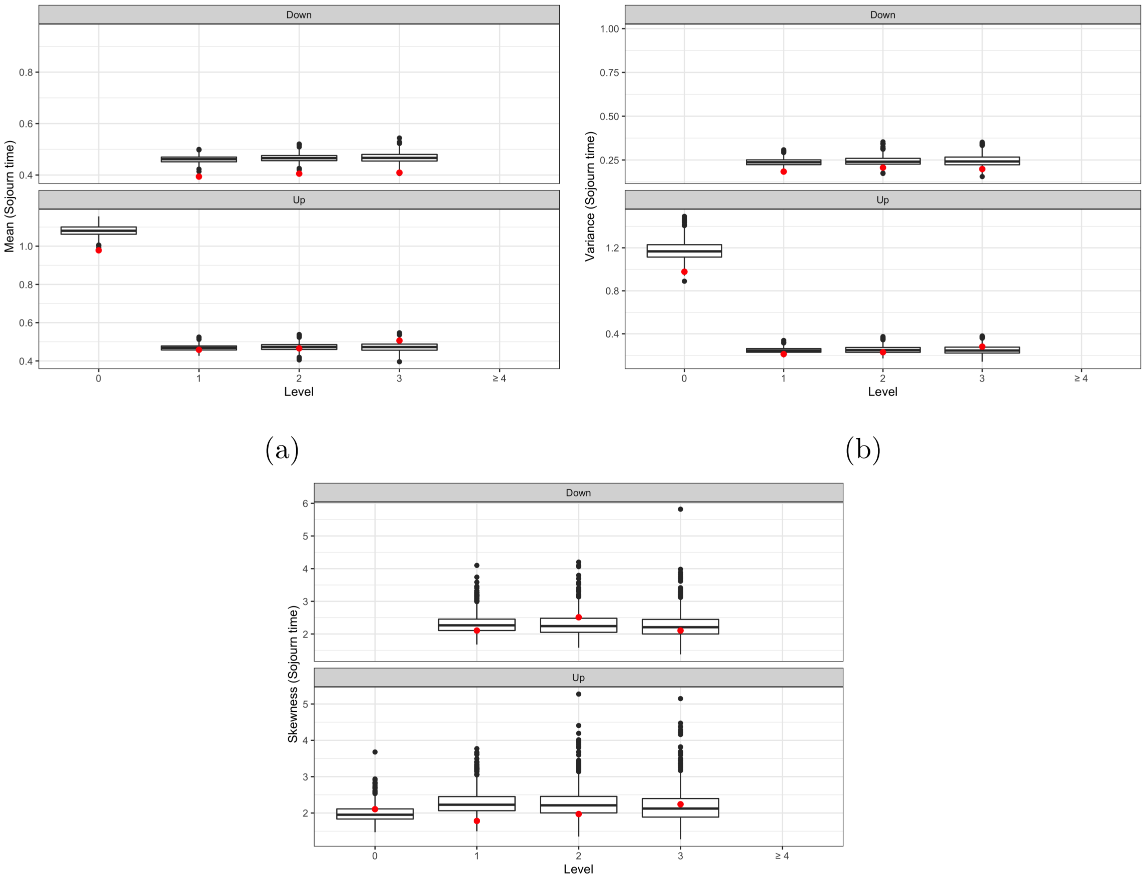


Figure 9.4.5: Comparison of the observed upward transition probabilities between levels (red points) to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process (box-plots). Note that the data for level 0 is omitted since the upward transition probability for level 0 is 1 by design and data associated with upward transitions from levels above 3 are omitted due to minimal data.

Conditional Sojourn Time

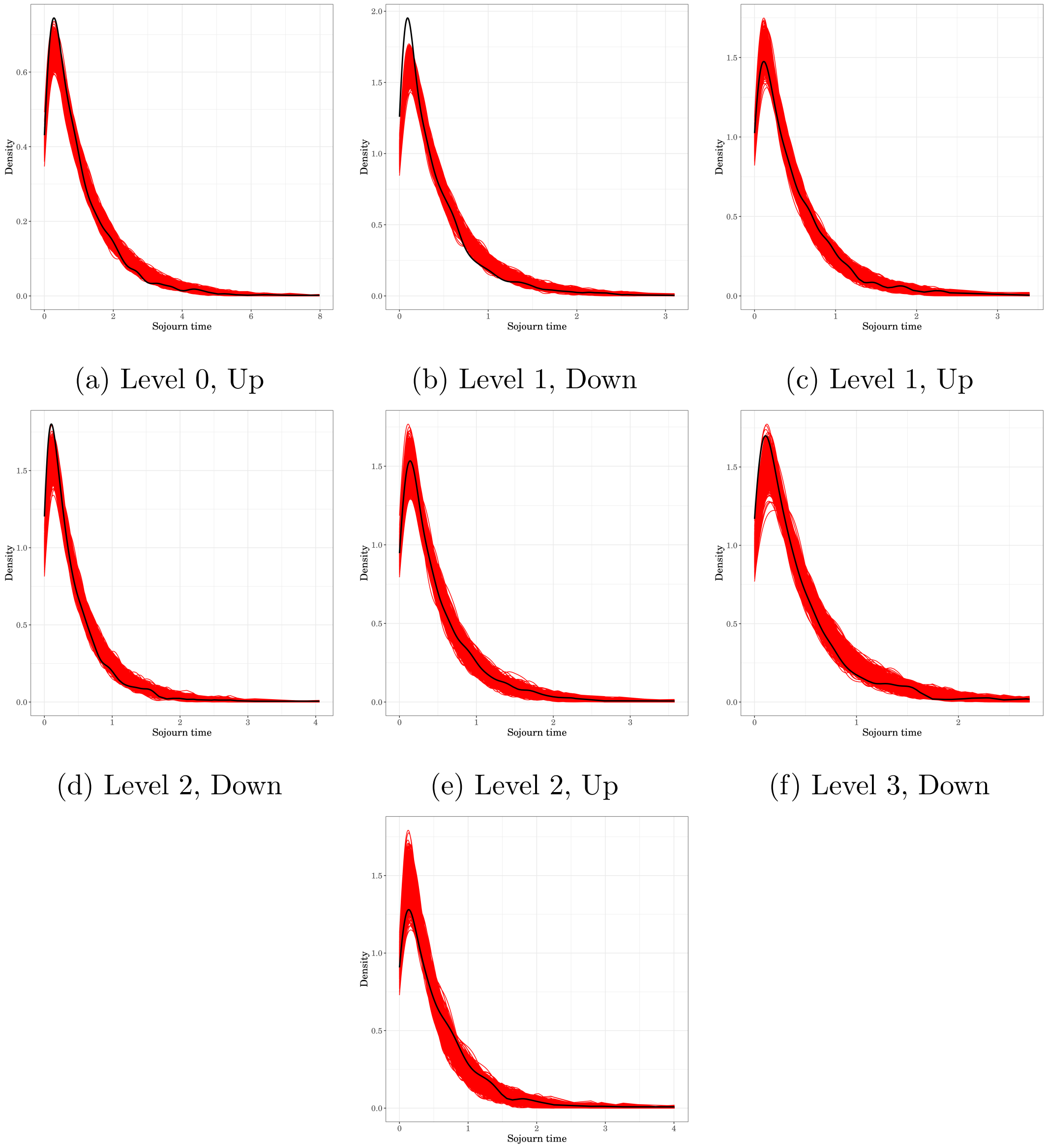
Figure 9.4.6a shows that the average conditional sojourn times are incorrectly estimated, particularly for the downward transitions. In addition to this, Figures 9.4.6b and 9.4.6c show that the second and third moments of the distribution of conditional sojourn times are also incorrectly estimated. Therefore, the fitted infinite level-independent QBD process is unable to model the conditional sojourn times associated with the finite level-independent QBD process. Visualisations of the observed versus estimated densities of conditional sojourn times for selected levels and transitions are provided in Figure 9.4.7.



(a) (b)

(c)

Figure 9.4.6: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each level based on observed data (red points) and the distribution of conditional sojourn times for each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process (box-plots). Note that the data associated with level 0 and a downward transition is omitted since the process cannot move below level 0 and data associated with upward transitions from levels above 3 are omitted due to minimal data.



(a) Level 0, Up

(b) Level 1, Down

(c) Level 1, Up

(d) Level 2, Down

(e) Level 2, Up

(f) Level 3, Down

(g) Level 3, Up

Figure 9.4.7: Comparison of the observed density (black line) of conditional sojourn times for each level and transition to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted infinite level-independent QBD process.

Discussion

To summarise, the fitted infinite level-independent QBD process did not capture the stationary and transient behaviour observed in the finite level-independent QBD process data. This was a result of the infinite level-independent QBD process allowing transitions to levels above the maximum level observed in the data.

The next three examples will now focus on level-dependent QBD process data, where we fit various level-independent QBD process and level-dependent QBD processes to data generated from a level-dependent QBD process with three phases.

9.4.3 Fitting a level-independent QBD process to level-dependent QBD process data

In this example, we consider a specific type of PH/M/∞ queue described by Baumann and Sandmann [3]. This level-dependent QBD process has the following properties.

- Phase-type inter-arrival times, PH($\boldsymbol{\alpha}$, T).
- Exponential service times with rate μ.
- Infinite number of servers.

The infinitesimal generator for this level-dependent QBD process is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where $A_0^{(n)} = T - n\mu I$, $A_-^{(n)} = n\mu I$, and $A_+^{(n)} = \mathbf{t}\boldsymbol{\alpha}$.

We use this example to show how our goodness of fit method performs when fitting level-independent QBD processes to level-dependent QBD process data. In particular, we fit an infinite level-independent QBD process with three phases to data generated from an infinite level-dependent QBD process with three phases, such that

- $\boldsymbol{\alpha} = (1, 0, 0)$,
- $T = \begin{bmatrix} -10 & 10 & 0 \\ 0 & -10 & 10 \\ 0 & 0 & -10 \end{bmatrix}$, and
- $\mu = 3$.

Goodness of fit test

In this example, we define the null and alternative hypotheses as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all infinite level-independent QBD processes with three phases.

Using the goodness of fit test developed in Section 9.2, we find that the observed Mahalanobis distance (red line) is larger than the expected Mahalanobis distances under the null hypothesis (histogram), as shown in Figure 9.4.8. As a result, we obtain an empirical p-value of 0 which suggests that the behaviour of the level-dependent QBD process is not captured by the infinite level-independent QBD process. We now consider the diagnostic plots to determine where the dissimilarities exist.

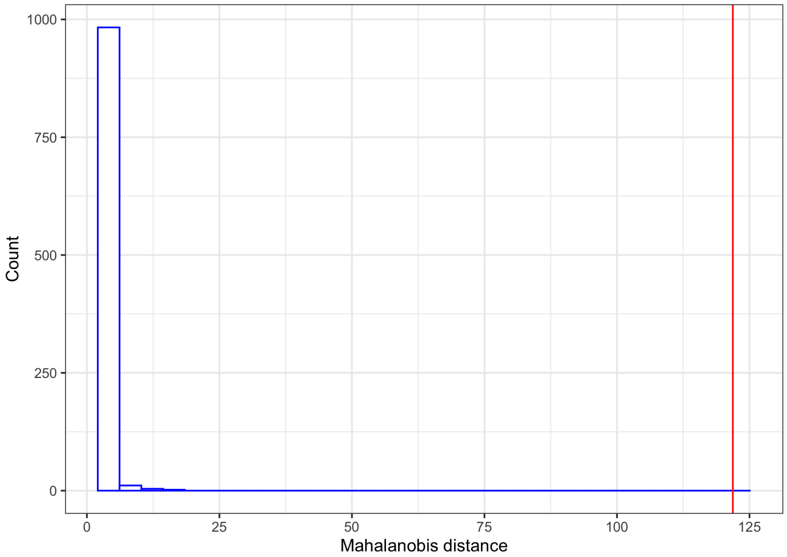


Figure 9.4.8: Comparison of the observed Mahalanobis distance (red line) against the expected Mahalanobis distances assuming an infinite level-independent QBD process (histogram).

Stationary Distribution

The rates of the infinite level-independent QBD process are independent of the level by design which can inhibit the ability to capture the stationary behaviour of data generated from a level-dependent QBD process. As a result, the observed proportion of time spent in each level differs to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process, as illustrated in Figure 9.4.9.

Similar to the previous example, the data used to generate the diagnostic plots for the transient behaviour will be truncated to ensure equal comparisons between observed and simulated data.

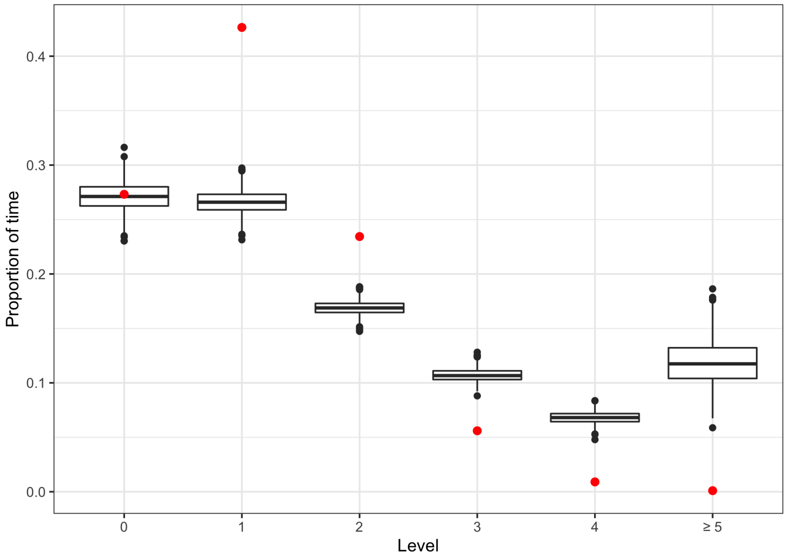


Figure 9.4.9: Comparison of the observed proportion of time spent in each level (red points) to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process (box-plots).

Upward transition Probabilities

Figure 9.4.10 compares the observed upward transition probabilities for each level to the upward transition probabilities for each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process. Similar to before, the transition probabilities of the infinite level-independent QBD process are independent of the level by design and are therefore unable to capture the transition behaviour between levels observed in the data.

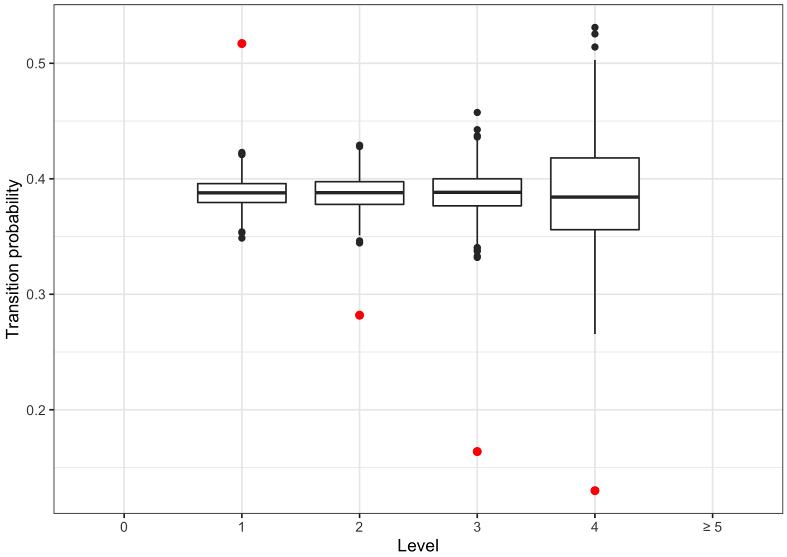
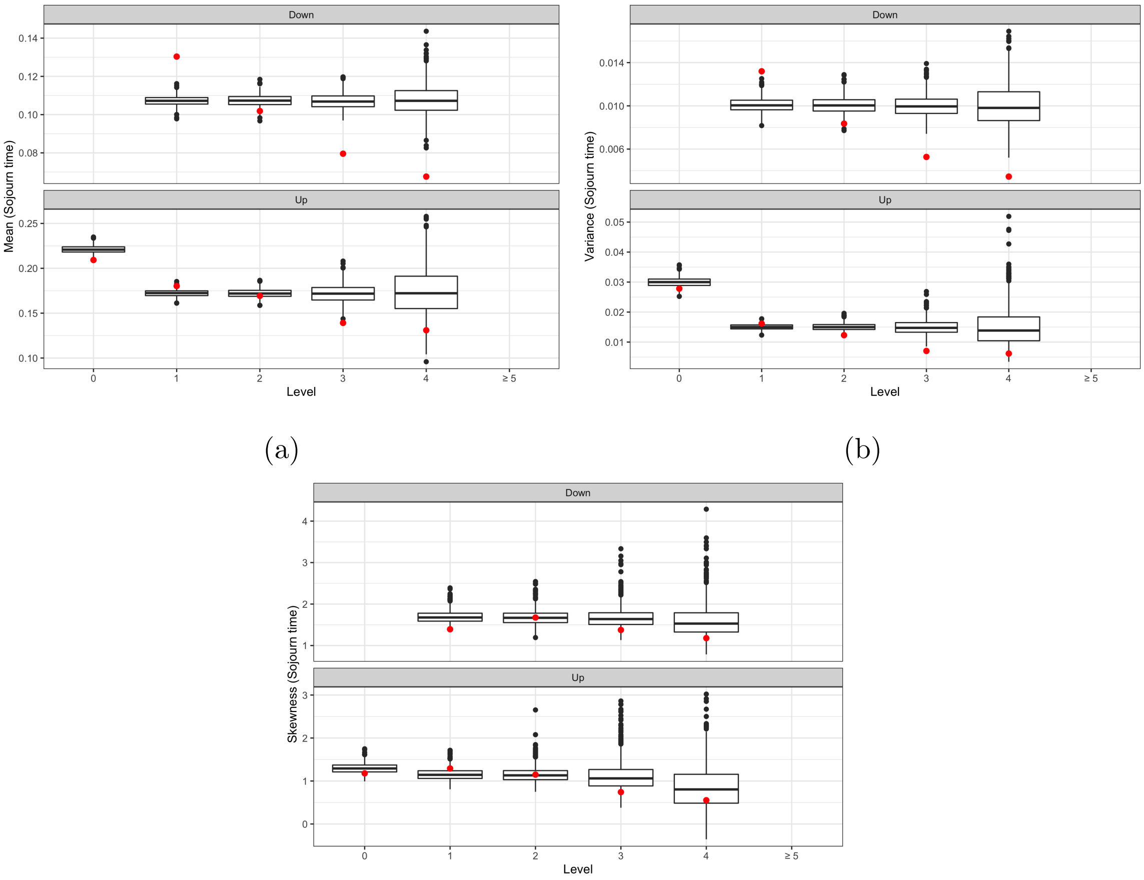


Figure 9.4.10: Comparison of the observed upward transition probabilities between levels (red points) to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted level-independent QBD process (box-plots). Note that the data for level 0 is omitted since the upward transition probability for level 0 is 1 by design and the data associated with upward transitions from level 5 are omitted due to only one data point in the observed data.

Conditional Sojourn Time

As illustrated in Figure 9.4.11, the first three moments of the distribution of conditional sojourn times are incorrectly estimated due to the level-dependent nature of the data, thus suggesting that the fitted infinite level-independent QBD process is incapable of modelling the sojourn times of a level-dependent QBD process. Plots of the observed versus estimated densities of conditional sojourn times are provided in Figure 9.4.12. Note that the densities for level 4 are estimated on minimal data since the QBD process spends more time in levels below 4.

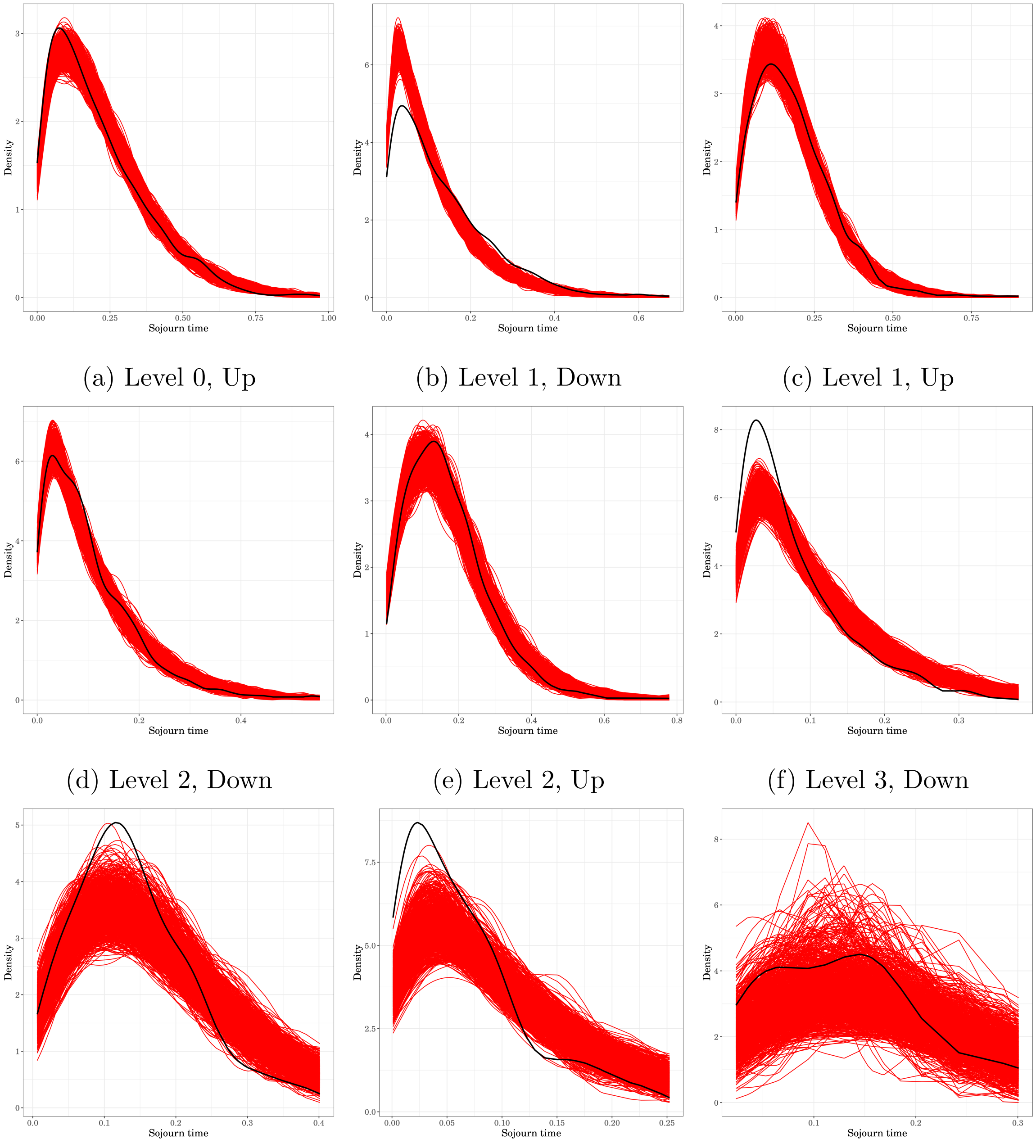


(a)

(b)

(c)

Figure 9.4.11: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each level based on observed data (red points) and the distribution of conditional sojourn times for each level based on 1000 simulated data sets generated from the fitted infinite level-independent QBD process (box-plots). Note that the data associated with level 0 and a downward transition is omitted since the process cannot move below level 0 and the data associated with level 5 are omitted due to minimal data.



(a) Level 0, Up

(b) Level 1, Down

(c) Level 1, Up

(d) Level 2, Down

(e) Level 2, Up

(f) Level 3, Down

(g) Level 3, Up

(h) Level 4, Down

(i) Level 4, Up

Figure 9.4.12: Comparison of the observed density (black line) of conditional sojourn times for each level and selected transitions to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted infinite level-independent QBD process.

Discussion

Overall, the level-dependent nature of the stationary and transient behaviour could not be captured by the fitted infinite level-independent QBD process.

The next two examples will show how our goodness of fit method performs when fitting QBD processes with fewer number of phases required.

9.4.4 Fitting a level-dependent QBD process with one phase to three-phase level-dependent QBD process data

In this example, we fit an infinite level-dependent QBD process with one phase to data generated from an infinite level-dependent QBD process with three phases, such that the infinitesimal generator matrix is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where $A_0^{(n)} = T - n\mu I$, $A_-^{(n)} = n\mu I$, and $A_+^{(n)} = \mathbf{t}\boldsymbol{\alpha}$. Similar to before, we fit an infinite level-dependent QBD process with one phase to data generated from an infinite level-dependent QBD process with three phases, such that

- $\boldsymbol{\alpha} = (1, 0, 0)$,
- $T = \begin{bmatrix} -10 & 10 & 0 \\ 0 & -10 & 10 \\ 0 & 0 & -10 \end{bmatrix}$, and
- $\mu = 3$.

Goodness of fit test

In this example, we define the null and alternative hypotheses as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all level-dependent QBD processes with one phase.

Given that the fitted QBD process has only one phase, there is limited variability in the transition probabilities estimated from the simulated data and hence are approximately equal to the estimated transition probabilities of the fitted QBD process. Hence, the covariance matrix is singular as the expected differences associated with transition probabilities are close to zero, as illustrated in Figure 9.4.13.

Therefore, we cannot use the goodness of fit test described in Section 9.2 as we would lose critical information relating to the transition probabilities. Instead, we use the diagnostic plots to explore what types of behaviour the level-dependent QBD process with one phase may or may not have captured.

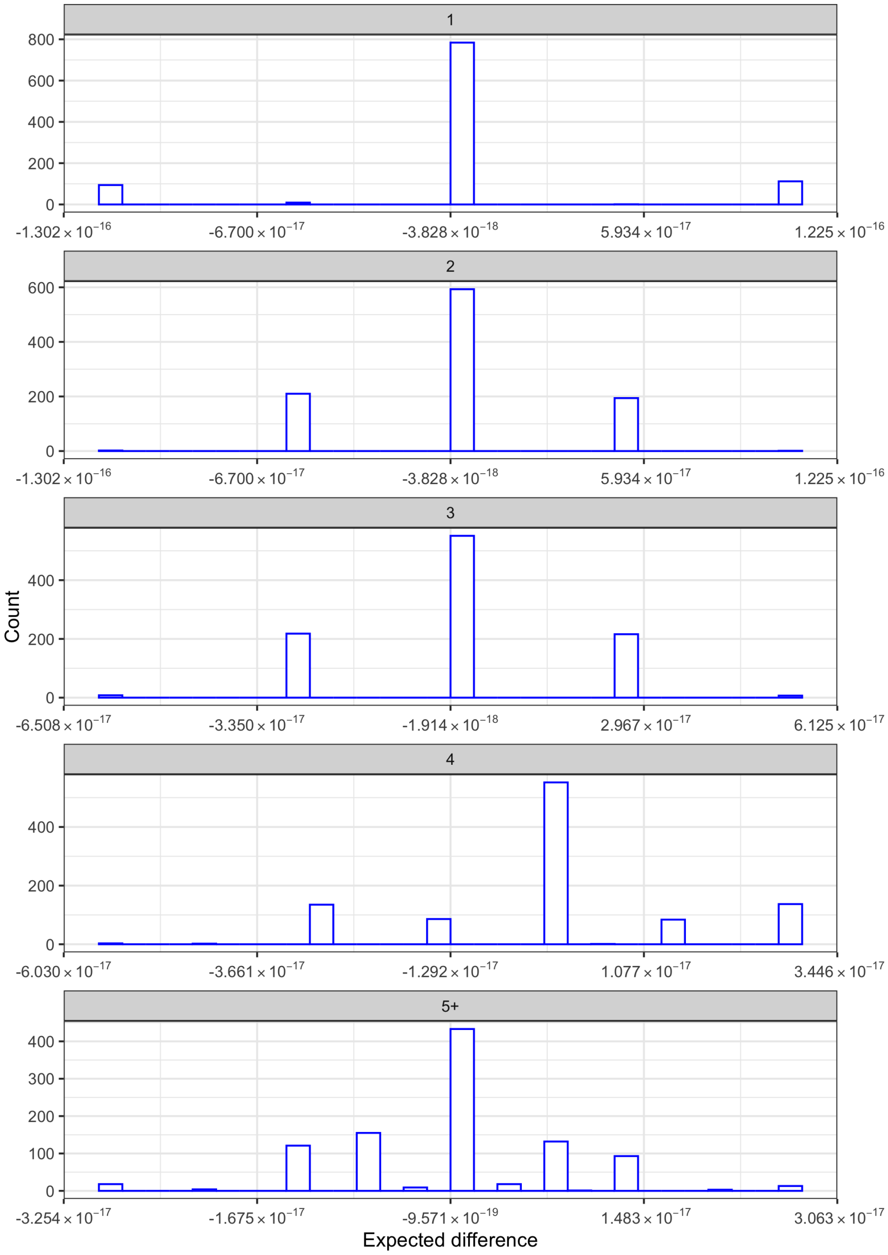


Figure 9.4.13: Expected differences associated with transition probabilities assuming a level-dependent QBD process with 1 phase.

Stationary Distribution

Given the level-dependent nature of the QBD process, the stationary behaviour of the data is captured by the level-dependent QBD process with one phase, as illustrated in Figure 9.4.14. Therefore, we look towards the transient behaviour for dissimilarities between the observed data and the fitted level-dependent QBD process with one phase.

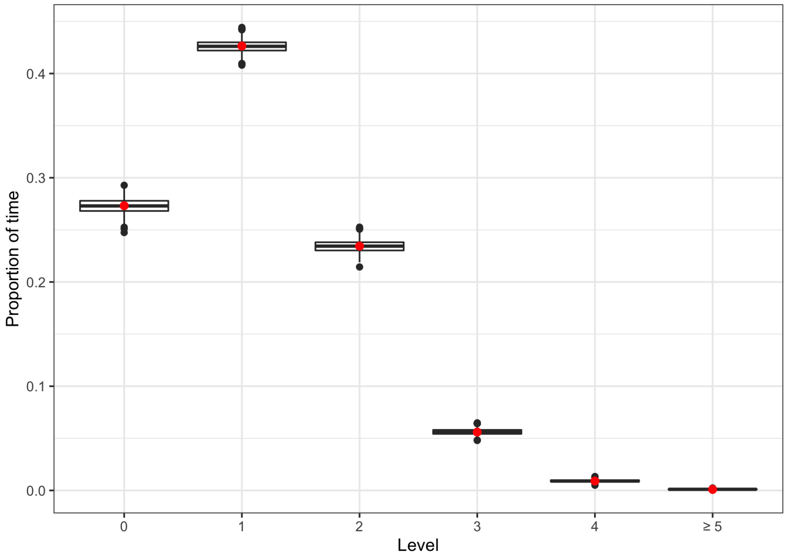


Figure 9.4.14: Comparison of the observed proportion of time spent in each level (red points) to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots).

Upward transition Probabilities

By a similar reasoning, there is no evidence of dissimilarity between the transition behaviour observed in the data compared to that expected under the null hypothesis. This is illustrated in Figure 9.4.15 which compares the observed upward transition probabilities between levels to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase. Let's now consider the conditional sojourn times.

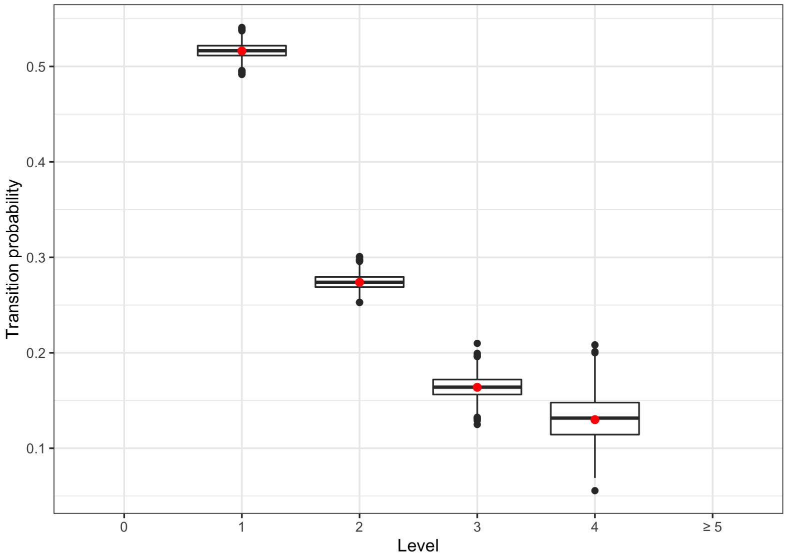
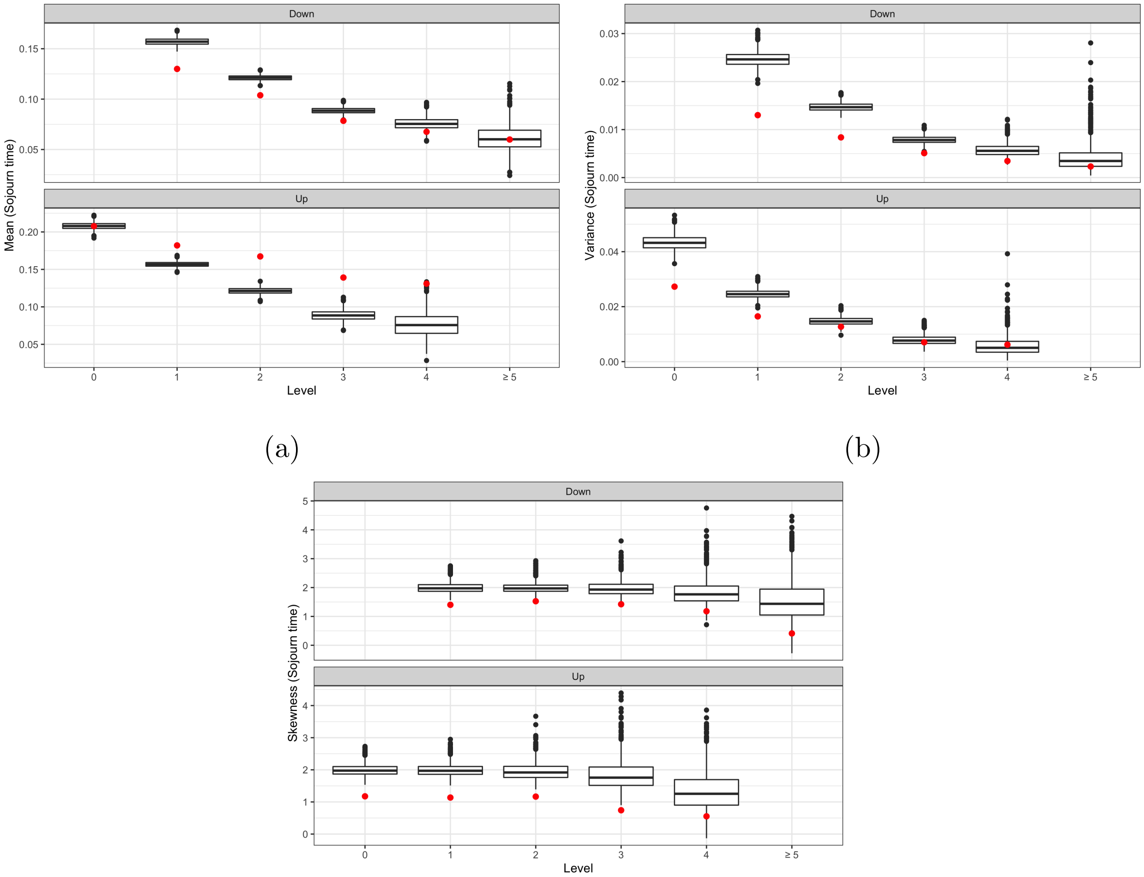


Figure 9.4.15: Comparison of the observed upward transition probabilities between levels (red points) to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots). Note that the data for level 0 is omitted since the upward transition probability for level 0 is 1 by design and the data associated with upward transitions from level 5 are omitted due to only one data point in the observed data.

Conditional Sojourn Time

Comparing the first three moments of the distribution of conditional sojourn times for each level based on observed data to those of the distribution of conditional sojourn times for each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase, we see that the sojourn times are poorly estimated for each level and transition, as illustrated in Figure 9.4.16.

Illustrations of the observed versus estimated densities of conditional sojourn times for selected levels and transitions are provided in Figures 9.4.17 and 9.4.18. Note that the densities for levels 4 and above are estimated on minimal data since the QBD process spends more time in levels below 4.

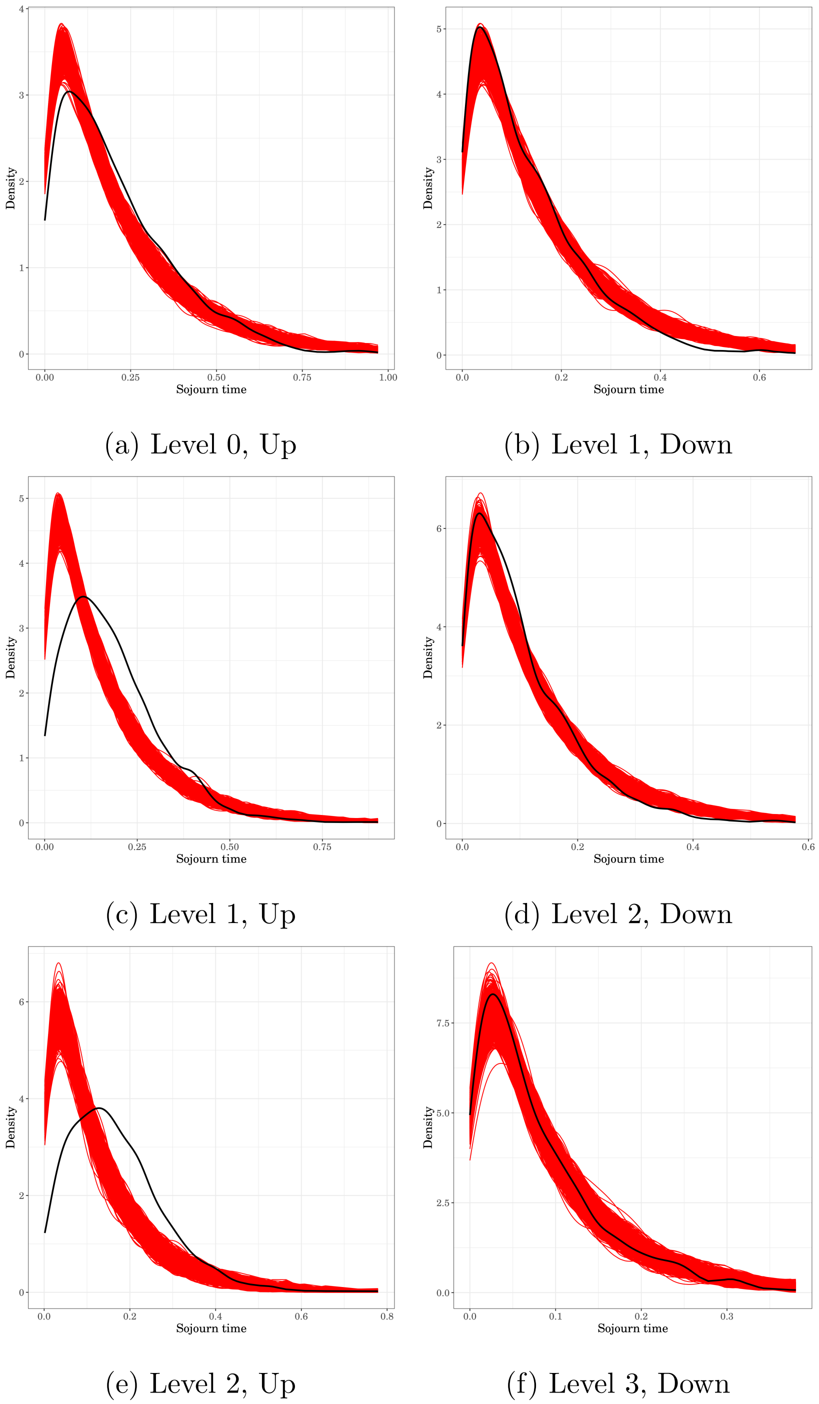


(a)

(b)

(c)

Figure 9.4.16: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each level based on observed data (red points) and the distribution of conditional sojourn times for each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots). Note that the data associated with level 0 and a downward transition is omitted since the process cannot move below level 0 and the data associated with upward transitions from level 5 are omitted due to only one data point in the observed data.



(a) Level 0, Up

(b) Level 1, Down

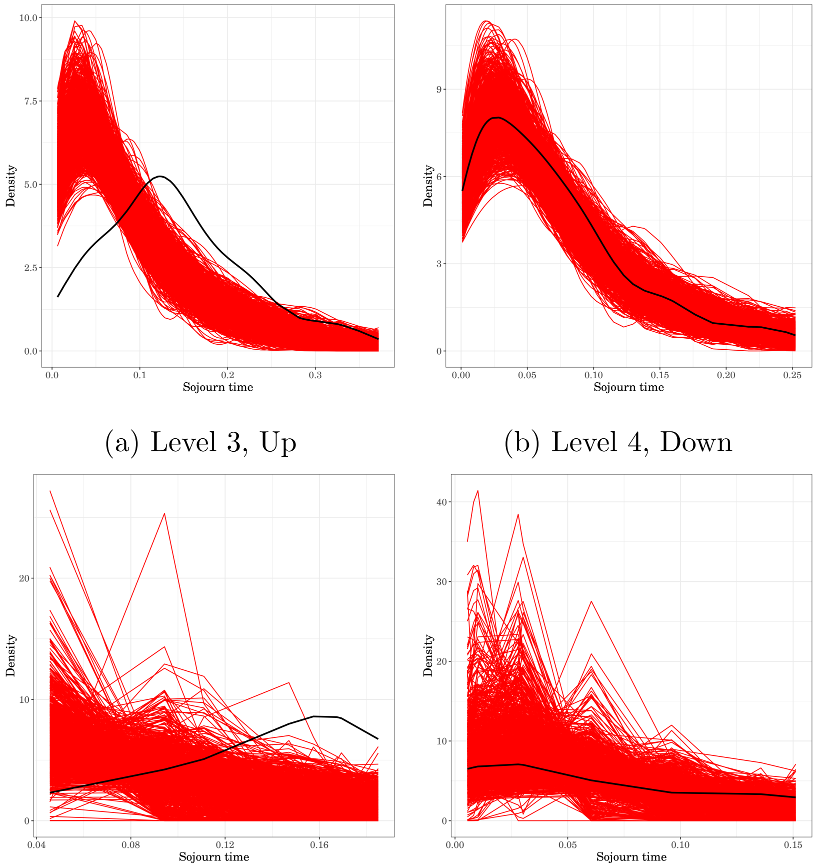
(c) Level 1, Up

(d) Level 2, Down

(e) Level 2, Up

(f) Level 3, Down

Figure 9.4.17: Comparison of the observed density (black line) of conditional sojourn times for levels 0, 1, 2, and 3 and selected transitions to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted level-dependent QBD process with one phase.



(a) Level 3, Up

(b) Level 4, Down

(c) Level 4, Up

(d) Level 5, Down

Figure 9.4.18: Comparison of the observed density (black line) of conditional sojourn times for levels 3, 4, and 5 and selected transitions to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted level-dependent QBD process with one phase.

Discussion

The fitted level-dependent QBD process with one phase was able to capture the stationary behaviour of the level-dependent QBD process with three phases, as well as the transition behaviour between levels. However, dissimilarities were observed in the conditional sojourn times which comes as a result of not enough phases.

We now explore the fit of a level-dependent QBD process with two phases to data generated from a level-dependent QBD process with three phases.

9.4.5 Fitting a level-dependent QBD process with two phases to three-phase level-dependent QBD process data

In this example, we fit an infinite level-dependent QBD process with two phases to data generated from an infinite level-dependent QBD process with three phases, such that the infinitesimal generator matrix is defined as

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \dots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \dots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where $A_0^{(n)} = T - n\mu I$, $A_-^{(n)} = n\mu I$, and $A_+^{(n)} = \mathbf{t}\boldsymbol{\alpha}$. Similar to the previous example, we fit an infinite level-dependent QBD process with two phases to data generated from an infinite level-dependent QBD process with three phases, such that

- $\boldsymbol{\alpha} = (1, 0, 0)$,
- $T = \begin{bmatrix} -10 & 10 & 0 \\ 0 & -10 & 10 \\ 0 & 0 & -10 \end{bmatrix}$, and
- $\mu = 3$.

Goodness of fit test

In this example, we define the null and alternative hypotheses as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all level-dependent QBD processes with two phases.

Using the goodness of fit test developed in Section 9.2, we obtain an empirical p-value of 0.038. Assuming a 5% significance level, there is evidence of dissimilarity between the behaviour of the level-dependent QBD process with three phases and that of the level-dependent QBD process with two phases. This result is also illustrated in Figure 9.4.19 which plots the observed Mahalanobis distance (red line) against the expected Mahalanobis distances under the null hypothesis (histogram).

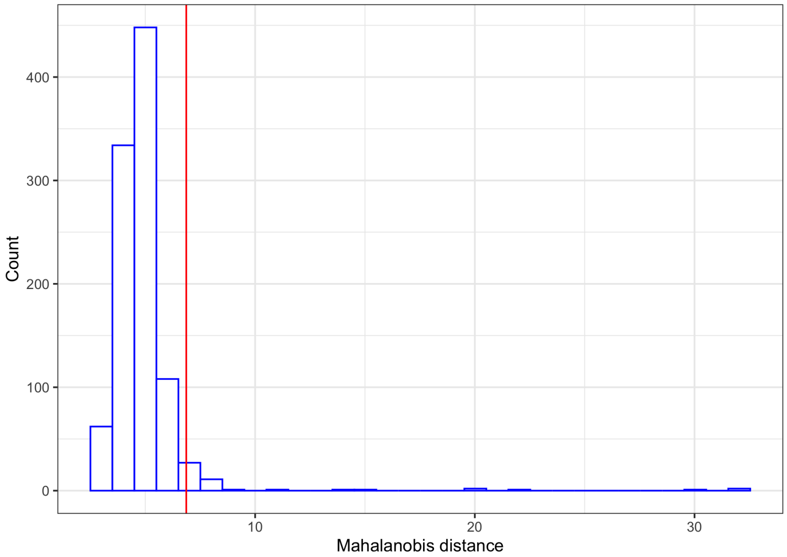


Figure 9.4.19: Comparison of the observed Mahalanobis distance (red line) against the expected Mahalanobis distances assuming a level-dependent QBD process with 2 phases (histogram).

Stationary Distribution

Figure 9.4.20 shows that the observed proportion of time spent in each level is similar to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases. Similar to the previous example, this was expected due to the level-dependent nature of the fitted QBD process.

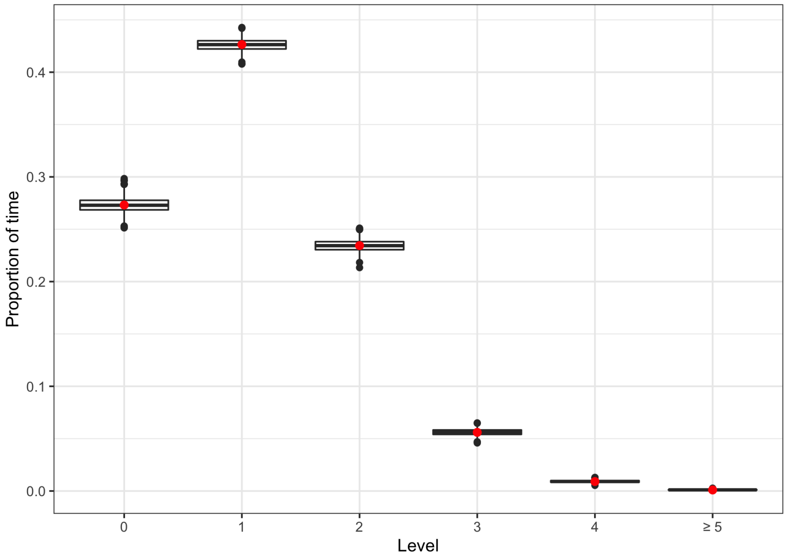


Figure 9.4.20: Comparison of the observed proportion of time spent in each level (red points) to the observed proportion of time spent in each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots).

Upward transition Probabilities

Similarly, Figure 9.4.21 compares the observed upward transition probabilities between levels to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases. As expected, the transition behaviour between levels is captured by the level-dependent QBD process with two phases.

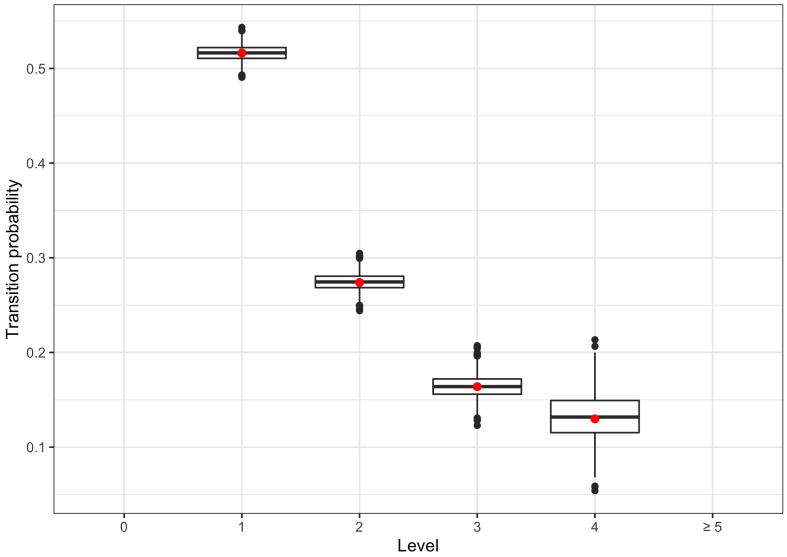
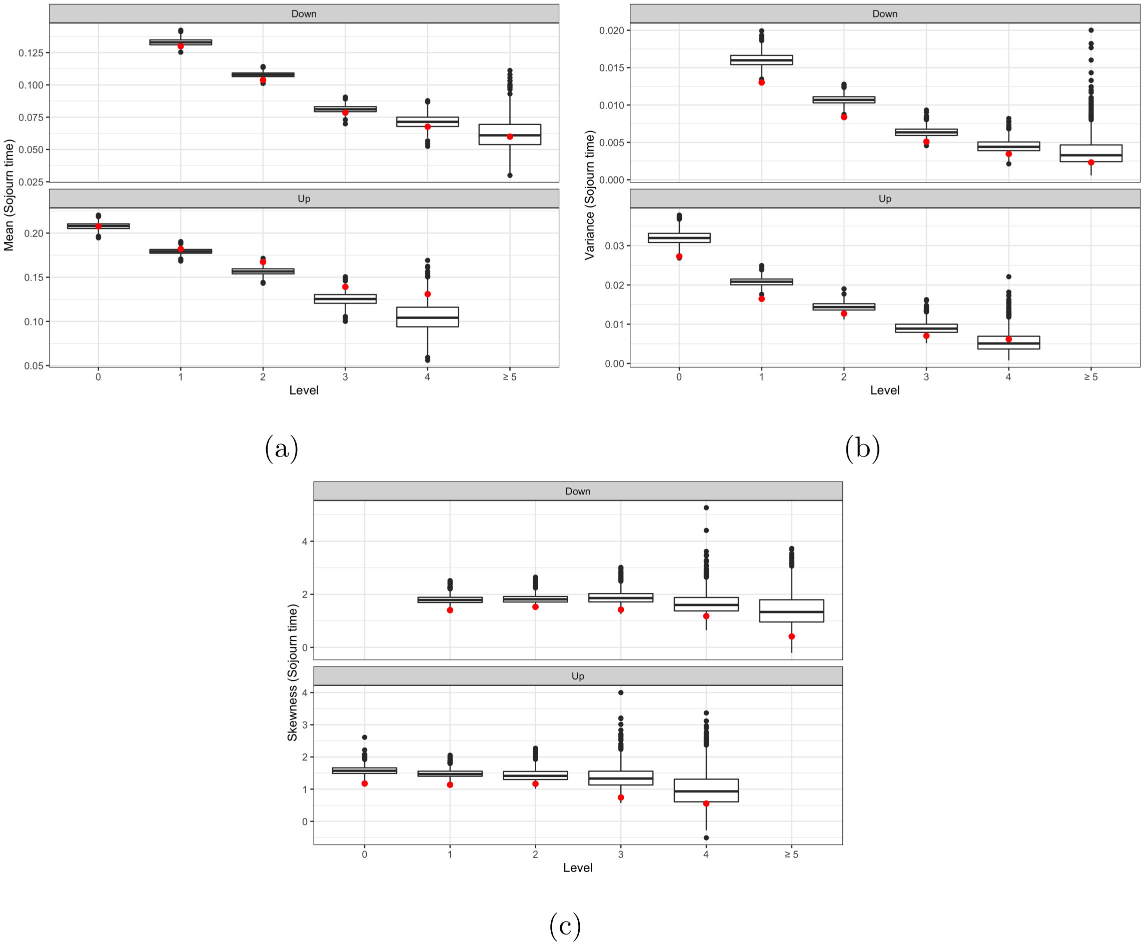


Figure 9.4.21: Comparison of the observed upward transition probabilities between levels (red points) to the observed upward transition probabilities between levels based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that the data for level 0 is omitted since the upward transition probability for level 0 is 1 by design and the data associated with upward transitions from level 5 are omitted due to only one data point in the observed data.

Conditional Sojourn Time

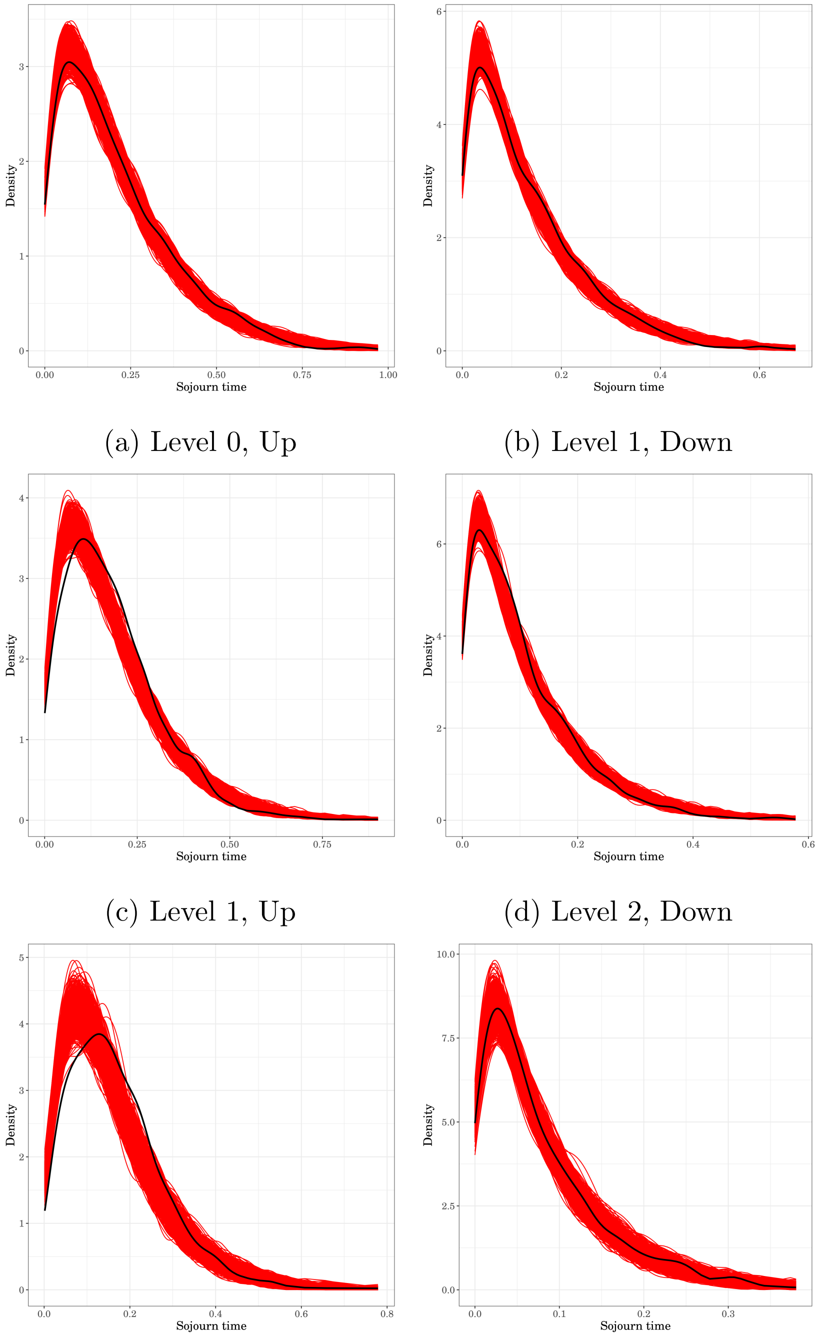
As illustrated in Figure 9.4.22a, the average sojourn time conditioned on moving up a level is slightly underestimated, and only slightly over-estimated for the downward transitions. Despite a slight improvement in capturing the mean of the conditional sojourn times, Figures 9.4.22b and 9.4.22c show that the second and third moments of the distribution of conditional sojourn times remain poorly estimated. Visualisations of the observed versus estimated densities of conditional sojourn times are provided in Figures 9.4.23 and 9.4.24. Similar to before, the densities for levels 4 and above are estimated on minimal data, thus demonstrating less precision compared to levels below 4 which are more frequently visited.



(a) (b)

(c)

Figure 9.4.22: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each level based on observed data (red points) and the distribution of conditional sojourn times for each level based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that the data associated with level 0 and a downward transition is omitted since the process cannot move below level 0 and the data associated with upward transitions from level 5 are omitted due to only one data point in the observed data.



(a) Level 0, Up

(b) Level 1, Down

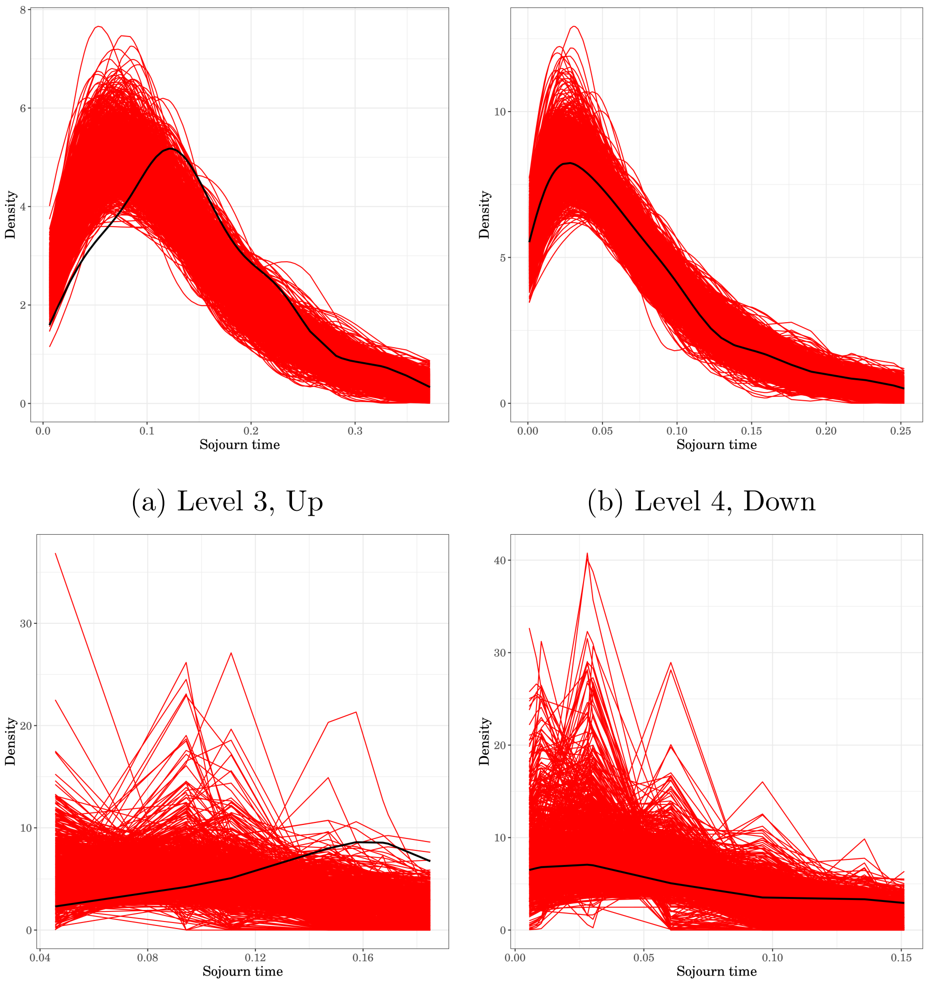
(c) Level 1, Up

(d) Level 2, Down

(e) Level 2, Up

(f) Level 3, Down

Figure 9.4.23: Comparison of the observed density (black line) of conditional sojourn times for levels 0, 1, 2, and 3 and selected transitions to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted level-dependent QBD process with two phases.



(a) Level 3, Up

(b) Level 4, Down

(c) Level 4, Up

(d) Level 5, Down

Figure 9.4.24: Comparison of the observed density (black line) of conditional sojourn times for levels 3, 4, and 5 and selected transitions to the observed density of conditional sojourn times for each level and transition based on 1000 simulated data sets (red lines) generated from the fitted level-dependent QBD process with two phases.

Discussion

Despite capturing the stationary behaviour and part of the transient behaviour, the two-phase level-dependent QBD process was not able to completely capture the behaviour relating to sojourn times within each level. We noticed an improvement between the fitted one-phase level-dependent QBD process and the fitted two-phase level- dependent QBD process but the dissimilarities were still strong enough to reject the null hypothesis.

9.5 Estimation of statistical significance and statistical power

With each statistical hypothesis test, there are four possible decisions, as described in Table 9.5.1.

Null hypothesis	Accept H_0	Reject H_0
True	Correct decision	Type I error
False	Type II error	Correct decision

Table 9.5.1: Statistical errors in hypothesis testing.

The Type I error is known as the statistical significance α of a hypothesis test and the statistical power of a hypothesis test is $1 - \beta$, where β is the Type II error.

In this section, we demonstrate the statistical significance and statistical power of our goodness of fit test by considering the level-dependent QBD process described in Section 9.4.3. Hence, in this section we consider the null and alternative hypotheses,

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all level-dependent QBD processes with three phases.

To estimate the statistical significance and power of our goodness of fit test, we use empirical estimation, implemented using R [43], as described in the following sections.

9.5.1 Statistical significance

The statistical significance of a goodness of fit test refers to the probability of rejecting H_0 when H_0 is true. Hence, in this section we fit several level-dependent QBD processes with three phases to the observed data to estimate the statistical significance. The empirical estimation method for the statistical significance of our goodness of fit test is described in Algorithm 6.

Algorithm 6: Empirical estimation of the statistical significance of our goodness of fit test.

```
Step 1: Fit a level-dependent QBD process with three phases to the
  observed data, $\tilde{X}$, and denote it $\widehat{\mathcal{Q}}$.
Step 2: Simulate the null distribution, as described in steps (a) and (b).
    (a) Simulate $K$ events from the fitted level-dependent QBD process with
  three phases, $\widehat{\mathcal{Q}}$. Denote the simulated data set as $X_d$, for $d = 1, \ldots, D$.
    (b) Fit a level-dependent QBD process with three phases to each
  simulated data set. Denote the fitted level-dependent QBD process with
  three phases as $\widehat{\mathcal{Q}}_d$, for $d = 1, \ldots, D$.
Step 3: Determine the fit of $\widehat{\mathcal{Q}}$ to the observed data $\tilde{X}$ using the goodness
  of fit test described in Section 9.2.
Step 4: Repeat Steps 1 to 3 a large number of times to estimate the
  statistical significance.
```

Typically, a large number of simulations are required to approximate the statistical significance of a goodness of fit test. For example, we would need to repeat steps 1 to 3 10,000 times, each with an empirical null distribution consisting of 10,000 points. However, this is computationally demanding as each individual step requires the estimation of a QBD process. Therefore, we only repeat steps 1 to 3 50 times, each with an empirical null distribution consisting of 50 points.

We note that this does not accurately estimate the significance of the goodness of fit test but it does provide an example of the performance of the goodness of fit test. Further analysis is therefore required to confidently approximate the statistical significance of this goodness of fit test.

As shown in Figure 9.5.1 and Table 9.5.2, we retained H_0 100% of the time which provides some evidence to suggest that this goodness of fit test is unlikely to reject the null hypothesis when it is true.

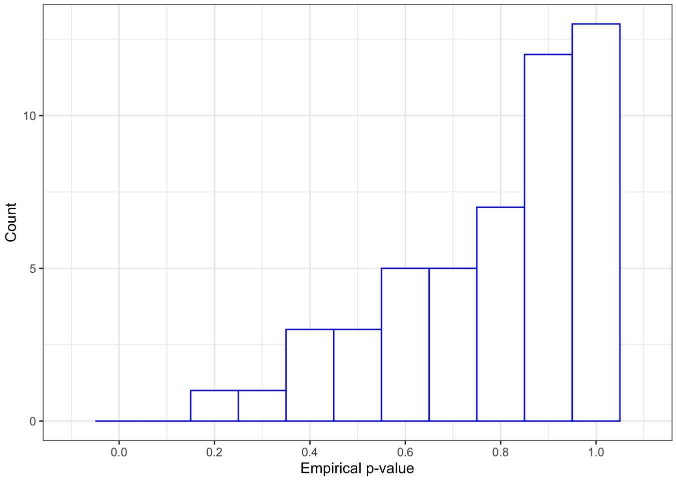


Figure 9.5.1: Histogram of the empirical p-values from the estimation of statistical significance of our goodness of fit test.

Reject H_0 when H_0 is true	Retain H_0 when H_0 is true
0	50

Table 9.5.2: Number of times we rejected/retained the null hypothesis when the null hypothesis is true from the estimation of statistical significance of our goodness of fit test.

9.5.2 Statistical power

The statistical power of a goodness of fit test refers to the probability of rejecting H_0 when H_0 is false. In terms of fitting QBD processes to queueing system data, the QBD process could be incorrect for any of the following reasons.

- Incorrectly fit a QBD process with too few or too many levels.
- Incorrectly fit a QBD process with too few or too many phases.
- Incorrectly fit a level-independent or level-dependent QBD process to observed data.

In this section we fit several level-dependent QBD processes with 1, 2, and 4 phases to the observed data to estimate the statistical power. In the case of the level-dependent QBD process with one and two phases, we expect to reject the null hypothesis as these QBD processes have fewer phases and thus will not capture the transient behaviour of the level-dependent QBD process with three phases. The level-dependent QBD process with four phases overfits the data, in that there are more phases than required to capture the stationary and transient behaviour. However, this does not mean that level-dependent QBD process with four phases is incorrect. Therefore, our goodness of fit test may not be able to detect over-fitting in this case.

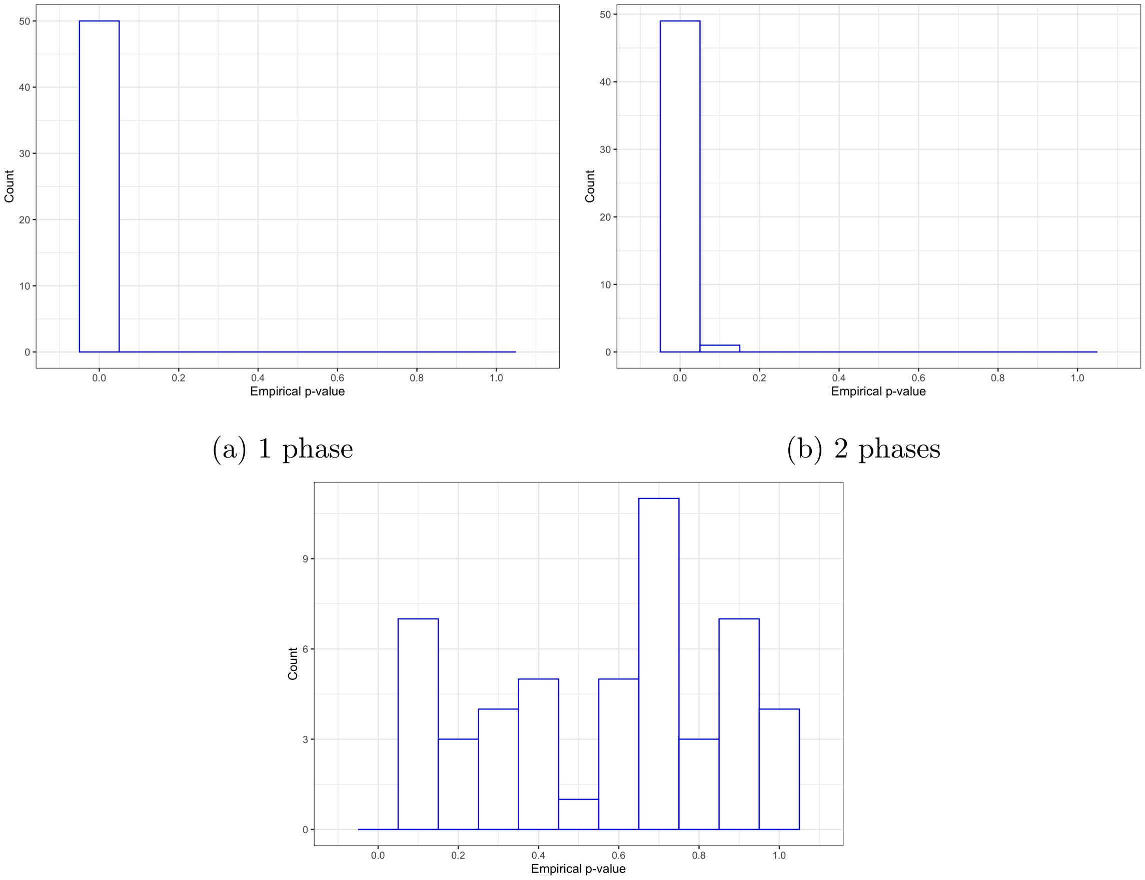
We use the empirical estimation method described in Algorithm 7 to estimate the statistical power of our goodness of fit test. The results of these estimations are summarised in Figure 9.5.2 and illustrated in Table 9.5.3. For similar reasoning to before, we only repeat steps 1 to 3 50 times, each with an empirical null distribution consisting of 50 points. Hence, further analysis is required to confidently approximate the statistical power of this goodness of fit test.

```
Algorithm 7: Empirical estimation of the statistical power of our goodness
of fit test.
  Step 1: Incorrectly fit a $j$ phase level-dependent QBD process to the
    observed data, $\tilde{X}$, and denote it $\widehat{\mathcal{Q}}$.
  Step 2: Simulate the null distribution, as described in steps (a) and (b).
      (a) Simulate $K$ events from the fitted $j$ phase level-dependent QBD
    process, $\widehat{\mathcal{Q}}$. Denote the simulated data set as $X_d$, for $d = 1, \dots, D$.
      (b) Fit a $j$ phase level-dependent QBD process to each simulated data
    set. Denote the fitted $j$ phase level-dependent QBD process as $\widehat{\mathcal{Q}}_d$, for
    $d = 1, \dots, D$.
  Step 3: Determine the fit of $\widehat{\mathcal{Q}}$ to the observed data $\tilde{X}$ using the goodness
    of fit test described in Section 9.2.
  Step 4: Repeat Steps 1 to 3 a large number of times to estimate the
    statistical power.
```

As expected, we correctly rejected the null hypothesis in the case of the level-dependent QBD process with one and two phases majority of the time, as illustrated in Figures 9.5.2a. However, our goodness of fit test was not able to detect over-fitting with a level-dependent QBD process with four phases as demonstrated by retaining the null hypothesis for every estimation, as shown in Figure 9.5.2c. Therefore, in a practical setting we recommend initially under-fitting the data and then increase the number of phases as required to avoid over-fitting the data.

Model	Reject H_0 when H_0 is false	Retain H_0 when H_0 is false
1 Phase	50	0
2 Phases	49	1
4 Phases	0	50

Table 9.5.3: Summary of the estimation of the statistical power of our goodness of fit test of our goodness of fit test.



(a) 1 phase

(b) 2 phases

(c) 4 phases

Figure 9.5.2: Histograms of the empirical p-values from the estimation of statistical power of our goodness of fit test.

9.6 Summary

The goodness of fit test developed in this chapter provides a way of determining whether a particular type of QBD process captures the stationary and transient behaviour observed in a queueing system data set. If the null hypothesis is rejected, the diagnostic method provides further insight into the discrepancies between the expected behaviours of the particular type of QBD process and the behaviours observed in the data.

Given the computational demand inherent from our goodness of fit test, we only provided a small simulation study to explore the statistical power and significance of our goodness of fit test. However, the results provided some evidence that our goodness of fit method correctly rejects or retains the null hypothesis where required. Further analysis is therefore needed to approximate the statistical power and significance of this goodness of fit test.

Chapter 10

Application: Royal Adelaide Hospital intensive care unit

The identification of a suitable queueing model for an intensive care unit is a challenging task due to the unpredictable nature of patient admissions and the complexity of patient treatment and bed management. Additionally, the dependence between the patient admission process and the distribution of patient length of stay renders standard queueing models invalid for modelling ICUs.

If we simply observe the input and output of an ICU, we obtain information about the changes in the number of patients in the ICU but minimal information is obtained about the servers and waiting spaces in an ICU. Hence, the definition of the number of servers and waiting spaces available for a standard queueing model remains unclear. Therefore, we use QBD processes to model the bed occupancy of an ICU due to the freedom in the distribution of sojourn times and the ability to model ICUs with a dependence between the patient admission process and the distribution of patient length of stay.

In this chapter, we fit various QBD processes to the Royal Adelaide Hospital intensive care unit data set using the methods developed in Chapters 7 and 8. First, we consider the class of level-dependent QBD processes to gain insight into the operation of the RAH ICU. Second, we identify and fit suitable structured QBD processes to fit to the bed occupancy data. We then assess the fit of each structured QBD process using the methods described in Chapter 9. Using the best fitting structured QBD process, we then predict the behaviour of the RAH ICU under various scenarios.

10.1 Insight into the RAH ICU

In this section, we look towards gaining insight into the operation of the Royal Adelaide Hospital ICU by means of level-dependent QBD processes. We consider two level-dependent QBD processes; a level-dependent QBD process with one phase (general birth-and-death process) and a level-dependent QBD process with two phases.

Recall from Section 5.1 that the minimum observed bed occupancy was 20 and the maximum observed bed occupancy was 45, as illustrated in Figure 10.1.1. Despite the physical lower limit of bed occupancy in the RAH ICU being 0, the minimum bed occupancy in our queueing models will remain at 20 since there is no data associated with bed occupancies less than 20. On the other hand, we have data up to and including the physical upper limit of 45 beds. Therefore, the QBD processes considered in this chapter will be finite.

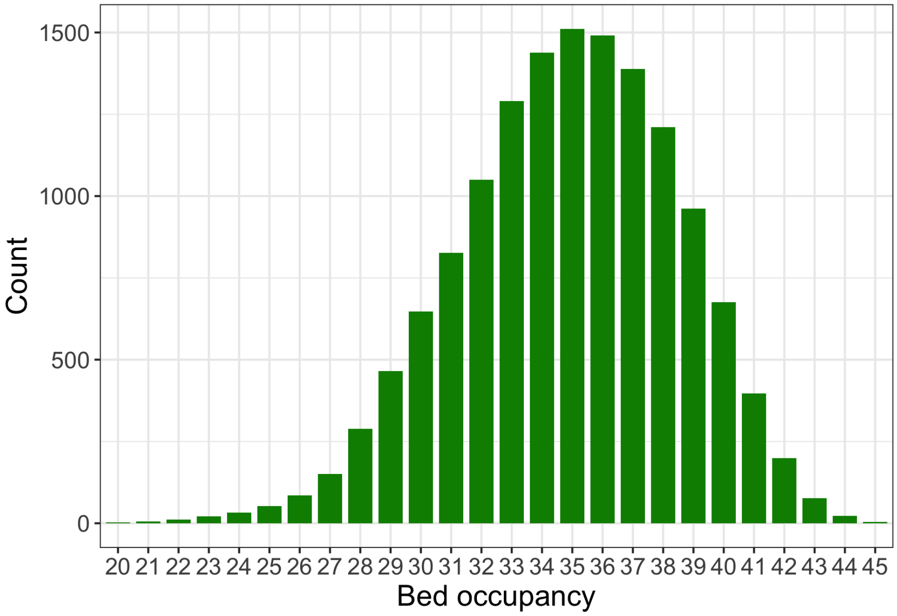


Figure 10.1.1: Bar graph of the observed bed occupancy in the RAH ICU.

10.1.1 Level-dependent QBD process with one phase

In this section, we fit a level-dependent QBD process with one phase to the RAH ICU data set, where the generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

Note that level 0 represents a bed occupancy of 20 and level 25 represents a bed occupancy of 45.

For a level-dependent QBD process with one phase, there are no hidden phase transitions. Hence, we can directly calculate the maximum likelihood estimates of the parameters for the level-dependent QBD process with one phase without the EM algorithm.

As shown in Figure 10.1.2, the admission rate of patients to the ICU tends to decrease as the number of patients in the ICU increases and the discharge rate of patients from the ICU remains relatively constant. The variability in the estimated rates for the lower and higher bed occupancies is due to minimal data.

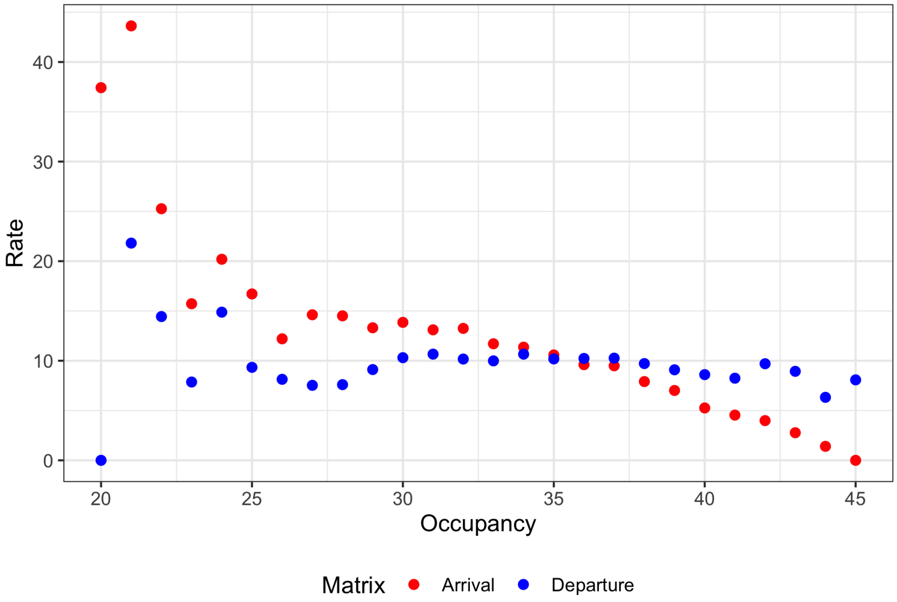


Figure 10.1.2: Maximum likelihood estimates for the level-dependent QBD process with one phase fitted to the RAH ICU data set.

The fitted level-dependent QBD process with one phase captures the long term proportion of time spent at each bed occupancy and the transition probabilities between bed occupancies, as illustrated in Figures 10.1.3 and 10.1.4. However, the distribution of sojourn times conditioned on moving up a bed occupancy are generally underestimated in terms of mean and variance, as demonstrated in Figure 10.1.5. Additionally, the mean sojourn time conditioned on moving down a bed occupancy is generally overestimated, whereas the variance is generally underestimated.

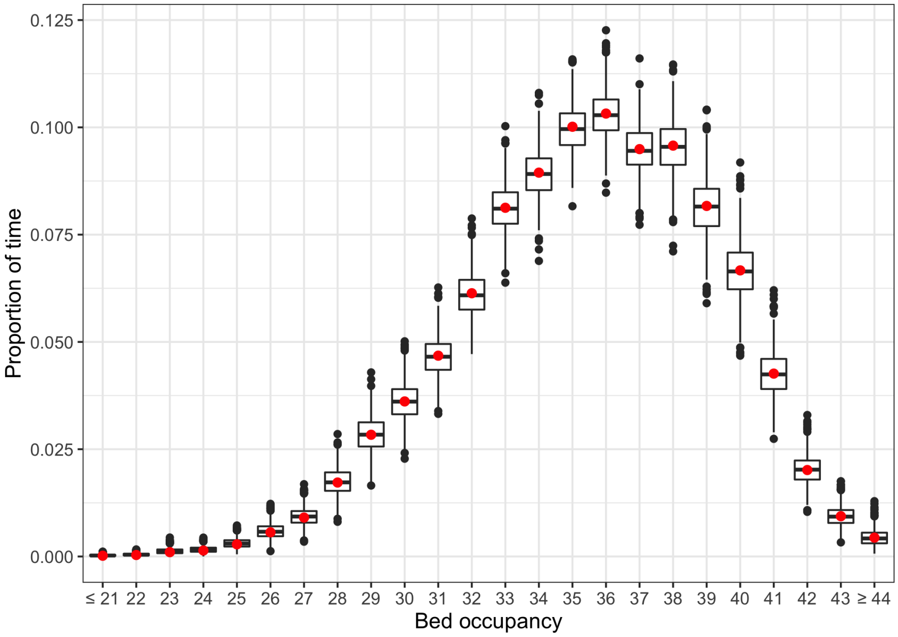


Figure 10.1.3: Comparison of the observed proportion of time spent at each bed occupancy (red points) to the observed proportion of time spent at each bed occupancy based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

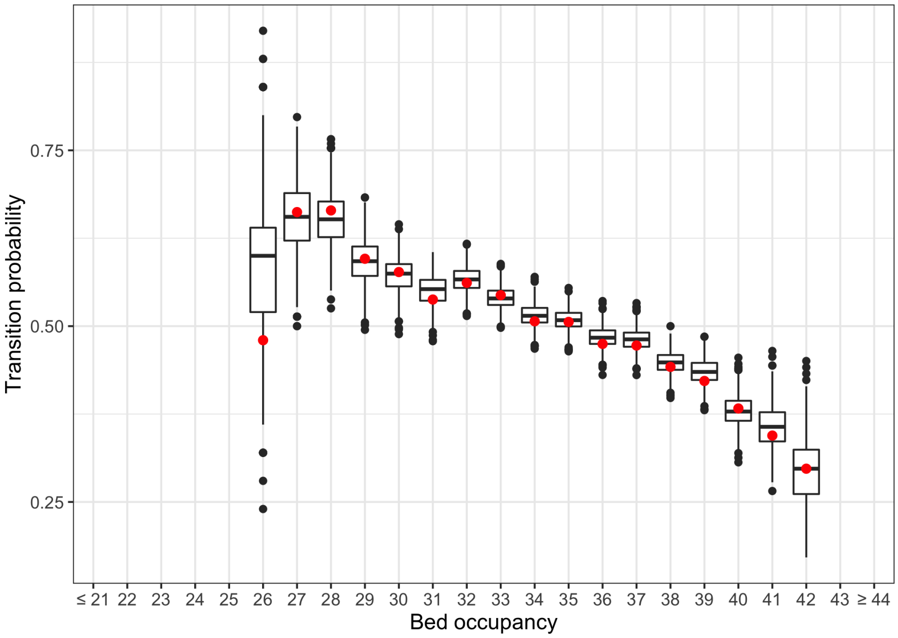
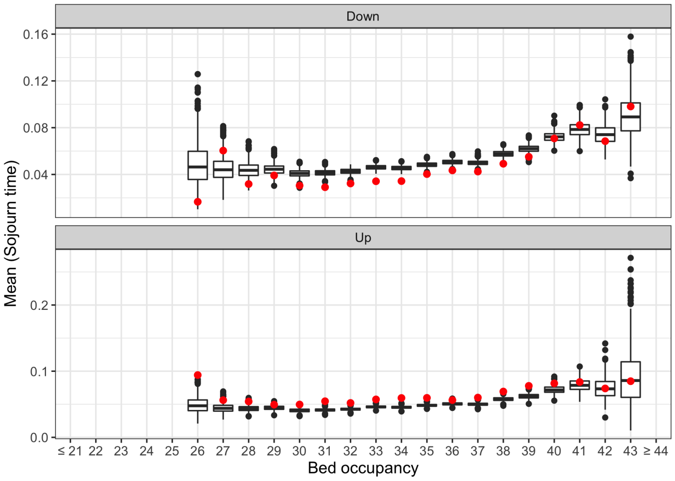
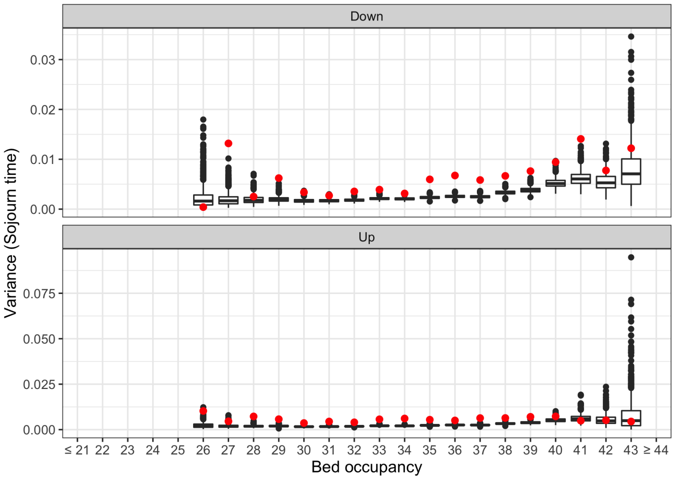


Figure 10.1.4: Comparison of the observed upward transition probabilities between bed occupancies (red points) to the observed upward transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots). Note that the data associated with upward transitions below 26 and above 42 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a)



(b)

Figure 10.1.5: Comparison between the first (a) and second (b) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted level-dependent QBD process with one phase (box-plots). Note that the data associated with transitions below 26 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

10.1.2 Level-dependent QBD process with two phases

Next, we fit a level-dependent QBD process with two phases to the RAH ICU data, where the generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+ & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

As before, level 0 represents a bed occupancy of 20 and level 25 represents a bed occupancy of 45.

Particularly for the phase transitions (1, 1) and (2, 2), the admission rate of patients to the ICU decreases as the number of patients in the ICU increases and the discharge rate of patients from the ICU remains relatively constant, as shown in Figure 10.1.6.

Further insight into the patient flow in the ICU is gained by considering the dominant transitions within each phase transition. Given that the discharge rates are the dominant rates for phase (1, 1) transitions and the admission rates are the dominant rates for phase (2, 2) transitions, we label phase 1 as the discharge phase and phase 2 as the admission phase. Hence, the ICU at the RAH is more likely to see an admission following an admission rather than an arrival following a discharge, and a discharge following a departure rather than a discharge following an admission.

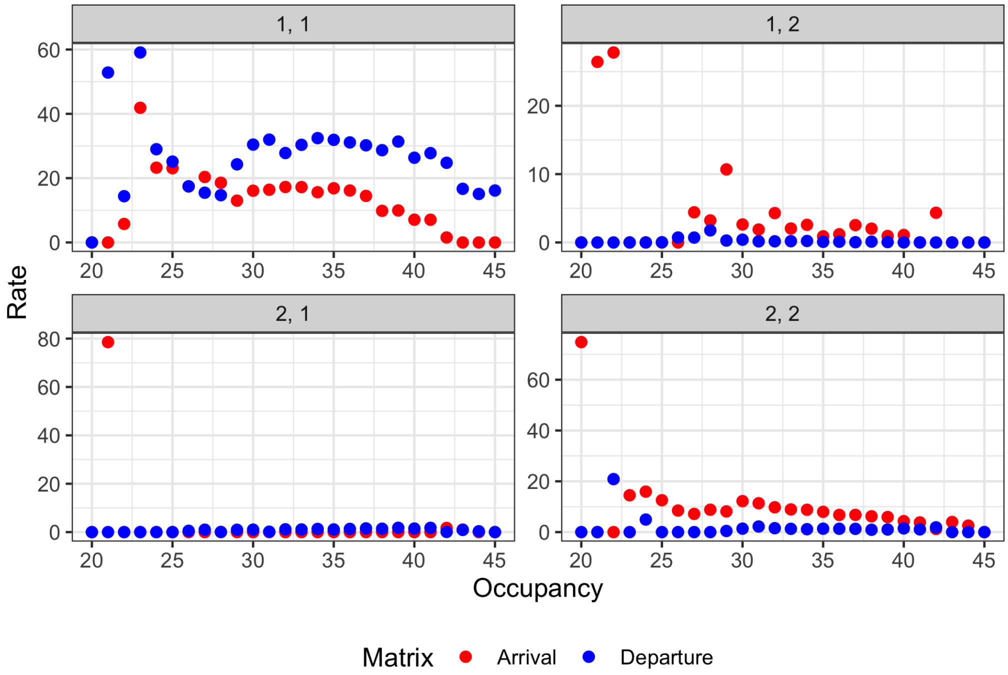


Figure 10.1.6: Maximum likelihood estimates for the level-dependent QBD process with two phases fitted to the RAH ICU data set, where a, b indicates a phase transition from a to b.

Similar to before, the fitted level-dependent QBD process with two phases captures the stationary behaviour observed in the bed occupancy data, as illustrated in Figure 10.1.7. On the other hand, Figures 10.1.8, and 10.1.9 demonstrate that the fitted level-dependent QBD process with two phases better captures the transient behaviour compared to the fitted level-dependent QBD process with one phase. Note that the variation towards the lower and upper bed occupancies is due to minimal data.

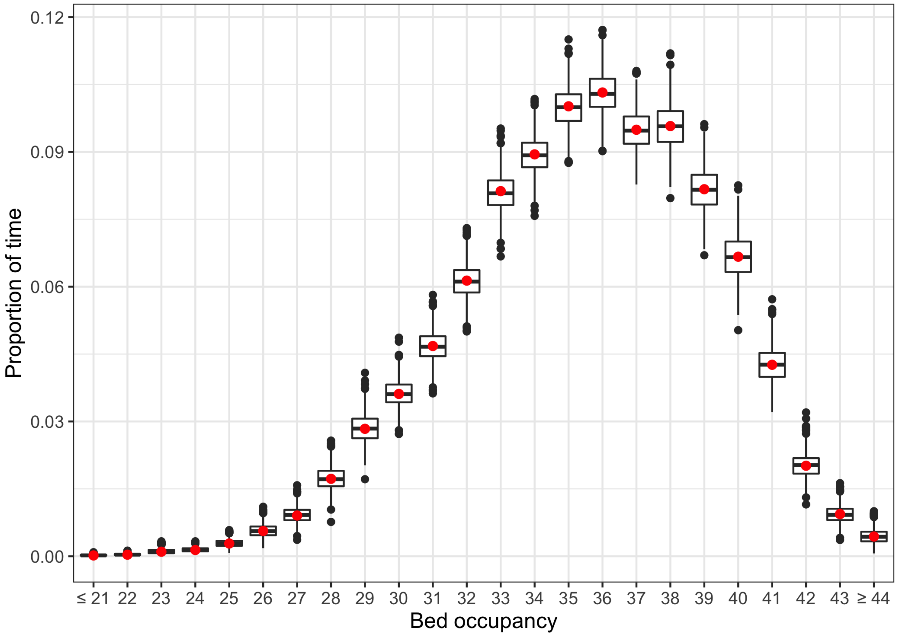


Figure 10.1.7: Comparison of the observed proportion of time spent at each bed occupancy (red points) to the observed proportion of time spent at each bed occupancy based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

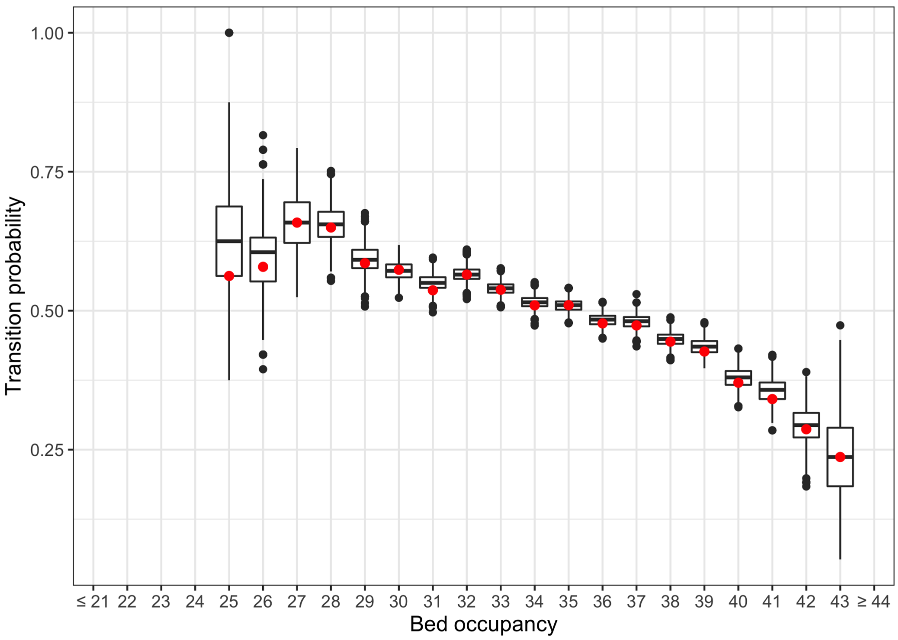
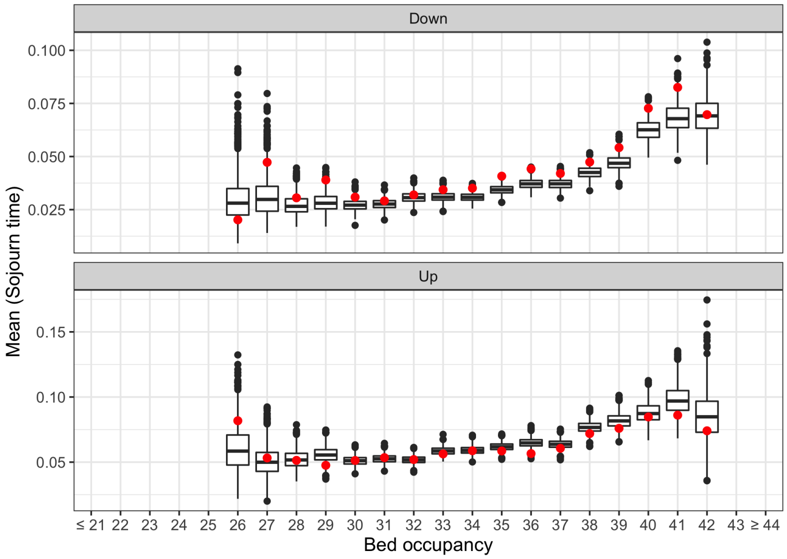
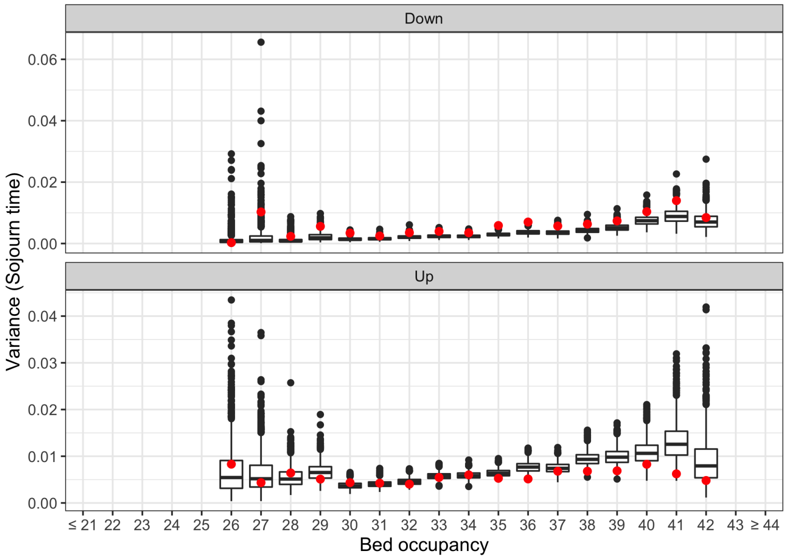


Figure 10.1.8: Comparison of the observed upward transition probabilities between bed occupancies (red points) to the observed upward transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that the data associated with upward transitions below 25 and above 42 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a)



(b)

Figure 10.1.9: Comparison between the first (a) and second (b) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that the data associated with transitions below 26 and above 42 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

10.1.3 Discussion

We have gained further insight into the operation of the RAH ICU by using level-dependent QBD processes. Evident in both fitted level-dependent QBD processes was a relatively constant discharge rate and a decreasing admission rate as the bed occupancy in the RAH ICU increases. Furthermore, the one-phase and two-phase level-dependent QBD processes captured the stationary behaviour of the RAH ICU. However, we saw an improvement in the transient behaviour when an additional phase was considered, thus emphasising the importance of phase structure in the fitted QBD process.

In addition to this, the two-phase level-dependent QBD process uncovered a dynamic relating to the admission and discharge of patients in the ICU, such that you are more likely to see a patient admission following another patient admission rather than a patient admission following a patient discharge, and a patient discharge following another patient discharge rather than a patient discharge following a patient admission.

Due to the level-dependent nature of such QBD processes, these queueing models were able to capture elements of the stationary and transient behaviour observed in the RAH ICU data set. However, level-dependent QBD processes are over-parameterised and hence cannot be used for prediction modelling. Therefore, we now look towards structured QBD processes to find a suitable model to use for prediction.

10.2 Identification of suitable models

In this section, we explore the class of structured QBD processes to find a suitable QBD process with which we can predict the behaviour of the RAH ICU under various conditions.

Within the structured QBD process framework, several forms of the infinitesimal generator matrix may be suitable for the RAH ICU. In the following sections, we first explore structured QBD processes with level-dependent scales and then extend our investigation to structured QBD processes with level and phase transition dependent scales. In doing so, we make an informative decision on the structure required within the infinitesimal generator matrix.

For simplicity, we begin our investigation by considering a one-phase structured QBD process with level-dependent polynomial forms. However, emphasis will be placed on structured QBD process with two or more phases as the general birth-and-death process was unable to capture the transient behaviour of the bed occupancy within the RAH ICU.

10.2.1 One-phase structured QBD process with level-dependent polynomial forms

First, we consider a one-phase structured QBD process with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where

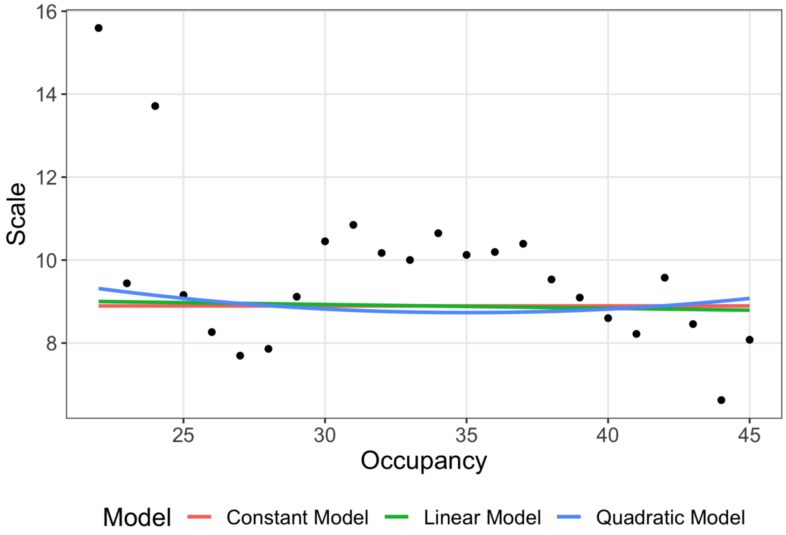
$$A_-^{(\ell)} = h_\ell A_-, \text{ for } 1 \le \ell \le 25,$$

$$A_+^{(\ell)} = f_\ell A_+, \text{ for } 0 \le \ell \le 24,$$

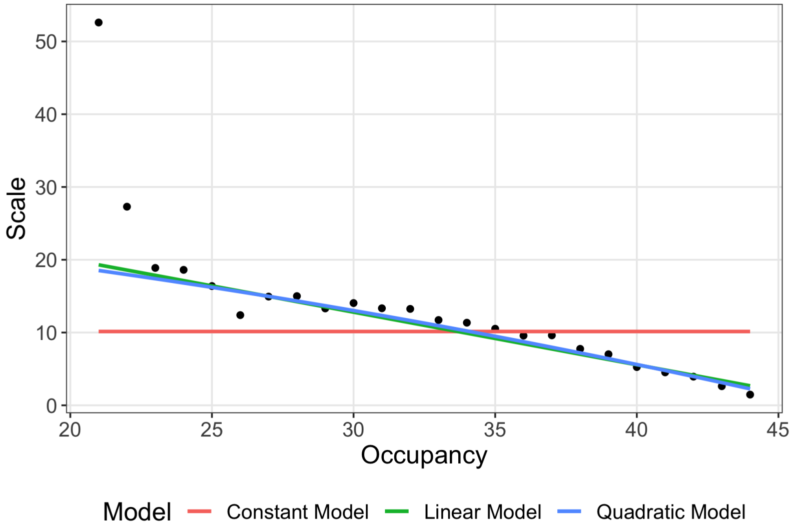
$A_0^{(\ell)}$ for $0 \le \ell \le 25$ are the negative of the relevant row sums, and f_ℓ and h_ℓ are non-negative, level-dependent scales. Given that there are no hidden phase transitions associated with a one-phase structured QBD process, we can calculate the maximum likelihood estimates of the parameters for the one-phase structured QBD process with level-dependent polynomial forms without the EM algorithm.

To determine whether polynomial forms of the level are required for a one-phase structured QBD process, we fitted a constant model, linear model, and a quadratic model to the maximum likelihood estimates of the level-dependent arrival and departure scales of the one-phase structured QBD process, as illustrated in Figure 10.2.1.

Note that we used weighted least squares regression since the lower and upper bed occupancies of the ICU are infrequently visited, which introduces uncertainty associated with those maximum likelihood estimates. In this case, the weights for each regression model considered are the normalised expected number of visits to each level under the best fitting one-phase structured QBD process with level-dependent scales.



(a) Departure



(b) Arrival

Figure 10.2.1: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level-dependent scales (points) for the fitted one-phase structured QBD process with level-dependent polynomial forms.

The level-dependent scales associated with arrivals are linearly decreasing, which suggests that the rate of admissions to the ICU decreases as the number of patients in the ICU increases. The results of the likelihood ratio test in Table 10.2.1 confirm that a quadratic term is not needed to explain the behaviour of the level-dependent scales associated with arrivals. Therefore, admissions to the RAH ICU decrease at a linear rate as the bed occupancy increases.

On the other hand, the level-dependent scales associated with departures are relatively constant across all levels which suggests that discharges from the ICU do not depend on the number of patients in the ICU. This result is also confirmed by the likelihood ratio tests presented in Table 10.2.1, which shows that neither a linear or quadratic term is needed to explain the behaviour of the level-dependent scales associated with departures.

Matrix	Model 1	Model 2	Statistic	Degrees of freedom	P-value
Departure	Constant	Linear	0.064	1	0.800
Departure	Linear	Quadratic	0.114	1	0.736
Arrival	Constant	Linear	53.987	1	2.018×10^{-13}
Arrival	Linear	Quadratic	0.230	1	0.631

Table 10.2.1: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models for the one-phase structured QBD process with level-dependent scales.

Therefore, a linear functional form of the level is most suitable for arrivals and a constant scale is most suitable for departures, such that the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

10.2.2 Two, three, and four-phase structured QBD processes with level-dependent polynomial forms

Next, we consider a structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where

$$A_-^{(\ell)} = h_\ell A_-, \text{ for } 1 \leq \ell \leq 25,$$

$$A_0^{(\ell)} = g_\ell A_0, \text{ for } 0 \leq \ell \leq 25,$$

$$A_+^{(\ell)} = f_\ell A_+, \text{ for } 0 \leq \ell \leq 24,$$

the diagonal entries of $A_0^{(\ell)}$ for $0 \leq \ell \leq 25$ are the negative of the relevant row sums, and h_ℓ, g_ℓ, and f_ℓ are non-negative, level-dependent scales.

By a similar method to Section 10.2.1, we determined suitable functional forms of the level for the two, three, and four-phase structured QBD processes with level-dependent scales.

Firstly, for a two-phase structured QBD process, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1),$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell + \beta_2^g \ell^2,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell + \beta_2^f \ell^2.$$

Secondly, for a structured QBD process with three phases, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2,$$

for $0 \leq \ell \leq 25$, the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell + \beta_2^g \ell^2,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

Lastly, for a four-phase structured QBD process, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2,$$

for $0 \leq \ell \leq 25$, the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

These results are also summarised in Table 10.2.2.

Phases	Matrix	Model
2	Departure	Linear
	Unchanged	Quadratic
	Arrival	Quadratic
3	Departure	Quadratic
	Unchanged	Quadratic
	Arrival	Linear
4	Departure	Quadratic
	Unchanged	Linear
	Arrival	Linear

Table 10.2.2: Summary of the chosen functional forms for the two, three, and four-phase structured QBD process with level-dependent scales.

See Appendices E.1.1, E.1.2, and E.1.3 for the relevant plots and likelihood ratio test results.

10.2.3 Two-phase structured QBD process with level and phase transition dependent polynomial forms

Next, we extend the EM algorithm described in Section 8.3 to allow the scales to depend on level and phase transitions. In this case, the infinitesimal generator matrix is of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where

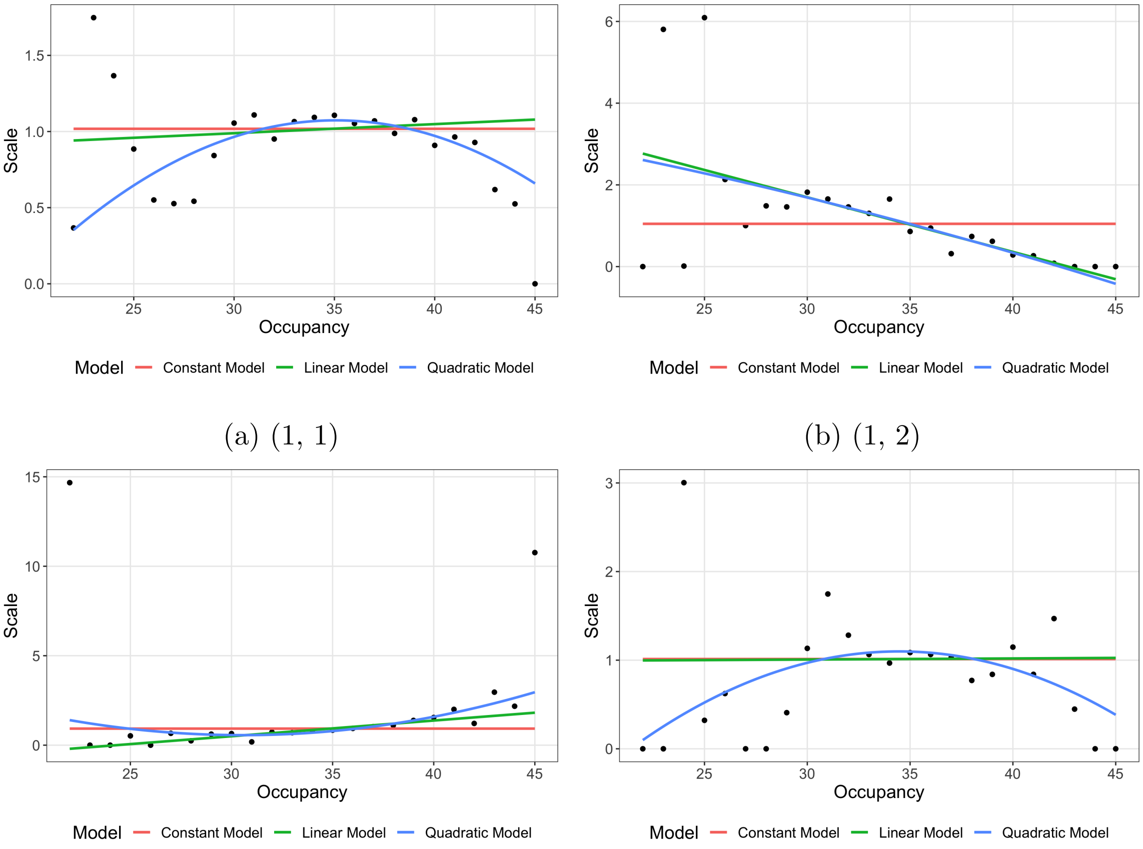
$$\left[A_-^{(\ell)}\right]_{i,j} = h_{\ell,i,j} \left[A_-\right]_{i,j}, \text{ for } 1 \leq \ell \leq 25 \text{ and } 1 \leq i, j, \leq 2,$$

$$\left[A_0^{(\ell)}\right]_{i,j} = g_{\ell,i,j} \left[A_0\right]_{i,j}, \text{ for } 0 \leq \ell \leq 25 \text{ and } 1 \leq i \neq j, \leq 2,$$

$$\left[A_+^{(\ell)}\right]_{i,j} = f_{\ell,i,j} \left[A_+\right]_{i,j}, \text{ for } 0 \leq \ell \leq 24 \text{ and } 1 \leq i, j, \leq 2,$$

the diagonal entries of $A_0^{(\ell)}$ for $0 \leq \ell \leq 25$ are the negative of the relevant row sums, and $h_{\ell,i,j}$, $g_{\ell,i,j}$, and $f_{\ell,i,j}$ are non-negative, level and phase transition dependent scales.

First, we look at the fitted constant, linear, and quadratic models of the maximum likelihood estimates of the level and phase transition dependent scales associated with a decrease in level of the fitted two-phase structured QBD process, as illustrated in Figure 10.2.2.



(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

Figure 10.2.2: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent departure scales for the fitted two-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

According to the likelihood ratio tests presented in Table 10.2.3, a quadratic term is needed to explain the behaviour of the level and phase transition dependent scales associated with the (1, 1) and (2, 1) phase transitions but it is not needed for the (1, 2) phase transition. For the (2, 2) phase transition, neither a linear or quadratic term is needed.

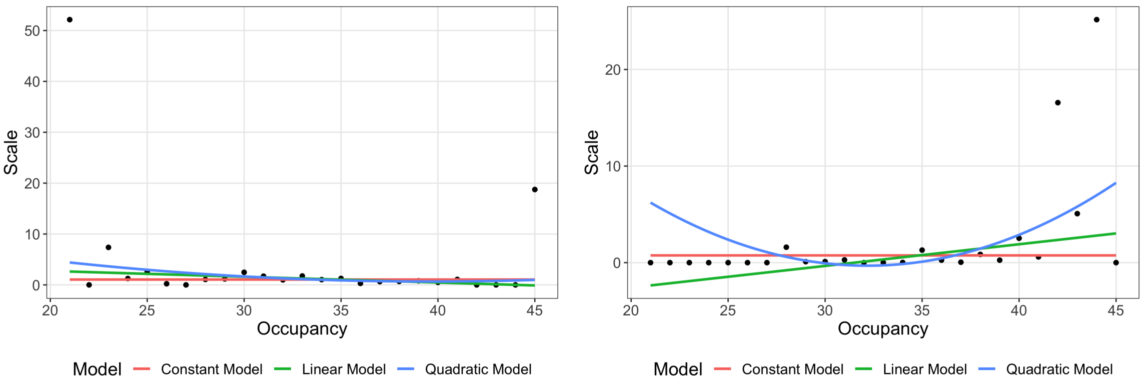
Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	0.703	1	0.402
(1, 1)	Linear	Quadratic	10.285	1	1.341×10^{-3}
(1, 2)	Constant	Linear	21.836	1	2.971×10^{-6}
(1, 2)	Linear	Quadratic	0.060	1	0.807
(2, 1)	Constant	Linear	8.673	1	3.230×10^{-3}
(2, 1)	Linear	Quadratic	4.662	1	3.083×10^{-2}
(2, 2)	Constant	Linear	0.004	1	0.947
(2, 2)	Linear	Quadratic	2.969	1	8.487×10^{-2}

Table 10.2.3: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with departures for the two-phase structured QBD process with level and phase transition dependent scales.

Next, we look at the fitted constant, linear, and quadratic models of the maximum likelihood estimates of the level and phase transition dependent scales associated with transitions within a level of the fitted two-phase structured QBD process, as illustrated in Figure 10.2.3.

As confirmed by the likelihood test results presented in Table 10.2.4, the level and phase transition dependent scales associated with the (2, 1) phase transition are better explained by a quadratic model.

There is an outlier in the plot of the maximum likelihood estimates associated with the (1, 2) phase transition which makes it difficult to visually assess the behaviour of the level and phase transition dependent scales. Including the outlier in the analysis, the results of the likelihood ratio test suggest that neither a linear or quadratic term is required to explain the behaviour of the level and phase transition dependent scales.



(a) (1, 2)

(b) (2, 1)

Figure 10.2.3: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent unchanged scales for the fitted two-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 2)	Constant	Linear	2.849	1	9.145×10^{-2}
(1, 2)	Linear	Quadratic	0.675	1	0.411
(2, 1)	Constant	Linear	3.317	1	6.858×10^{-2}
(2, 1)	Linear	Quadratic	4.718	1	2.986×10^{-2}

Table 10.2.4: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with transitions within a level for the two-phase structured QBD process with level and phase transition dependent scales.

Figure 10.2.4 plots the constant model, linear model, and a quadratic model that were fitted to the maximum likelihood estimates of the level and phase transition dependent scales that have a value below 20. Here, it appears that the level and phase transition dependent scales associated with the (1, 2) phase transition are linearly decreasing.

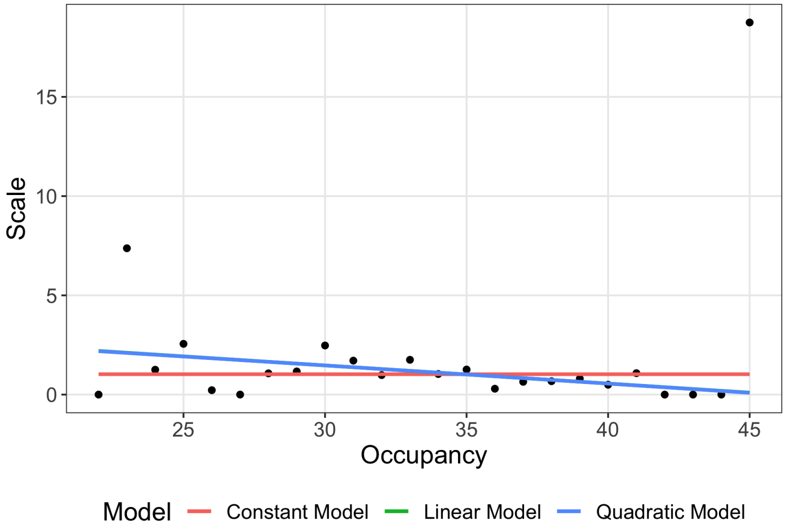


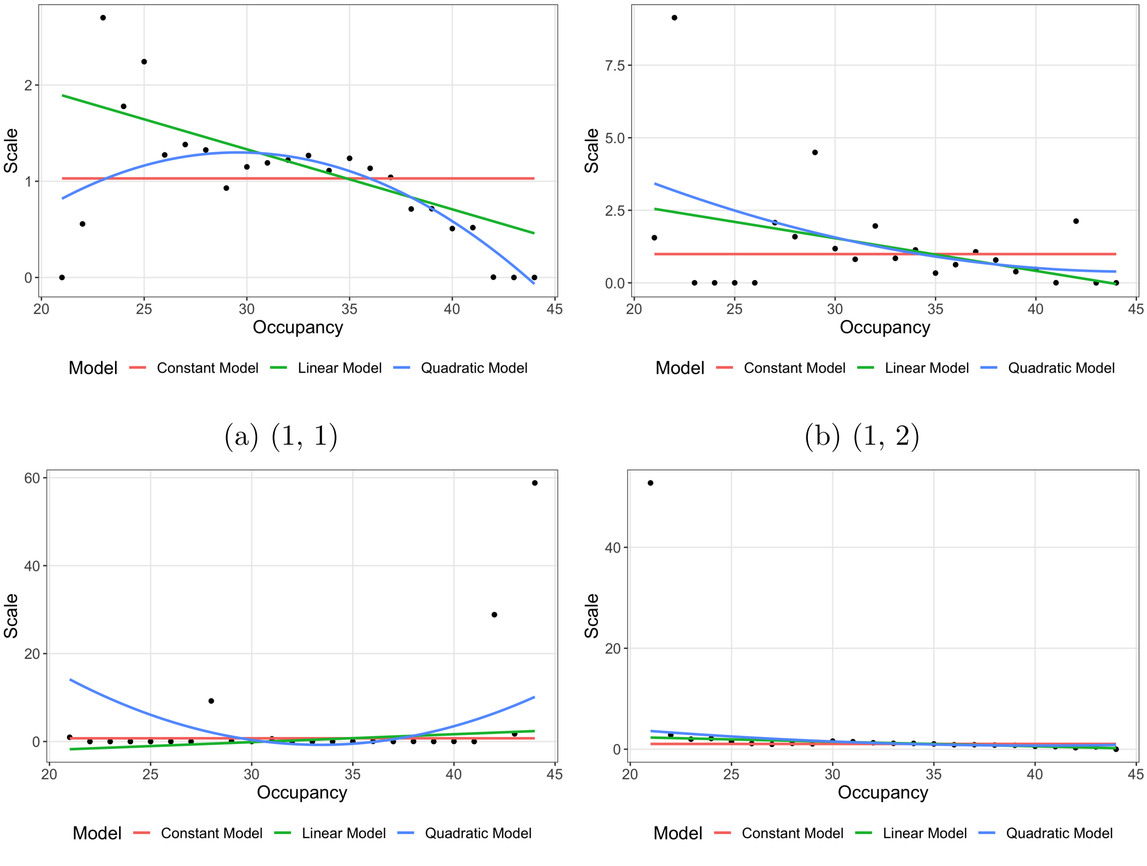
Figure 10.2.4: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent unchanged scales with a value below 20 for the fitted two-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Using the new likelihood test results presented in Table 10.2.5, a linear term is required to explain the behaviour of the level and phase transition dependent scales associated with the (1, 2) phase transition.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 2)	Constant	Linear	6.436	1	1.118×10^{-2}
(1, 2)	Linear	Quadratic	0.000	1	0.988

Table 10.2.5: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with transitions within a level for the two-phase structured QBD process with level and phase transition dependent scales with a value below 20.

Lastly, we fitted a constant model, linear model, and a quadratic model to the maximum likelihood estimates of the level and phase transition dependent scales associated with an increase in level using weighted least squares regression, as illustrated in Figure 10.2.5.



(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

Figure 10.2.5: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales for the fitted two-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

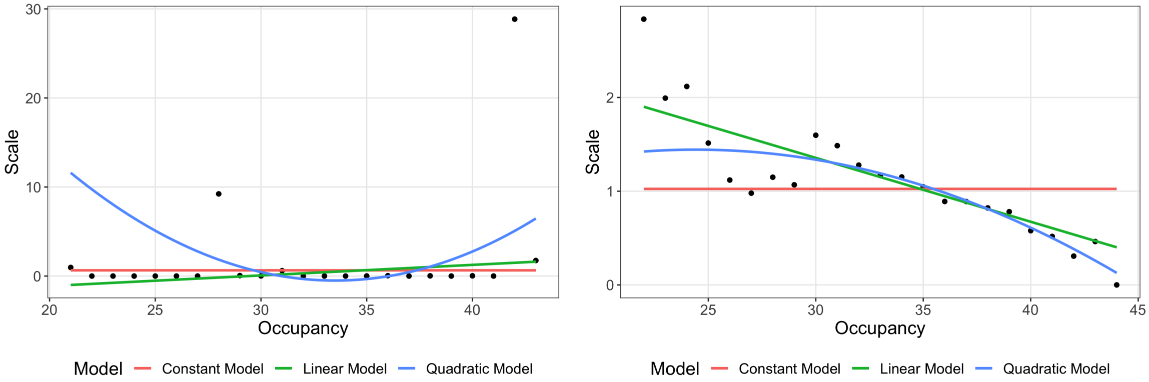
Based on the results of the likelihood ratio tests shown in Table 10.2.6, the linear model best describes the level and phase transition dependent scales associated with the (1, 2) phase transition, while the quadratic model best describes the level and phase transition dependent scales associated with the (1, 1) phase transition.

For the (2, 1) and (2, 2) phase transitions, the outliers make it difficult to visually assess the behaviour of the level and phase transition dependent scales. Including the outliers in the analysis, the results of the likelihood ratio test suggest that neither a linear or quadratic term is required to explain the behaviour of the level and phase transition dependent scales associated with the (2, 2) phase transition, but a quadratic term is needed for the (2, 1) phase transition.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	20.710	1	5.344×10^{-6}
(1, 1)	Linear	Quadratic	10.969	1	9.264×10^{-4}
(1, 2)	Constant	Linear	6.421	1	1.128×10^{-2}
(1, 2)	Linear	Quadratic	0.405	1	0.525
(2, 1)	Constant	Linear	0.569	1	0.451
(2, 1)	Linear	Quadratic	4.468	1	3.454×10^{-2}
(2, 2)	Constant	Linear	2.232	1	0.135
(2, 2)	Linear	Quadratic	0.431	1	0.511

Table 10.2.6: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals for the two-phase structured QBD process with level and phase transition dependent scales.

We further explored the behaviour associated with the (2, 1) and (2, 2) phase transitions by identifying values above 40 as outliers and removing them from the data. Figure 10.2.6 plots the constant model, linear model, and quadratic model that have been fitted to the maximum likelihood estimates of the level and phase transition dependent scales that have a value below 40. Visually, it is much easier to see that the level and phase transition dependent scales appear to be relatively constant for the (2, 1) phase transition but the level and phase transition dependent scales may decrease at a non-linear rate for the (2, 2) phase transition.



(a) (2, 1)

(b) (2, 2)

Figure 10.2.6: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales with a value below 40 for the fitted two-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

The results of the likelihood ratio test for these new models confirm that the level and phase transition dependent scales associated with the (2, 1) phase transition are constant but suggest that a quadratic term is required to explain the behaviour of the level and phase transition dependent scales associated with the (2, 2) phase transition, as shown in Table 10.2.7.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(2, 1)	Constant	Linear	0.329	1	0.566
(2, 1)	Linear	Quadratic	3.614	1	5.728×10^{-2}
(2, 2)	Constant	Linear	31.689	1	1.810×10^{-8}
(2, 2)	Linear	Quadratic	4.462	1	3.466×10^{-2}

Table 10.2.7: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals for the two-phase structured QBD process with level and phase transition dependent scales. Note that the scales used in these tests have values below 40.

To summarise, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(2,2)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1), & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1) + \beta_2^{h,i,j}(\ell-1)^2, & \text{if } (i,j) \in \{(1,1),(2,1)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(2,1)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(2,1)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1),(2,2)\}. \end{cases}$$

The results of the chosen functional forms for the two-phase structured QBD process with level and phase transition dependent scales are also summarised in Table 10.2.8.

Matrix	Phase transition	Model
Departure	(1, 1)	Quadratic
	(1, 2)	Linear
	(2, 1)	Quadratic
	(2, 2)	Constant
Unchanged	(1, 2)	Linear
	(2, 1)	Quadratic
Arrival	(1, 1)	Quadratic
	(1, 2)	Linear
	(2, 1)	Constant
	(2, 2)	Quadratic

Table 10.2.8: Summary of the chosen functional forms for the two-phase structured QBD process with level and phase transition dependent scales.

10.2.4 Three and four-phase structured QBD processes with level and phase transition dependent polynomial forms

Continuing on from the previous section, we now consider an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where

$$\left[A_-^{(\ell)}\right]_{i,j} = h_{\ell,i,j} \left[A_-\right]_{i,j}, \text{ for } 1 \leq \ell \leq 25 \text{ and } 1 \leq i, j, \leq k,$$

$$\left[A_0^{(\ell)}\right]_{i,j} = g_{\ell,i,j} \left[A_0\right]_{i,j}, \text{ for } 0 \leq \ell \leq 25 \text{ and } 1 \leq i \neq j, \leq k,$$

$$\left[A_+^{(\ell)}\right]_{i,j} = f_{\ell,i,j} \left[A_+\right]_{i,j}, \text{ for } 0 \leq \ell \leq 24 \text{ and } 1 \leq i, j, \leq k,$$

k is the number of phases, the diagonal entries of $A_0^{(\ell)}$ for $0 \leq \ell \leq 25$ are the negative of the relevant row sums, and $h_{\ell,i,j}$, $g_{\ell,i,j}$, and $f_{\ell,i,j}$ are non-negative, level and phase transition dependent scales.

Using a similar method to Section 10.2.3, we determined suitable functional forms of the level for the three and four-phase structured QBD processes with level and phase transition dependent scales.

Firstly, for the three-phase structured QBD process, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,1),\ (1,3),\\ & (2,1),\ (3,2),\ (3,3)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1), & \text{if } (i,j) \in \{(2,3)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1) + \beta_2^{h,i,j}(\ell - 1)^2, & \text{if } (i,j) \in \{(1,2),\ (2,2),\\ & (3,1)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (2,1),\ (2,3),\ (3,2)\}, \\ 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(3,1)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (2,1),\ (3,1),\ (3,2)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(1,2),(2,2),(2,3)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1),(3,3)\}. \end{cases}$$

These results are also summarised in Table 10.2.9.

Matrix	Model	Phase transitions
Departure	Constant	$(1,1)$, $(1,3)$, $(2,1)$, $(3,2)$, $(3,3)$
	Linear	$(2,3)$
	Quadratic	$(1,2)$, $(2,2)$, $(3,1)$
Unchanged	Constant	$(1,3)$, $(2,1)$, $(2,3)$, $(3,2)$
	Linear	$(1,2)$
	Quadratic	$(3,1)$
Arrival	Constant	$(1,2)$, $(1,3)$, $(2,1)$, $(3,1)$, $(3,2)$
	Linear	$(2,2)$, $(2,3)$
	Quadratic	$(1,1)$, $(3,3)$

Table 10.2.9: Summary of the chosen functional forms for the three-phase structured QBD process with level and phase transition dependent scales.

Lastly, for a structured QBD process with four phases, the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (1,4),\ (2,1),\ (2,4),\ (3,1),\ (3,4)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1), & \text{if } (i,j) \in \{(3,2),\ (4,1)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1) + \beta_2^{h,i,j}(\ell-1)^2, & \text{if } (i,j) \in \{(1,1),\ (1,2),\ (2,2),\ (2,3),\ (3,3),\ (4,2),\ (4,3),\ (4,4)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (1,4),\ (2,1),\ (2,3),\ (2,4),\ (3,1),\ (3,2),\ (3,4),\ (4,2),\ (4,3)\}, \\ 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(4,1)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(1,2)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3), (1,4), (2,1), (2,4), (3,1), \\ & (3,2), (3,3), (3,4), (4,1), (4,2), (4,3)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1), (2,2), (2,3), (4,4)\}. \end{cases}$$

These results are also summarised in Table 10.2.10.

Matrix	Model	Phase transitions
Departure	Constant	$(1,3)$, $(1,4)$, $(2,1)$, $(2,4)$, $(3,1)$, $(3,4)$
	Linear	$(3,2)$, $(4,1)$
	Quadratic	$(1,1)$, $(1,2)$, $(2,2)$, $(2,3)$, $(3,3)$, $(4,2)$, $(4,3)$, $(4,4)$
Unchanged	Constant	$(1,3)$, $(1,4)$, $(2,1)$, $(2,3)$, $(2,4)$, $(3,1)$, $(3,2)$, $(3,4)$, $(4,2)$, $(4,3)$
	Linear	$(4,1)$
	Quadratic	$(1,2)$
Arrival	Constant	$(1,3)$, $(1,4)$, $(2,1)$, $(2,4)$, $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(4,1)$, $(4,2)$, $(4,3)$
	Linear	$(1,2)$
	Quadratic	$(1,1)$, $(2,2)$, $(2,3)$, $(4,4)$

Table 10.2.10: Summary of the chosen functional forms for the four-phase structured QBD process with level and phase transition dependent scales.

See Appendices E.1.4 and E.1.5 for the relevant plots and likelihood ratio test results.

10.2.5 Discussion

The analysis and modelling presented above assumes that the phases do not interchange within the structured QBD process. However, this is an invalid assumption as there is no direct meaning of a phase in the context of a queueing system and phases can interchange as needed to explain the behaviour of the queueing process. Hence, the functional forms of a structured QBD process should not be pre-specified and instead the selection of the best fitting functional forms should be incorporated in the EM algorithm itself.

For completeness, we continue with the investigation of finding the best fitting structured QBD process for the RAH ICU. However, we suggest the development of a statistical method for structured QBD process that correctly incorporates the selection of functional forms as an area of future research.

10.3 Structured QBD process estimation and goodness of fit

In this section, we find the most suitable model for the RAH ICU by fitting and assessing the fit of each structured QBD processes described in Section 10.2. In particular, we use the methods developed in Chapters 8 and 9 to assess the fit of each fitted structured QBD process and then identify the most suitable model for the RAH ICU.

10.3.1 Four-phase structured QBD process with level and phase transition dependent polynomial forms

In this section, we consider a four-phase structured QBD process, such that the null and alternative hypotheses are defined as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all four-phase structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (1,4), \\ & (2,1),\ (2,4),\ (3,1),\ (3,4)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1), & \text{if } (i,j) \in \{(3,2),\ (4,1)\}, \\ 1 + \beta_1^{h,i,j}(\ell-1) + \beta_2^{h,i,j}(\ell-1)^2, & \text{if } (i,j) \in \{(1,1),\ (1,2), \\ & (2,2),\ (2,3),\ (3,3),\ (4,2), \\ & (4,3),\ (4,4)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (1,4),\ (2,1),\ (2,3),\ (2,4), \\ & (3,1),\ (3,2),\ (3,4),\ (4,2),\ (4,3)\}, \\ 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(4,1)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(1,2)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\ (1,4),\ (2,1),\ (2,4),\ (3,1), \\ & (3,2),\ (3,3),\ (3,4),\ (4,1),\ (4,2),\ (4,3)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1),\ (2,2),\ (2,3),\ (4,4)\}. \end{cases}$$

Using the hypothesis test described in Section 9.2, we obtain an empirical p-value of 0.02 which suggests that a four-phase structured QBD process with the functional forms defined above does not completely capture the elements of the stationary and transient behaviour observed in the RAH ICU data. We now further explore this result by considering the diagnostic plots relating to the stationary and transient behaviour.

Figure 10.3.1 illustrates that the fitted four-phase structured QBD process closely estimates the proportion of time spent in the most frequently visited bed occupancies, but generally poorly estimates the other bed occupancies.

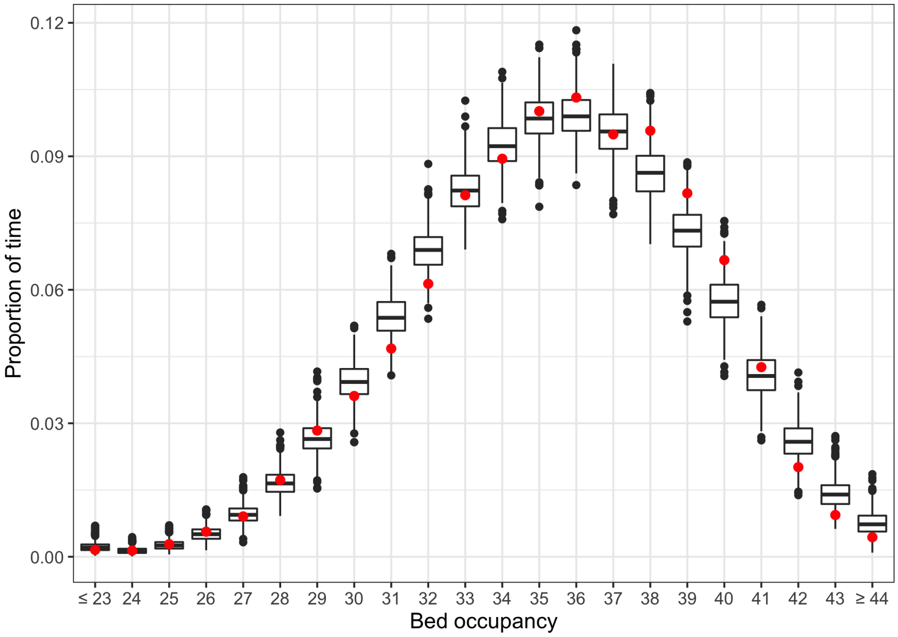


Figure 10.3.1: Comparison of the observed proportion of time spent at each bed occupancy (red points) to the observed proportion of time spent at each bed occupancy based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level and phase transition polynomial forms (boxplots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

With the exception of the less frequently visited bed occupancies, the transition probabilities of the fitted four-phase structured QBD process are similar to those observed in the RAH ICU data set, as demonstrated in Figure 10.3.2. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

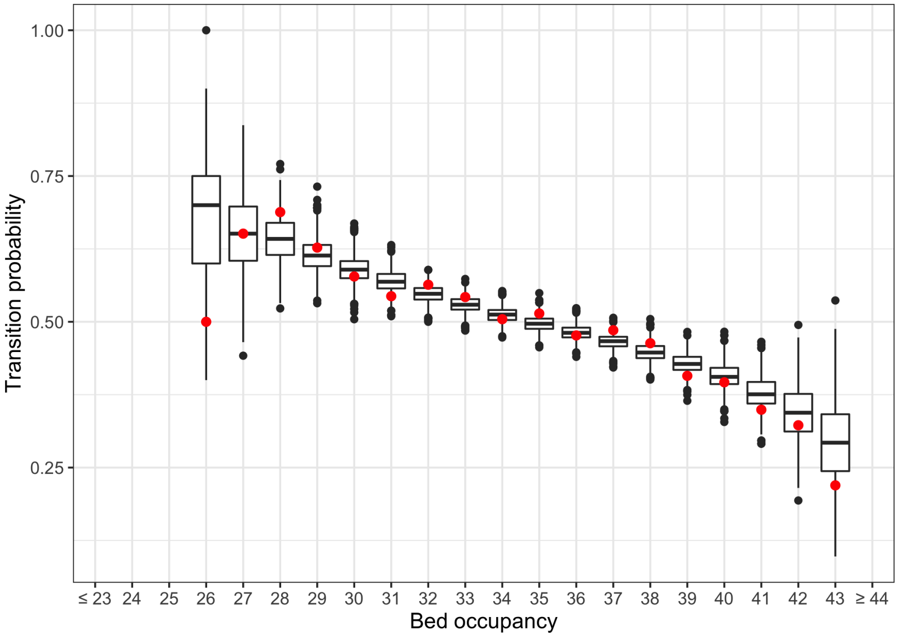
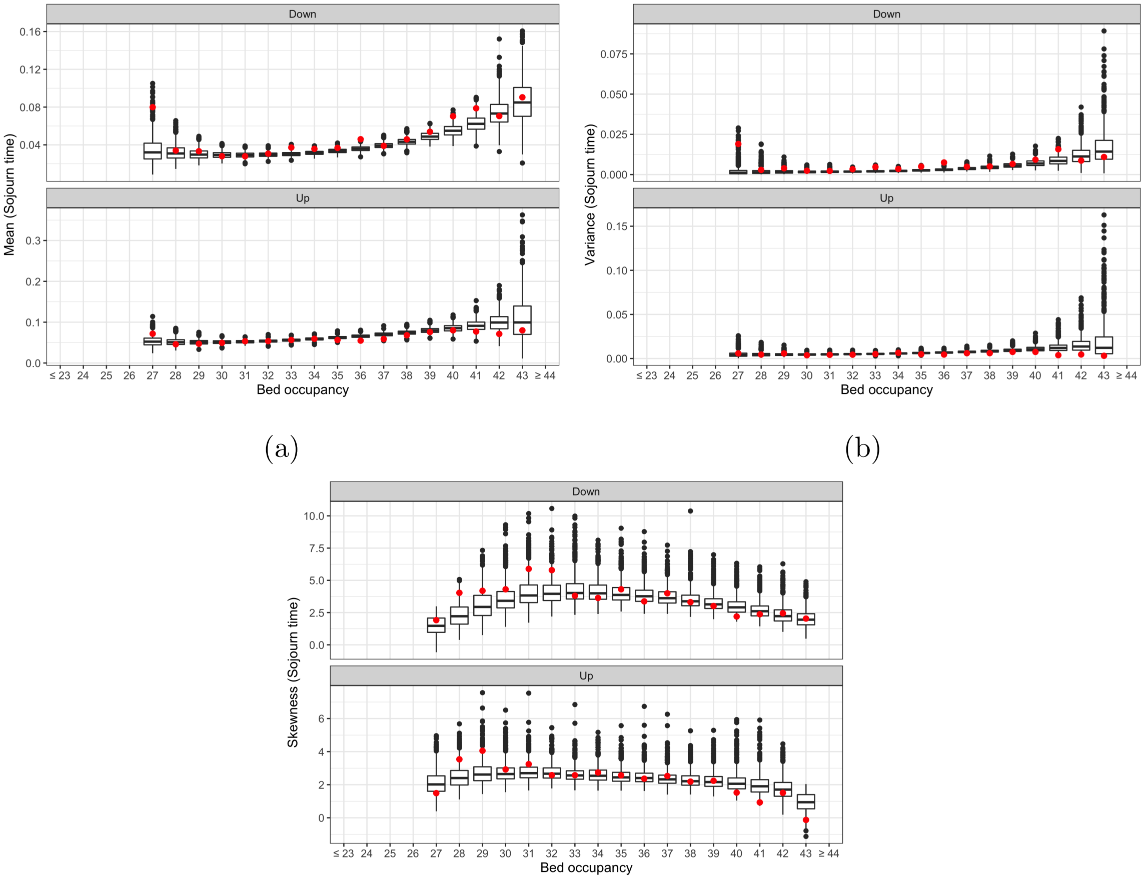


Figure 10.3.2: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level and phase transition polynomial forms (boxplots). Note that the transition probabilities for bed occupancies below 26 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Particularly for the most frequently visited bed occupancies, the fitted four-phase structured QBD process closely models the sojourn time within each bed occupancy, as illustrated in Figures 10.3.3a, 10.3.3b, and 10.3.3c. These results are further illustrated in Figures 10.3.4 and 10.3.5, which visually compare the observed distributions of conditional sojourn times compared to the simulated distributions of conditional sojourn times from the fitted four-phase structured QBD process.

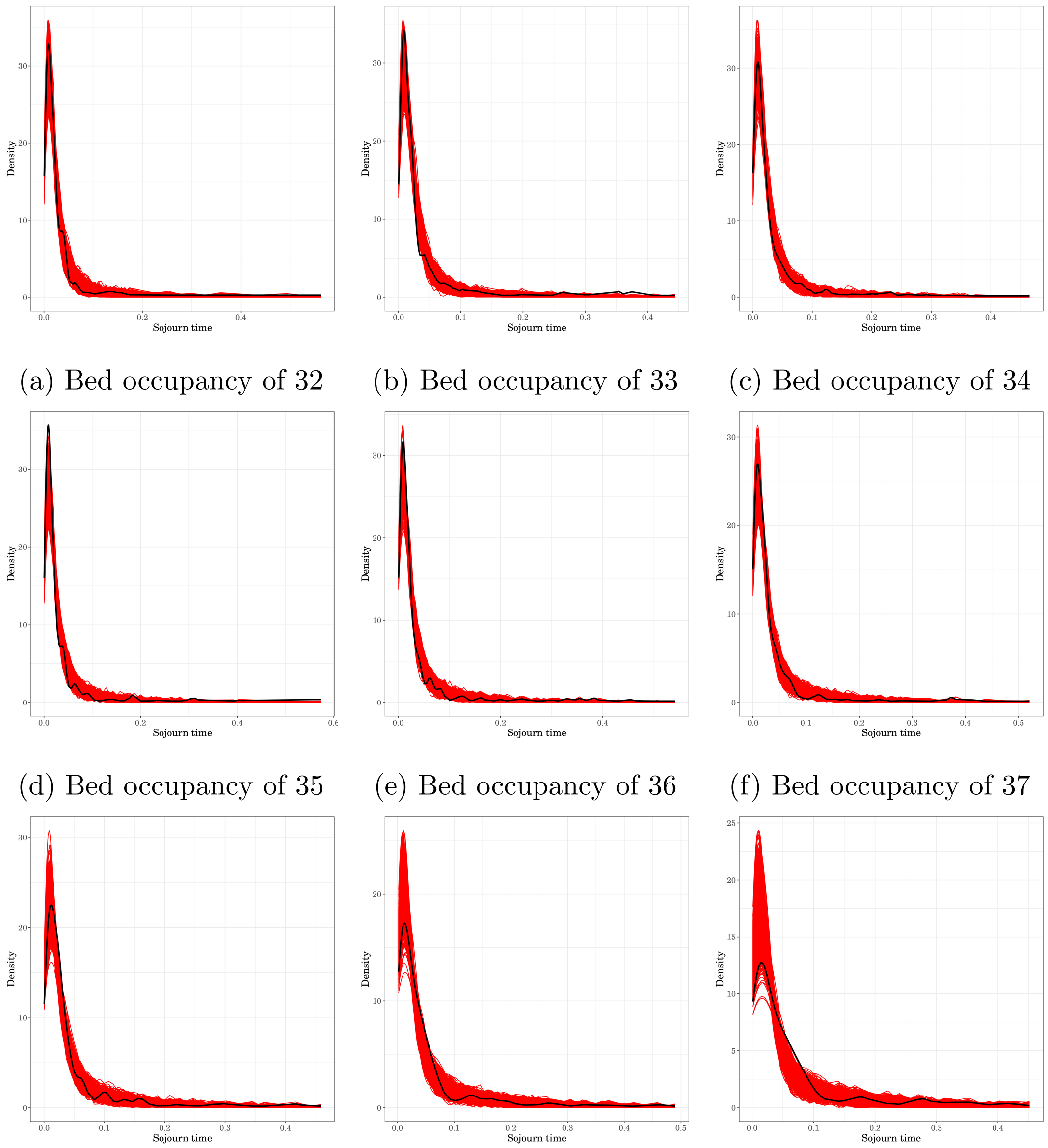


(a)

(b)

(c)

Figure 10.3.3: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level and phase transition polynomial forms (box-plots). Note that the conditional sojourn times for bed occupancies below 27 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

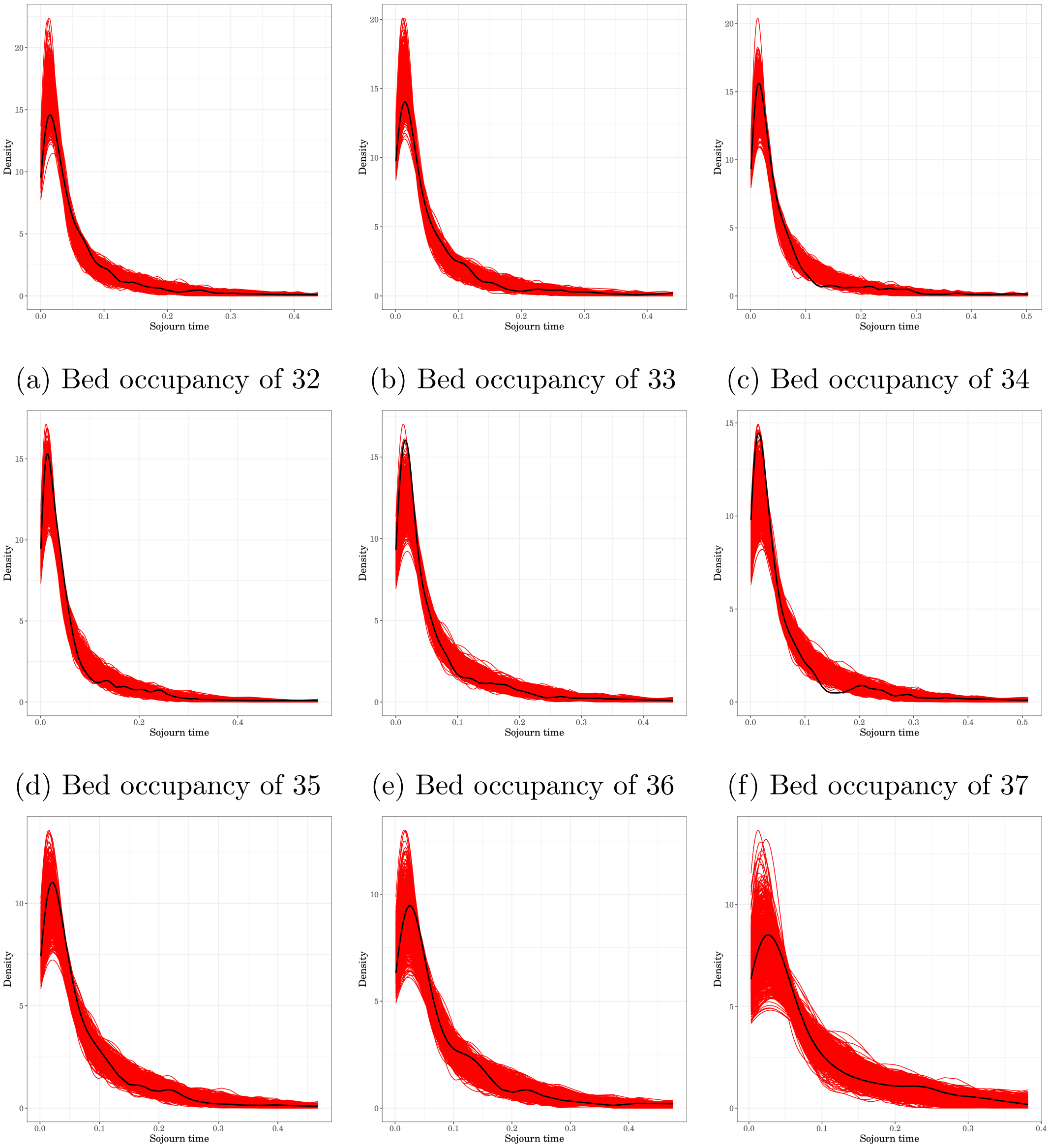


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure 10.3.4: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms (red lines). Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure 10.3.5: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level and phase transition polynomial forms (red lines). Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

10.3.2 Other structured QBD processes

Given that the method of assessing the fit of each structured QBD process is similar to above, we summarise the results of the goodness of fit test for each type of structured QBD process considered in Table 10.3.1. We refer the reader to Appendix E.2 for a detailed discussion of the diagnostic plots associated with each goodness of fit test performed.

Number of phases	Polynomial forms	Section	Empirical p-value
1	Level only	10.2.2	0
2	Level only	10.2.2	0.002
3	Level only	10.2.2	0.001
4	Level only	10.2.2	0
2	Level and phase	10.2.3	0
3	Level and phase	10.2.4	0.006

Table 10.3.1: Summary of the empirical p-values obtained from the goodness of fit test performed on each type of structured QBD process considered. Note that the Section column refers to the section in which the structured QBD process was defined.

10.3.3 Discussion

Based on the results presented in Table 10.3.1, the four-phase structured QBD process with level and phase transition dependent polynomial forms is the most suitable model for the RAH ICU. While the four-phase structured QBD process captured some elements of the stationary and transient behaviour observed in the RAH ICU data set, improvements to the model could still be made.

In any case, the best fitting model of those considered in this chapter is the four-phase structured QBD process with level and phase transition dependent polynomials forms. In the following sections, we first demonstrate how this structured QBD process compares to the standard queueing models often used to model ICUs. We then predict future behaviour of the RAH ICU under various scenarios.

10.4 Comparison to standard queueing models

In this section, we use the diagnostic plots developed in Section 9.3 to illustrate how the four-phase structured QBD process with level and phase transition dependent polynomial forms compares to the $M/M/\bullet$ queueing models that are often used to model intensive care units [2].

In terms of the long term proportion of time spent at each bed occupancy, the four-phase structured QBD process provides a much better fit than the $M/M/\bullet$ queueing models, as illustrated in Figure 10.4.1. Additionally, the $M/M/\infty$ and $M/M/42$ queueing models allow transitions above the maximum bed occupancy of 45 as indicated by the predicted proportion of time spent above a bed occupancy of 45 being positive. However, such transitions do not exist within an intensive care unit with a finite number of beds and resources, thus highlighting the invalidity of such queueing models.

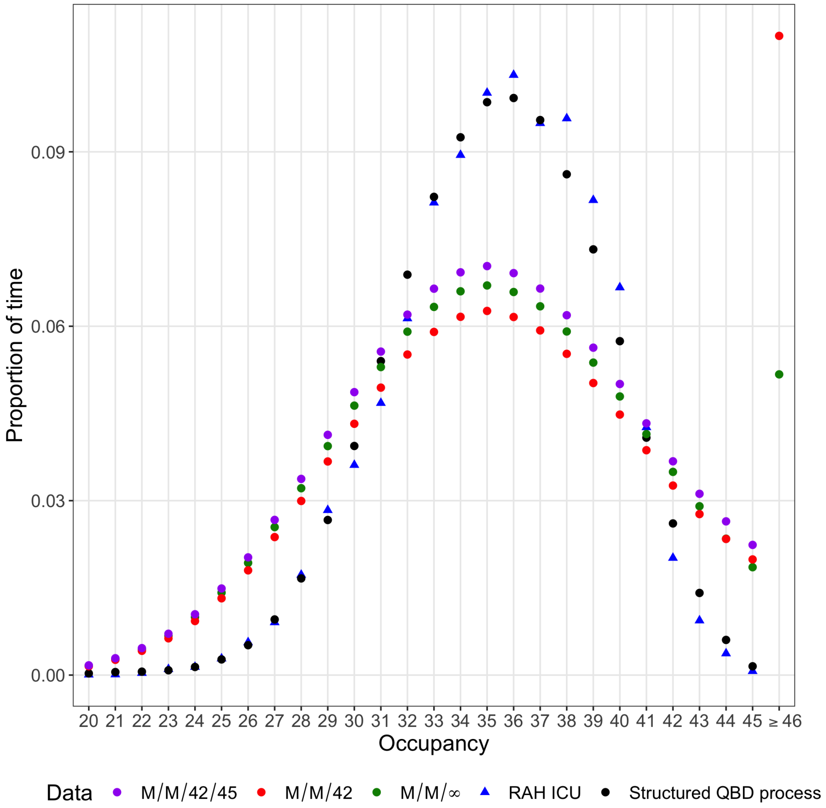


Figure 10.4.1: Comparison of the observed proportion of time spent at each bed occupancy (blue) versus the proportion of time spent at each bed occupancy assuming an $M/M/42/45$ queueing model (purple), assuming an $M/M/42$ queueing model (red), assuming an $M/M/\infty$ queueing model (green), and assuming a four-phase structured QBD process with level and phase transition dependent polynomial forms (black).

Given that the stationary behaviour of the bed occupancy of the RAH ICU is incorrectly estimated by the $M/M/\bullet$ queueing models, no further comparison between the $M/M/\bullet$ queueing models and the fitted structured QBD process is required here. We therefore refer the reader to Appendix F for a discussion of the transient behaviour.

10.4.1 Discussion

From the analysis and modelling presented above, we have developed a queueing model that provides a better fitting model compared to the standard $M/M/\bullet$ queueing models. In addition to this, we have captured elements of the stationary and transient behaviour of the bed occupancy of the RAH ICU without the use of an over-parameterised queueing model. Therefore, structured QBD processes can provide more accurate and meaningful insights into the behaviour of the RAH ICU compared to $M/M/\bullet$ queueing models.

10.5 Predicting future behaviour under various scenarios

Patient admissions to an intensive care unit are known for being unpredictable and highly variable [2, 46]. For example, admissions to an ICU may change suddenly due to an accident resulting in several emergency admissions, or the overall trend in admissions may change as a result of a decline in the demand of elective surgical procedures.

Insight into how the stationary and transient behaviour of the RAH ICU changes as the rate of patient admissions increases can be useful in situations where there is an expected surge in elective surgeries or a series of unexpected emergency admissions. On the other hand, a reduction in the rate of admissions (e.g. fewer elective surgical procedures) will alleviate the pressure on ICU staff and resources but the extent to which that occurs can be answered through the use of a fitted structured QBD process.

In addition to this, patient length of stay within an ICU also varies considerably and often depends on the state of the ICU, the health status of the patient, and the complexity of the medial condition [2, 46]. However, policy changes may change the patient discharge process from the RAH ICU, and so an exploration into how the behaviour of the RAH ICU is affected as the rate of discharges changes may also prove useful.

As the COVID-19 pandemic unfolded across the world, the demand for intensive care resources increased dramatically. As a result, the need for more informative models of intensive care units increased [35, 48]. Through the enforcement of worldwide lockdowns, intensive care units saw a reduction in the number of non-COVID-19 related admissions due to the cancellation of elective procedures and an overall reduction in human movement [23, 44]. However, non-COVID-19 related intensive care was still required for some patients and so only a proportion of intensive care resources were reserved for COVID-19 patients who required critical care.

Without an appropriate model of an ICU, we cannot predict how an ICU will respond to changes in the admission and discharge processes or in the number of intensive care resources available. Therefore, in this section we use the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms to predict the behaviour of the RAH ICU under the following scenarios:

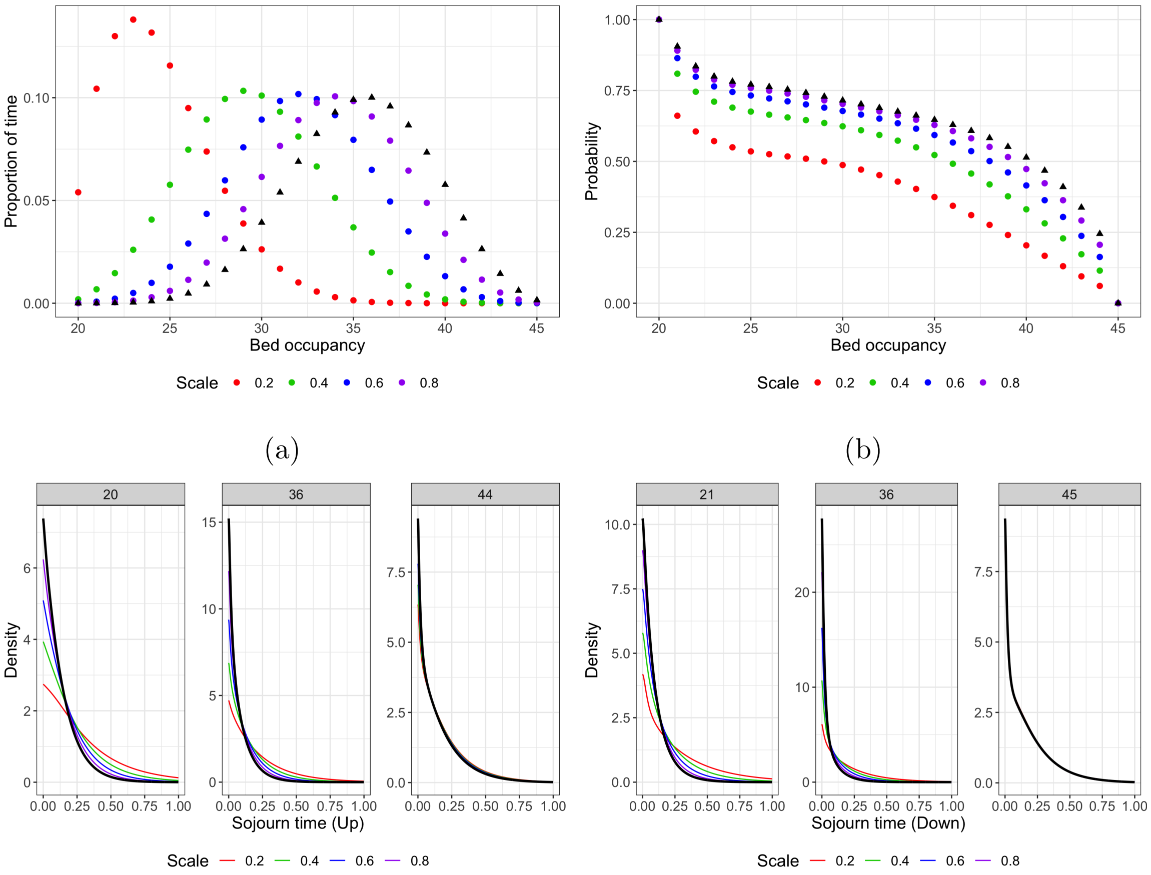
- the rate of patient admissions decreases/increases,
- the rate of patient discharges decreases/increases, and
- a proportion of the number of beds within the ICU are reserved for a certain type of patient.

10.5.1 Decrease in patient admissions

First, we explore the effect of decreasing the rate of patient admissions to the ICU by scaling the $A_+^{(\ell)}$ block matrices of the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms. That is, we reduce the rate of admissions to the RAH ICU by $(1-\gamma)100\%$ through scaling the $A_+^{(\ell)}$ block matrices by $\gamma = \{0.2, 0.4, 0.6, 0.8\}$, such that

$$Q = \begin{bmatrix} A_0^{(0)} & \gamma A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & \gamma A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & \gamma A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & \gamma A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

Figure 10.5.1 illustrates how the stationary and transient behaviour change as the rate of patient admissions to the RAH ICU decreases. In particular, we notice a decrease in proportion of time spent above 85% capacity as the rate of admissions decreases, as well as an increase in the amount of time the ICU spends without a patient admission or discharge. Note that 85% capacity is recognised by healthcare professionals as a key performance measure and the capacity beyond which a healthcare system may begin to struggle with demand [27, 30].



(a) (b)

(c) (d)

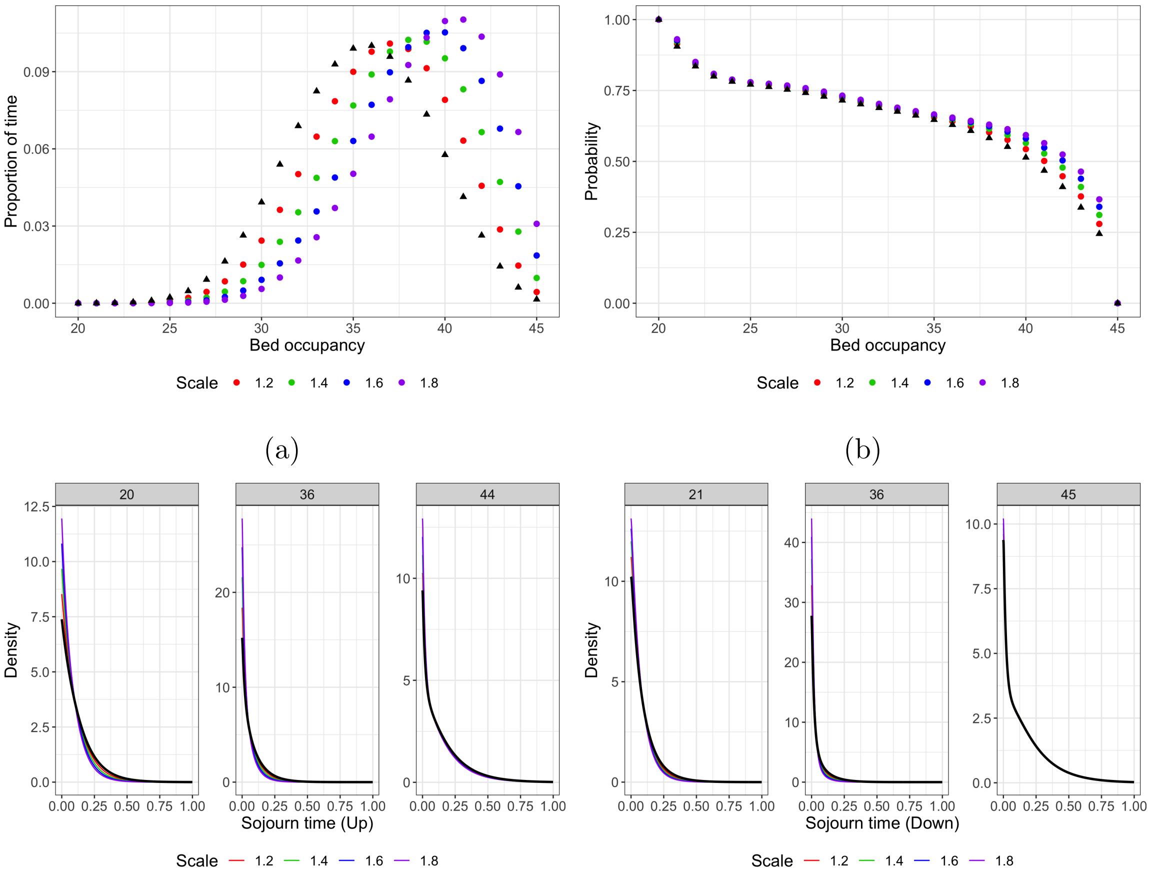
Figure 10.5.1: Illustrations of how the stationary and transient behaviour of the RAH ICU changes as the rate of patient admissions decreases. (a) Comparison of the bed occupancy stationary distributions for each reduction in arrival rate. (b) Comparison of the upward bed occupancy transition probabilities for each reduction in arrival rate. (c) Comparison of the probability density functions of the upward sojourn times for selected bed occupancies and for each reduction in arrival rate. (d) Comparison of the probability density functions of the downward sojourn times for selected bed occupancies and for each reduction in arrival rate. Note that the black triangles indicate the stationary distribution without a reduction in arrival rate.

10.5.2 Increase in patient admissions

Next, we explore the effect of increasing the rate of patient admissions to the ICU by $(\gamma - 1)100\%$ through scaling the $A_+^{(\ell)}$ block matrices by $\gamma = \{1.2, 1.4, 1.6, 1.8\}$, such that

$$Q = \begin{bmatrix} A_0^{(0)} & \gamma A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & \gamma A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & \gamma A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & \gamma A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

As shown in Figure 10.5.2a, the proportion of time spent above 85% capacity increases as the rate of admissions increases. Despite increasing the rate of patient admissions, the upward transition probabilities remain relatively the same as those of the fitted structured QBD process, as illustrated in Figure 10.5.2b. However, the time between a patient admission or discharge generally decreases as the rate of admissions increases, as demonstrated in Figures 10.5.2c and 10.5.2d.



(a)

(b)

(c)

(d)

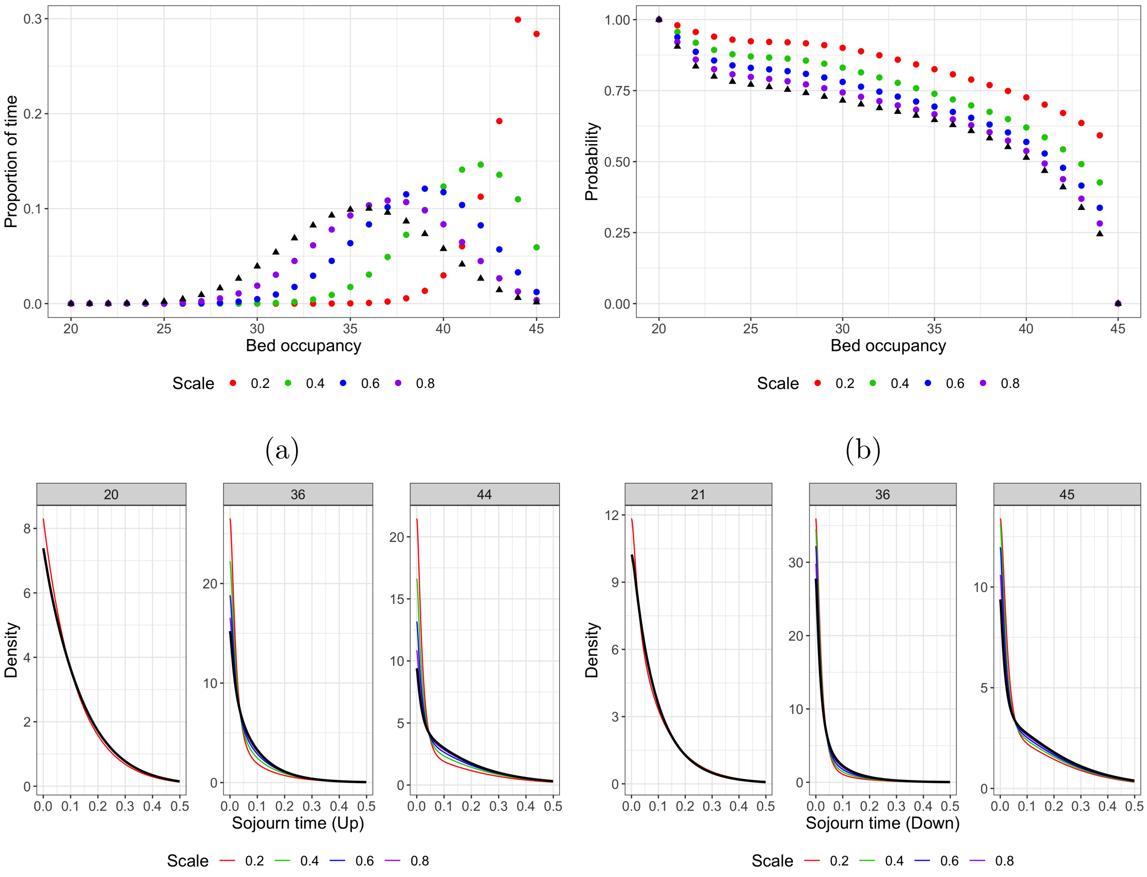
Figure 10.5.2: Illustrations of how the stationary and transient behaviour of the RAH ICU changes as the rate of patient admissions increases. (a) Comparison of the bed occupancy stationary distributions for each increase in arrival rate. (b) Comparison of the upward bed occupancy transition probabilities for each increase in arrival rate. (c) Comparison of the probability density functions of the upward sojourn times for selected bed occupancies and for each increase in arrival rate. (d) Comparison of the probability density functions of the downward sojourn times for selected bed occupancies and for each increase in arrival rate. Note that the black triangles indicate the stationary distribution without an increase in arrival rate.

10.5.3 Decrease in patient discharges

Now, let's consider the effect of decreasing the rate of patient discharges from the ICU by scaling the $A_-^{(\ell)}$ block matrices of the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms. In this case, we reduce the rate of discharges from the RAH ICU by $(1-\gamma)100\%$ through scaling the $A_-^{(\ell)}$ block matrices by $\gamma = \{0.2, 0.4, 0.6, 0.8\}$, such that

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ \gamma A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & \gamma A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & \gamma A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & \gamma A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

As you slow the discharge process of the RAH ICU, you expectedly see a significant increase in the proportion of time spent above 85% capacity and an increase in the upward transition probabilities, as shown in Figures 10.5.3a and 10.5.3b. Generally, Figures 10.5.3c and 10.5.3d show that the ICU spends less time without a patient admission or discharge as the discharge process is slowed.



(a)

(b)

(c)

(d)

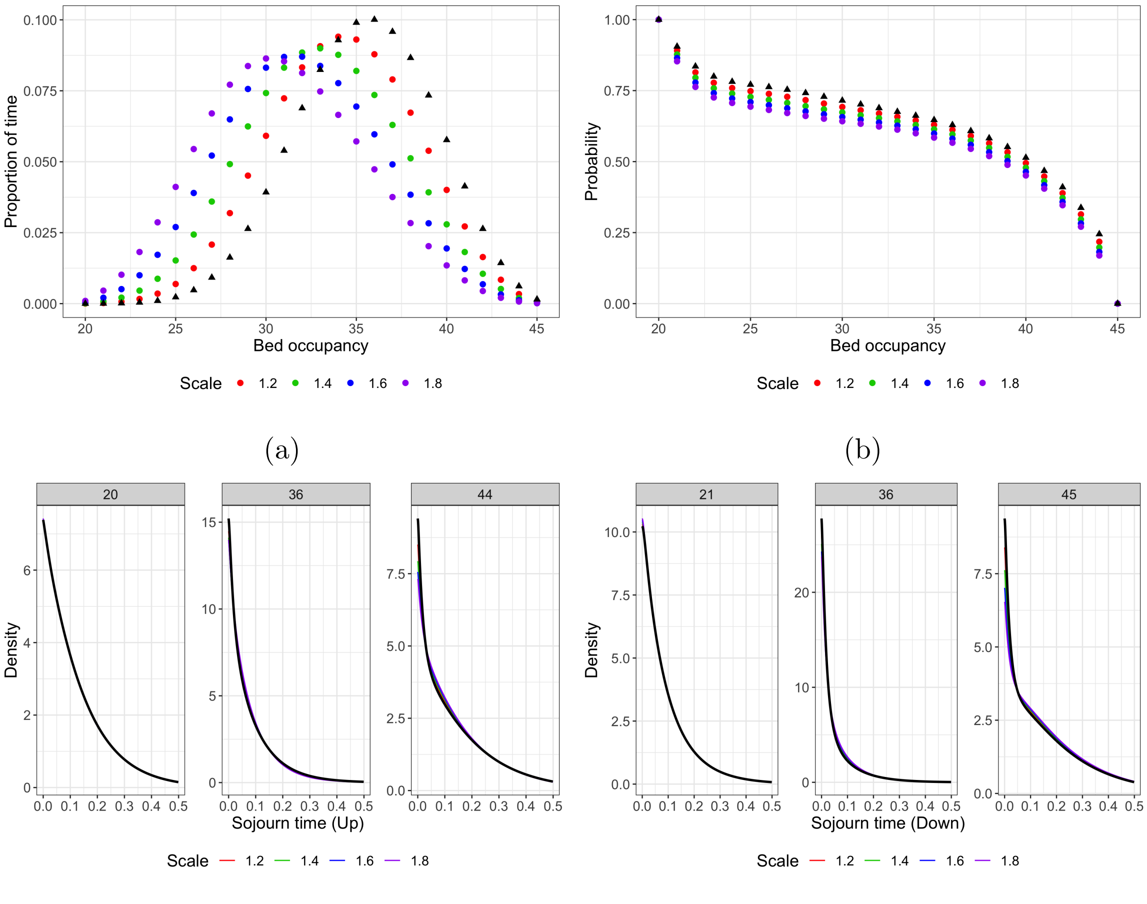
Figure 10.5.3: Illustrations of how the stationary and transient behaviour of the RAH ICU changes as the rate of patient discharges decreases. (a) Comparison of the bed occupancy stationary distributions for each reduction in departure rate. (b) Comparison of the upward bed occupancy transition probabilities for each reduction in departure rate. (c) Comparison of the probability density functions of the upward sojourn times for selected bed occupancies and for each reduction in departure rate. (d) Comparison of the probability density functions of the downward sojourn times for selected bed occupancies and for each reduction in departure rate. Note that the black triangles indicate the stationary distribution without a reduction in departure rate.

10.5.4 Increase in patient discharges

Lastly, we explore the effect of speeding up the patient discharge process from the RAH ICU by scaling the $A_-^{(\ell)}$ block matrices of the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms. Here, we increase the rate of discharges from the RAH ICU by $(\gamma - 1)100\%$ through scaling the $A_-^{(\ell)}$ block matrices by $\gamma = \{1.2, 1.4, 1.6, 1.8\}$, such that

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ \gamma A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & \gamma A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & \gamma A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & \gamma A_-^{(25)} & A_0^{(25)} \end{bmatrix}.$$

As expected, the proportion of time spent above 85% capacity is considerably reduced as the discharge rate of patients from the RAH ICU increases, as shown in Figures 10.5.4a. As shown in Figure 10.5.4b, the upward transition probabilities only slightly decrease as the rate of patient discharges increase. However, little change is observed in the time between patient admissions and discharges as the discharge process speeds up, as shown in Figures 10.5.4c and 10.5.4d. Despite an investigation into these results, the cause of this unexpected observation was not found and therefore requires further investigation.



(a)

(b)

(c)

(d)

Figure 10.5.4: Illustrations of how the stationary and transient behaviour of the RAH ICU changes as the rate of patient discharges increases. (a) Comparison of the bed occupancy stationary distributions for each increase in departure rate. (b) Comparison of the upward bed occupancy transition probabilities for each increase in departure rate. (c) Comparison of the probability density functions of the upward sojourn times for selected bed occupancies and for each increase in departure rate. (d) Comparison of the probability density functions of the downward sojourn times for selected bed occupancies and for each increase in departure rate. Note that the black triangles indicate the stationary distribution without an increase in departure rate.

10.5.5 Reservation of beds within an ICU

Due to the COVID-19 pandemic, the number of non-COVID-19 related admissions reduced as a result of elective procedure cancellations and an overall reduction in human movement [23, 44]. However, intensive care services were still needed for some patients and so only a proportion of intensive care resources were reserved for COVID-19 patients. Given the large degree of uncertainty during a pandemic, the proportion of intensive care resources ICU staff should reserve for COVID-19 patients became an important area of research.

Hence, we now consider the problem of predicting the long-term behaviour of a split ICU, such that a certain number of intensive care beds are reserved for COVID-19 patients (infectious disease ICU) and the remaining intensive care beds form a smaller ICU for non-COVID-19 patients (non-infectious disease ICU). In this section, we divide the 45 intensive care beds between the infectious disease ICU and non-infectious disease ICU in several ways as described in Table 10.5.1.

	Infectious disease ICU	Non-infectious disease ICU
Split 1	9	36
Split 2	18	27
Split 3	27	18
Split 4	36	9

Table 10.5.1: Details of how many beds are in the infectious disease ICU and non-infectious disease ICU for each split considered.

Non-infectious disease ICU

While we are reducing the capacity of the RAH ICU for non-infectious disease patients by means of the fitted structured QBD process, we must still respect the admission and discharge processes observed in the RAH ICU data. That is, we must reduce the capacity of the non-infectious disease ICU in such a way that we minimise the risk of indirectly speeding up or slowing down the admission and discharge processes.

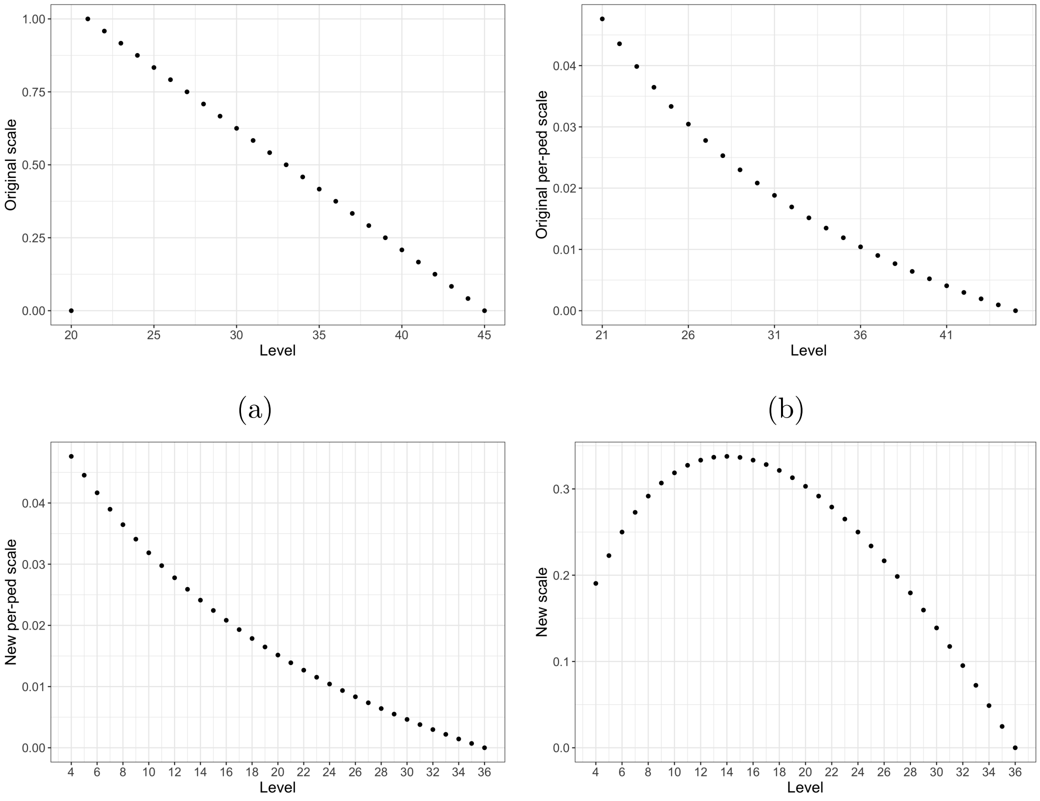
Firstly, we look at how we adjust the block matrices associated with a decrease in level of the fitted structured QBD process to model a non-infectious disease ICU with a reduced capacity of N_{max}, such that the infinitesimal generator matrix is of the form

$$\tilde{Q} = \begin{bmatrix} \tilde{A}_0^{(0)} & \tilde{A}_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ \tilde{A}_-^{(1)} & \tilde{A}_0^{(1)} & \tilde{A}_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & \tilde{A}_-^{(2)} & \tilde{A}_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \tilde{A}_0^{(C_N-2)} & \tilde{A}_+^{(C_N-2)} & 0 \\ 0 & 0 & 0 & \dots & \tilde{A}_-^{(C_N-1)} & \tilde{A}_0^{(C_N-1)} & \tilde{A}_+^{(C_N-1)} \\ 0 & 0 & 0 & \dots & 0 & \tilde{A}_-^{(C_N)} & \tilde{A}_0^{(C_N)} \end{bmatrix}.$$

Note that level C_N corresponds to a new maximum capacity of N_{max} and level 0 corresponds to a predicted observed minimum, N_{min}, which has a similar long-term proportion to that of level 20 in the structured QBD process fitted to the RAH ICU data.

Given the nature of the ICU, we adjust the block matrices associated with a decrease in level such that we maintain the per-bed behaviour of the RAH ICU. For example, consider the level and phase transition dependent scales $h_{\ell,i,j}$ of the fitted structured QBD process.

- If $h_{\ell,i,j} = 1$ for all $\ell \in [1, 25]$, then
$$\left[\tilde{A}_{-}^{(\ell^*)}\right]_{i,j} = \left[A_{-}\right]_{i,j}, \text{ for } 1 \leq \ell^* \leq C_N.$$
- If $h_{\ell,i,j} = 1 + \beta_1^{h,i,j}(\ell - 1)$ for all $\ell \in [1, 25]$, we first scale $h_{\ell,i,j}$ to reflect the per-bed behaviour, such that
$$\psi_{\ell,i,j} = \frac{h_{\ell,i,j}}{\ell + 20}, \text{ for } \ell \in [1, 25],$$
as illustrated in Figure 10.5.5b. Next, we translate the scales based on per-bed behaviour from $[1, 25]$ to $[1, C_N]$ by using the transformation
$$\ell = \frac{(25 - 1)(\ell^* - 1)}{(C_N - 1)} + 1,$$
to obtain the scales based on per-bed behaviour for the reduced capacity ICU, $\widetilde{\psi}_{\ell^*,i,j}$, as shown in Figure 10.5.5c. We then multiply $\widetilde{\psi}_{\ell^*,i,j}$ by the bed occupancies of the reduced capacity ICU, such that
$$\tilde{h}_{\ell^*,i,j} = \widetilde{\psi}_{\ell^*,i,j} \times (\ell^* + N_{min}), \text{ for } \ell^* \in [1, C_N],$$
as illustrated in Figure 10.5.5d. From this, we obtain the adjusted block matrices associated with a decrease in level, such that
$$\left[\tilde{A}_{-}^{(\ell^*)}\right]_{i,j} = \tilde{h}_{\ell^*,i,j}\left[A_{-}\right]_{i,j}, \text{ for } 1 \leq \ell^* \leq C_N.$$
- If $h_{\ell,i,j} = 1 + \beta_1^{h,i,j}(\ell - 1) + \beta_2^{h,i,j}(\ell - 1)^2$ for all $\ell \in [1, 25]$, we transform the domain of the scales in a similar way to before, such that
$$\left[\tilde{A}_{-}^{(\ell^*)}\right]_{i,j} = \tilde{h}_{\ell^*,i,j}\left[A_{-}\right]_{i,j}, \text{ for } 1 \leq \ell^* \leq C_N.$$



(a) (b)

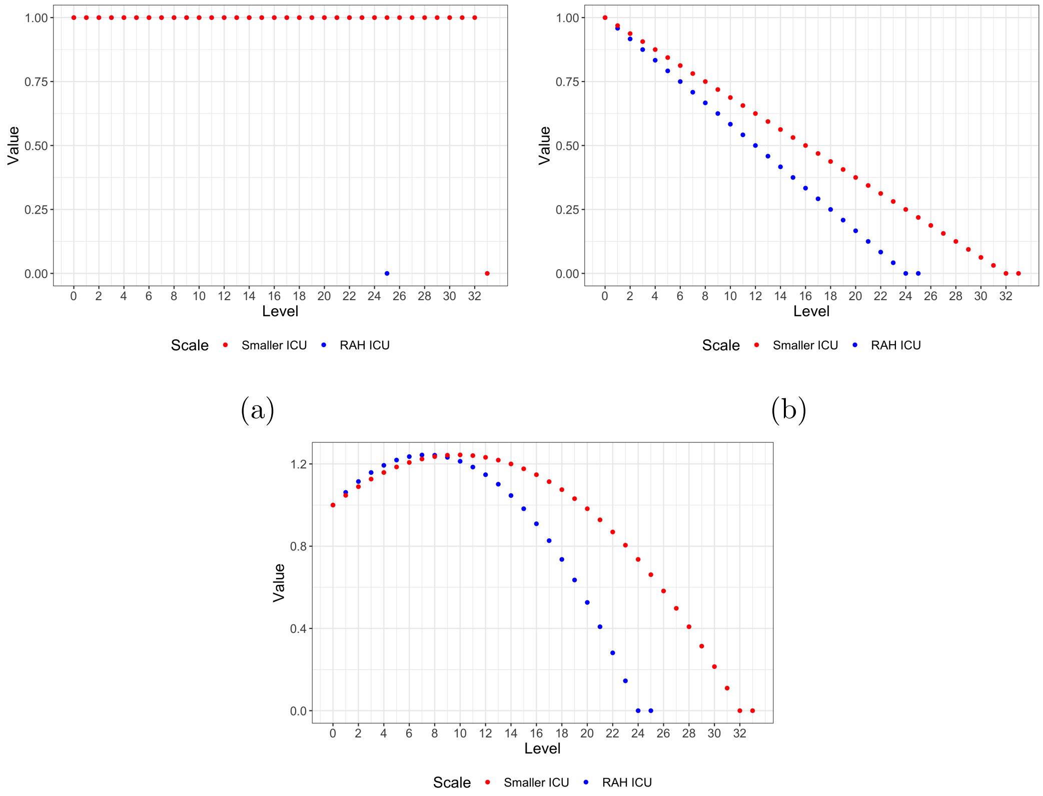
(c) (d)

Figure 10.5.5: Illustration of the process of transforming the scales associated with the structured QBD process fitted to the RAH ICU data set to account for fewer beds in the adjusted structured QBD process with a capacity of 36 patients. (a) Plot of the scales from the structured QBD process fitted to the RAH ICU data set. (b) Plot of the per-bed scales from the structured QBD process fitted to the RAH ICU data set. (c) Plot of the new per-bed scales from the adjusted structured QBD process to account for fewer beds in the ICU. (d) Plot of the new scales from the adjusted structured QBD process to account for fewer beds in the ICU.

Secondly, we adjust the block matrices associated with within level transitions to reflect the quiet and busy behaviour of the ICU. For example, consider the level and phase transition dependent scales $f_{\ell,i,j}$ of the fitted structured QBD process.

- If $f_{\ell,i,j} = 1$ for all $\ell \in [0, 24]$, then
$$\left[\tilde{A}_+^{(\ell^*)}\right]_{i,j} = [A_+]_{i,j}\,, \text{ for } 0 \le \ell^* \le C_N - 1.$$
- If $f_{\ell,i,j} = 1 + \beta_1^{f,i,j}\ell$ for all $\ell \in [0, 24]$, we adjust the slope of the linear function to capture both the quiet and busy behaviour of the ICU. That is,
$$\left[\tilde{A}_+^{(\ell^*)}\right]_{i,j} = \tilde{f}_{\ell^*,i,j}\,[A_+]_{i,j}\,, \text{ for } 0 \le \ell^* \le C_N - 1,$$
where
$$\tilde{f}_{\ell^*,i,j} = 1 + \frac{f_{44,i,j} - f_{20,i,j}}{(N_{max} - 1) - N_{min}}\ell^*, \text{ for } 0 \le \ell^* \le C_N - 1.$$
- If $f_{\ell,i,j} = 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2$ for all $\ell \in [0, 24]$, we transform the domain of the quadratic function to capture both the quiet and busy behaviour of the ICU, such that
$$\left[\tilde{A}_+^{(\ell^*)}\right]_{i,j} = \tilde{f}_{\ell^*,i,j}\,[A_+]_{i,j}\,, \text{ for } 0 \le \ell^* \le C_N - 1,$$
where $\tilde{f}_{\ell^*,i,j}$ is the resulting quadratic after transforming the domain from $[0, 24]$ to $[0, C_N - 1]$ using the transformation
$$\ell = \frac{24\ell^*}{C_N - 1}.$$

Illustrations of each type of transformation are provided in Figure 10.5.6, which compare the scales associated with the structured QBD process fitted to the RAH ICU data set (blue points) to the scales associated with the adjusted structured QBD process to account for fewer beds (red points).



(a)

(b)

(c)

Figure 10.5.6: Illustrations of how the scales of the adjusted structured QBD process (red points) compare to the scales associated with the structured QBD process fitted to the RAH ICU data set (blue points). (a) Comparison of the constant scales associated with an increase in level and the phase transition (1, 3). (b) Comparison of the linear scales associated with an increase in level and the phase transition $(1, 2)$. (c) Comparison of the quadratic scales associated with an increase in level and the phase transition (4, 4).

Lastly, we consider two methods of adjusting the block matrices associated with within level transitions; one method based on per-bed behaviour and the other method reflecting the quiet and busy behaviour of the ICU. These methods are similar to those described above.

Let's now explore the long-term behaviour of a non-infectious disease ICU with various capacities of 9, 18, 27, and 36 beds, such that the scales associated with within level changes and a decrease in level are adjusted on a per-bed basis. Note that we also consider the effect of reducing the rate of patient admissions to reflect a decline in the number of elective surgeries and emergency admissions.

Figure 10.5.7 illustrates the long-term proportion of time spent at each bed occupancy alongside various reductions in patient admissions to the non-infectious disease ICU. In doing so, we not only predict how much time the non-infectious disease ICU spends above 85% capacity (refer to Figure 10.5.8), but we also gain insight into how a decline in the rate of patient admissions caused by a drop in the number of elective surgical procedures and emergency admissions will effect the ICU.

To achieve a proportion of time spent above 85% capacity similar to that observed in the RAH ICU data set, admissions would need to be reduced by around 60% for a maximum capacity of 18 and roughly 50% for a maximum capacity of 27 beds, as shown in Figure 10.5.8. For a maximum capacity of 9 beds, a reduction greater than 80% is required to achieve a similar proportion of time spent above 85% capacity to that observed in the RAH ICU data set. However, there are no reductions necessary for a maximum capacity of 36 beds as the proportion of time spent above 85% capacity is always less than that of the RAH ICU. This suggests that the non-infectious disease ICU operates better than the RAH ICU despite no change in admissions and a reduced capacity, therefore demonstrating the invalidity of assuming that the scales associated with within level changes are dependent on per-bed behaviour

It is also worth noting that in terms of a functional capacity of 85%, an ICU with a maximum capacity of 9 beds and an 80% reduction in the rate of patient admissions is roughly similar to that of an ICU with a maximum capacity of 27 beds and a 40% reduction in the rate of patient admissions. Additionally, similar comments can be made regarding the comparison of an ICU with a maximum capacity of 18 beds and a 60% reduction in the rate of patient admissions and an ICU with a maximum capacity of 36 beds and no reduction in the rate of patient admissions.

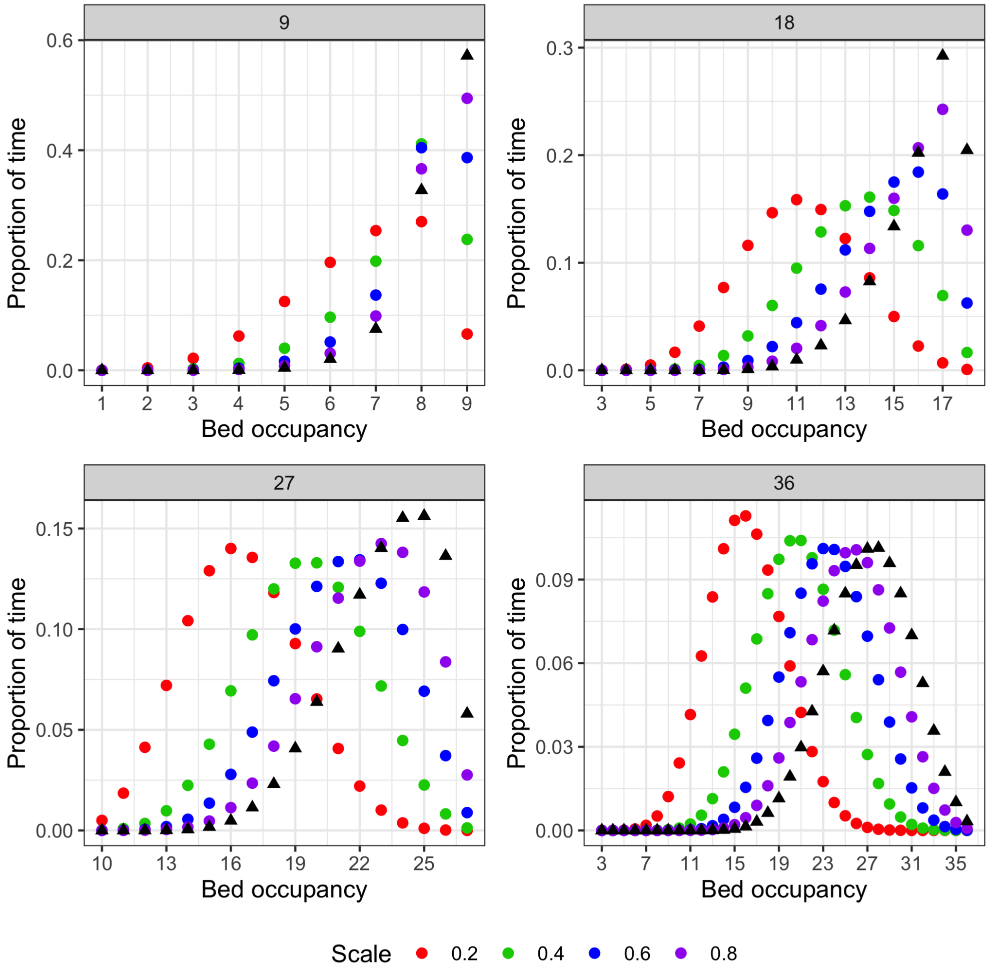


Figure 10.5.7: Comparison of how the long-term proportion of time the non-infectious disease ICU spends at each bed occupancy changes as the rate of patient admissions decreases for a maximum capacity of 9 (a), 18 (b), 27 (c), and 36 (d). Note that the black triangles represent the predicted long-term proportion of time the non-infectious disease ICU spends at each bed occupancy assuming no change in the rate of patient admissions. Also note that these predictions are based on models which assume the scales associated with block matrices describing a decrease in level and within level changes are based on per-bed behaviour.

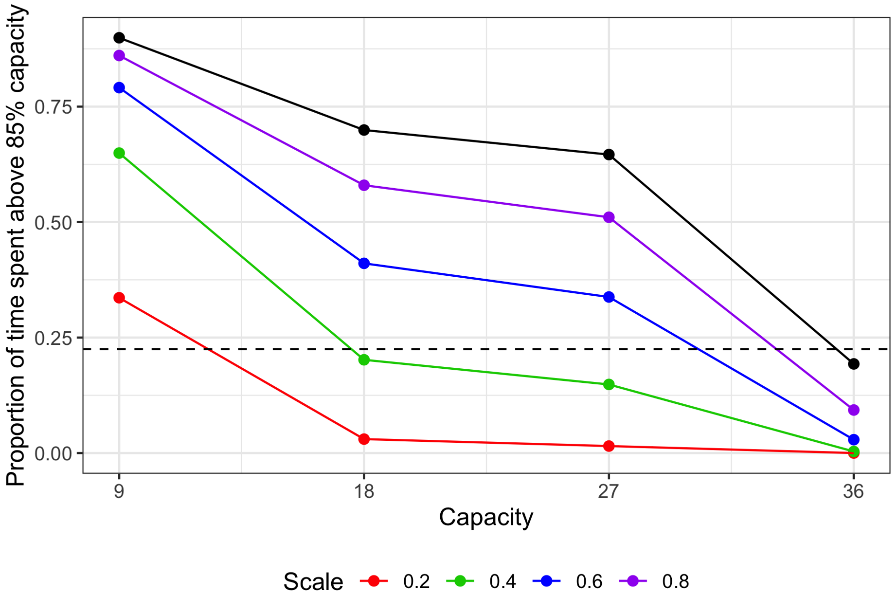


Figure 10.5.8: Comparison of how the proportion of time spent above 85% capacity changes as the rate of patient admissions decreases for a maximum capacity of 9, 18, 27 and 36, where the black line indicates the proportion of time spent above 85% capacity without a reduction in the rate of patient admissions and the dashed line indicates the proportion of time the RAH ICU operates above 85% capacity. Note that these predictions are based on models which assume the scales associated with block matrices describing a decrease in level and within level changes are based on per-bed behaviour.

Alternatively, let's now explore the long-term behaviour of a non-infectious disease ICU with various capacities of 9, 18, 27, and 36 beds, such that only the scales associated with a decrease in level are adjusted on a per-bed basis. That is, the scales associated with within level changes are adjusted to reflect both the empty and full behaviours of the RAH ICU.

Similar to above, Figure 10.5.9 plots the long-term proportion of time spent at each bed occupancy for each considered reduction in patient admissions to the non-infectious disease ICU and Figure 10.5.10 plots the associated proportion of time spent above 85% capacity.

In this scenario, admissions would need to be reduced by around 70% for a maximum capacity of 9 beds, around 40% for a maximum capacity of 18 beds and around 30% for a maximum capacity of 27 beds to achieve a proportion of time spent above 85% capacity similar to that of the RAH ICU. However, there are no reductions necessary for a maximum capacity of 36 beds as the proportion of time spent above 85% capacity is similar to that of the RAH ICU.

Additionally, an ICU with a maximum capacity of 18 beds and a 60% reduction in the rate of patient admissions is roughly similar to that of an ICU with a maximum capacity of 36 beds and an 20% reduction in the rate of patient admissions in terms of a functional capacity of 85%. Similar comments can be made about an ICU with a maximum capacity of 27 beds and a 40% reduction in the rate of patient admissions and an ICU with a maximum capacity of 36 beds and no reduction in the rate of patient admissions.

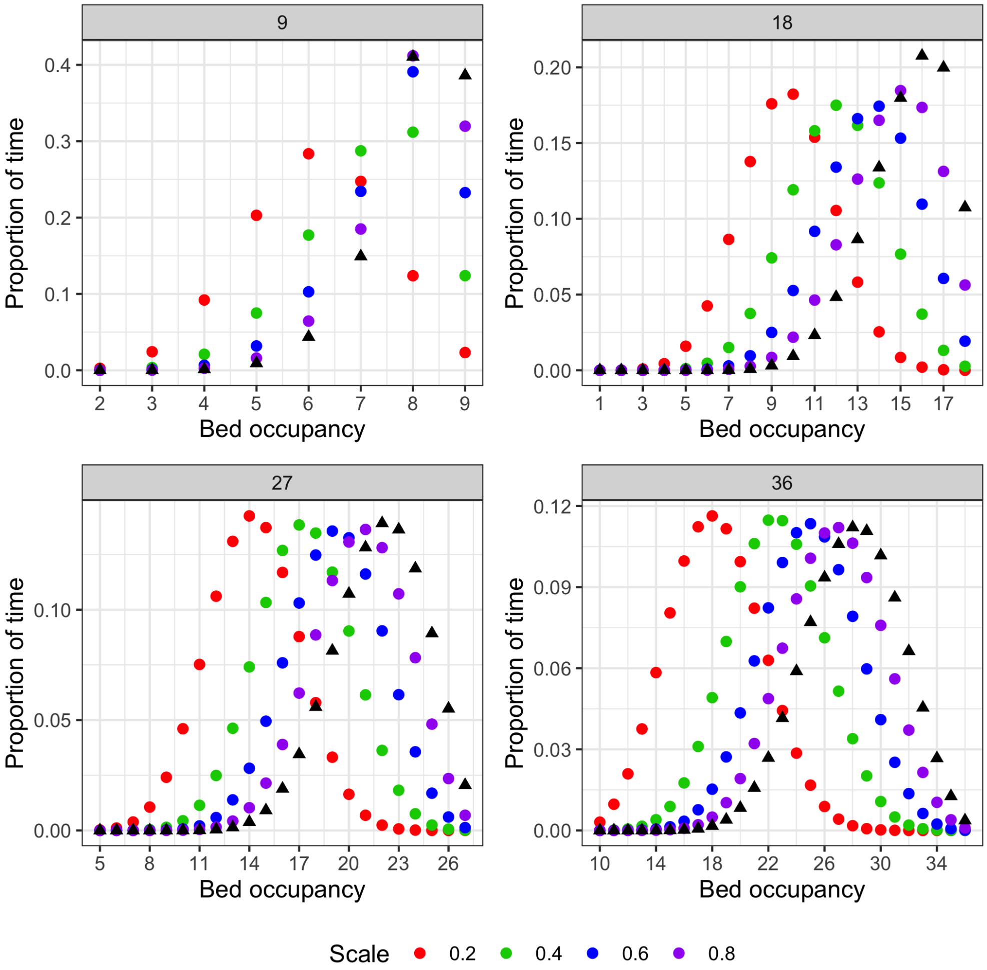


Figure 10.5.9: Comparison of how the long-term proportion of time the non-infectious disease ICU spends at each bed occupancy changes as the rate of patient admissions decreases for a maximum capacity of 9 (a), 18 (b), 27 (c), and 36 (d). Note that the black triangles represent the predicted long-term proportion of time the non-infectious disease ICU spends at each bed occupancy assuming no change in the rate of patient admissions. Also note that these predictions are based on models which assume the scales associated with block matrices describing a decrease in level are based on per-bed behaviour.

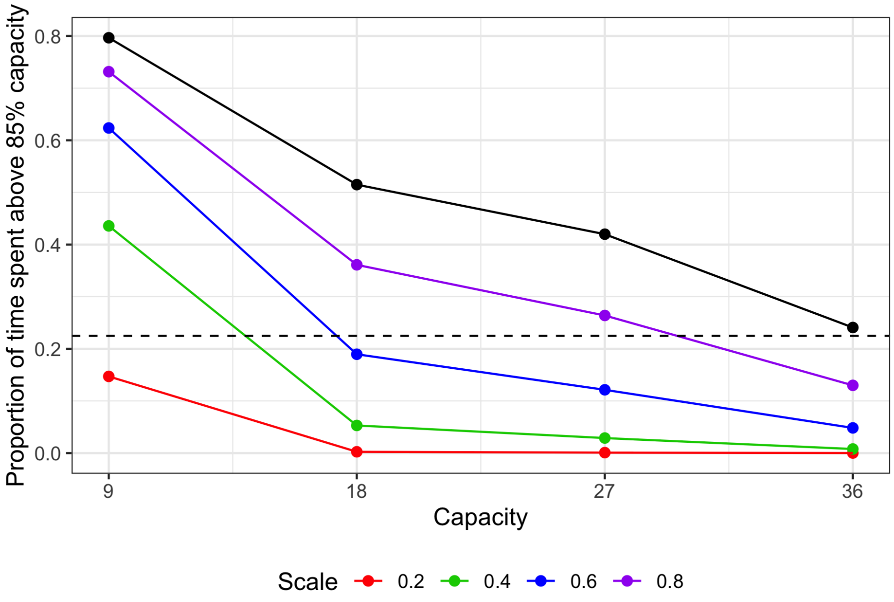


Figure 10.5.10: Comparison of how the proportion of time spent above 85% capacity changes as the rate of patient admissions decreases for a maximum capacity of 9, 18, 27 and 36, where the black line indicates the proportion of time spent above 85% capacity without a reduction in the rate of patient admissions and the dashed line indicates the proportion of time the RAH ICU operates above 85% capacity. Note that these predictions are based on models that assume the scales associated with block matrices describing a decrease in level are based on per-bed behaviour.

Overall, the approach which maintains the per-bed behaviour only for the downward level transitions provides more sensible models compared to the models which maintain the per-bed behaviour for the within level and downward level transitions. Therefore, the admission process of a reduced capacity ICU should reflect both the quiet and busy nature of the original ICU and the patient discharges of a reduced capacity ICU should reflect the per-bed nature of the original ICU.

Infectious disease ICU

To protect non-infectious disease patients within the ICU, intensive care beds are reserved to treat those with COVID-19. The demand for COVID-19 related intensive care services depends on a number of factors; four of which include

- the number of positive COVID-19 cases per day,
- the percentage of COVID-19 patients directed to each hospital or medical facility,
- the percentage of COVID-19 cases that require intensive care, and
- the duration of time COVID-19 patients require intensive care.

For completeness, we now explore how the bed occupancy of the infectious disease ICU could behave under various scenarios. We note that this is not a detailed model of how the COVID-19 pandemic has impacted or could impact the RAH ICU, as the pandemic is still unfolding and thus remains an open area of research.

For simplicity, we assume an $M/M/c/c$ queueing model for the infectious disease ICU such that

$$\lambda_k = \begin{cases} N_{COVID-19} P_{ICU} P_{RAH}, & k < c, \\ 0, & k \geq c, \end{cases}$$

and

$$\mu_k = \frac{k}{LOS_{COVID-19}}, \quad k = 1, 2, \ldots, c,$$

where

- c is the number of beds in the infectious disease ICU,
- $N_{COVID-19}$ is the number of positive COVID-19 cases reported per day,
- P_{ICU} is the percentage of COVID-19 cases that will need intensive care,
- P_{RAH} is the percentage of COVID-19 patients requiring intensive care being directed to the RAH infectious disease ICU, and
- $LOS_{COVID-19}$ is the average number of days COVID-19 patients require intensive care.

This queueing model is also known as the Erlang-loss queueing system as there are no waiting spaces available for patients to queue in [32]. Hence, any patients who find all beds occupied upon admission to the infectious disease ICU are blocked. While this does not reflect the true behaviour of health care services, the $M/M/c/c$ queueing model will provide insight as to the proportion of blocked patients who required COVID-19 related intensive care, P_c, which is defined as

$$P_c = \frac{\frac{\gamma^c}{c!}}{\sum\limits_{k=0}^{c} \frac{\gamma^k}{k!}},$$

where $\gamma = \lambda/\mu$ is the traffic intensity.

In this modelling, we explore how the proportion of time spent above 85% capacity changes as we vary the values of the above parameters, such that

1. $c = 9$, 18, 27, and 36 beds,
2. $N_{COVID-19} = 50$, 100, 500, 1000, 5000, and 10000 cases per day,
3. P_{ICU} is held fixed at 5% [35, 48],
4. $P_{RAH} = 25$, 50%, 75%, and 100%, and
5. $LOS_{COVID-19} = 2$, 7, 12, 14 days [35, 48].

As expected, the infectious disease ICU becomes overloaded as the number of positive COVID-19 cases per day increases, as well as when the percentage of COVID-19 patients directed to the RAH infectious disease ICU increases and the average number of days COVID-19 patients require intensive care increases, as shown in Figure 10.5.11.

This result is also reflected in the blocking probabilities shown in Figure 10.5.12 which illustrates that there is an increasing proportion of COVID-19 patients being blocked from the infectious disease ICU as the number of positive COVID-19 cases per day increases, the percentage of COVID-19 patients directed to the RAH infectious disease ICU increases, and the average number of days COVID-19 patients require intensive care increases.

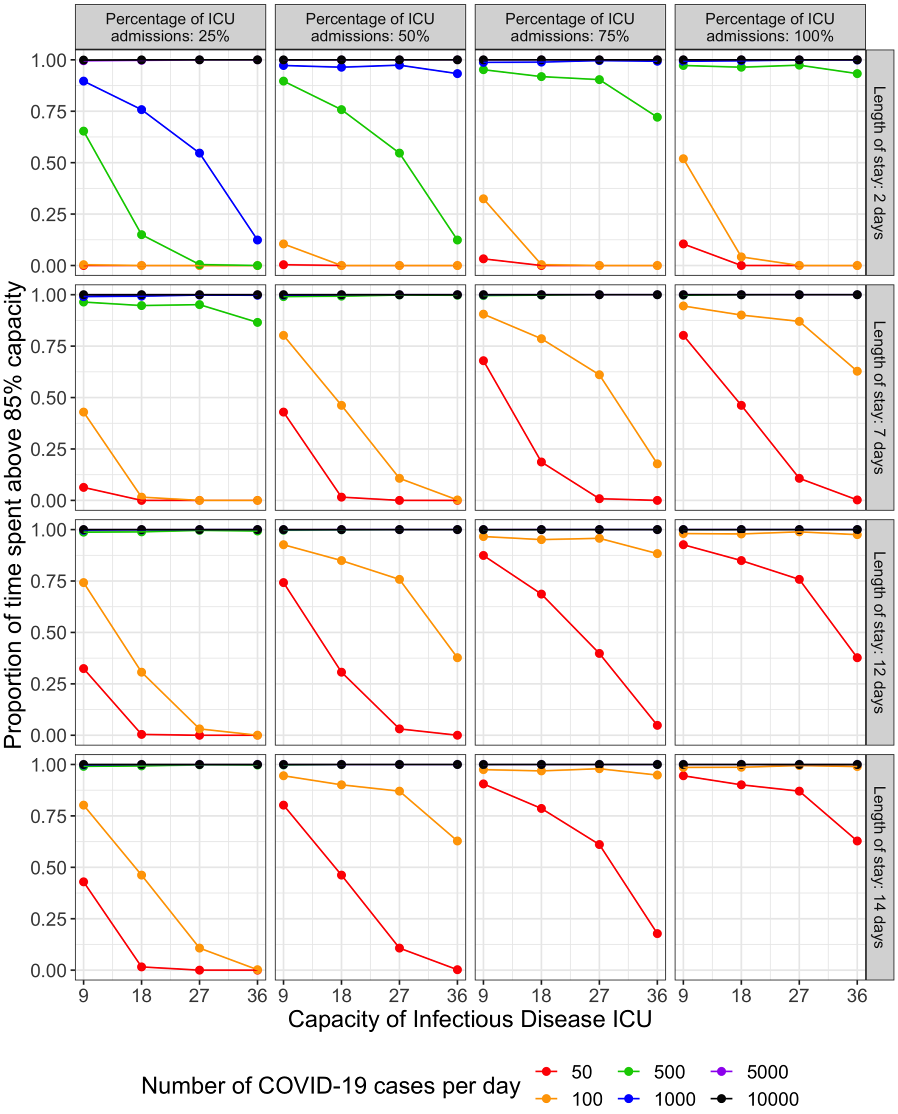


Figure 10.5.11: Comparison of how the proportion of time spent above 85% capacity (y-axis) changes as the number of positive COVID-19 cases per day (coloured lines) increases, the average number of days COVID-19 patients require intensive care (rows) increases, and the accepted percentage of COVID-19 ICU admissions the RAH ICU (columns) increases for a maximum capacity (x-axis) of 9, 18, 27 and 36.

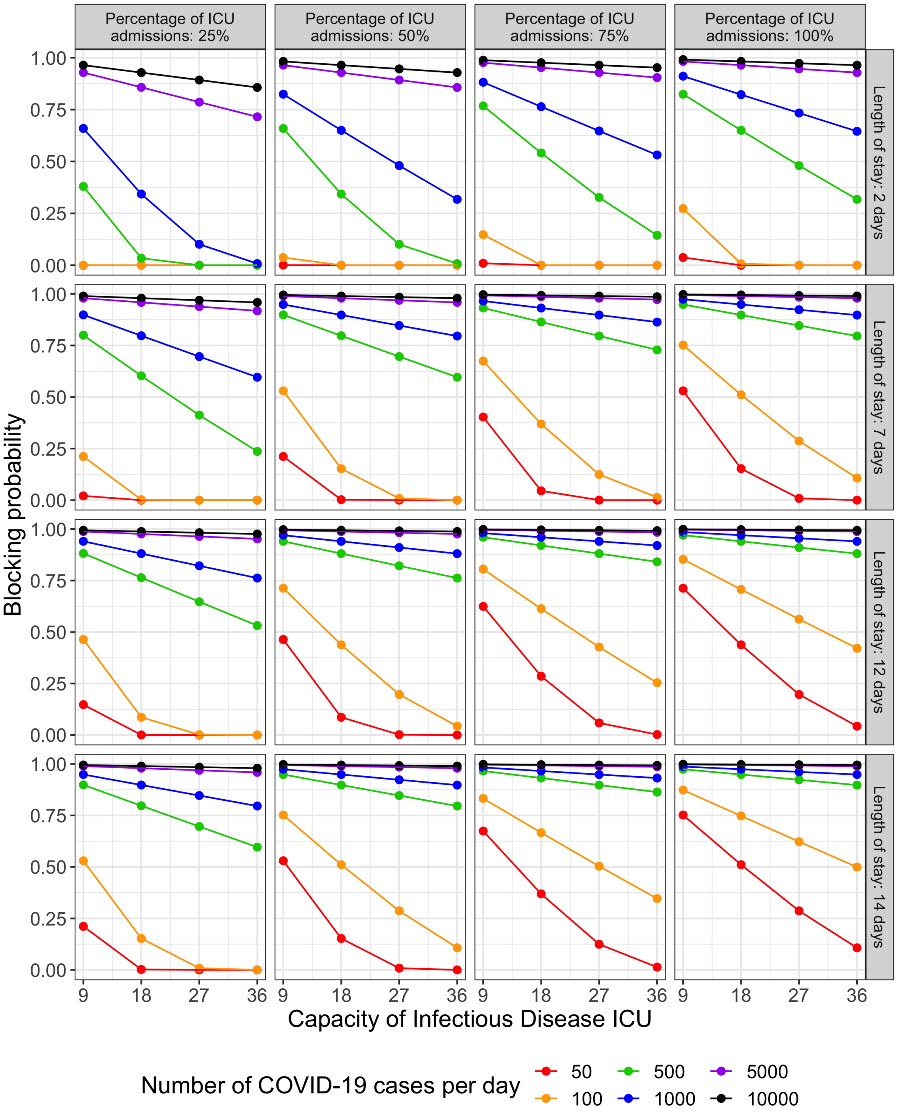


Figure 10.5.12: Comparison of how the blocking probability (y-axis) changes as the number of positive COVID-19 cases per day (coloured lines) increases, the average number of days COVID-19 patients require intensive care (rows) increases, and the accepted percentage of COVID-19 ICU admissions the RAH ICU (columns) increases for a maximum capacity (x-axis) of 9, 18, 27 and 36.

With public health measures such as social distancing, lockdowns, and mask wearing, we can reduce the transmission of COVID-19. Additionally, vaccinations and early-intervention treatments will reduce hospitalisation rates and hence reduce the rate of COVID-19 related ICU admissions. However, the duration of intensive care required for each COVID-19 patient depends on the patient's condition and the availability of COVID-19 specific treatments. Lastly, the percentage of COVID-19 patients who require intensive care sent to each hospital depends on the operational status and the capacity to care and treat more patients.

10.6 Summary

By means of level-dependent QBD processes, we gained valuable insight into the operation of the RAH ICU. In particular, we discovered that the RAH ICU is more likely to have a patient admission following a patient admission rather than a patient admission following a patient discharge and is also more likely to have a patient discharge following a patient discharge rather than a patient discharge following a patient admission. In addition to this, we also discovered that as the number of occupied beds increased, the rate of patient admissions decreased but the rate of patient discharges remained relatively constant.

Through the use of structured QBD processes, we demonstrated the ability to predict the behaviour of the RAH ICU under various scenarios which may reflect changes in hospital policies, a short-term surge in emergency admissions, or a decline in the number of elective surgical procedures. Furthermore, the exploration of a split ICU to accomodate both non-infectious disease infectious disease ICU admissions revealed how often each ICU would operate above 85% capacity. In doing so, health care professionals are able to make informed decisions on the number of beds to reserve for infectious disease patients based on the state of the healthcare system and the demand for intensive care services.

Chapter 11

Conclusion

In this thesis, we have developed statistical methods to fit quasi-birth-and-death processes to queueing system data, irrespective of the possibility of dependence between the arrival process and the distribution of service times. In particular, we have developed various statistical estimation methods and a goodness of fit test to suit level-dependent, level-independent, and structured quasi-birth-and-death processes. The following discussion provides a summary of the research presented in this thesis, the main results from each estimation method and the goodness of fit method developed in this thesis, as well as some directions for further research in this area.

11.1 Summary

An intensive care unit is a crucial and limited resource in a hospital which provides high-level care to critically ill patients. The performance of intensive care units is continuously monitored and improvements to patient care and bed occupancy management are an important area of research, as discussed in Chapter 2. Of particular interest is the modelling of bed occupancy in intensive care units using queueing theory.

Stochastic processes and Markov chains form a basis for queueing theory and the modelling of queueing systems, such as telecommunication and healthcare systems. In Chapter 3, we reviewed the Poisson process which is often used to model arrivals to a queueing system. Markov processes are another type of stochastic process that are used to model random phenomena, such as changes in population size or changes in the stock market. In Chapter 4, we reviewed the fundamentals of queueing theory and then explored how Markov processes are used to derive the maximum likelihood estimates of the parameters of various elementary queueing models.

Following the review of stochastic processes and queueing theory, we modelled the bed occupancy of the Royal Adelaide Hospital intensive care unit using elementary Markovian queueing models. Through a discussion of the modelling assumptions in Chapter 5, the issue of dependence between patient admissions and length of stay became apparent. In particular, the queueing models typically used to model intensive care units assume independence between the arrival process and the distribution of service times. However, studies have found evidence of a dependence between the patient arrival process and the distribution of length of stay, thus rendering standard queueing models invalid for this research problem.

We began Chapter 6 with a review of phase-type distributions which form a basis for the development of quasi-birth-and-death processes. QBD processes are a generalisation of the birth-and-death process and present themselves as an alternative to modelling queueing systems with a dependence between the arrival process and the distribution of service times. QBD processes are either level-dependent or level-independent in nature, such that the transition rates will either depend on the level the process is currently in or remain independent of the level. In either case, all that is observed while watching a QBD process evolve are the changes in level and the times at which those changes occurred. That is, the phase process remains hidden. For example, we observe the changes in bed occupancy in an intensive care unit but we cannot observe the change in environment such as patient care or bed management.

A natural approach to modelling situations with unobserved data is the expectation-maximisation algorithm. Asmussen *et al.* [1] developed an EM algorithm to fit phase-type distributions to data by using the connection between continuous-time Markov chains and phase-type distributions to calculate the maximum likelihood estimates. Building upon the work completed by Asmussen, in Chapter 7 we developed two statistical model fitting methods based on the EM algorithm that fit level-dependent and level-independent QBD processes to queueing system data, respectively. We assessed the accuracy of each method by comparing the stationary and transient behaviour of a known QBD process to that of several QBD processes fitted to data simulated from the known QBD process. In particular, we considered stationary and transient behaviour, such as the long term proportion of time spent in each level, transitions between levels, and the sojourn time before the process moves up or down from each level.

Despite level-dependent QBD processes being able to provide insight into the operation of a queueing system, such models cannot be used for predicting future behaviour due to the possibility of over-parameterisation. Although level-independent QBD processes have fewer parameters than that of a level-dependent QBD process, not all queueing systems are level-independent in nature. With this motivation, in Chapter 8 we developed a new type of QBD process called a structured QBD process which is a blend between level-dependent and level-independent QBD processes. That is, the infinitesimal generator matrix of a structured QBD process is defined in such a way as to explain the behaviour of the queueing system through the use of functional forms of the level and phase transition. Examples of such behaviours include queueing systems with linear or quadratic trends in the arrival or departure rates, as well as queueing systems that exhibit constant arrival or departure trends regardless of the state of the queueing system.

Following the development of structured QBD processes, we then extended the EM algorithm developed in Chapter 7 to structured QBD processes. Similar to before, we assessed the accuracy of this method by using simulated data from several structured QBD processes that exhibit combinations of constant, linear, and quadratic behaviour in the transition rates. For each structured QBD process considered, we compared several estimations of each structured QBD process to the respective known structured QBD process by considering several components of stationary and transient behaviour.

As with any statistical modelling method, the question of how well the fitted model explains the behaviour observed in the data is generally answered by means of a goodness of fit test. Goodness of fit tests typically measure the distance between the observed data and what is expected under the fitted model. Hence, we began Chapter 9 by reviewing distance measures which formed a basis for the development of a goodness of fit test that quantifies how well a fitted QBD process explains the stationary and transient behaviour observed in queueing system data. We then developed a diagnostic method which illustrates where, if at all, a fitted QBD process inadequately models the stationary and transient behaviour. To demonstrate the performance of our goodness of fit test, we considered various numerical examples and completed a small simulation study to illustrate the statistical power and significance.

In Chapter 10, we applied the methods developed in this thesis to a data set obtained from the Royal Adelaide Hospital intensive care unit. While a level-dependent QBD process provides meaningful insight into the operation of the Royal Adelaide Hospital intensive care unit, such a model cannot be used to predict the behaviour of the intensive care unit as it is over-parameterised. Therefore, an investigation into suitable types of structured QBD processes was completed. However, this investigation brought to light the issue of phase interchangeability thus suggesting that pre-specification of the functional forms of a structured QBD process is not always a suitable approach.

Despite this finding, it was found that the stationary and transient behaviour of the bed occupancy in the RAH ICU was better explained by the best-fitting structured QBD process than the standard Markovian queueing models used in practice. We then used the best-fitting structured QBD process to predict the state of the intensive care unit as the rate of admissions and departures changed. We also considered a simulation study which explored the effect of reserving beds for infectious disease patients, thus limiting the number of beds available to provide intensive care for non-infectious disease patients.

The queueing models used in practice to model the bed occupancy of intensive care units do not allow for dependence between patient admissions and the distribution of patient length of stay. These queueing models also have limited freedom in the distribution of time spent at each bed occupancy. Therefore, analyses and future predictions of intensive care units using such models may provide inaccurate results. With this motivation, we have developed statistical modelling methods that fit various types of QBD processes to queueing system data that allow for dependence between the arrival process and the distribution of service times. We have also developed a goodness of fit test to assess the fit of any QBD process to queueing system data to ensure the reliability of future predictions.

11.2 Future work

In Section 8.9, we described a method of choosing the functional forms of a structured QBD process which is based on data visualisation and model comparisons. However, this method ignores the possibility of phases interchanging within the structured QBD process. Therefore, an improvement to the methods developed in Chapter 8 would be to incorporate the selection of the best fitting functional form into the EM algorithm.

The Mahalanobis distance defined in our goodness of fit test does not follow a multivariate normal distribution, and hence the squared Mahalanobis distance is not chi-square distributed. As a result, we empirically estimated the statistical power and significance of the goodness of fit test developed in Chapter 9. However, only a small number of simulations were performed due to the computational demand of estimating the power and significance of the goodness of fit test. Further research into the distribution of non-multivariate normal Mahalanobis distances is therefore required.

In Chapter 10, we assumed that there is only one class of patient admitted to the Royal Adelaide Hospital intensive care unit. However, several patient classes exist within the intensive care unit; an example of which is elective and emergency patients. In addition to this, the QBD processes presented in this thesis are time-homogeneous which therefore assumes that the ICU operates no differently at night compared to during the day. However, the assumption of time-homogeneity is incorrect as there is less movement during the night shifts compared to day and afternoon shifts, as explored in Chapter 5. Therefore, extending this work to two-dimensional QBD processes or time-inhomogeneous QBD processes could provide more informative models of an intensive care unit, and hence is another direction for future research.

Appendix A

Proof of Theorem 8.3.1

By taking the appropriate partial derivatives of the log-likelihood, we can obtain the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (\{A_0^{(\ell)}; 0 \leq \ell < \infty\}, \{A_-^{(\ell)}; 1 \leq \ell < \infty\}, \{A_+^{(\ell)}; 0 \leq \ell < \infty\})$ for a structured QBD process. Note that in the following proofs, j_* and J_* denote the appropriate lower and upper limits of the summation where required.

Lemma A.0.1.

$$\widehat{A}^{(\ell)}_{\bullet_{i,j}} = \frac{M_{\ell,i,j}}{Z_{\ell,i}}, \textit{ for } \ell < S_L \textit{ and } \ell > S_U,$$

where

$$M_{\ell,i,j} = \begin{cases} S_{\ell,i,j}, & \textit{if } \bullet = 0, \\ D_{\ell,i,j}, & \textit{if } \bullet = -, \\ U_{\ell,i,j}, & \textit{if } \bullet = +, \end{cases}$$

and,

$$(S_L, S_U) = \begin{cases} (L_0, U_0) & \textit{if } \bullet = 0, \\ (L_-, U_-), & \textit{if } \bullet = -, \\ (L_+, U_+) & \textit{if } \bullet = +. \end{cases}$$

Proof:

$$
\begin{aligned}
\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A^{(\ell)}_{\bullet_{i,j}}} &= \frac{\partial}{\partial A^{(\ell)}_{\bullet_{i,j}}} \left(\sum_{i=1}^{J_\ell} A^{(\ell)}_{0_{i,i}} Z_{\ell,i} + \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} \log\left(A^{(\ell)}_{\bullet_{i,j}}\right) M_{\ell,i,j} \right), \\
&\qquad \textit{by selecting the relevant terms from the constrained log-likelihood,} \\
&= \frac{\partial}{\partial A^{(\ell)}_{\bullet_{i,j}}} \left(-\sum_{i=1}^{J_\ell} \left(\sum_{\substack{j=1\\ j\neq i}}^{J_\ell} A^{(\ell)}_{0_{i,j}} + \sum_{j=1}^{J_{\ell-1}} A^{(\ell)}_{-_{i,j}} + \sum_{j=1}^{J_{\ell+1}} A^{(\ell)}_{+_{i,j}} \right) Z_{\ell,i} + \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} \log\left(A^{(\ell)}_{\bullet_{i,j}}\right) M_{\ell,i,j} \right), \\
&\qquad \textit{since the diagonal entries of } A^{(\ell)}_0 \textit{ are the negative of the relevant row sums,} \\
&= -Z_{\ell,i} + \frac{M_{\ell,i,j}}{A^{(\ell)}_{\bullet_{i,j}}}.
\end{aligned}
$$

By rearranging, we get

$$
\widehat{A}^{(\ell)}_{\bullet_{i,j}} = \frac{M_{\ell,i,j}}{Z_{\ell,i}}.
$$

Lemma A.0.2.

$$
\widehat{A}_{\bullet_{i,j}} = \frac{\displaystyle\sum_{\ell=S_L}^{S_U} M_{\ell,i,j}}{\displaystyle\sum_{\ell=S_L}^{S_U} \phi(\ell) Z_{\ell,i}}, \textit{ for } S_L \leq \ell \leq S_U.
$$

where

$$
M_{\ell,i,j} = \begin{cases} S_{\ell,i,j}, & \textit{if } \bullet = 0, \\ D_{\ell,i,j}, & \textit{if } \bullet = -, \\ U_{\ell,i,j}, & \textit{if } \bullet = +, \end{cases}
$$

and,

$$
(S_L, S_U, \phi(\ell)) = \begin{cases} (L_0, U_0, g_\ell), & \textit{if } \bullet = 0, \\ (L_-, U_-, h_\ell), & \textit{if } \bullet = -, \\ (L_+, U_+, f_\ell), & \textit{if } \bullet = +. \end{cases}
$$

Proof:

$$\begin{aligned}\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A_{\bullet_{i,j}}} &= \frac{\partial}{\partial A_{\bullet_{i,j}}}\left(\sum_{\ell=S_L}^{S_U}\sum_{i=1}^{J_\ell} A^{(\ell)}_{0_{i,i}} Z_{\ell,i} + \sum_{\ell=S_L}^{S_U}\sum_{i=1}^{J_\ell}\sum_{j=j_*}^{J_*}\log\left(\phi(\ell)A_{\bullet_{i,j}}\right)M_{\ell,i,j}\right),\\
&\qquad \textit{by selecting the relevant terms from the constrained log-likelihood,}\\
&= \frac{\partial}{\partial A_{\bullet_{i,j}}}\left(-\sum_{\ell=S_L}^{S_U}\sum_{i=1}^{J_\ell}\left(\sum_{\substack{j=1\\j\neq i}}^{J_\ell} g_\ell A_{0_{i,j}} + \sum_{j=1}^{J_{\ell-1}} h_\ell A_{-_{i,j}} + \sum_{j=1}^{J_{\ell+1}} f_\ell A_{+_{i,j}}\right) Z_{\ell,i} + \sum_{\ell=S_L}^{S_U}\sum_{i=1}^{J_\ell}\sum_{j=j_*}^{J_*}\log\left(\phi(\ell)A_{\bullet_{i,j}}\right)M_{\ell,i,j}\right),\\
&\qquad \textit{since the diagonal entries of } A_0 \textit{ are the negative of the relevant row sums,}\\
&= -\sum_{\ell=S_L}^{S_U}\phi(\ell)Z_{\ell,i} + \sum_{\ell=S_L}^{S_U}\frac{M_{\ell,i,j}}{A_{\bullet_{i,j}}}.\end{aligned}$$

By rearranging, we get $\bullet$

$$\widehat{A}_{\bullet_{i,j}} = \frac{\displaystyle\sum_{\ell=S_L}^{S_U} M_{\ell,i,j}}{\displaystyle\sum_{\ell=S_L}^{S_U}\phi(\ell)Z_{\ell,i}}.$$

Lemma A.0.3.

$$\widehat{\phi}_\ell = \frac{\displaystyle\sum_{\ell=S_L}^{S_U} M_{\ell,i,j}}{\displaystyle\sum_{\ell=S_L}^{S_U} A_{\bullet_{i,j}} Z_{\ell,i}}, \textit{ for } S_L \le \ell \le S_U.$$

where

$$M_{\ell,i,j} = \begin{cases} S_{\ell,i,j}, & \textit{if } \bullet = 0,\\ D_{\ell,i,j}, & \textit{if } \bullet = -,\\ U_{\ell,i,j}, & \textit{if } \bullet = +,\end{cases}$$

and,

$$(S_L, S_U, \bullet) = \begin{cases} (L_0, U_0, 0), & \textit{if } \phi(\ell) = g(\ell),\\ (L_-, U_-, -), & \textit{if } \phi(\ell) = h(\ell),\\ (L_+, U_+, +), & \textit{if } \phi(\ell) = f(\ell).\end{cases}$$

Proof:

$$\frac{\partial \ell_C(\boldsymbol{\theta}|\mathbf{d})}{\partial \phi(\ell)} = \frac{\partial}{\partial \phi(\ell)} \left(\sum_{\ell=S_L}^{S_U} \sum_{i=1}^{J_\ell} A^{(\ell)}_{0_{i,i}} Z_{\ell,i} + \sum_{\ell=S_L}^{S_U} \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} \log\left(\phi(\ell) A_{\bullet_{i,j}}\right) M_{\ell,i,j} \right),$$

by selecting the relevant terms from the constrained log-likelihood,

$$= \frac{\partial}{\partial \phi(\ell)} \left(- \sum_{\ell=S_L}^{S_U} \sum_{i=1}^{J_\ell} \left(\sum_{\substack{j=1 \\ j \neq i}}^{J_\ell} g_\ell A_{0_{i,j}} + \sum_{j=1}^{J_{\ell-1}} h_\ell A_{-_{i,j}} + \sum_{j=1}^{J_{\ell+1}} f_\ell A_{+_{i,j}} \right) Z_{\ell,i} + \sum_{\ell=S_L}^{S_U} \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} \log\left(\phi(\ell) A_{\bullet_{i,j}}\right) M_{\ell,i,j} \right),$$

since the diagonal entries of A_0 *are the negative of the relevant row sums,*

$$= - \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} A_{\bullet_{i,j}} Z_{\ell,i} + \sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} \frac{M_{\ell,i,j}}{\phi(\ell)}.$$

By rearranging, we get

$$\widehat{\phi}_\ell = \frac{\displaystyle\sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} M_{\ell,i,j}}{\displaystyle\sum_{i=1}^{J_\ell} \sum_{j=j_*}^{J_*} A_{\bullet_{i,j}} Z_{\ell,i}}.$$

Appendix B

Proof of Corollary 8.3.1.1

By taking the appropriate partial derivatives of the constrained log-likelihood, we can obtain the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (B_0, B_-, B_+, A_0, A_-, A_+, C_0, C_-, C_+)$ for a finite level-independent QBD process with boundaries at level 0 and level C.

Lemma B.0.1.

$$\widehat{B}_- = \widehat{A}_{-_{i,j}}^{(L_- -1)} = \frac{D_{L_- -1,i,j}}{Z_{L_- -1,i} - \lambda_i^{(L)}}, \; \textit{for}\ 1 \leq i \leq J_{L_- -1}, 1 \leq j \leq J_{L_- -2}.$$

Proof:

$$\begin{aligned}
\frac{\partial \Delta_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A^{(L_- -1)}_{-_{i,j}}} = \frac{\partial}{\partial A^{(L_- -1)}_{-_{i,j}}} & \left(\sum_{i=1}^{J_{L_- -1}} A^{(L_- -1)}_{0_{i,i}} Z_{L_- -1,i} + \sum_{i=1}^{J_{L_- -1}} \sum_{j=1}^{J_{L_- -2}} \log\left(A^{(L_- -1)}_{-_{i,j}}\right) D_{L_- -1,i,j} \right. \\
& \left. + \sum_{i=1}^{J_{L_- -1}} \lambda_i^{(L)} \sum_{j=1}^{J_{L_- -2}} A^{(L_- -1)}_{-_{i,j}} \right),
\end{aligned}$$

by selecting the relevant terms from the constrained log-likelihood,

$$\begin{aligned}
= \frac{\partial}{\partial A^{(L_- -1)}_{-_{i,j}}} & \left(- \sum_{i=1}^{J_{L_- -1}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{J_{L_- -1}} A^{(L_- -1)}_{0_{i,j}} + \sum_{j=1}^{J_{L_- -2}} A^{(L_- -1)}_{-_{i,j}} + \sum_{j=1}^{J_{L_-}} A^{(L_- -1)}_{+_{i,j}} \right) Z_{L_- -1,i} \right. \\
& + \sum_{i=1}^{J_{L_- -1}} \sum_{j=1}^{J_{L_- -2}} \log\left(A^{(L_- -1)}_{-_{i,j}}\right) D_{L_- -1,i,j} \\
& \left. + \sum_{i=1}^{J_{L_- -1}} \lambda_i^{(L)} \sum_{j=1}^{J_{L_- -2}} A^{(L_- -1)}_{-_{i,j}} \right),
\end{aligned}$$

since the diagonal entries of $A_0^{(L_- -1)}$ *are the negative of the relevant row sums,*

$$= -Z_{L_- -1,i} + \frac{D_{L_- -1,i,j}}{A^{(L_- -1)}_{-_{i,j}}} + \lambda_i^{(L)}.$$

By rearranging, we get

$$\widehat{B}_- = \widehat{A}^{(L_- -1)}_{-_{i,j}} = \frac{D_{L_- -1,i,j}}{Z_{L_- -1,i} - \lambda_i^{(L)}}, \text{ \textit{for} } 1 \le i \le J_{L_- -1}, 1 \le j \le J_{L_- -2}.$$

Lemma B.0.2.

$$\widehat{A}_{-} = \widehat{A}^{(L_-)}_{-i,j} = \frac{D_{L_-,i,j}}{Z_{L_-,i} - \displaystyle\sum_{l=L_-+1}^{U_-} \lambda^{(-)}_{l,i,j} + \lambda^{(L)}_{i}}, \ \textit{for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_- -1}.$$

Proof:

$$\begin{aligned}\frac{\partial \Delta_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A^{(L_-)}_{-i,j}} &= \frac{\partial}{\partial A^{(L_-)}_{-i,j}} \left(\sum_{i=1}^{J_{L_-}} A^{(L_-)}_{0_{i,i}} Z_{L_-,i} + \sum_{i=1}^{J_{L_-}} \sum_{j=1}^{J_{L_-}-1} \log\left(A^{(L_-)}_{-i,j}\right) D_{L_-,i,j} + \sum_{l=L_-+1}^{U_-} \sum_{i=1}^{J_l} \sum_{j=1}^{J_{l-1}} \lambda^{(-)}_{l,i,j} A^{(L_-)}_{-i,j} \right. \\ &\qquad \left. - \sum_{i=1}^{J_{L_-}} \lambda^{(L)}_{i} \sum_{j=1}^{J_{L_-}-1} A^{(L_-)}_{-i,j} \right),\end{aligned}$$

by selecting the relevant terms from the constrained log-likelihood,

$$\begin{aligned}&= \frac{\partial}{\partial A^{(L_-)}_{-i,j}} \left(- \sum_{i=1}^{J_{L_-}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{J_{L_-}} A^{(L_-)}_{0_{i,j}} + \sum_{j=1}^{J_{L_-}-1} A^{(L_-)}_{-i,j} + \sum_{j=1}^{J_{L_-}+1} A^{(L_-)}_{+i,j} \right) Z_{L_-,i} + \sum_{i=1}^{J_{L_-}} \sum_{j=1}^{J_{L_-}-1} \log\left(A^{(L_-)}_{-i,j}\right) D_{L_-,i,j} \right. \\ &\qquad \left. + \sum_{l=L_-+1}^{U_-} \sum_{i=1}^{J_l} \sum_{j=1}^{J_{l-1}} \lambda^{(-)}_{l,i,j} A^{(L_-)}_{-i,j} - \sum_{i=1}^{J_{L_-}} \lambda^{(L)}_{i} \sum_{j=1}^{J_{L_-}-1} A^{(L_-)}_{-i,j} \right),\end{aligned}$$

since the diagonal entries of $A^{(L_-)}_0$ are the negative of the relevant row sums,

$$= -Z_{L_-,i} + \frac{D_{L_-,i,j}}{A^{(L_-)}_{-i,j}} + \sum_{l=L_-+1}^{U_-} \lambda^{(-)}_{l,i,j} - \lambda^{(L)}_{i}.$$

By rearranging, we get

$$\widehat{A}_{-} = \widehat{A}^{(L_-)}_{-i,j} = \frac{D_{L_-,i,j}}{Z_{L_-,i} - \displaystyle\sum_{l=L_-+1}^{U_-} \lambda^{(-)}_{l,i,j} + \lambda^{(L)}_{i}}, \ \textit{for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_- -1}.$$

Lemma B.0.3.

$$\widehat{A}_+ = \widehat{A}^{(U_+)}_{-i,j} = \frac{U_{U_+,i,j}}{Z_{U_+,i} + \lambda^{(+)}_{U_+,i,j} - \lambda^{(U)}_i}, \text{ for } 1 \leq i \leq J_{U_+}, 1 \leq j \leq J_{U_+ + 1}.$$

Proof:

$$\begin{aligned}
\frac{\partial \Delta_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A^{(U_+)}_{+i,j}} &= \frac{\partial}{\partial A^{(U_+)}_{+i,j}} \left(\sum_{i=1}^{J_{U_+}} A^{(U_+)}_{0_{i,i}} Z_{U_+,i} + \sum_{i=1}^{J_{U_+}} \sum_{j=1}^{J_{U_+ + 1}} \log\left(A^{(U_+)}_{+i,j}\right) D_{U_+,i,j} \right. \\
&\qquad \left. - \sum_{i=1}^{J_l} \sum_{j=1}^{J_{l-1}} \lambda^{(+)}_{l,i,j} A^{(U_+)}_{+i,j} + \sum_{i=1}^{J_{U_+}} \lambda^{(U)}_i \sum_{j=1}^{J_{U_+ + 1}} A^{(U_+)}_{+i,j} \right),
\end{aligned}$$

by selecting the relevant terms from the constrained log-likelihood,

$$\begin{aligned}
&= \frac{\partial}{\partial A^{(U_+)}_{+i,j}} \left(- \sum_{i=1}^{J_{U_+}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{J_{U_+}} A^{(U_+)}_{0_{i,j}} + \sum_{j=1}^{J_{U_+ -1}} A^{(U_+)}_{-i,j} + \sum_{j=1}^{J_{U_+ + 1}} A^{(U_+)}_{+i,j} \right) Z_{U_+,i} \right. \\
&\qquad + \sum_{i=1}^{J_{U_+}} \sum_{j=1}^{J_{U_+ + 1}} \log\left(A^{(U_+)}_{+i,j}\right) D_{U_+,i,j} \\
&\qquad \left. - \sum_{i=1}^{J_l} \sum_{j=1}^{J_{l+1}} \lambda^{(+)}_{l,i,j} A^{(U_+)}_{+i,j} + \sum_{i=1}^{J_{U_+}} \lambda^{(U)}_i \sum_{j=1}^{J_{U_+ + 1}} A^{(U_+)}_{+i,j} \right),
\end{aligned}$$

since the diagonal entries of $A^{(U_+)}_0$ are the negative of the relevant row sums,

$$= -Z_{U_+,i} + \frac{D_{U_+,i,j}}{A^{(U_+)}_{+i,j}} - \lambda^{(+)}_{U_+,i,j} + \lambda^{(U)}_i.$$

By rearranging, we get

$$\widehat{A}_+ = \widehat{A}^{(U_+)}_{+i,j} = \frac{D_{U_+,i,j}}{Z_{U_+,i} + \lambda^{(+)}_{l,i,j} - \lambda^{(U)}_i}, \text{ for } 1 \leq i \leq J_{U_+}, 1 \leq j \leq J_{U_+ + 1}.$$

Lemma B.0.4.

$$\widehat{C}_+ = \widehat{A}^{(U_++1)}_{-_{i,j}} = \frac{U_{U_++1,i,j}}{Z_{U_++1,i} + \lambda_i^{(U)}}, \ \textit{for } 1 \leq i \leq J_{U_++1}, \leq j \leq J_C.$$

Proof:

$$\begin{aligned}
\frac{\partial \Delta_C(\boldsymbol{\theta}|\mathbf{d})}{\partial A^{(U_++1)}_{+_{i,j}}} &= \frac{\partial}{\partial A^{(U_++1)}_{+_{i,j}}} \left(\sum_{i=1}^{J_{U_++1}} A^{(U_++1)}_{0_{i,i}} Z_{U_++1,i} + \sum_{i=1}^{J_{U_++1}} \sum_{j=1}^{J_C} \log\left(A^{(U_++1)}_{+_{i,j}}\right) D_{U_++1,i,j} \right. \\
&\qquad \left. - \sum_{i=1}^{J_C} \lambda_i^{(U)} \sum_{j=1}^{J_C} A^{(U_++1)}_{+_{i,j}} \right), \\
&\qquad \textit{by selecting the relevant terms from the constrained log-likelihood,} \\
&= \frac{\partial}{\partial A^{(U_++1)}_{+_{i,j}}} \left(- \sum_{i=1}^{J_{U_++1}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{J_{U_++1}} A^{(U_++1)}_{0_{i,j}} + \sum_{j=1}^{J_{U_+}} A^{(U_++1)}_{-_{i,j}} + \sum_{j=1}^{J_C} A^{(U_++1)}_{+_{i,j}} \right) Z_{U_++1,i} \right. \\
&\qquad \left. + \sum_{i=1}^{J_{U_++1}} \sum_{j=1}^{J_C} \log\left(A^{(U_++1)}_{+_{i,j}}\right) D_{U_++1,i,j} - \sum_{i=1}^{J_C} \lambda_i^{(U)} \sum_{j=1}^{J_C} A^{(U_++1)}_{+_{i,j}} \right), \\
&\qquad \textit{since the diagonal entries of } A_0^{(U_++1)} \textit{ are the negative of the relevant row sums,} \\
&= -Z_{U_++1,i} + \frac{D_{U_++1,i,j}}{A^{(U_++1)}_{+_{i,j}}} - \lambda_i^{(U)}.
\end{aligned}$$

By rearranging, we get

$$\widehat{C}_+ = \widehat{A}^{(U_++1)}_{+_{i,j}} = \frac{D_{U_++1,i,j}}{Z_{U_++1,i} + \lambda_i^{(U)}}, \ \textit{for } 1 \leq i \leq J_{U_++1}, \leq j \leq J_C.$$

Next, we shall use Lemmas A.0.1, A.0.2, B.0.1, B.0.2, B.0.3, and B.0.4 to derive the maximum likelihood estimates of the parameters $\boldsymbol{\theta} = (B_0, B_-, B_+, A_0, A_-, A_+, C_0, C_-, C_+)$ for a finite level-independent QBD process with boundaries at level 0 and level C. Note that this is the proof of Corollary 8.3.1.1.

Given that the proofs for each maximum likelihood estimate are similar, we only show the proofs for the maximum likelihood estimates of B_- and A_-.

Using Lemma B.0.1, we have that

$$\widehat{B}_{-} = \widehat{A}^{(L_{-}-1)}_{-i,j} = \frac{D_{L_{-}-1,i,j}}{Z_{L_{-}-1,i} - \lambda_i^{(L)}}, \text{ for } 1 \leq i \leq J_1, 1 \leq j \leq J_0. \tag{B.0.1}$$

Notice that we have not solved for $\lambda_i^{(L)}$. Recall that if $f_\ell = g_\ell = h_\ell = 1$, $\forall \ell$, then we have the additional constraint

$$\sum_{i=1}^{J_{L_B}} \lambda_i^{(L)} \left(\sum_{j=1}^{J_{L_{-}-2}} A^{(L_B-1)}_{-i,j} - \sum_{j=1}^{J_{L_B-1}} A^{(L_B)}_{-i,j} \right).$$

In order to solve for $\lambda_i^{(L)}$, we need to find an expression for the maximum likelihood estimate of $A^{(L_{-})}_{-i,j}$. Using Lemma B.0.2, we have that

$$A^{(L_{-})}_{-i,j} = \frac{D_{L_{-},i,j}}{Z_{L_{-},i} - \sum_{\ell=L_{-}+1}^{U_{-}} \lambda^{(-)}_{L_{-},i,j} + \lambda_i^{(L)}}, \text{ for } 1 \leq i \leq J_{L_{-}}, 1 \leq j \leq J_{L_{-}-1}. \tag{B.0.2}$$

To find the maximum likelihood estimate of $A^{(L_{-})}_{-i,j}$, we need to solve for $\lambda^{(-)}_{\ell,i,j}$, for $L_{-} + 1 \leq \ell \leq U_{-}$. Using Lemma A.0.2, we have that

$$A^{(\ell)}_{-i,j} = \frac{D_{\ell,i,j}}{Z_{\ell,i} + \lambda^{(-)}_{\ell,i,j}}, \text{ for } L_{-} + 1 \leq \ell \leq U_{-}, \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1},$$

which tells us that

$$\lambda^{(-)}_{\ell,i,j} = \frac{D_{\ell,i,j} - A^{(\ell)}_{-i,j} Z_{\ell,i}}{A^{(\ell)}_{-i,j}}, \text{ for } L_{-} + 1 \leq \ell \leq U_{-}, \text{ and } 1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1}.$$

By substituting the expressions for $\lambda^{(-)}_{\ell,i,j}$, $L_{-} + 1 \leq \ell \leq U_{-}$ into Equation (B.0.2), we obtain an expression for the maximum likelihood estimate of $A^{(L_{-})}_{-i,j}$.

$$A^{(L_-)}_{-i,j} = \frac{D_{L_-,i,j}}{Z_{L_-,i} - \displaystyle\sum_{\ell=L_-+1}^{U_-} \lambda^{(-)}_{L_-,i,j} + \lambda^{(L)}_i}$$

$$A^{(L_-)}_{-i,j} Z_{L_-,i} - A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} \lambda^{(-)}_{\ell,i,j} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j}$$

$$A^{(L_-)}_{-i,j} Z_{L_-,i} - A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} \frac{D_{\ell,i,j} - A^{(\ell)}_{-i,j} Z_{\ell,i}}{A^{(\ell)}_{-i,j}} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j}$$

substituting the expressions for $\lambda^{(-)}_{\ell,i,j}, L_- + 1 \leq \ell \leq U_-$ *into the equation,*

$$A^{(L_-)}_{-i,j} Z_{L_-,i} - A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} \frac{D_{\ell,i,j}}{A^{(\ell)}_{-i,j}} + A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} Z_{\ell,i} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j}$$

$$A^{(L_-)}_{-i,j} \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i} \right) - A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} \frac{D_{\ell,i,j}}{A^{(\ell)}_{-i,j}} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j}$$

collecting the terms involving $Z_{\ell,i}$

$$A^{(L_-)}_{-i,j} \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i} \right) - A^{(L_-)}_{-i,j} \sum_{\ell=L_-+1}^{U_-} \frac{D_{\ell,i,j}}{A^{(L_-)}_{-i,j}} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j},$$

since $A^{(L_-)}_{-i,j} = A^{(\ell)}_{-i,j}$ *for* $L_- + 1 \leq \ell \leq U_-$ *and* $1 \leq i \leq J_\ell, 1 \leq j \leq J_{\ell-1}$.

$$A^{(L_-)}_{-i,j} \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i} \right) - \sum_{\ell=L_-+1}^{U_-} D_{\ell,i,j} + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = D_{L_-,i,j}$$

$$A^{(L_-)}_{-i,j} \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i} \right) + A^{(L_-)}_{-i,j} \lambda^{(L)}_i = \sum_{\ell=L_-}^{U_-} D_{\ell,i,j}$$

$$A^{(L_-)}_{-i,j} \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i} + \lambda^{(L)}_i \right) = \sum_{\ell=L_-}^{U_-} D_{\ell,i,j}$$

By rearranging, we have that

$$A^{(L_-)}_{-i,j} = \frac{\displaystyle\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\displaystyle\sum_{\ell=L_-}^{U_-} Z_{\ell,i} + \lambda^{(L)}_i}, \text{ for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_- -1}.$$

Let's now solve for $\lambda_i^{(L)}$. Recall the additional constraint,

$$\sum_{i=1}^{J}\lambda_i^{(L)}\left(\sum_{j=1}^{J_{L_--2}} A^{(L_--1)}_{-i,j} - \sum_{j=1}^{J_{L_--1}} A^{(L_-)}_{-i,j}\right).$$

By substitution, we have the following.

$$\sum_{j=1}^{J_{L_--2}} A^{(L_--1)}_{-i,j} = \sum_{j=1}^{J_{L_--1}} A^{(L_-)}_{-i,j}$$

$$\frac{\displaystyle\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j}}{\displaystyle\sum_{\ell=1}^{L_--1} Z_{\ell,i} - \lambda_i^{(L)}} = \frac{\displaystyle\sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\displaystyle\sum_{\ell=L_-}^{U_-} Z_{\ell,i} + \lambda_i^{(L)}}$$

substituting the expressions for $A^{(L_--1)}_{-i,j}$ and $A^{(L_-)}_{-i,j}$ into the equation,

$$\left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i}\right)\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j} + \lambda_i^{(L)}\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j} = \left(\sum_{\ell=1}^{L_--1} Z_{\ell,i}\right)\sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j} - \lambda_i^{(L)}\sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}$$

$$\lambda_i^{(L)}\left(\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j} + \sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}\right) = \left(\sum_{\ell=1}^{L_--1} Z_{\ell,i}\right)\sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j} - \left(\sum_{\ell=L_-}^{U_-} Z_{\ell,i}\right)\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j}$$

collecting terms with $\lambda_i^{(L)}$.

By rearranging, we have that

$$\lambda_i^{(L)} = \frac{\left(\displaystyle\sum_{\ell=1}^{L_--1} Z_{\ell,i}\right)\displaystyle\sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j} - \left(\displaystyle\sum_{\ell=L_-}^{U_-} Z_{\ell,i}\right)\displaystyle\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j}}{\displaystyle\sum_{j=1}^{J_{L_--2}}\sum_{\ell=1}^{L_--1} D_{\ell,i,j} + \sum_{j=1}^{J_{L_--1}}\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}. \tag{B.0.3}$$

Hence, the maximum likelihood estimate of A_- is

$$\widehat{A}_- = \widehat{A}^{(L_-)}_{-_{i,j}} = \frac{\displaystyle\sum_{\ell=L_-}^{U_-} D_{\ell,i,j}}{\displaystyle\sum_{\ell=L_-}^{U_-} Z_{\ell,i} + \lambda_i^{(L)}}, \text{ for } 1 \leq i \leq J_{L_-}, 1 \leq j \leq J_{L_- -1},$$

where $\lambda_i^{(L)}$ is given in Equation (B.0.3).

The proofs for the remaining maximum likelihood estimates follow a similar structure to that presented above.

Appendix C

Summary of the dependence on initial values for Section 8.10

C.1 Structured QBD process with linearly increasing scalars associated with a decrease in level

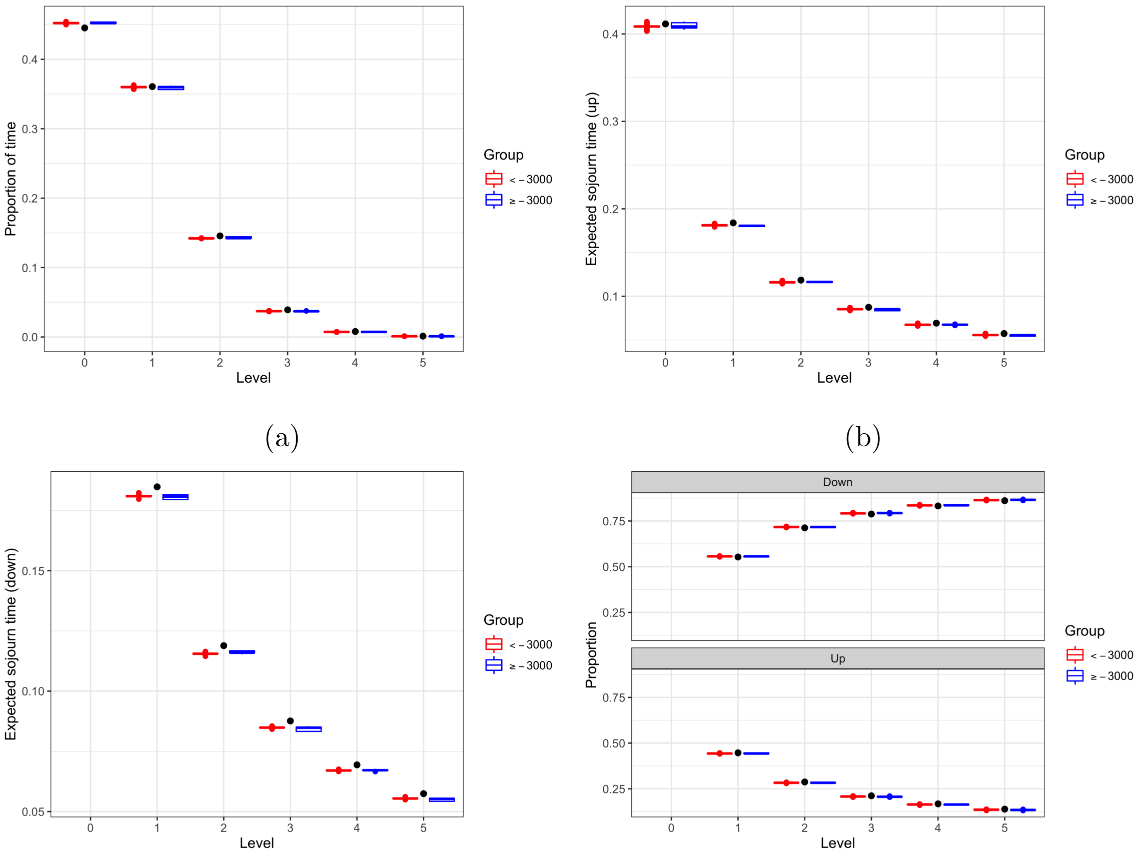
In this section, we explore the dependence on the initial values of the parameters in the example described in Section 8.10.1. Recall that we simulated a single sample of the assumed QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters to create the histogram of observed log-likelihoods of the data shown in Figure 8.10.2.

Of particular interest is whether there is any difference in the stationary and transient behaviours between the estimated structured QBD processes with observed log-likelihoods of the data less than -3000 (labelled Group A) compared to those with observed log-likelihoods of the data greater than or equal to -3000 (labelled Group B). Note that we expect to see a difference between the true and expected behaviours as the 1000 estimations are based on a single sample of data.

Figure C.1.1 shows that there is very little difference between the stationary and transient behaviours of the estimated structured QBD processes in Group A compared to the estimated structured QBD processes in Group B. These results are also demonstrated by the similar MSE values presented in Table C.1.1.

Behaviour	Group	Mean	Variance
Stationary distribution	A	9.295×10^{-6}	1.067×10^{-12}
	B	1.030×10^{-5}	8.919×10^{-12}
Sojourn time (Up)	A	5.156×10^{-6}	2.604×10^{-12}
	B	6.866×10^{-6}	2.114×10^{-12}
Sojourn time (Down)	A	7.760×10^{-6}	5.773×10^{-13}
	B	8.308×10^{-6}	3.759×10^{-12}
Transition probabilities	A	2.013×10^{-3}	1.688×10^{-11}
	B	2.017×10^{-3}	5.318×10^{-11}

Table C.1.1: Mean and variance of the MSE for each type of behaviour and each group of structured QBD processes with linearly increasing scalars associated with a decrease in level.



(a)

(b)

(c)

(d)

Figure C.1.1: Comparison of the behaviour of the estimated structured QBD processes with log-likelihoods less than -3000 to the behaviour of the estimated structured QBD process with log-likelihoods greater than or equal to -3000. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (black points) for the first six levels. (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (black points) for the first six levels. (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (black points) for the first six levels. (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (black points) for the first six levels. By design, the sojourn times conditioned on moving down from level 0 are 0 and the transition probabilities for level 0 are omitted as they are either 0 or 1.

C.2 Structured QBD process with linearly decreasing scalars associated with an increase in level

In this section, we explore the dependence on the initial values of the parameters in the example described in Section 8.10.2. The histogram of observed log-likelihoods of the data shown in Figure 8.10.4 was created by fitting 1000 QBD processes, each with a different set of randomly generated initial values of the parameters, to a single sample of data with 10,000 changes in level.

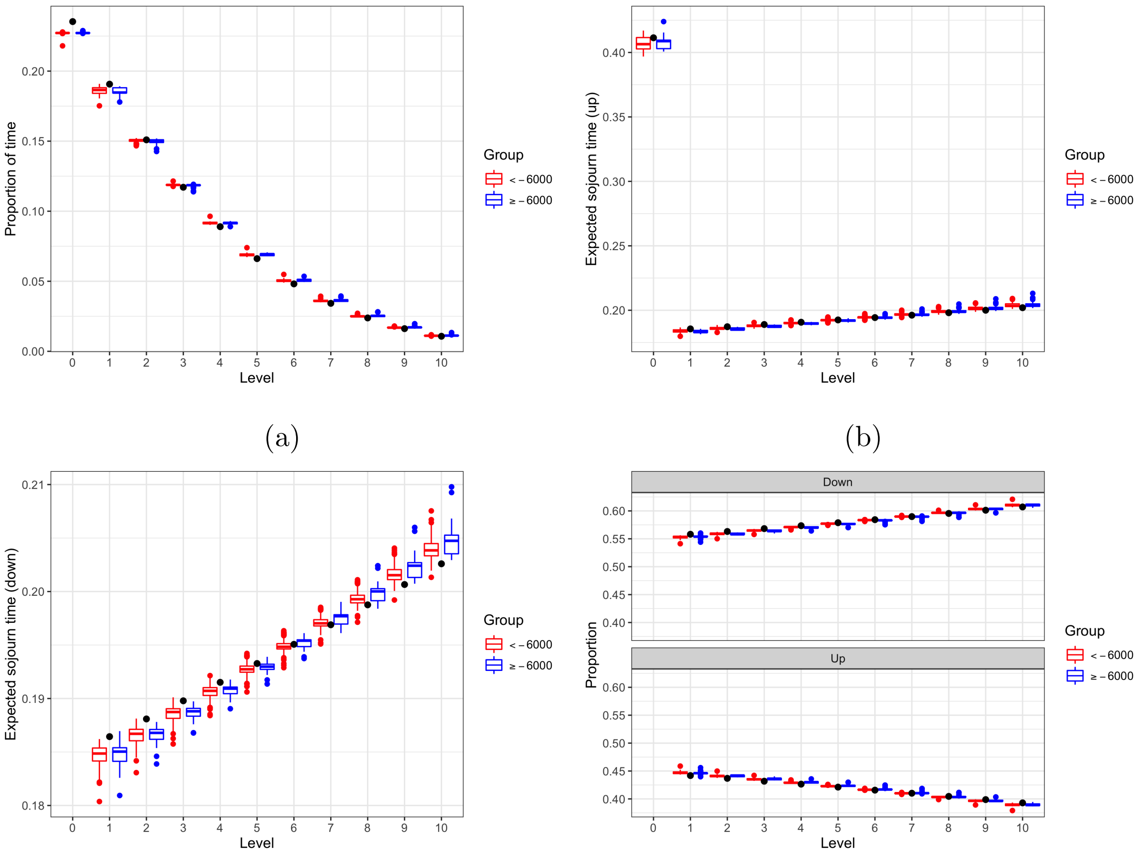
Here, we are interested in whether there is any difference in the stationary and transient behaviours between the estimated structured QBD processes with observed log-likelihoods of the data less than -6000 (labelled Group A) compared to those with observed log-likelihoods of the data greater than or equal to -6000 (labelled Group B). Similar to before, we expect to see a difference between the true and expected behaviours as the 1000 estimations are based on a single sample of data.

Figures C.2.1a and C.2.1d show that there is very little difference between the stationary distributions and transition probabilities between levels of the estimated structured QBD processes in Group A compared to those of the estimated structured QBD processes in Group B, respectively. These results are also demonstrated by the similar MSE values presented in Table C.2.1.

Additionally, the expected sojourn times of the estimated structured QBD processes in Group A are similar to the expected sojourn times of the estimated structured QBD processes in Group B, as illustrated in Figures C.2.1b and C.2.1c.

Behaviour	Group	Mean	Variance
Stationary distribution	A	6.930×10^{-6}	5.977×10^{-12}
	B	7.855×10^{-6}	9.409×10^{-12}
Sojourn time (Up)	A	1.015×10^{-5}	6.397×10^{-11}
	B	1.677×10^{-5}	7.181×10^{-10}
Sojourn time (Down)	A	7.661×10^{-4}	2.436×10^{-9}
	B	8.019×10^{-4}	2.064×10^{-8}
Transition probabilities	A	6.774×10^{-3}	7.485×10^{-10}
	B	6.782×10^{-3}	7.793×10^{-10}

Table C.2.1: Mean and variance of the MSE for each type of behaviour and each group of structured QBD processes with linearly decreasing scalars associated with an increase in level.



(a)

(b)

(c)

(d)

Figure C.2.1: Comparison of the behaviour of the estimated structured QBD processes with log-likelihoods less than -6000 to the behaviour of the estimated structured QBD process with log-likelihoods greater than or equal to -6000. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (black points) for the first eleven levels. (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (black points) for the first eleven levels. (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (black points) for the first eleven levels. (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (black points) for the first eleven levels. By design, the sojourn times conditioned on moving up from level 10 are 0, the sojourn times conditioned on moving down from level 0 are 0, and the transition probabilities for levels 0 and 10 are omitted as they are either 0 or 1.

C.3 Structured QBD process with linearly decreasing scalars associated with an increase in level and quadratic scalars associated with a decrease in level

Lastly, we explore the dependence on the initial values of the parameters in the example described in Section 8.10.3. Again, recall that we simulated a single sample of data from the assumed QBD process with 10,000 changes in level and then started the EM algorithm with 1000 different sets of randomly generated initial values of the parameters. From this process, we obtain the histogram of observed log-likelihoods of the data shown in Figure 8.10.7.

In this example, we are interested in whether there is any difference in the stationary and transient behaviours between the estimated structured QBD processes with observed log-likelihoods of the data less than -6000 (labelled Group A) compared to those with observed log-likelihoods of the data greater than or equal to -6000 (labelled Group B). Similar to before, we expect to see a difference between the true and expected behaviours as the 1000 estimations are based on a single sample of data.

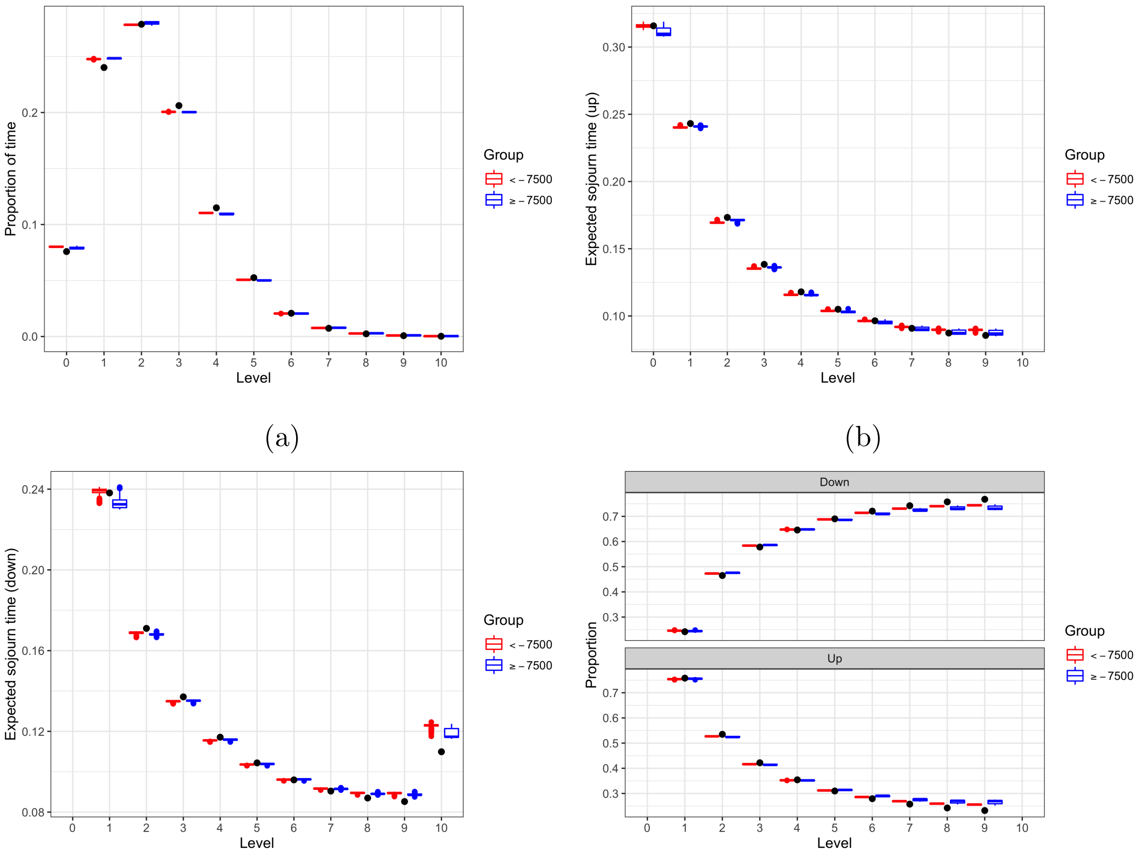
Firstly, Figure C.3.1a shows that there is very little difference between the stationary distributions of the estimated structured QBD processes in Group A compared to the stationary distributions of the estimated structured QBD processes in Group B.

Secondly, Figures C.3.1b and C.3.1c show that the expected sojourn times of the estimated structured QBD processes in Group A are slightly different to the expected sojourn times of the estimated structured QBD processes in Group B but not enough to suggest that estimated structured QBD processes with higher observed log-likelihoods better estimate the conditional sojourn times.

Lastly, Figure C.3.1d shows that there is very little difference between the transition probabilities between levels of the estimated structured QBD processes in Group A compared to the transition probabilities between levels of the estimated structured QBD processes in Group B. These results are also demonstrated by the similar MSE values presented in Table C.3.1.

Behaviour	Group	Mean	Variance
Stationary distribution	A	1.188×10^{-5}	2.246×10^{-14}
	B	1.341×10^{-5}	1.049×10^{-12}
Sojourn time (Up)	A	5.300×10^{-6}	5.170×10^{-13}
	B	5.817×10^{-6}	1.859×10^{-12}
Sojourn time (Down)	A	1.780×10^{-5}	2.703×10^{-12}
	B	1.282×10^{-5}	1.109×10^{-11}
Transition probabilities	A	1.039×10^{-4}	2.519×10^{-10}
	B	2.287×10^{-4}	6.793×10^{-9}

Table C.3.1: Mean and variance of the MSE for each type of behaviour and each group of structured QBD processes with linearly decreasing scalars associated with an increase in level and quadratic scalars associated with a decrease in level.



(a) (b)

(c) (d)

Figure C.3.1: Comparison of the behaviour of the estimated structured QBD processes with log-likelihoods less than -6000 to the behaviour of the estimated structured QBD process with log-likelihoods greater than or equal to -6000. (a) Box-plots of the estimated level-stationary distribution and the true level-stationary distribution (black points). (b) Box-plots of the expected sojourn times conditioned on moving up from a level and the true expected sojourn times conditioned on moving up a level (black points). (c) Box-plots of the expected sojourn times conditioned on moving down from a level and the true expected sojourn times conditioned on moving down a level (black points). (d) Box-plots of the estimated transition probabilities between each level and the true transition probabilities between each level (black points). By design, the sojourn times conditioned on moving up from level 10 are 0, the sojourn times conditioned on moving down from level 0 are 0, and the transition probabilities for levels 0 and 10 are omitted as they are either 0 or 1.

Appendix D

Normal QQ-plots for Section 9.2.3

In this section, we present an illustrative example of why we cannot assume multivariate normality in our goodness of fit test, as discussed in Section 9.2.3.

Consider the example described in Section 9.4.2, which involved fitting an infinite level-independent QBD process to data generated from a finite level-independent QBD process. To generate the Mahalanobis distances for the empirical null distribution, we fitted an infinite level-independent QBD process to each simulated set of data generated from the infinite level-independent QBD process.

Figure D.0.1 plots the QQ-plots which compare the observed differences associated with each level and transition of interest (points) to the expected differences assuming a normal distribution (line). As observed in each QQ-plot, the data do not appear to be normally distributed. Therefore, the distribution of $\mathbf{V}(X)$ is not multivariate normal.

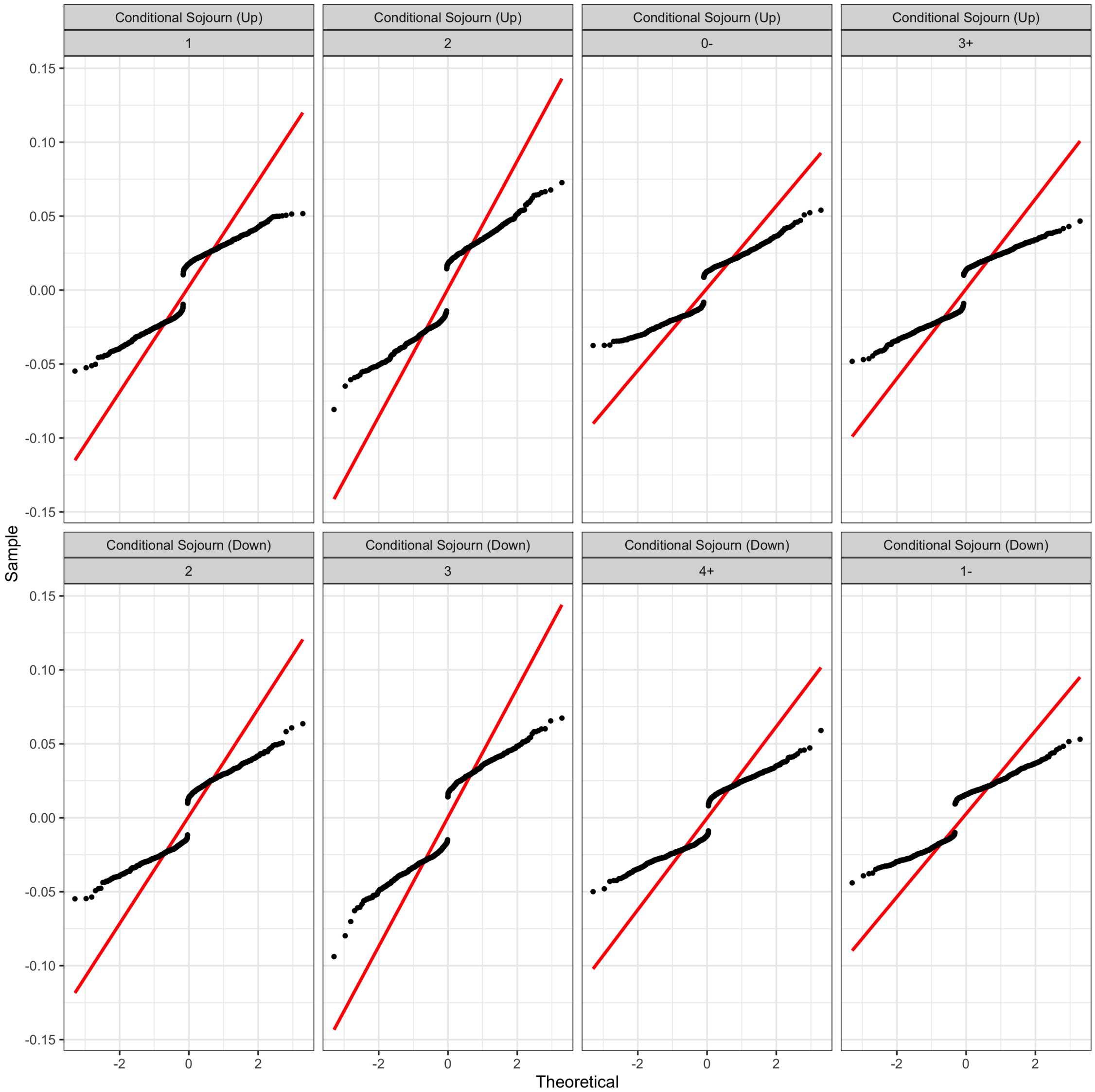


Figure D.0.1: Comparison of the observed differences associated with each level and transition of interest (points) to the expected differences assuming a normal distribution (line). Note that this information is taken from the null distribution of the example described in Section 9.4.2.

Appendix E

Summary of the structured QBD processes from Chapter 10

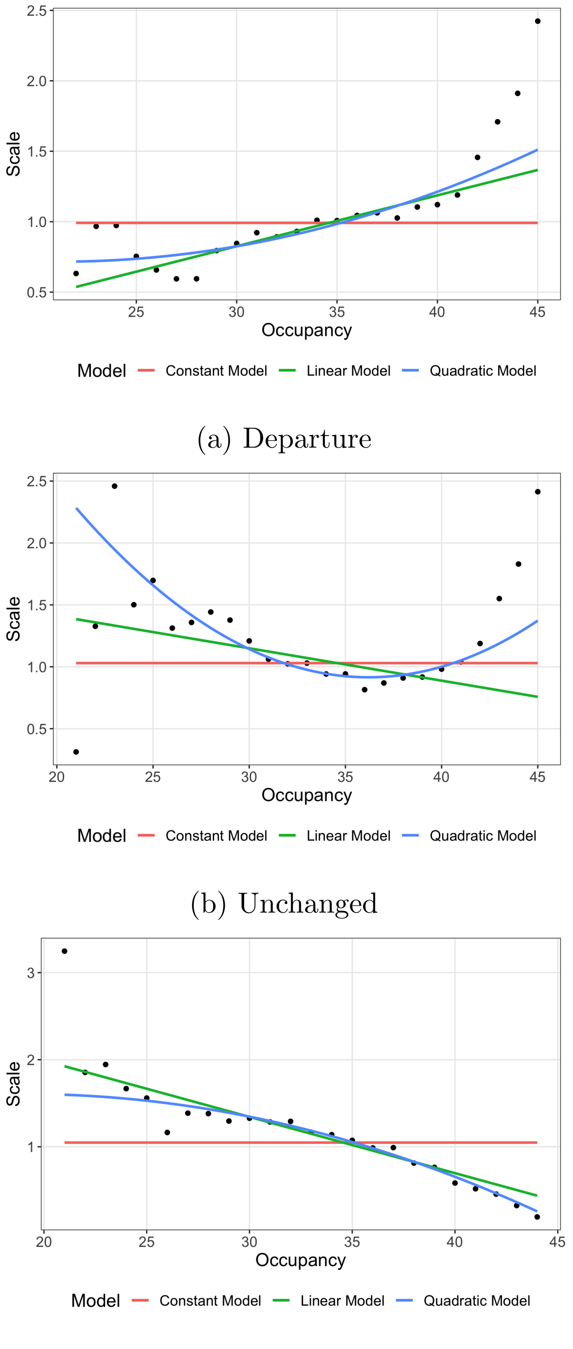
E.1 Identifying suitable models

In this section, we review the relevant plots and likelihood ratio test results used to determine the most suitable polynomial forms for each of the remaining types of structured QBD processes considered in Chapter 10.

E.1.1 Two-phase structured QBD process with level-dependent polynomial forms

Using the EM algorithm described in Section 8.3, we fitted several two-phase structured QBD processes with level-dependent scales to the RAH ICU data and chose the model with the highest log-likelihood.

We used weighted least squares regression to fit a constant, linear, and quadratic model to the maximum likelihood estimates of the departure, unchanged, and arrival level-dependent scales of the two-phase structured QBD process, as illustrated in Figure E.1.1.



(a) Departure

(b) Unchanged

(c) Arrival

Figure E.1.1: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level-dependent scales (points) for the fitted two-phase structured QBD process.

Based on the results of the likelihood-ratio tests presented in Table E.1.1, the behaviour of the level-dependent scales associated with arrivals and within level transitions are better explained using a quadratic model. However, a quadratic term is not needed to explain the behaviour of the level-dependent scales associated with departures.

Taking into consideration that an additional phase captures more of the variability in the bed occupancy data, these results suggest that admissions to the ICU are not necessarily decreasing at a linear rate and the rate of discharge depends on the number of patients in the ICU.

Matrix	Model 1	Model 2	Statistic	Degrees of freedom	P-value
Departure	Constant	Linear	32.275	1	1.339×10^{-8}
Departure	Linear	Quadratic	3.032	1	8.164×10^{-2}
Unchanged	Constant	Linear	7.953	1	4.800×10^{-3}
Unchanged	Linear	Quadratic	20.304	1	6.608×10^{-6}
Arrival	Constant	Linear	52.205	1	5.001×10^{-13}
Arrival	Linear	Quadratic	6.949	1	8.388×10^{-3}

Table E.1.1: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models for the two-phase structured QBD process with level-dependent scales.

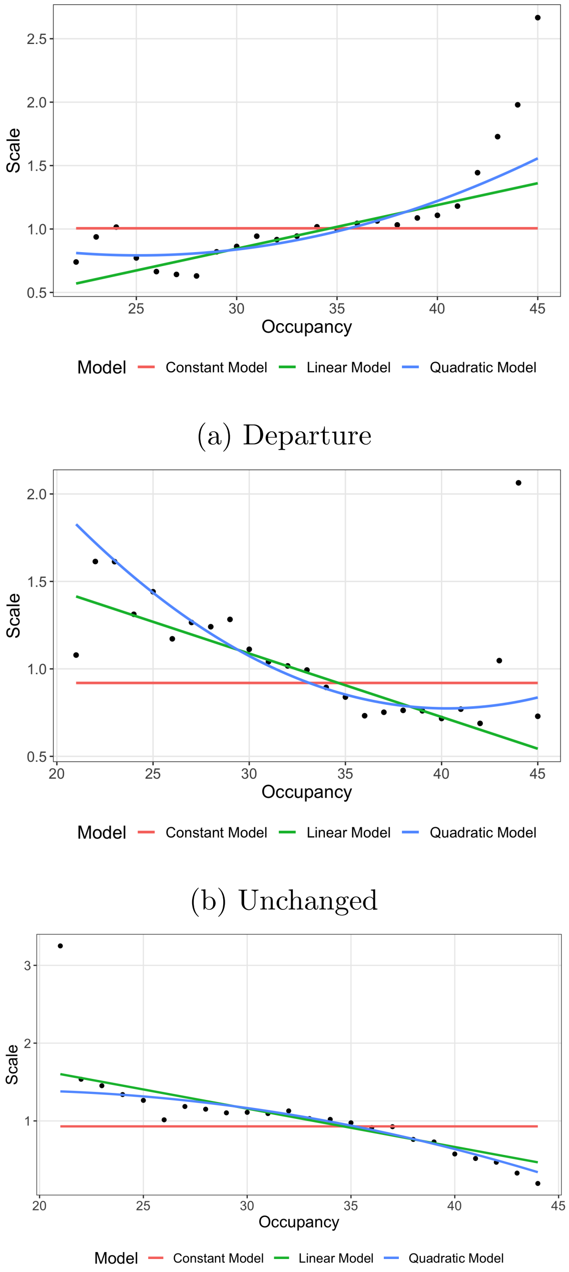
E.1.2 Three-phase structured QBD process with level-dependent polynomial forms

Using the EM algorithm described in Section 8.3, we fitted several three-phase structured QBD processes with level-dependent scales to the RAH ICU data and chose the model with the highest log-likelihood.

Using weighted least squares regression modelling, we fitted a constant, linear, and quadratic model to the maximum likelihood estimates of the level-dependent scales, as illustrated in Figure E.1.2.

The forms of the level-dependent scales associated with arrivals and departures are very similar to those of the two-phase structured QBD process with level-dependent scales. However, an additional phase has changed the behaviour of the level-dependent scales associated with within level changes.

Overall, a quadratic term is not needed to explain the behaviour of the level-dependent scales associated with arrivals but it is required to explain the behaviour of the level-dependent scales associated with departures and within level transitions, as demonstrated by the likelihood-ratio test results presented in Table E.1.2.



(a) Departure

(b) Unchanged

(c) Arrival

Figure E.1.2: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level-dependent scales (points) for the fitted three-phase structured QBD process.

Matrix	Model 1	Model 2	Statistic	Degrees of freedom	P-value
Departure	Constant	Linear	26.677	1	2.404×10^{-7}
Departure	Linear	Quadratic	5.208	1	2.248×10^{-2}
Unchanged	Constant	Linear	22.200	1	2.456×10^{-6}
Unchanged	Linear	Quadratic	8.204	1	4.181×10^{-3}
Arrival	Constant	Linear	36.197	1	1.784×10^{-9}
Arrival	Linear	Quadratic	2.738	1	9.801×10^{-2}

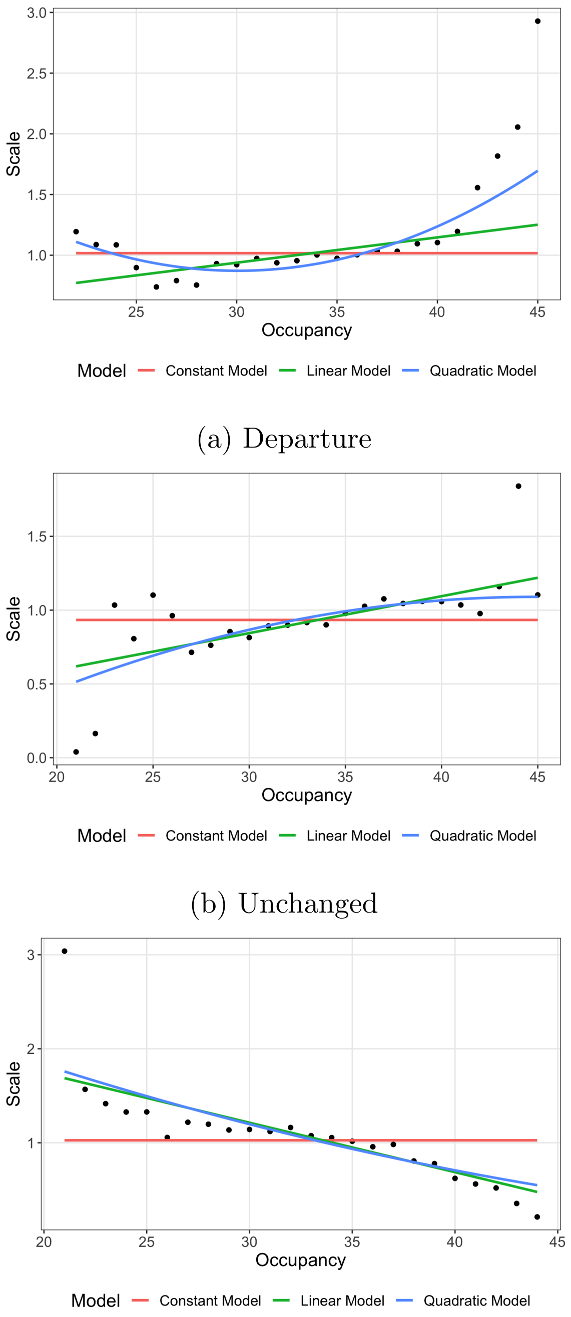
Table E.1.2: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models for the three-phase structured QBD process with level-dependent scales.

E.1.3 Four-phase structured QBD process with level-dependent polynomial forms

Using the EM algorithm described in Section 8.3, we fitted several four-phase structured QBD processes with level-dependent scales to the RAH ICU data and chose the model with the highest log-likelihood.

Using weighted least squares regression modelling, we fitted a constant, linear, and quadratic model to the maximum likelihood estimates of the level-dependent scales, as illustrated in Figure E.1.3.

The level-dependent scales are very similar to those of the three-phase structured QBD process with level-dependent scales, except for the level-dependent scales associated with within level changes which now appear to be more constant.



(a) Departure

(b) Unchanged

(c) Arrival

Figure E.1.3: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level-dependent scales (points) for the fitted four-phase structured QBD process.

Again, a quadratic term is not needed to explain the behaviour of the level-dependent scales associated with arrivals and is no longer needed to explain the behaviour of the level-dependent scales associated with within level changes. However, a quadratic term is now required to explain the behaviour of the level-dependent scales associated with departures, as demonstrated by the likelihood-ratio test results presented in Table E.1.3.

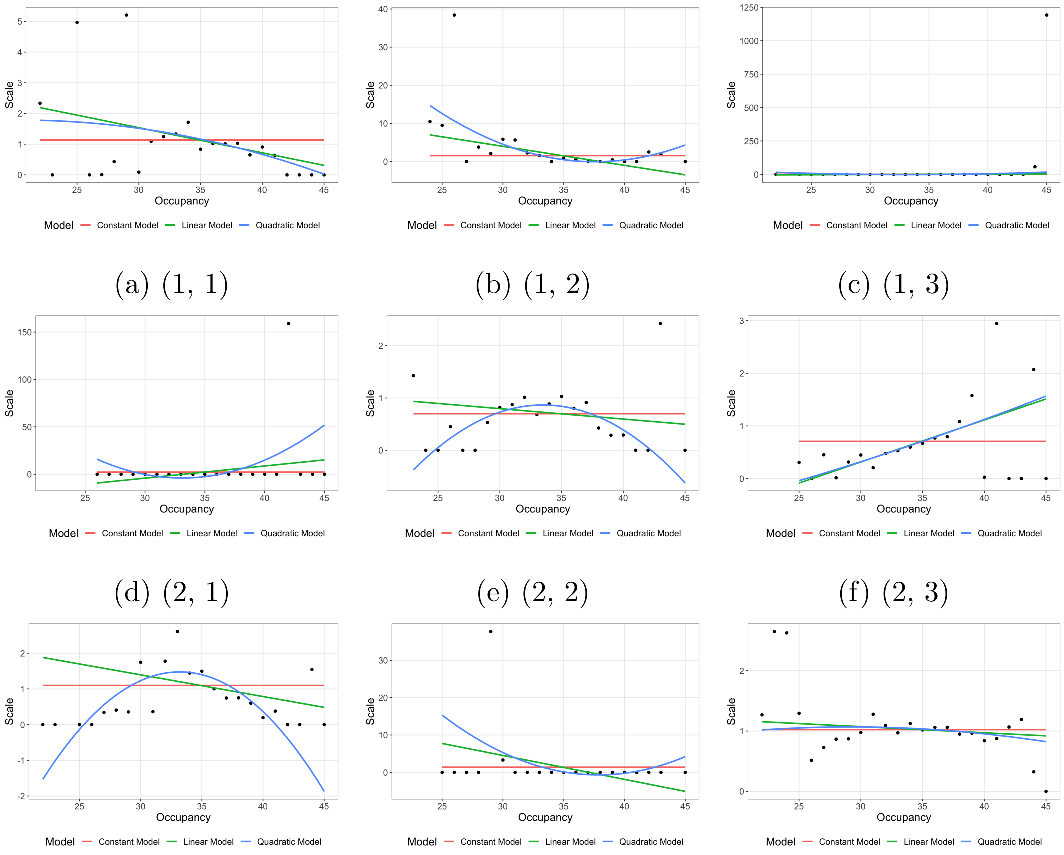
Matrix	Model 1	Model 2	Statistic	Degrees of freedom	P-value
Departure	Constant	Linear	9.972	1	1.589×10^{-3}
Departure	Linear	Quadratic	18.608	1	1.606×10^{-5}
Unchanged	Constant	Linear	16.166	1	5.803×10^{-5}
Unchanged	Linear	Quadratic	1.420	1	0.233
Arrival	Constant	Linear	26.254	1	2.993×10^{-7}
Arrival	Linear	Quadratic	0.305	1	0.581

Table E.1.3: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models for the four-phase structured QBD process with level-dependent scales.

E.1.4 Three-phase structured QBD process with level and phase transition dependent polynomial forms

Extending the EM algorithm described in Section 8.3, we fitted several three-phase structured QBD processes with level and phase transition dependent scales to the RAH ICU data and chose the model with the highest log-likelihood.

First, we consider the constant, linear, and quadratic models fitted to the maximum likelihood estimates of the level and phase transition dependent scales associated with a decrease in level of the fitted three-phase structured QBD process, as illustrated in Figure E.1.4. Note that zero valued level and phase transition dependent scales are omitted in the case where the corresponding block matrix element is also zero.



(a) (1, 1)

(b) (1, 2)

(c) (1, 3)

(d) (2, 1)

(e) (2, 2)

(f) (2, 3)

(g) (3, 1)

(h) (3, 2)

(i) (3, 3)

Figure E.1.4: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent departure scales for the fitted three-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

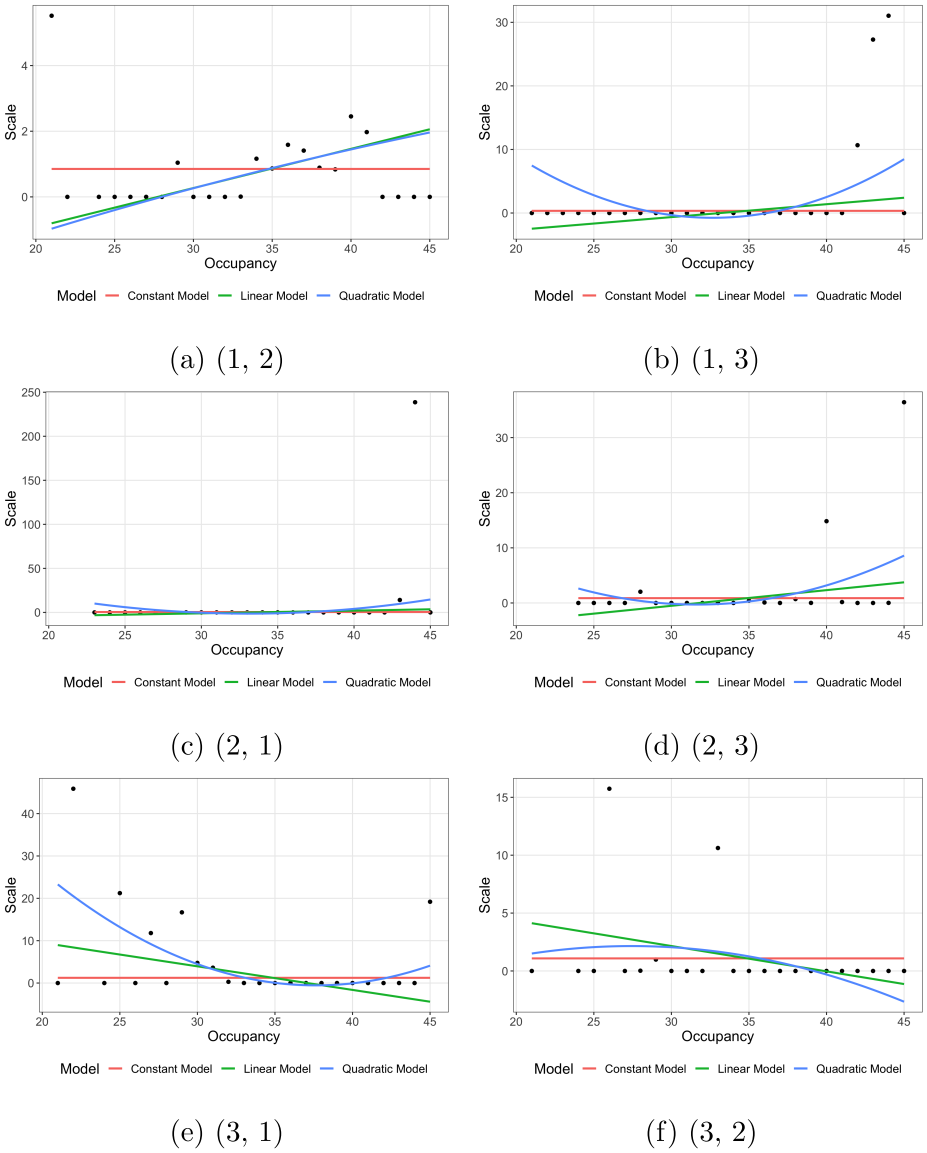
Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	2.693	1	0.101
(1, 1)	Linear	Quadratic	0.082	1	0.775
(1, 2)	Constant	Linear	6.684	1	9.726×10^{-3}
(1, 2)	Linear	Quadratic	4.721	1	2.980×10^{-2}
(1, 3)	Constant	Linear	0.077	1	0.781
(1, 3)	Linear	Quadratic	0.311	1	0.577
(2, 1)	Constant	Linear	1.189	1	0.276
(2, 1)	Linear	Quadratic	2.297	1	0.130
(2, 2)	Constant	Linear	1.023	1	0.312
(2, 2)	Linear	Quadratic	9.146	1	2.493×10^{-3}
(2, 3)	Constant	Linear	6.979	1	8.247×10^{-3}
(2, 3)	Linear	Quadratic	0.008	1	0.927
(3, 1)	Constant	Linear	2.386	1	0.122
(3, 1)	Linear	Quadratic	11.165	1	8.336×10^{-4}
(3, 2)	Constant	Linear	2.421	1	0.120
(3, 2)	Linear	Quadratic	1.181	1	0.277
(3, 3)	Constant	Linear	1.419	1	0.234
(3, 3)	Linear	Quadratic	0.320	1	0.571

Table E.1.4: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with departures for the three-phase structured QBD process with level and phase transition dependent scales.

Using the results of the likelihood-ratio tests presented in Table E.1.4, neither linear or quadratic terms are needed to explain the behaviour of the level and phase transition dependent scales associated with the (1, 1), (1, 3), (2, 1), (3, 2), and (3, 3) phase transitions. However, further investigation is needed for the (1, 3) and (2, 1) phase transitions. For the remaining phase transitions, the quadratic functional form best fits the level and phase transition dependent scales associated with the (1, 2), (2, 2), and (3, 1) phase transitions, whereas a linear functional form best fits the level and phase transition dependent scales associated with the (2, 3) phase transition.

Given the presence of outliers for the (1, 3) and (2, 1) phase transitions, we further explored the behaviour of the level and phase transition dependent scales by removing data with a value above 50. Using a similar process to before, the results of the updated likelihood ratio tests confirmed that neither a linear term or a quadratic term is needed to explain the behaviour of the level and phase transition dependent scales associated with the (1, 3) and (2, 1) phase transitions.

Next, the estimated constant, linear, and quadratic models of the maximum likelihood estimates of the level and phase transition dependent scales associated with transitions within a level of the fitted three-phase structured QBD process are plotted in Figure E.1.5. Note that zero valued level and phase transition dependent scales are omitted in the case where the corresponding block matrix element is also zero.



(a) (1, 2)

(b) (1, 3)

(c) (2, 1)

(d) (2, 3)

(e) (3, 1)

(f) (3, 2)

Figure E.1.5: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent unchanged scales for the fitted three-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

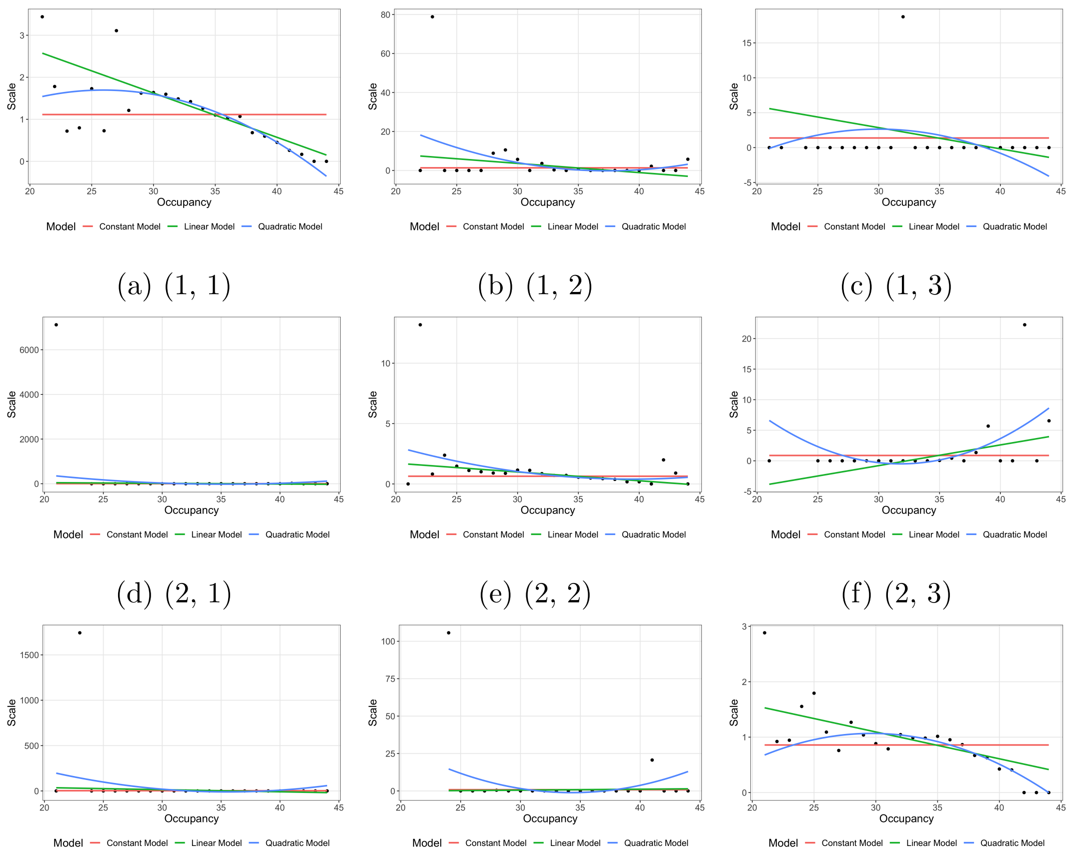
Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 2)	Constant	Linear	11.080	1	8.725×10^{-4}
(1, 2)	Linear	Quadratic	0.022	1	0.881
(1, 3)	Constant	Linear	2.049	1	0.152
(1, 3)	Linear	Quadratic	5.125	1	2.358×10^{-2}
(2, 1)	Constant	Linear	0.316	1	0.574
(2, 1)	Linear	Quadratic	1.015	1	0.314
(2, 3)	Constant	Linear	2.361	1	0.124
(2, 3)	Linear	Quadratic	1.712	1	0.191
(3, 1)	Constant	Linear	8.055	1	4.539×10^{-3}
(3, 1)	Linear	Quadratic	6.001	1	1.430×10^{-2}
(3, 2)	Constant	Linear	1.397	1	0.237
(3, 2)	Linear	Quadratic	0.163	1	0.686

Table E.1.5: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with transitions within each level for the three-phase structured QBD process with level and phase transition dependent scales.

The linear form best describes the behaviour observed in the level and phase transition dependent scales associated with the (1, 2) phase transition and the quadratic form best describes the behaviour observed in the level and phase transition dependent scales associated with the (1, 3) and (3, 1) phase transitions, as shown in Table E.1.5. For the remaining phase transitions, neither a linear or quadratic term is needed to explain the behaviour. However, further investigation is needed for the (2, 1) phase transition.

We further explored the behaviour of the level and phase transition dependent scales associated with the (2, 1) phase transition by removing values above 50. The results of the updated likelihood ratio test confirmed that behaviour of the level and phase transition dependent scales associated with the (2, 1) phase transition are independent of the level.
Lastly, using weighted least squares regression, we fitted a constant, linear, and quadratic model to the maximum likelihood estimates of the level and phase transition dependent scales associated with an increase in level of the fitted three-phase structured QBD process, as illustrated in Figure E.1.6.

Note that zero valued level and phase transition dependent scales are omitted in the case where the corresponding block matrix element is also zero.



(a) (1, 1)

(b) (1, 2)

(c) (1, 3)

(d) (2, 1)

(e) (2, 2)

(f) (2, 3)

(g) (3, 1)

(h) (3, 2)

(i) (3, 3)

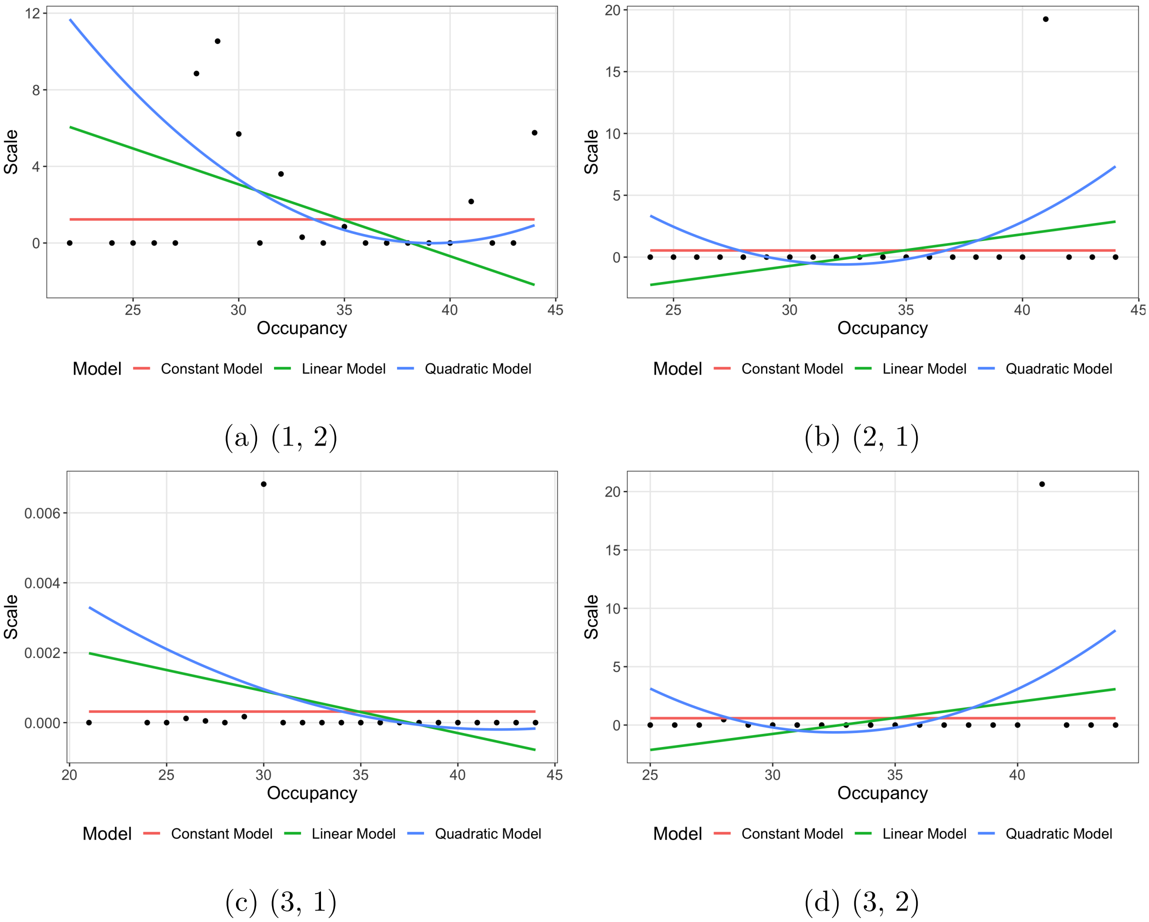
Figure E.1.6: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales for the fitted three-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	31.949	1	1.583×10^{-8}
(1, 1)	Linear	Quadratic	6.596	1	1.022×10^{-2}
(1, 2)	Constant	Linear	4.919	1	2.657×10^{-2}
(1, 2)	Linear	Quadratic	4.919	1	5.239×10^{-2}
(1, 3)	Constant	Linear	1.185	1	0.276
(1, 3)	Linear	Quadratic	0.362	1	0.548
(2, 1)	Constant	Linear	0.118	1	0.731
(2, 1)	Linear	Quadratic	1.002	1	0.317
(2, 2)	Constant	Linear	8.175	1	4.247×10^{-3}
(2, 2)	Linear	Quadratic	2.418	1	0.120
(2, 3)	Constant	Linear	3.880	1	4.888×10^{-2}
(2, 3)	Linear	Quadratic	3.121	1	7.730×10^{-2}
(3, 1)	Constant	Linear	0.367	1	0.545
(3, 1)	Linear	Quadratic	1.569	1	0.210
(3, 2)	Constant	Linear	0.028	1	0.867
(3, 2)	Linear	Quadratic	3.800	1	5.125×10^{-2}
(3, 3)	Constant	Linear	18.827	1	1.431×10^{-5}
(3, 3)	Linear	Quadratic	9.684	1	1.859×10^{-3}

Table E.1.6: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals for the three-phase structured QBD process with level and phase transition dependent scales.

Using the results of the likelihood-ratio tests presented in Table E.1.6, the linear form best describes the level and phase transition dependent scales associated with the (1, 2), (2, 2), and (2, 3) phase transitions and the quadratic form best describes the level and phase transition dependent scales associated with the (1, 1) and $(3, 3)$ phase transitions. Neither a linear or quadratic term is needed to explain the behaviour of the level and phase transition dependent scales for the (1, 3), (2, 1), (3, 1), and (3, 2) phase transitions. However, further investigation for the (1, 2), (2, 1), (3, 1), and (3, 2) phase transitions is required given the presence of outliers.

Figure E.1.7 plots the constant model, linear model, and a quadratic model fitted to the maximum likelihood estimates of the level and phase transition dependent scales that have a value below 50. This confirms that the level and phase transition dependent scales associated with the (2, 1), (3, 1), and (3, 2) phase transitions are constant and that the level and phase transition dependent scales associated with the (1, 2) phase transition may also be constant, rather than linearly decreasing.



(a) (1, 2)

(b) (2, 1)

(c) (3, 1)

(d) (3, 2)

Figure E.1.7: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales with a value below 50 for the fitted three-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

The likelihood ratio tests presented in Table E.1.7 now suggest that the level and phase transition dependent scales associated with the (1, 2) phase transition are linear, and the level and phase transition dependent scales associated with the $(2, 1)$, (3, 1), and (3, 2) phase transitions are constant.

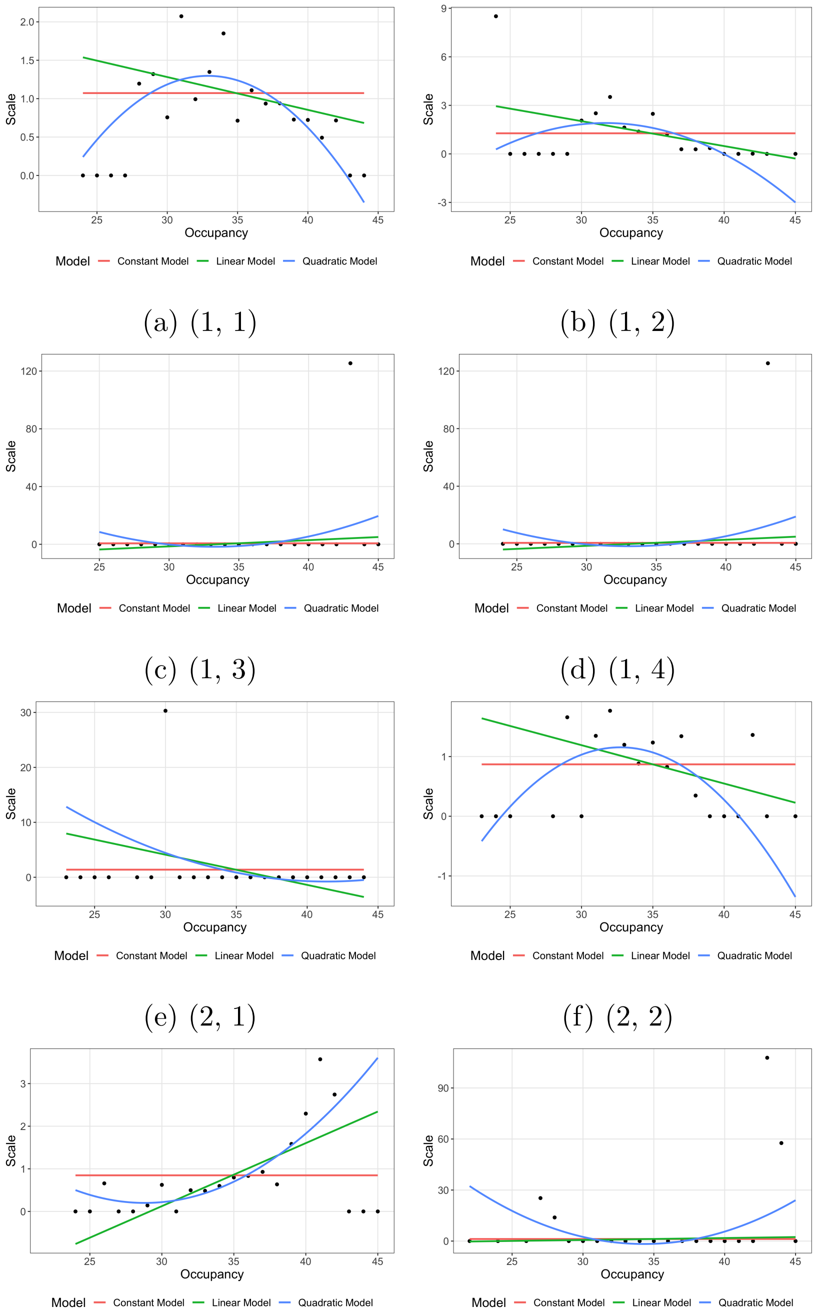
Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 2)	Constant	Linear	7.399	1	6.527×10^{-3}
(1, 2)	Linear	Quadratic	2.331	1	0.127
(2, 1)	Constant	Linear	1.835	1	0.176
(2, 1)	Linear	Quadratic	2.188	1	0.139
(3, 1)	Constant	Linear	2.144	1	0.143
(3, 1)	Linear	Quadratic	0.213	1	0.645
(3, 2)	Constant	Linear	1.705	1	0.192
(3, 2)	Linear	Quadratic	2.201	1	0.138

Table E.1.7: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals for the three-phase structured QBD process with level and phase transition dependent scales with a value below 50.

E.1.5 Four-phase structured QBD process with level and phase transition dependent polynomial forms

Similar to before, we extended the EM algorithm described in Section 8.3 to fit several four-phase structured QBD processes with level-dependent scales to the RAH ICU data and chose the model with the highest log-likelihood.

First, we consider the constant, linear, and quadratic models fitted to the maximum likelihood estimates of the level and phase transition dependent scales associated with a decrease in level of the fitted four-phase structured QBD process, as illustrated in Figures E.1.8 and E.1.9.



(a) (1, 1)

(b) (1, 2)

(c) (1, 3)

(d) (1, 4)

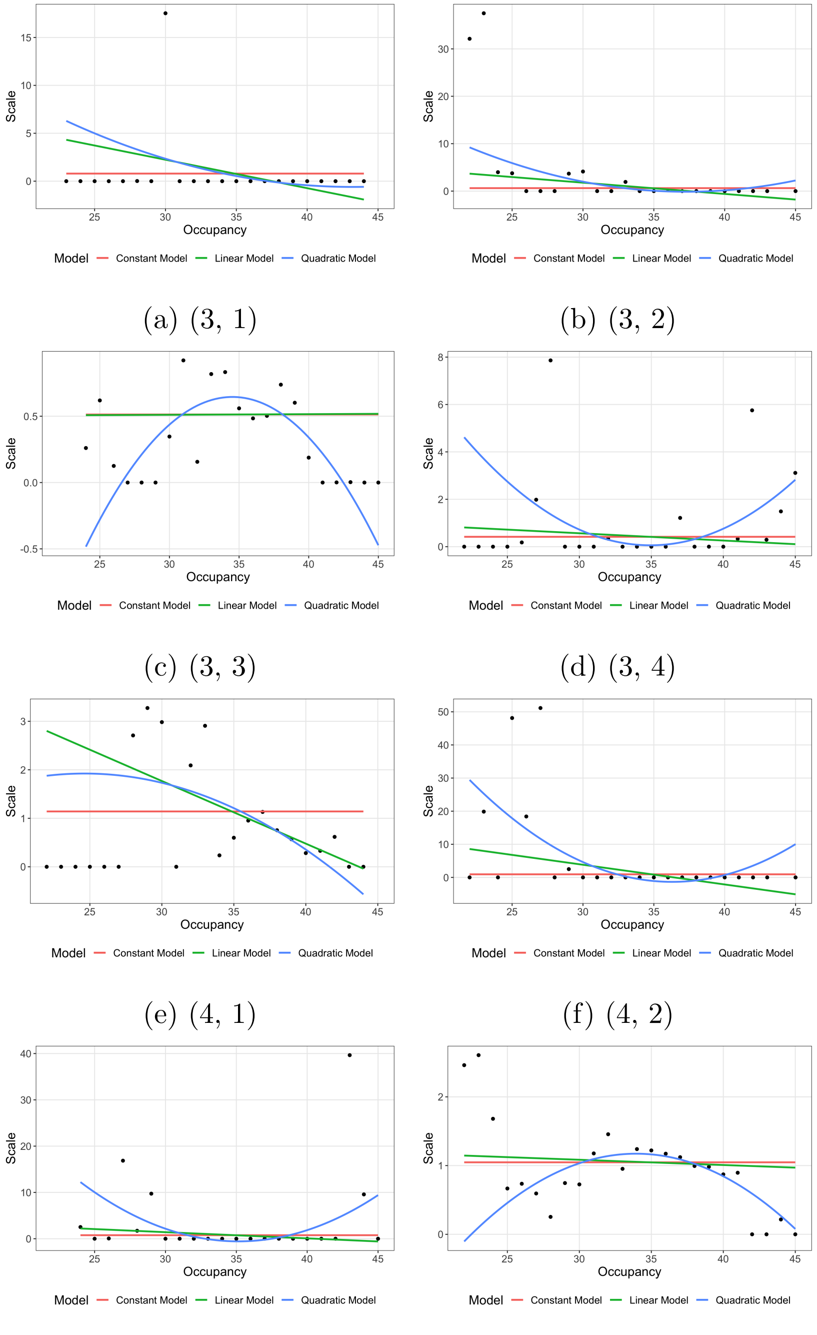
(e) (2, 1)

(f) (2, 2)

(g) (2, 3)

(h) (2, 4)

Figure E.1.8: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent departure scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.



(a) (3, 1)

(b) (3, 2)

(c) (3, 3)

(d) (3, 4)

(e) (4, 1)

(f) (4, 2)

(g) (4, 3)

(h) (4, 4)

Figure E.1.9: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent departure scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	2.497	1	0.114
(1, 1)	Linear	Quadratic	6.423	1	1.127×10^{-2}
(1, 2)	Constant	Linear	5.656	1	1.739×10^{-2}
(1, 2)	Linear	Quadratic	4.921	1	2.654×10^{-2}
(1, 3)	Constant	Linear	0.595	1	0.441
(1, 3)	Linear	Quadratic	1.590	1	0.207
(1, 4)	Constant	Linear	0.614	1	0.434
(1, 4)	Linear	Quadratic	1.600	1	0.206
(2, 1)	Constant	Linear	2.077	1	0.150
(2, 1)	Linear	Quadratic	0.263	1	0.608
(2, 2)	Constant	Linear	3.170	1	7.502×10^{-2}
(2, 2)	Linear	Quadratic	5.477	1	1.927×10^{-2}
(2, 3)	Constant	Linear	17.270	1	3.242×10^{-5}
(2, 3)	Linear	Quadratic	4.846	1	2.770×10^{-2}
(2, 4)	Constant	Linear	0.047	1	0.829
(2, 4)	Linear	Quadratic	4.346	1	3.71×10^{-2}
(3, 1)	Constant	Linear	1.990	1	0.158
(3, 1)	Linear	Quadratic	0.152	1	0.697
(3, 2)	Constant	Linear	4.494	1	3.402×10^{-2}
(3, 2)	Linear	Quadratic	3.518	1	6.071×10^{-2}
(3, 3)	Constant	Linear	0.001	1	0.977
(3, 3)	Linear	Quadratic	9.812	1	1.734×10^{-3}
(3, 4)	Constant	Linear	0.172	1	0.678
(3, 4)	Linear	Quadratic	3.468	1	6.257×10^{-2}
(4, 1)	Constant	Linear	5.536	1	1.862×10^{-2}
(4, 1)	Linear	Quadratic	0.392	1	0.531
(4, 2)	Constant	Linear	2.983	1	8.416×10^{-2}
(4, 2)	Linear	Quadratic	5.284	1	2.152×10^{-2}
(4, 3)	Constant	Linear	0.360	1	0.548
(4, 3)	Linear	Quadratic	5.209	1	2.247×10^{-2}
(4, 4)	Constant	Linear	0.237	1	0.627
(4, 4)	Linear	Quadratic	9.392	1	2.180×10^{-3}

Table E.1.8: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with departures for the four-phase structured QBD process with level and phase transition dependent scales.

Based on the results of the likelihood-ratio tests presented in Table E.1.8, the level and phase transition dependent scales associated with the (1, 1), (1, 2), $(2, 2)$, (2, 3), (2, 4), (3, 3), (4, 2), (4, 3), and (4, 4) phase transitions behave in a quadratic manner, whereas only the level and phase transition dependent scales associated with the (3, 2) and (4, 1) phase transitions behave linearly. The behaviour of the level and phase transition dependent scales for the remaining phase transitions remains relatively constant.

After removing outliers from the above analysis, only the level and phase transition dependent scales associated with the (2, 4) phase transition behaved differently. In this case, the suggested behaviour of the level and phase transition dependent scales associated with the (2, 4) phase transition changed from quadratic to constant, as demonstrated in Figure E.1.10 and Table E.1.9.

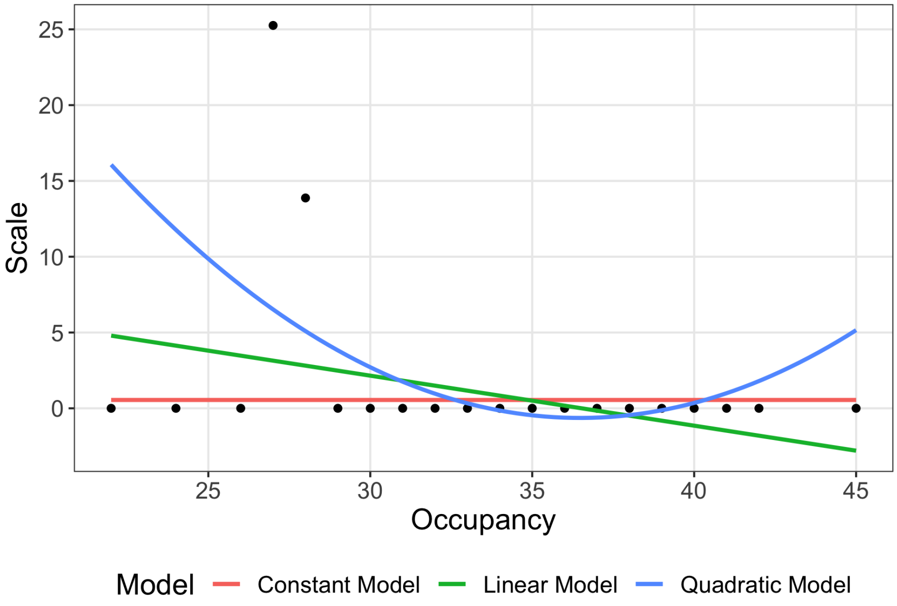
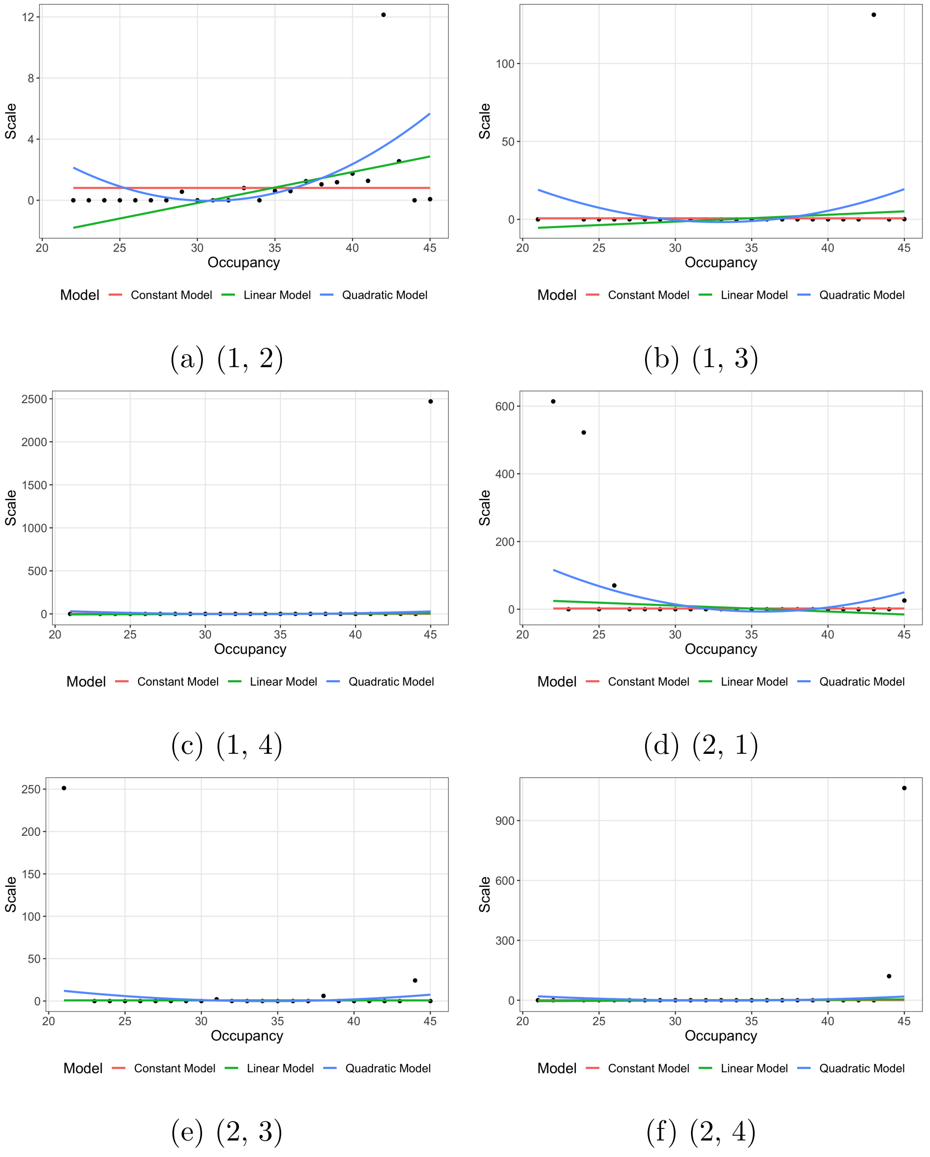


Figure E.1.10: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent deaprture scales associated with the (2, 4) phase transition and a value below 50 for the fitted four-phase structured QBD process.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(2, 4)	Constant	Linear	2.750	1	9.725×10^{-2}
(2, 4)	Linear	Quadratic	3.755	1	5.264×10^{-2}

Table E.1.9: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with departures, the (2, 4) phase transition, and a value below 50 for the four-phase structured QBD process with level and phase transition dependent scales.

Next, we consider the constant, linear, and quadratic models fitted to the maximum likelihood estimates of the level and phase transition dependent scales associated with no changes in level of the fitted four-phase structured QBD process, as illustrated in Figures E.1.11 and E.1.12.



(a) (1, 2)

(b) (1, 3)

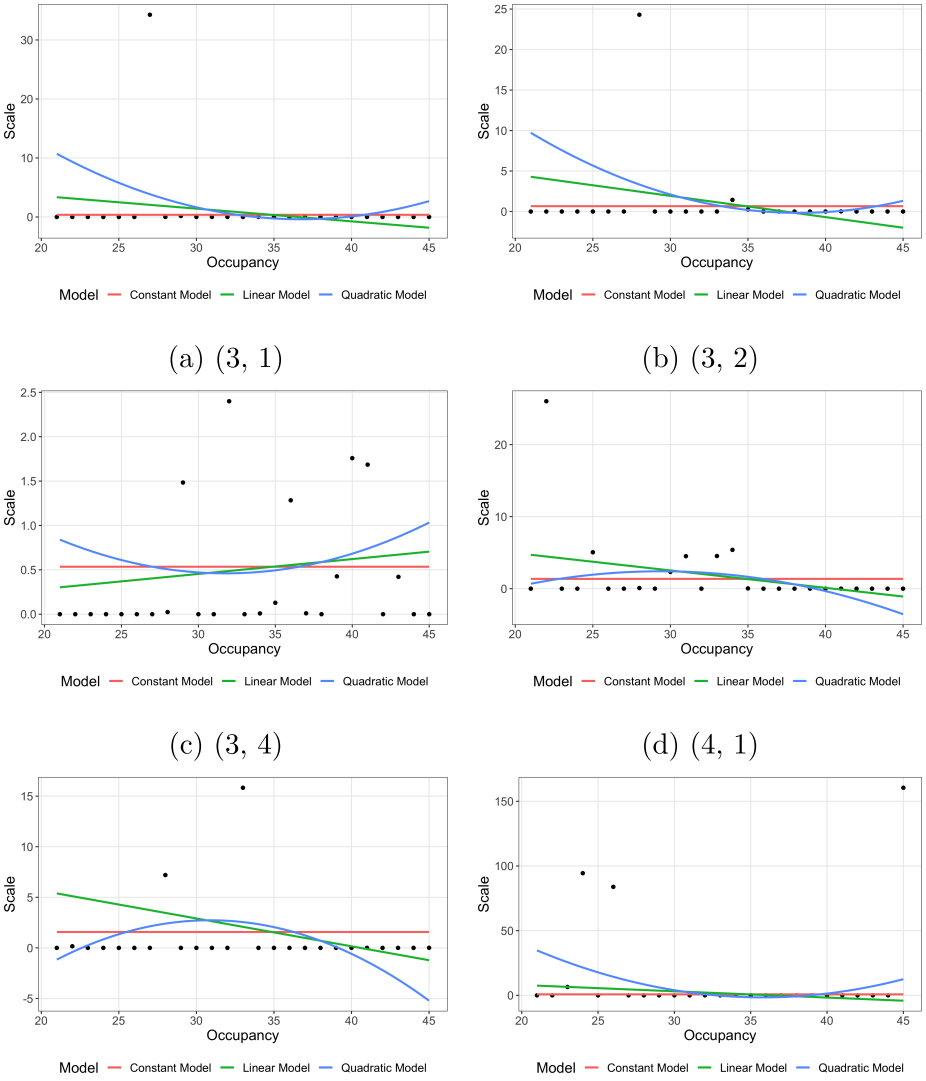
(c) (1, 4)

(d) (2, 1)

(e) (2, 3)

(f) (2, 4)

Figure E.1.11: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent unchanged scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.



(a) (3, 1)

(b) (3, 2)

(c) (3, 4)

(d) (4, 1)

(e) (4, 2)

(f) (4, 3)

Figure E.1.12: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent unchanged scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 2)	Constant	Linear	7.114	1	7.648×10^{-3}
(1, 2)	Linear	Quadratic	4.136	1	4.198×10^{-2}
(1, 3)	Constant	Linear	0.638	1	0.424
(1, 3)	Linear	Quadratic	1.630	1	0.202
(1, 4)	Constant	Linear	0.053	1	0.817
(1, 4)	Linear	Quadratic	0.219	1	0.640
(2, 1)	Constant	Linear	1.029	1	0.310
(2, 1)	Linear	Quadratic	3.849	1	4.979×10^{-2}
(2, 3)	Constant	Linear	0.000	1	0.995
(2, 3)	Linear	Quadratic	1.108	1	0.293
(2, 4)	Constant	Linear	0.121	1	0.728
(2, 4)	Linear	Quadratic	0.461	1	0.497
(3, 1)	Constant	Linear	1.293	1	0.255
(3, 1)	Linear	Quadratic	1.449	1	0.229
(3, 2)	Constant	Linear	2.029	1	0.154
(3, 2)	Linear	Quadratic	0.838	1	0.360
(3, 4)	Constant	Linear	0.150	1	0.698
(3, 4)	Linear	Quadratic	0.141	1	0.707
(4, 1)	Constant	Linear	4.226	1	3.981×10^{-2}
(4, 1)	Linear	Quadratic	1.183	1	0.277
(4, 2)	Constant	Linear	1.214	1	0.271
(4, 2)	Linear	Quadratic	0.646	1	0.422
(4, 3)	Constant	Linear	1.157	1	0.282
(4, 3)	Linear	Quadratic	3.636	1	5.653×10^{-2}

Table E.1.10: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with within level changes for the four-phase structured QBD process with level and phase transition dependent scales.

From the results of the likelihood-ratio tests presented in Table E.1.10, only the level and phase transition dependent scales associated with the (1, 2) and (2, 1) phase transitions require a quadratic term, and only the level and phase transition dependent scales associated with the (4, 1) phase transition requires a linear term. All remaining phase transitions have level and phase transition dependent scales that remain relatively constant.

Similar to before, we further explored the behaviour of level and phase transition dependent scales when values above 50 were removed. In this case, only the level and phase transition dependent scales associated with the (2, 1) phase transition behaved differently. That is, the suggested behaviour of the level and phase transition dependent scales associated with the (2, 1) phase transition changed from quadratic to constant, as demonstrated in Figure E.1.13 and Table E.1.11.

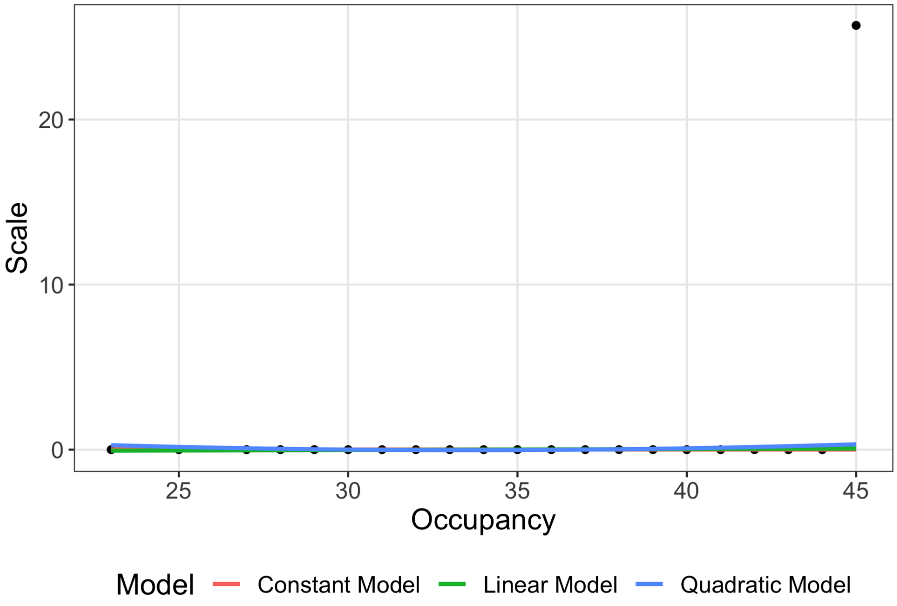
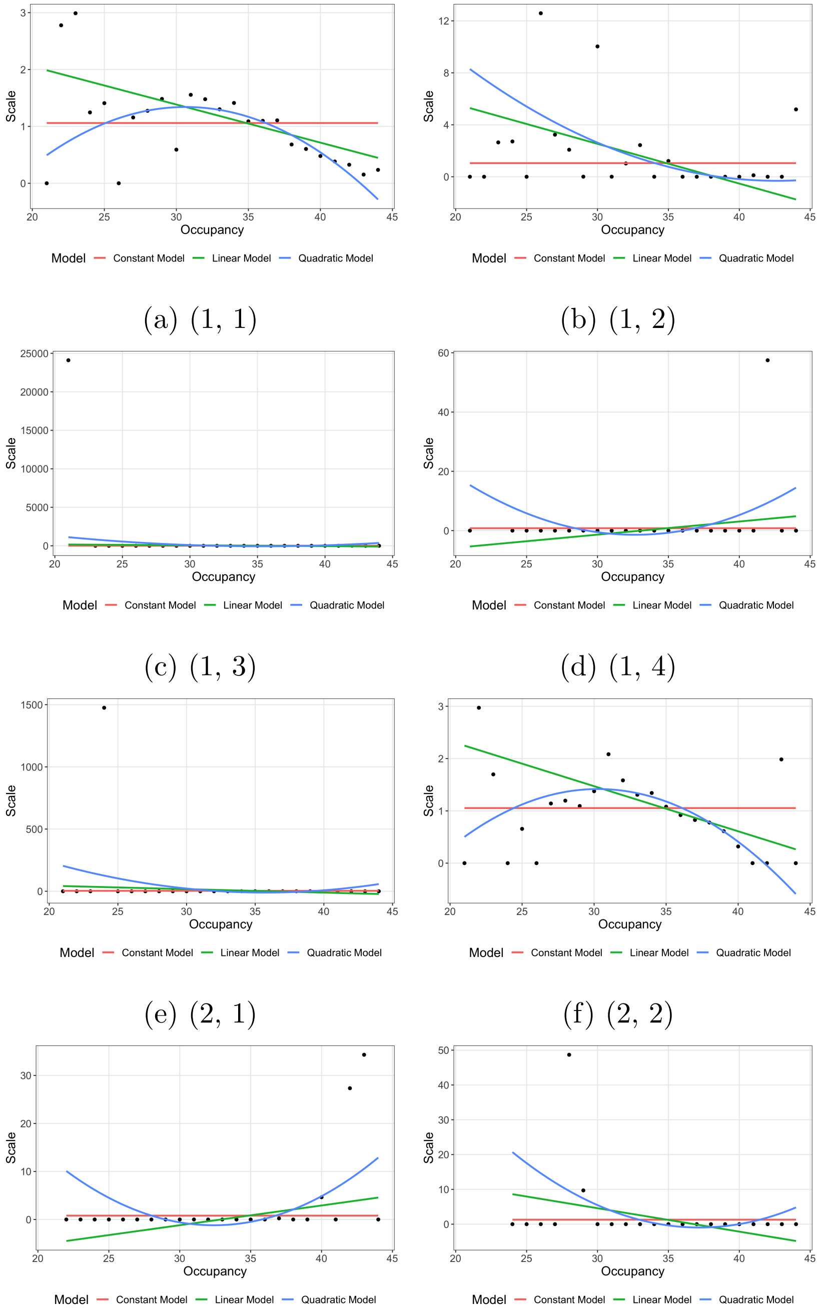


Figure E.1.13: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales associated with the (2, 1) phase transition and a value below 50 for the fitted four-phase structured QBD process.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(2, 1)	Constant	Linear	0.049	1	0.826
(2, 1)	Linear	Quadratic	0.211	1	0.646

Table E.1.11: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with within level changes, the (2, 1) phase transition, and a value below 50 for the four-phase structured QBD process with level and phase transition dependent scales.

Finally, using weighted least squares regression, we fitted a constant, linear, and quadratic model to the maximum likelihood estimates of the level and phase transition dependent scales associated with an increase in level of the fitted four-phase structured QBD process, as illustrated in Figures E.1.14 and E.1.15. Note that zero valued level and phase transition dependent scales are omitted in the case where the corresponding block matrix element is also zero.



(a) (1, 1)

(b) (1, 2)

(c) (1, 3)

(d) (1, 4)

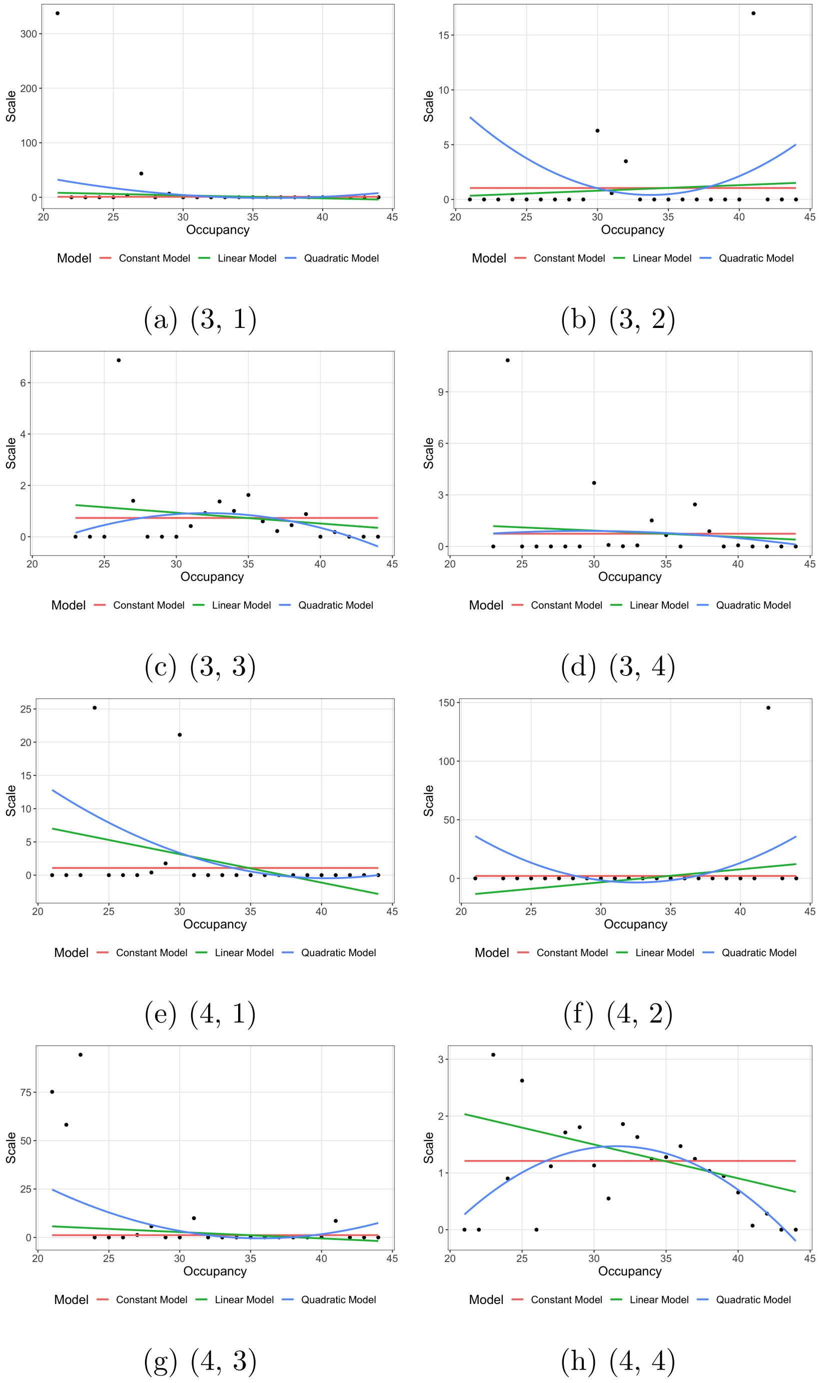
(e) (2, 1)

(f) (2, 2)

(g) (2, 3)

(h) (2, 4)

Figure E.1.14: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent arrival scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.



(a) (3, 1)

(b) (3, 2)

(c) (3, 3)

(d) (3, 4)

(e) (4, 1)

(f) (4, 2)

(g) (4, 3)

(h) (4, 4)

Figure E.1.15: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of selected level and phase transition dependent arrival scales for the fitted four-phase structured QBD process. Note that (a, b) indicates a phase transition from phase a to phase b.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(1, 1)	Constant	Linear	12.587	1	3.884×10^{-4}
(1, 1)	Linear	Quadratic	8.909	1	2.837×10^{-3}
(1, 2)	Constant	Linear	6.208	1	1.272×10^{-2}
(1, 2)	Linear	Quadratic	0.625	1	0.429
(1, 3)	Constant	Linear	0.145	1	0.704
(1, 3)	Linear	Quadratic	0.944	1	0.331
(1, 4)	Constant	Linear	1.261	1	0.261
(1, 4)	Linear	Quadratic	2.349	1	0.125
(2, 1)	Constant	Linear	0.491	1	0.483
(2, 1)	Linear	Quadratic	1.593	1	0.207
(2, 2)	Constant	Linear	14.213	1	1.632×10^{-4}
(2, 2)	Linear	Quadratic	8.518	1	3.516×10^{-3}
(2, 3)	Constant	Linear	3.196	1	7.382×10^{-2}
(2, 3)	Linear	Quadratic	5.932	1	1.487×10^{-2}
(2, 4)	Constant	Linear	2.634	1	0.105
(2, 4)	Linear	Quadratic	2.158	1	0.142
(3, 1)	Constant	Linear	1.320	1	0.251
(3, 1)	Linear	Quadratic	2.603	1	0.107
(3, 2)	Constant	Linear	0.085	1	0.770
(3, 2)	Linear	Quadratic	1.586	1	0.208
(3, 3)	Constant	Linear	1.038	1	0.308
(3, 3)	Linear	Quadratic	1.174	1	0.279
(3, 4)	Constant	Linear	0.316	1	0.574
(3, 4)	Linear	Quadratic	0.068	1	0.794
(4, 1)	Constant	Linear	3.004	1	8.307×10^{-2}
(4, 1)	Linear	Quadratic	0.540	1	0.462
(4, 2)	Constant	Linear	1.304	1	0.253
(4, 2)	Linear	Quadratic	2.389	1	0.122
(4, 3)	Constant	Linear	1.394	1	0.238
(4, 3)	Linear	Quadratic	4.787	1	2.868×10^{-2}
(4, 4)	Constant	Linear	6.383	1	1.152×10^{-2}
(4, 4)	Linear	Quadratic	6.676	1	9.775×10^{-3}

Table E.1.12: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals for the four-phase structured QBD process with level and phase transition dependent scales.

Firstly, the results of the likelihood-ratio tests presented in Table E.1.12 suggest that a quadratic form is needed to explain the behaviour of the level and phase transition dependent scales associated with the (1, 1), (2, 2), (2, 3), (4, 3), and $(4, 4)$ phase transitions. These results also suggest that only a linear term is required to describe the level and phase transition dependent scales associated with the $(1, 2)$ phase transition, and neither a linear or quadratic term is required for the remaining phase transitions.

Further exploration into the behaviour of level and phase transition dependent scales when outliers were removed revealed that only the (4, 3) phase transition behaved differently without the presence of outliers. That is, the suggested behaviour of the level and phase transition dependent scales associated with the (4, 3) phase transition changed from quadratic to constant, as demonstrated in Figure E.1.16 and Table E.1.13.

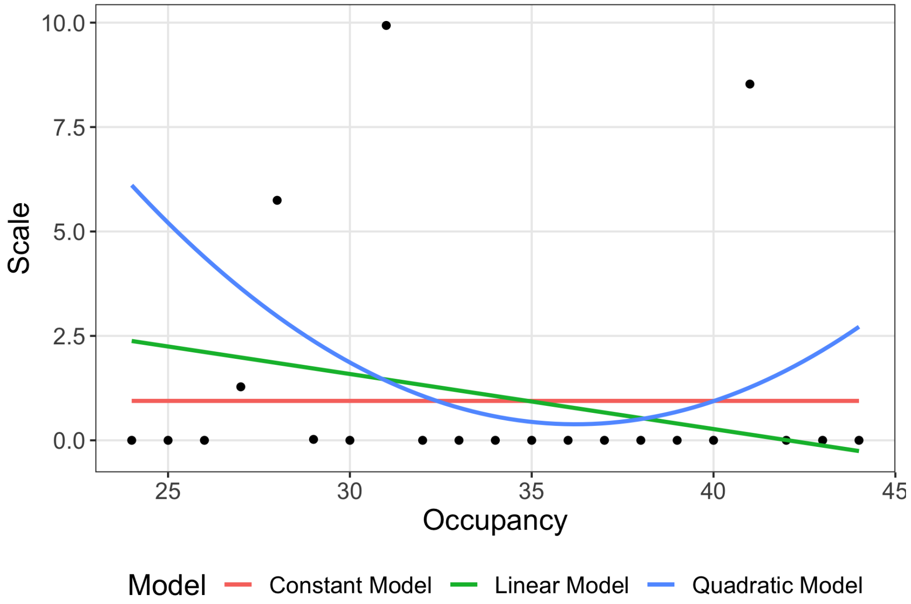


Figure E.1.16: Estimated constant model (red), linear model (green), and quadratic model (blue) of the maximum likelihood estimates of the level and phase transition dependent arrival scales associated with the (4, 3) phase transition and a value below 50 for the fitted four-phase structured QBD process.

Phase Transition	Model 1	Model 2	Statistic	Degrees of freedom	P-value
(4, 3)	Constant	Linear	0.630	1	0.428
(4, 3)	Linear	Quadratic	1.192	1	0.275

Table E.1.13: Summary of the likelihood-ratio tests which compare the fitted constant, linear, and quadratic models associated with arrivals, the (4, 3) phase transition, and a value below 50 for the four-phase structured QBD process with level and phase transition dependent scales.

E.2 Goodness of fit

In this section, we review the diagnostic plots associated with each goodness of fit test summarised in Table 10.3.1.

E.2.1 One-phase structured QBD process with level-dependent polynomial forms

First, we consider the one-phase structured QBD process with level-dependent polynomial forms, such that the null and alternative hypotheses are defined as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all one-phase structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

such that the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

Using the hypothesis test described in Section 9.2, we obtain an empirical p-value of 0 which suggests that this particular type of QBD process does not capture the stationary and transient behaviour observed in the RAH ICU data. Let's now consider the diagnostic plots for the stationary and transient behaviour.

Figure E.2.1 illustrates that the fitted one-phase structured QBD process over-estimates the proportion of time spent in bed occupancies 12 to 18, and generally under-estimates the proportion of time spent in the remaining bed occupancies.

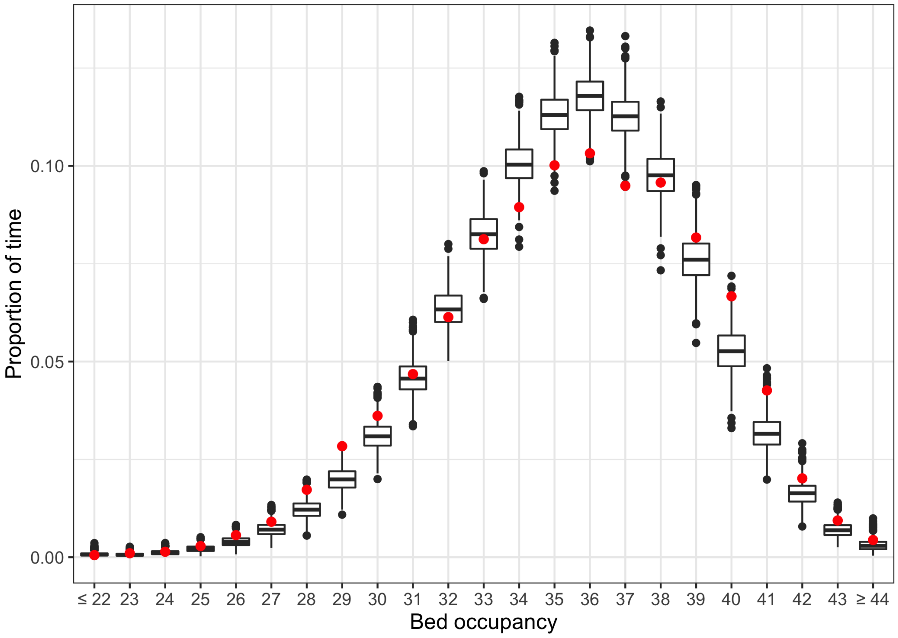


Figure E.2.1: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted one-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite not accurately capturing the stationary behaviour, the transition probabilities of the fitted one-phase structured QBD process are similar to those observed in the RAH ICU data set, as demonstrated in Figure E.2.2. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

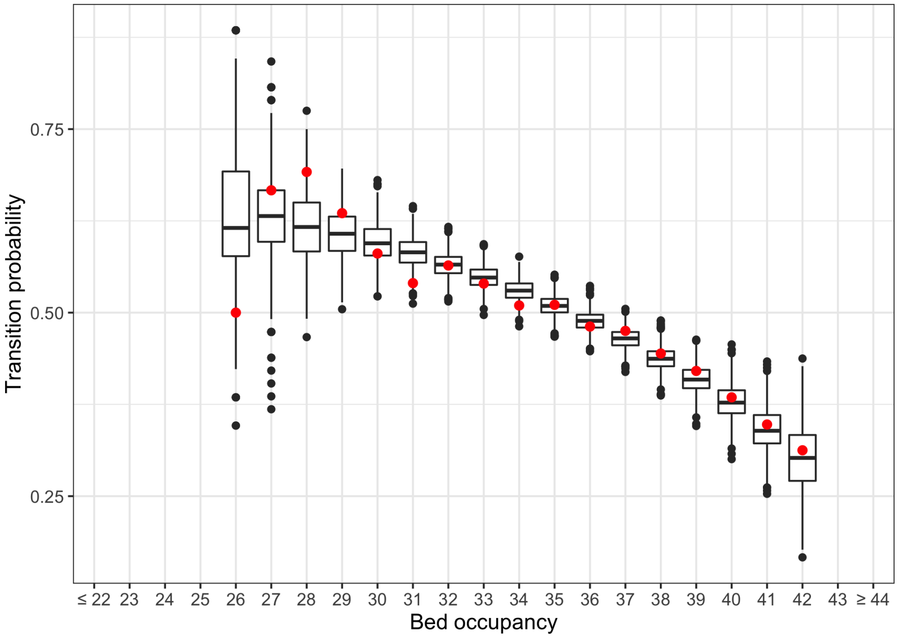
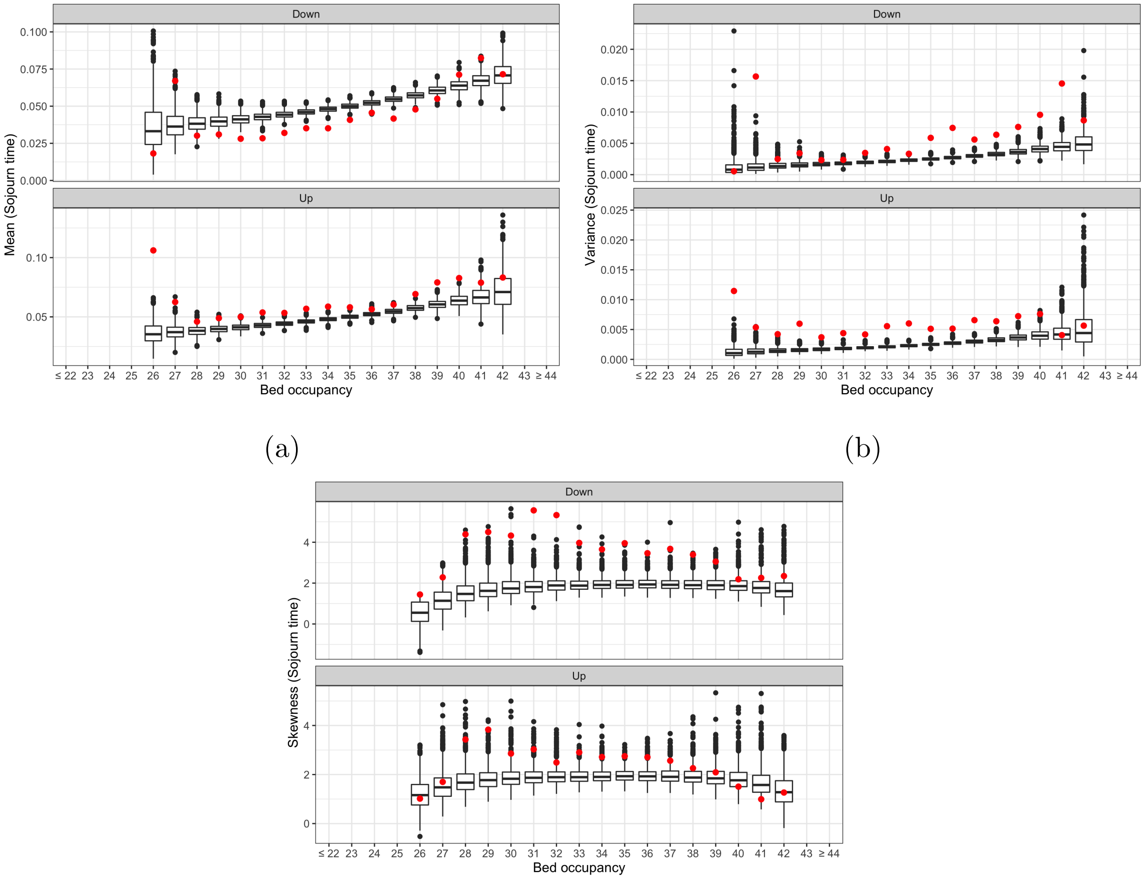


Figure E.2.2: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted one-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 26 and above 42 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Figures E.2.3a, E.2.3b, and E.2.3c demonstrate that the distributions of conditional sojourn times are inaccurately represented by the fitted one-phase structured QBD process. Furthermore, Figures E.2.4 and E.2.5 illustrate the difference in the observed distributions of conditional sojourn times compared to those of the data simulated from the fitted one-phase structured QBD process, drawing attention to the differences observed in the downward transitions.

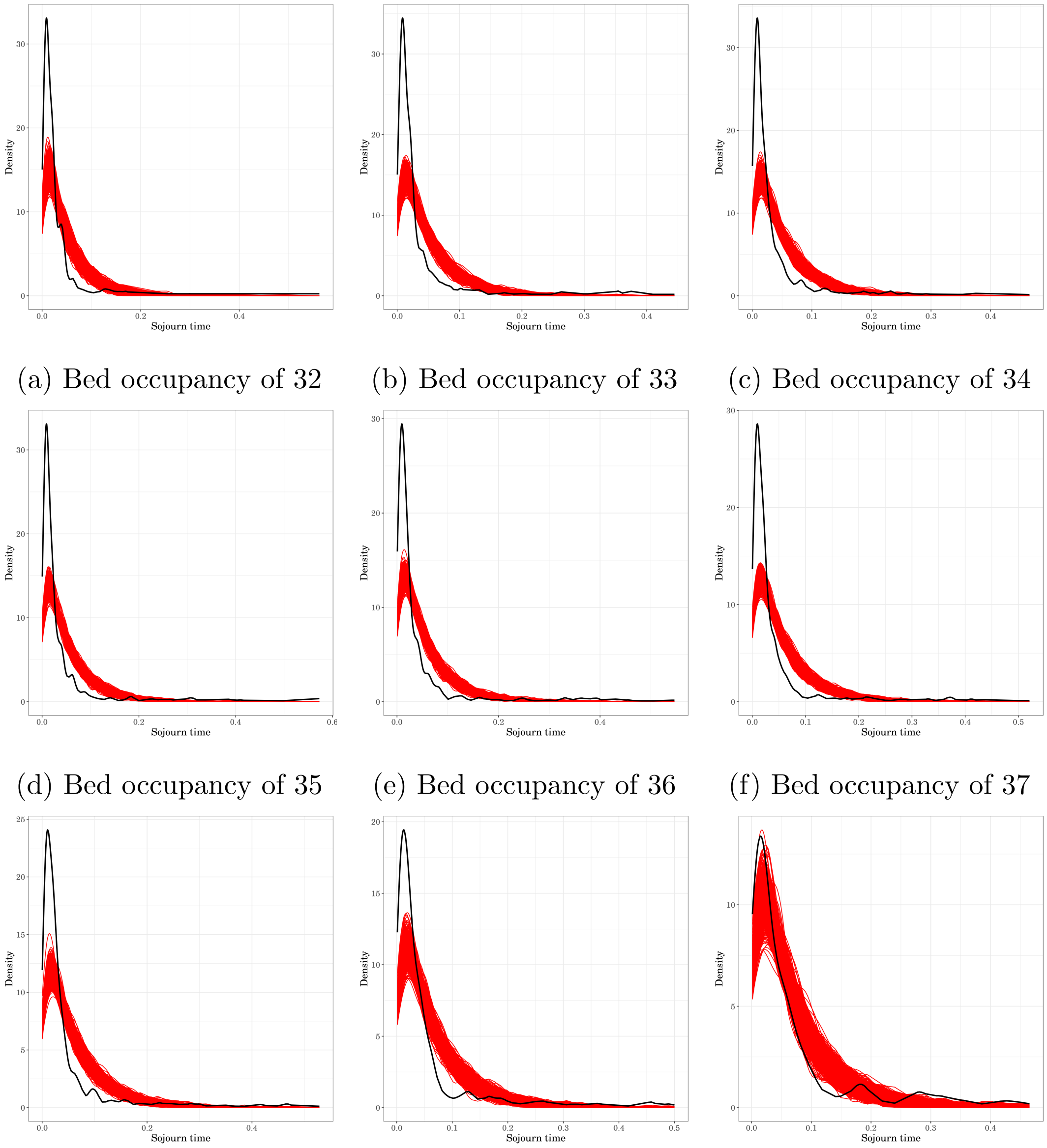


(a)

(b)

(c)

Figure E.2.3: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted level-dependent QBD process with two phases (box-plots). Note that the conditional sojourn times for bed occupancies below 26 and above 42 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

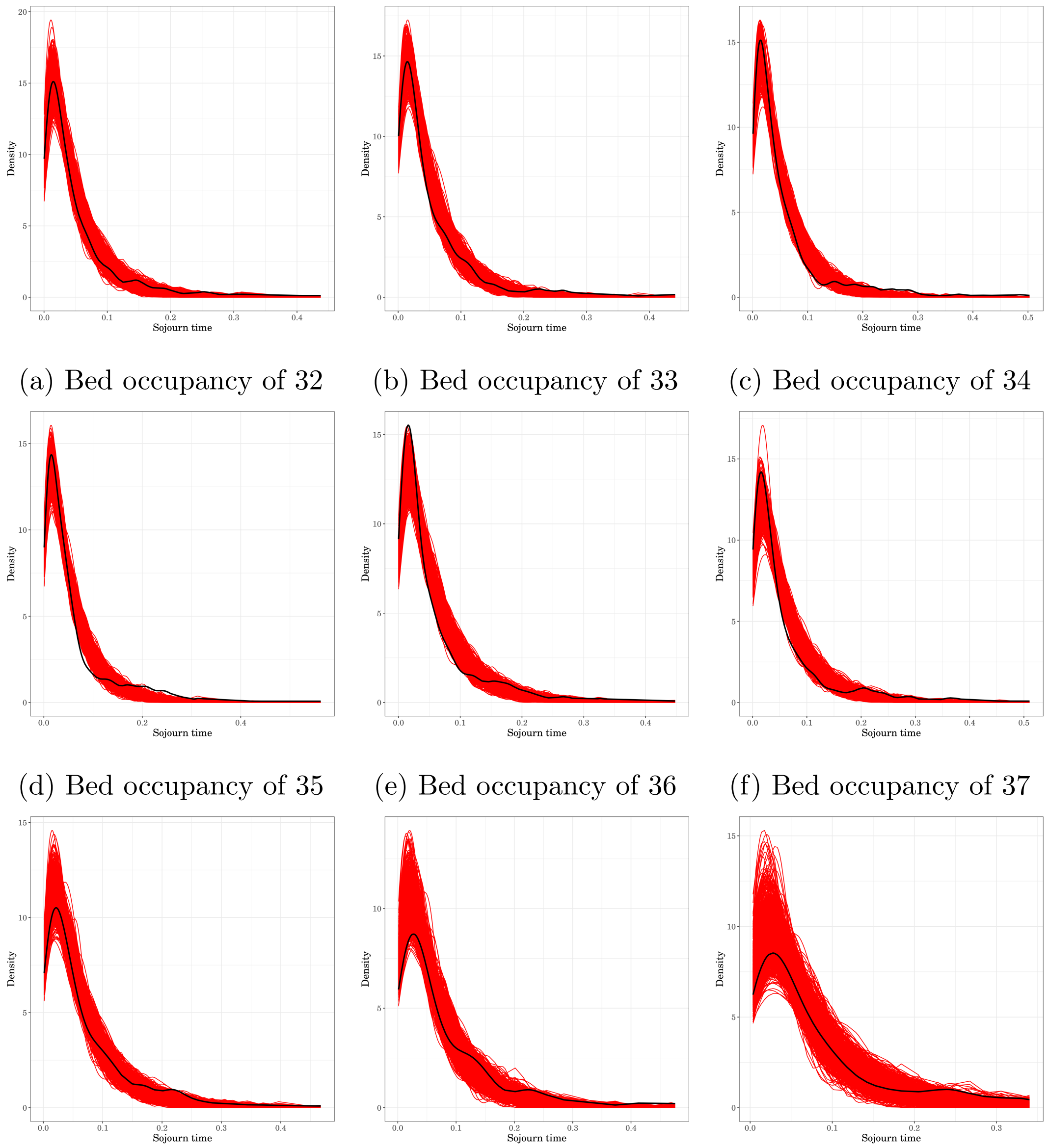


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.4: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted one-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.5: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted one-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

E.2.2 Two-phase structured QBD process with level-dependent polynomial forms

Here, we consider the two-phase structured QBD process, such that the null and alternative hypotheses are defined as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all two-phase structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1),$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell + \beta_2^g \ell^2,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell + \beta_2^f \ell^2.$$

Figure E.2.6 shows that despite introducing an extra phase into the structured QBD process, the proportion of time spent in the most frequently visited bed occupancies remains under-estimated.

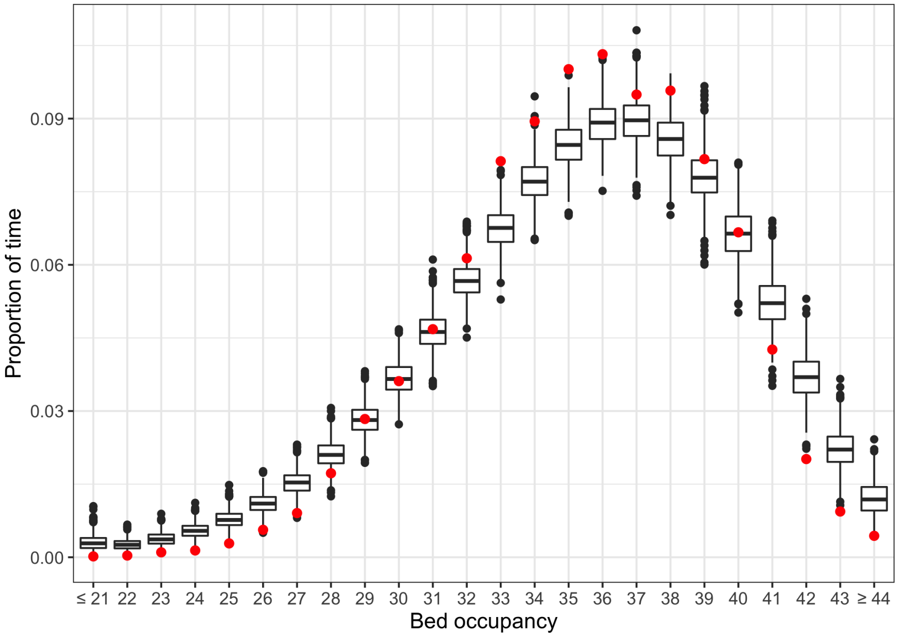


Figure E.2.6: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

There is less variability in the transition probabilities of the two-phase structured QBD process compared to those of the one-phase structured QBD process, but evidence of dissimilarity still remains, as shown in Figure E.2.7.

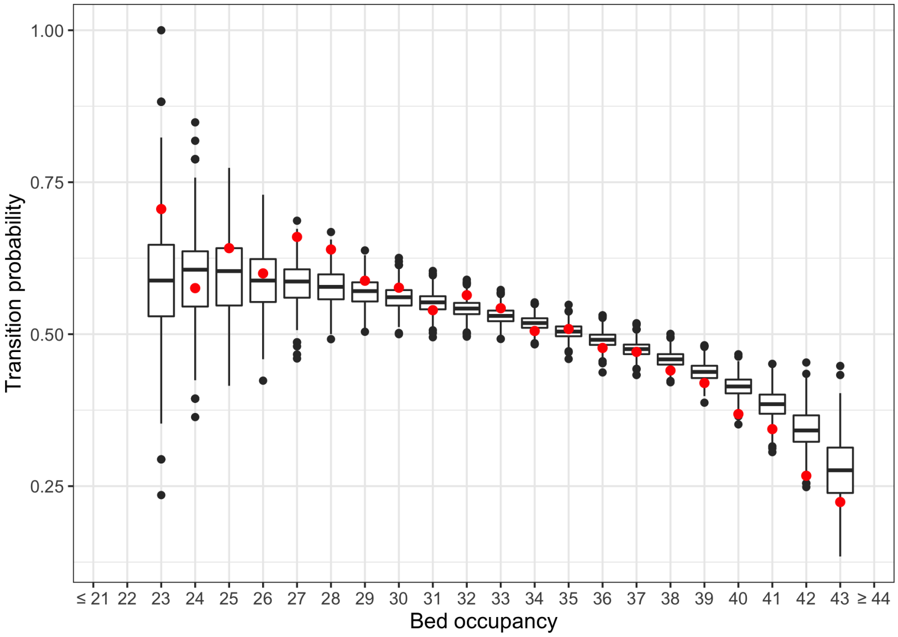
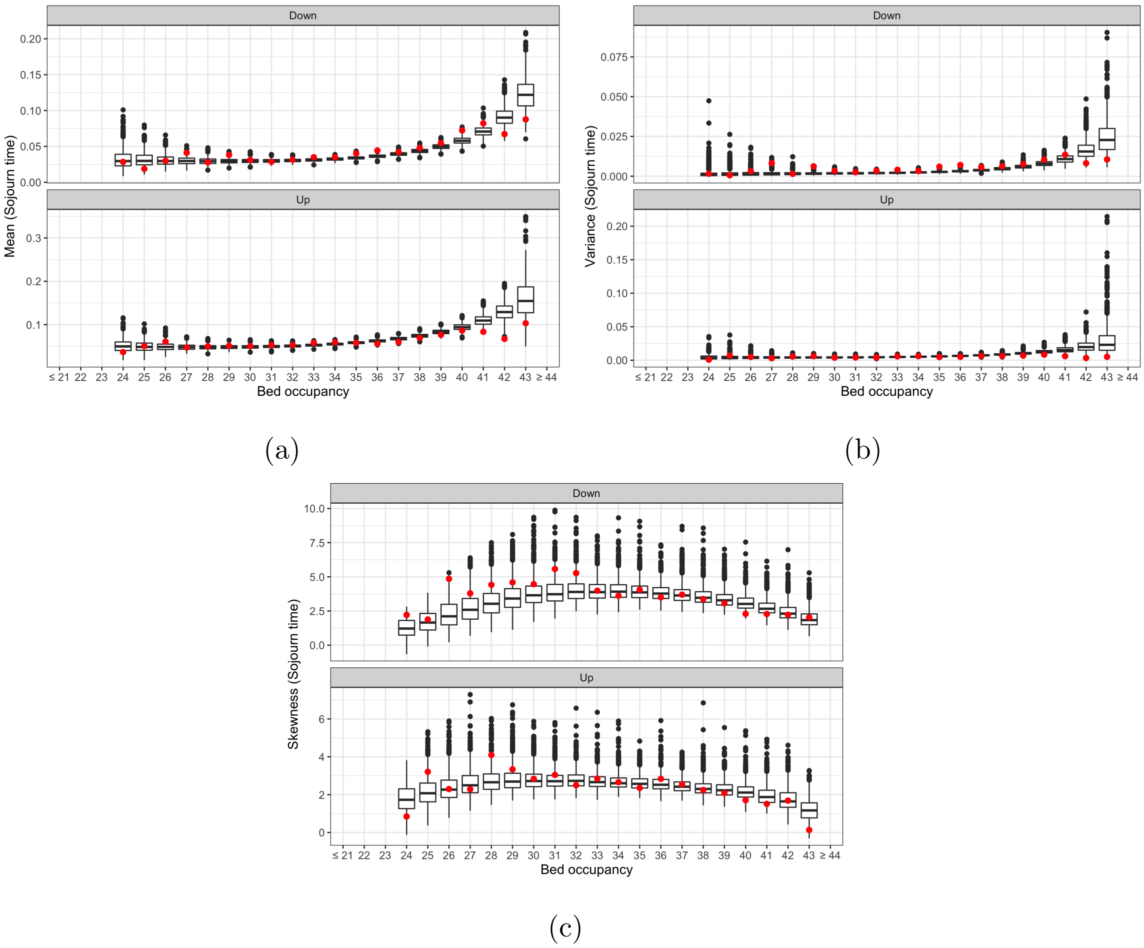


Figure E.2.7: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 23 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite the previous differences, the addition of an extra phase has improved the modelling of the conditional sojourn times, as illustrated in Figures E.2.8a, E.2.8b, and E.2.8c. However, dissimilarities remain as further illustrated in Figures E.2.9 and E.2.10.

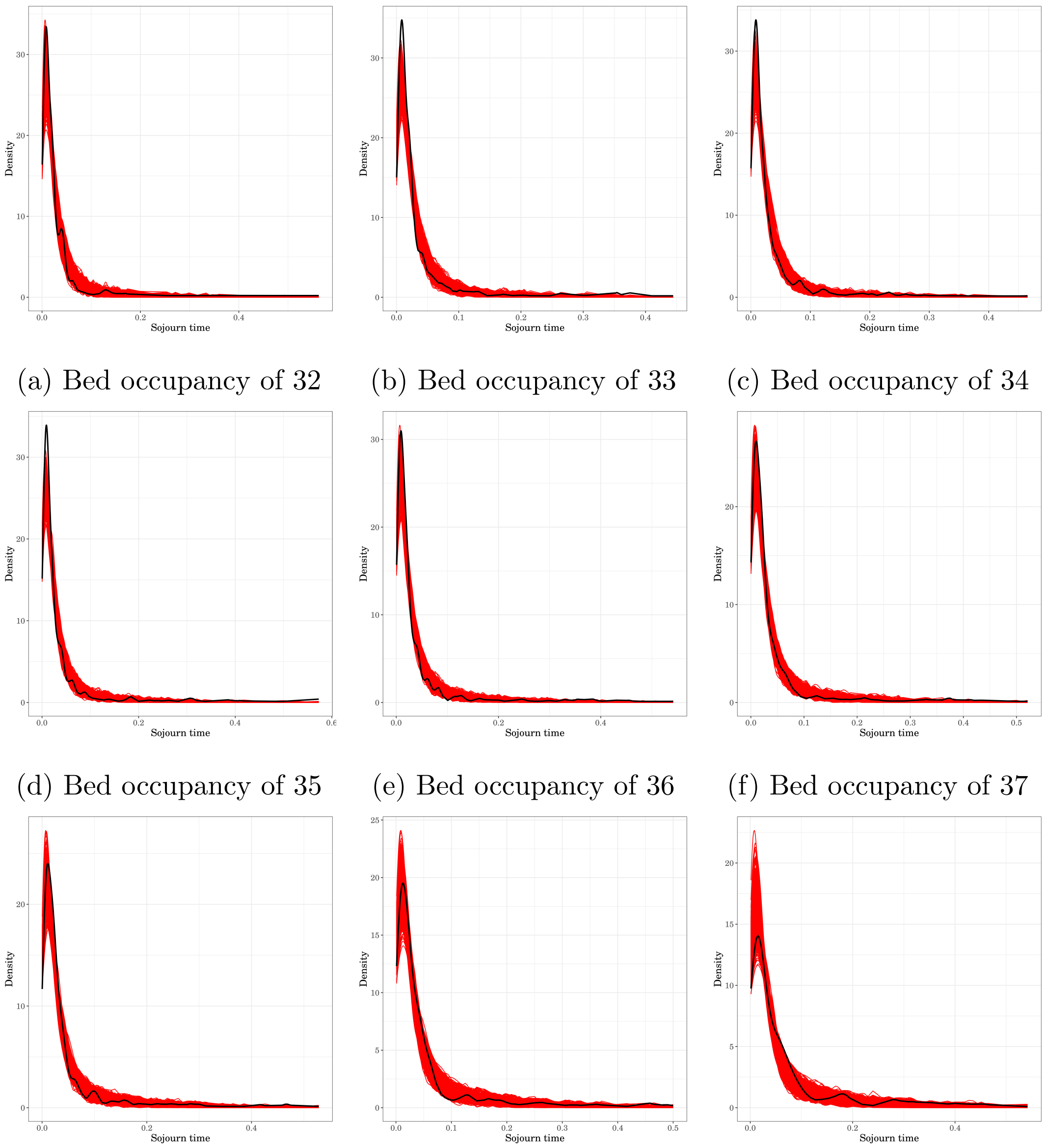


(a)

(b)

(c)

Figure E.2.8: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level-dependent polynomial forms (boxplots). Note that the conditional sojourn times for bed occupancies below 24 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

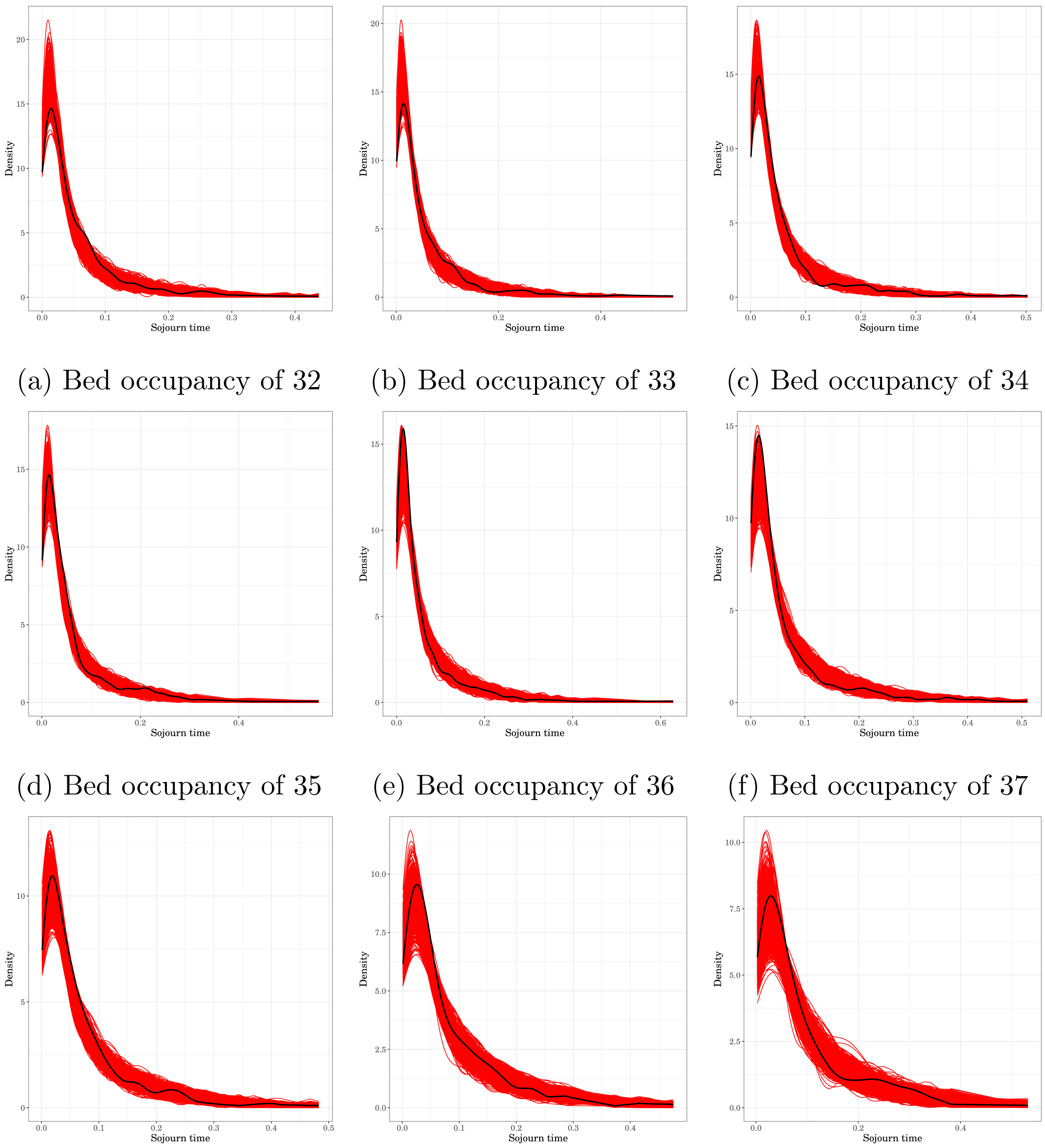


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.9: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.10: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

E.2.3 Three-phase structured QBD process with level-dependent polynomial forms

Next, we consider a three-phase structured QBD process which has an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2,$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell + \beta_2^g \ell^2,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

Hence, we define the null and alternative hypotheses as

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all structured QBD processes of this particular form.

Introducing another phase into the structured QBD process has further improved the fit in terms of stationary behaviour but there is still some under-estimation and over-estimation of bed occupancies, as shown in Figure E.2.11.

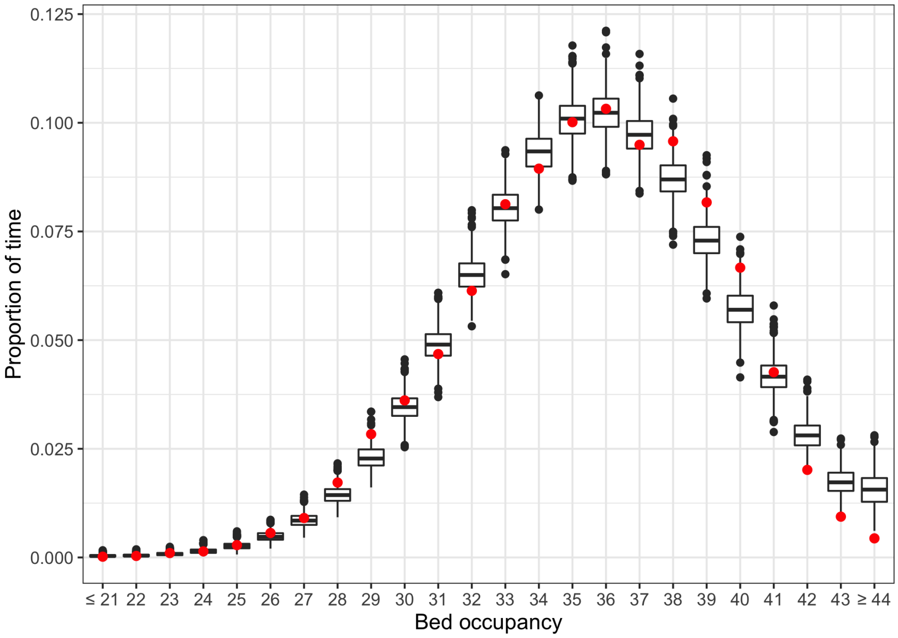


Figure E.2.11: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite seeing an improvement in terms of stationary behaviour, there remains some variability and evidence of dissimilarity in the transition probabilities of the fitted three-phase structured QBD process compared to that observed in the RAH ICU data set, as shown in Figure E.2.12. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

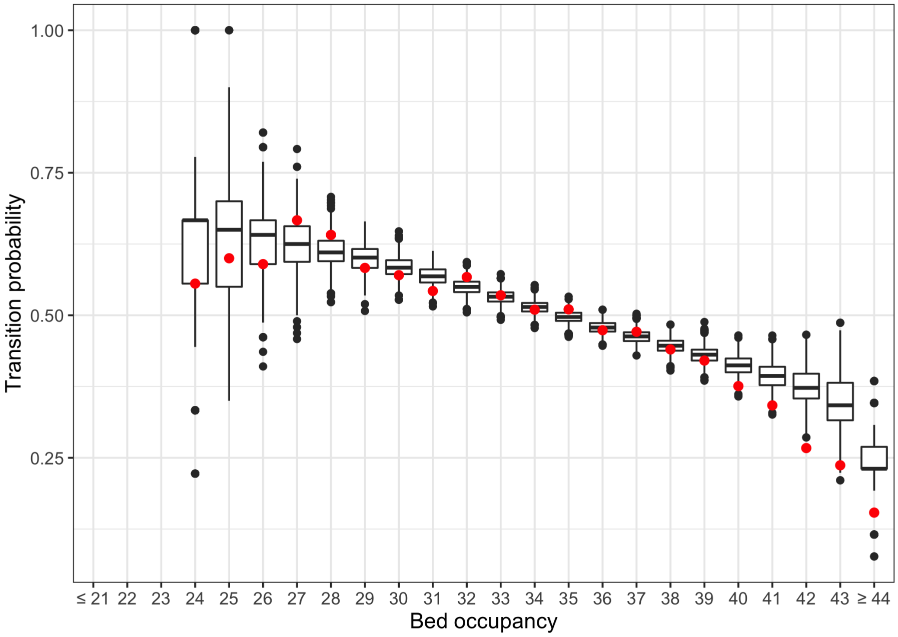
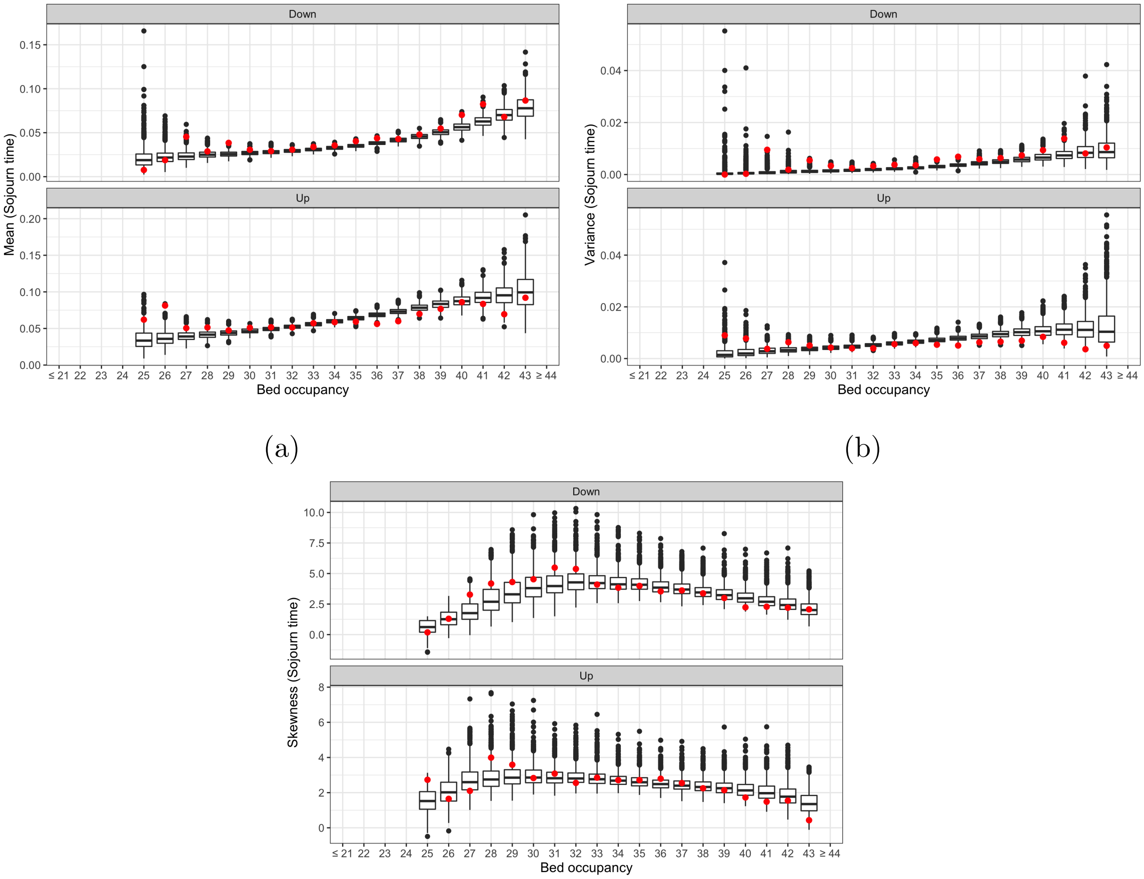


Figure E.2.12: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 24 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

The addition of another phase slightly improved the fit in terms of the distribution of sojourn time for each bed occupancy, as demonstrated in Figures E.2.13a, E.2.13b, and E.2.13c. However, dissimilarities still remain as further illustrated in Figures E.2.14 and E.2.15.

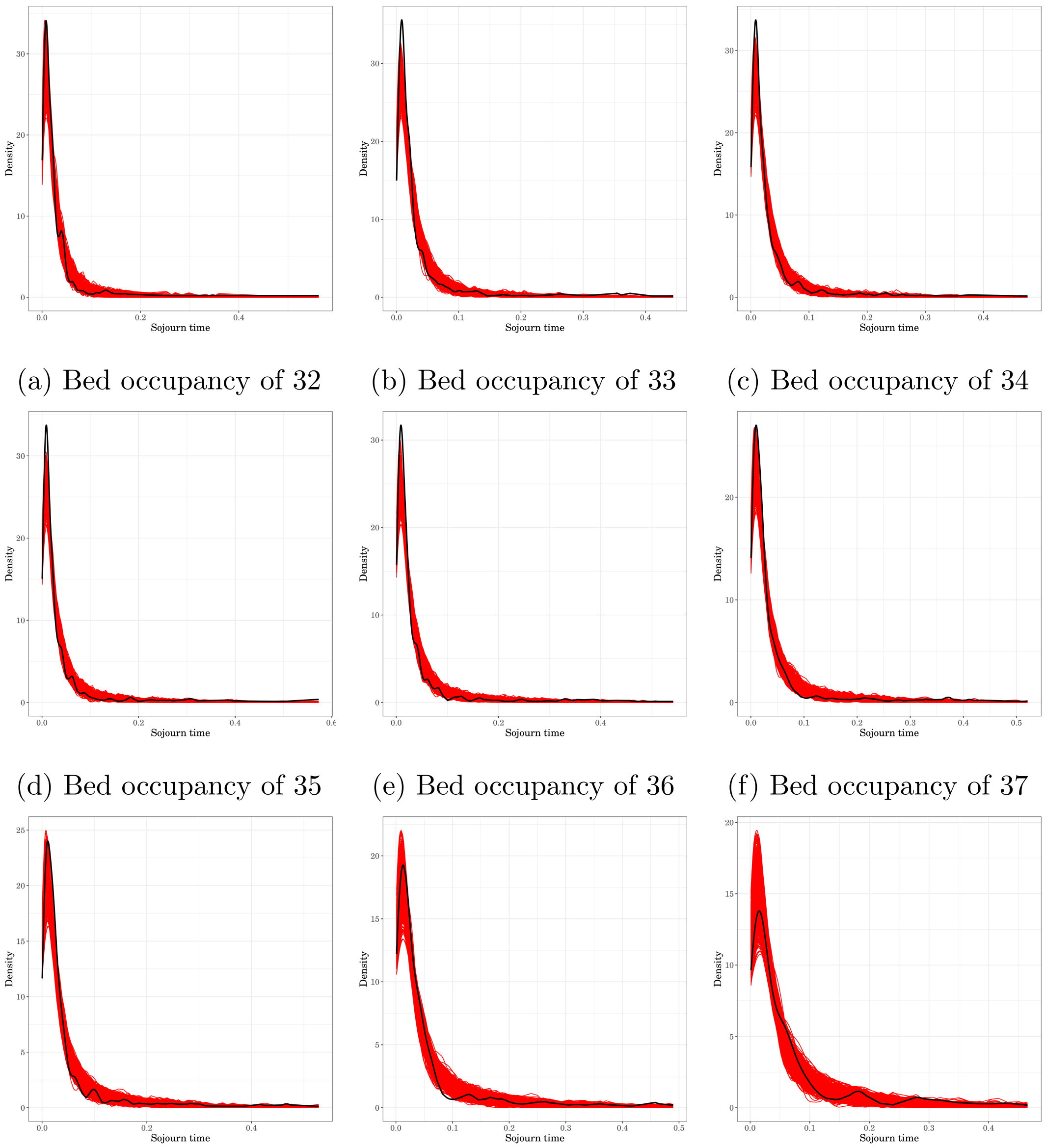


(a)

(b)

(c)

Figure E.2.13: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level-dependent polynomial forms (boxplots). Note that the conditional sojourn times for bed occupancies below 25 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

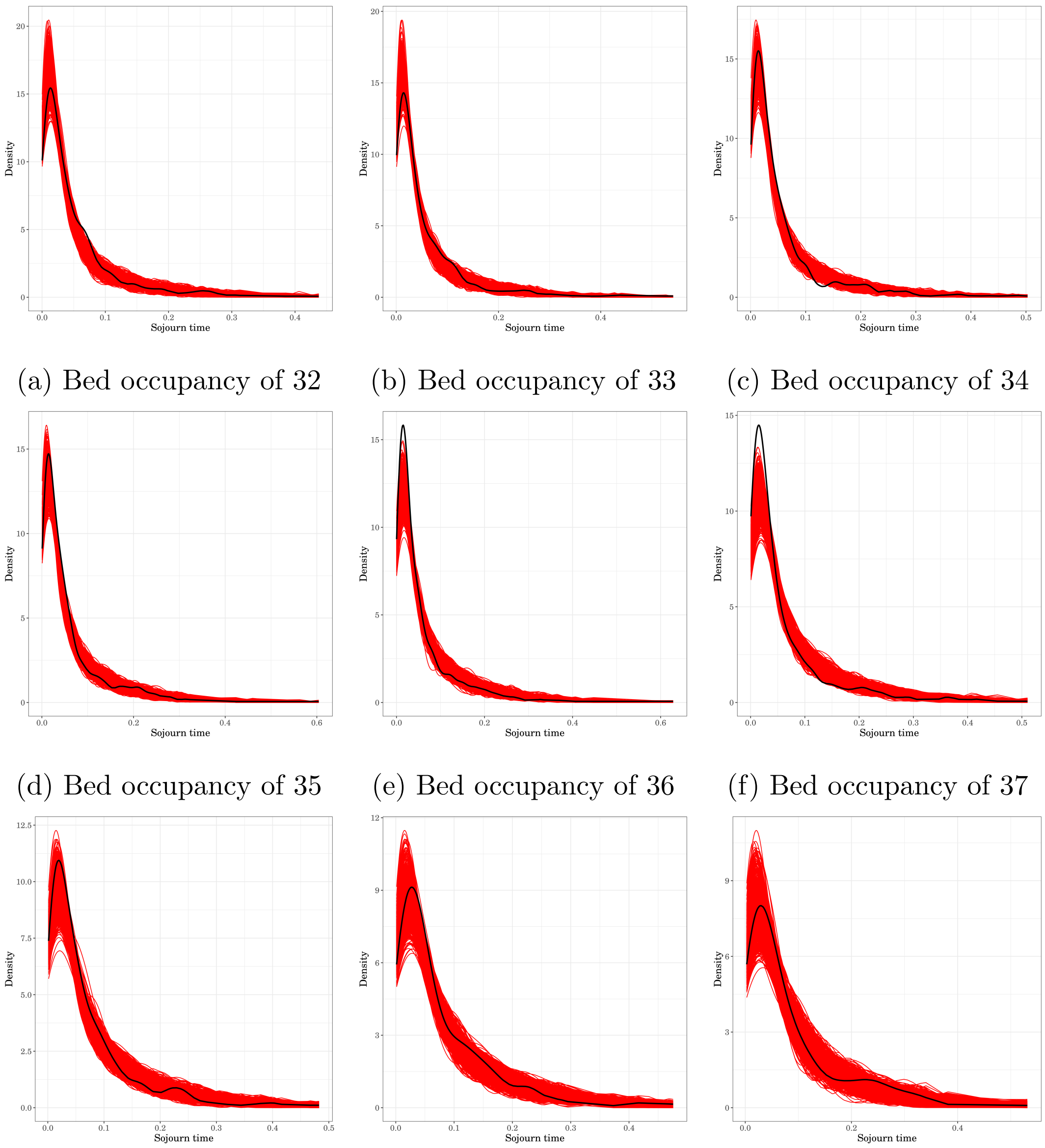


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.14: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.15: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

E.2.4 Four-phase structured QBD process with level-dependent polynomial forms

Now, let's consider a four-phase structured QBD process which has an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$h_\ell = 1 + \beta_1^h(\ell - 1) + \beta_2^h(\ell - 1)^2,$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$g_\ell = 1 + \beta_1^g \ell,$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$f_\ell = 1 + \beta_1^f \ell.$$

Hence, we define the null and alternative hypotheses as

$$H_0 : g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$
$$H_a : g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\},$$

where Ω is the set of all structured QBD processes of this particular form.

Once again, introducing another phase has improved the fit in terms of stationary behaviour but there is still some under-estimation and over-estimation of higher bed occupancies, as shown in Figure E.2.16.

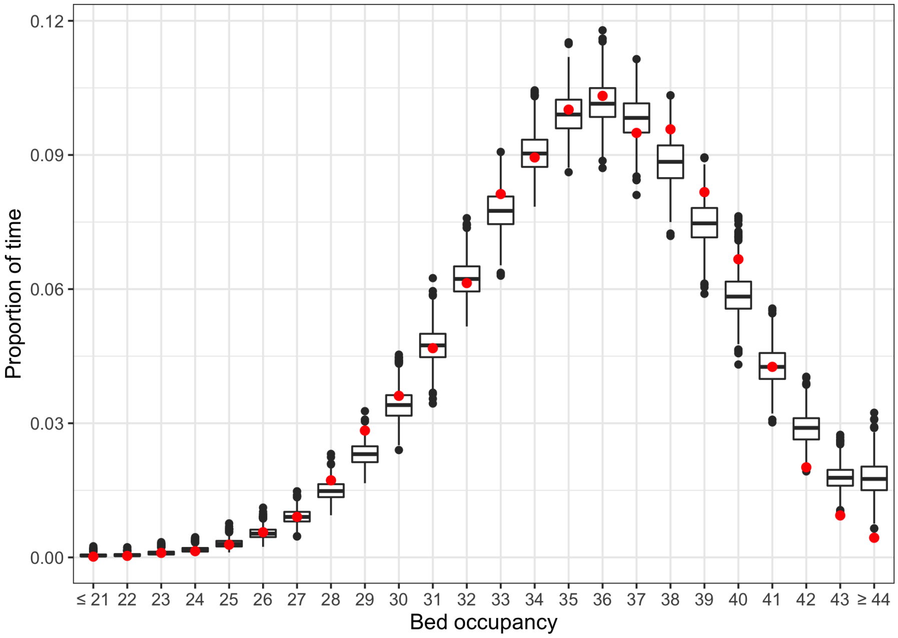


Figure E.2.16: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Similar to before, there remains some variability and evidence of dissimilarity in the transition probabilities of the fitted four-phase structured QBD process compared to that observed in the RAH ICU data set, as shown in Figure E.2.17. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

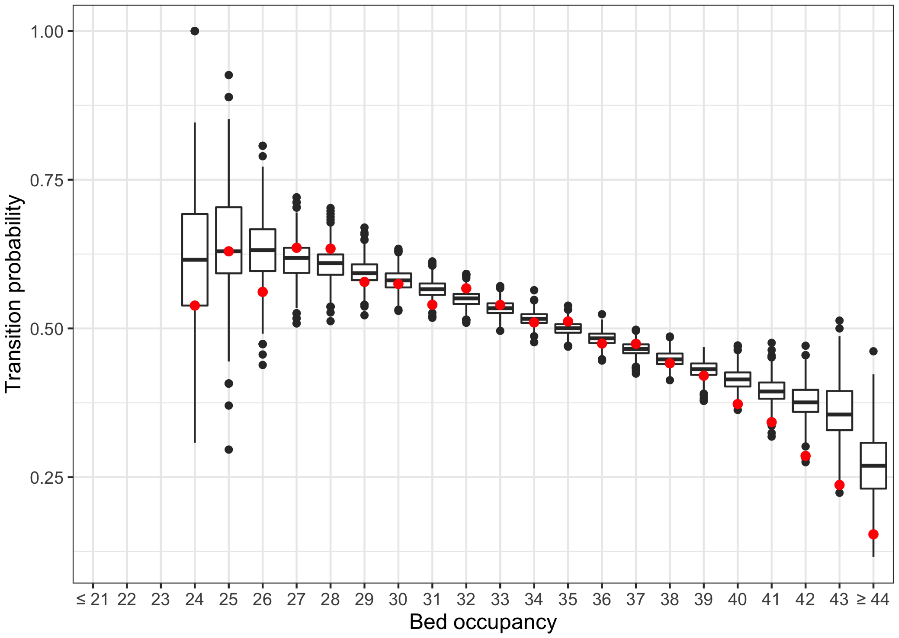
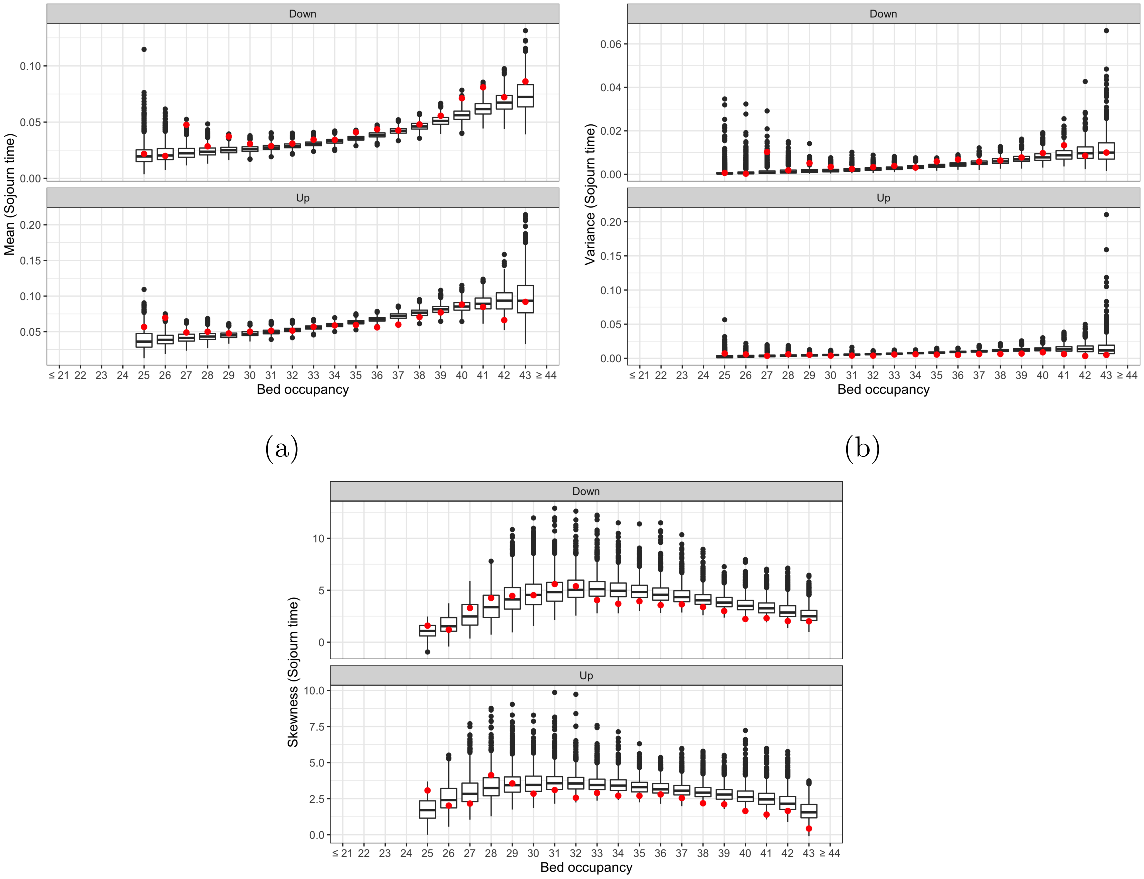


Figure E.2.17: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level-dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 24 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite the structured QBD process having four phases, dissimilarities still exist between the observed distributions of conditional sojourn times for each bed occupancy and the that of the fitted four-phase structured QBD process, as illustrated in Figures E.2.18a, E.2.18b, and E.2.18c, as well as in Figures E.2.19 and E.2.20.

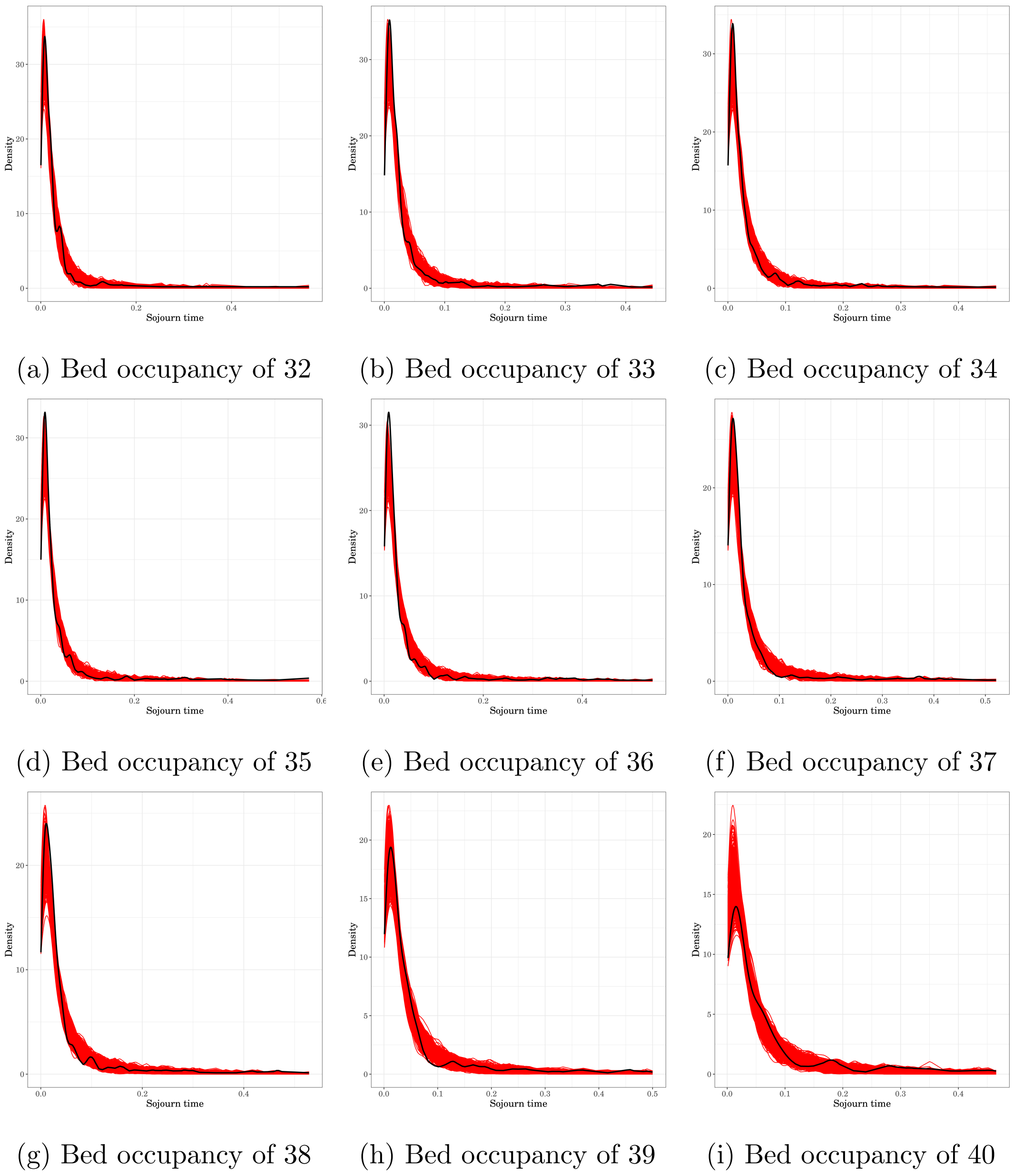


(a)

(b)

(c)

Figure E.2.18: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level-dependent polynomial forms (boxplots). Note that the conditional sojourn times for bed occupancies below 25 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

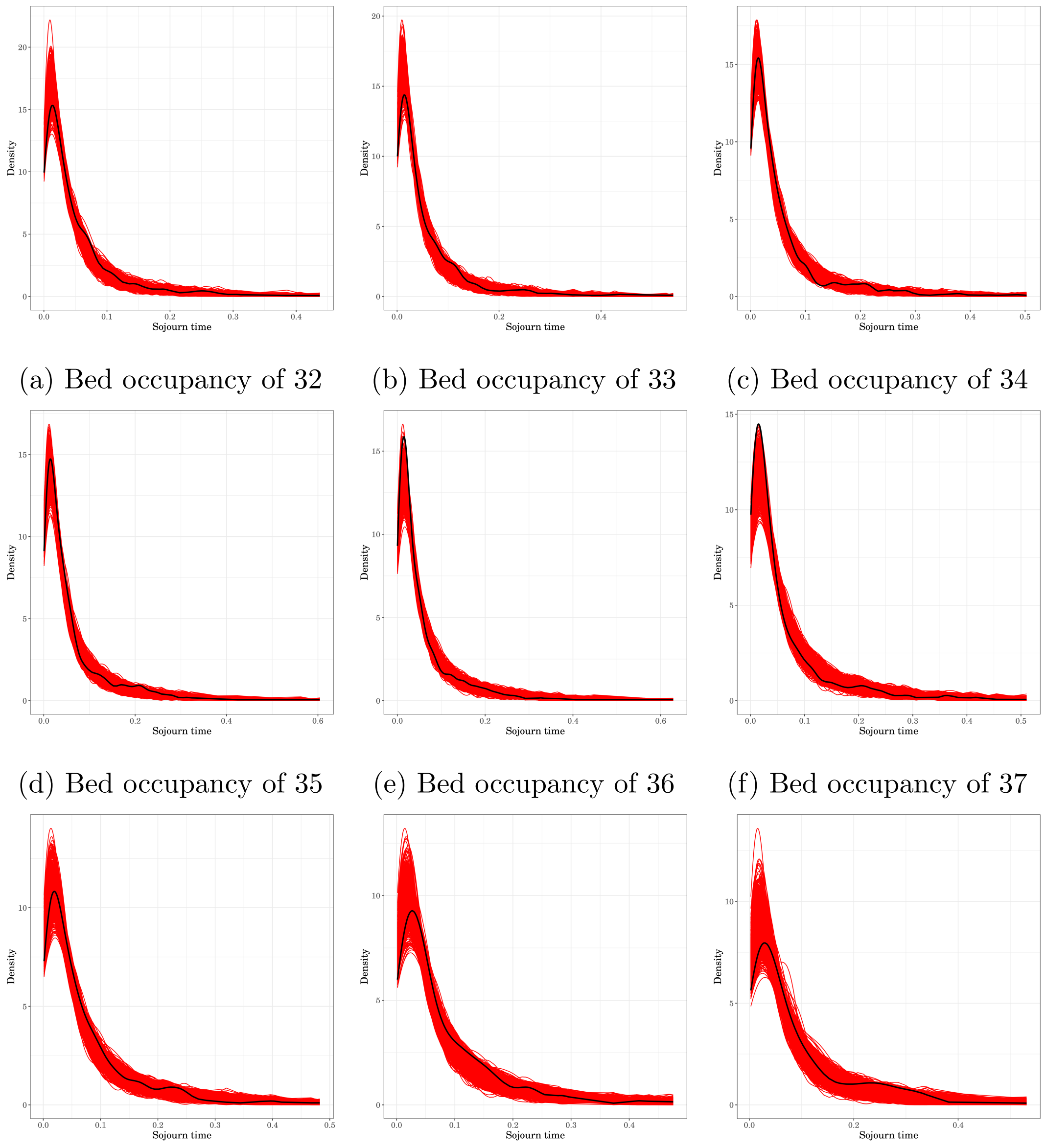


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.19: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.20: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted four-phase structured QBD process with level-dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

E.2.5 Two-phase structured QBD process with level and phase transition dependent polynomial forms

In this section, we now consider a two-phase structured QBD process, such that the null and alternative hypotheses are defined as

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all two-phase structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(2,2)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1), & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1) + \beta_2^{h,i,j}(\ell - 1)^2, & \text{if } (i,j) \in \{(1,1),(2,1)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(2,1)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(2,1)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1),(2,2)\}. \end{cases}$$

Figure E.2.21 illustrates that the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted two-phase structured QBD process is similar to the observed proportion of time spent in each bed occupancy from the RAH ICU data set.

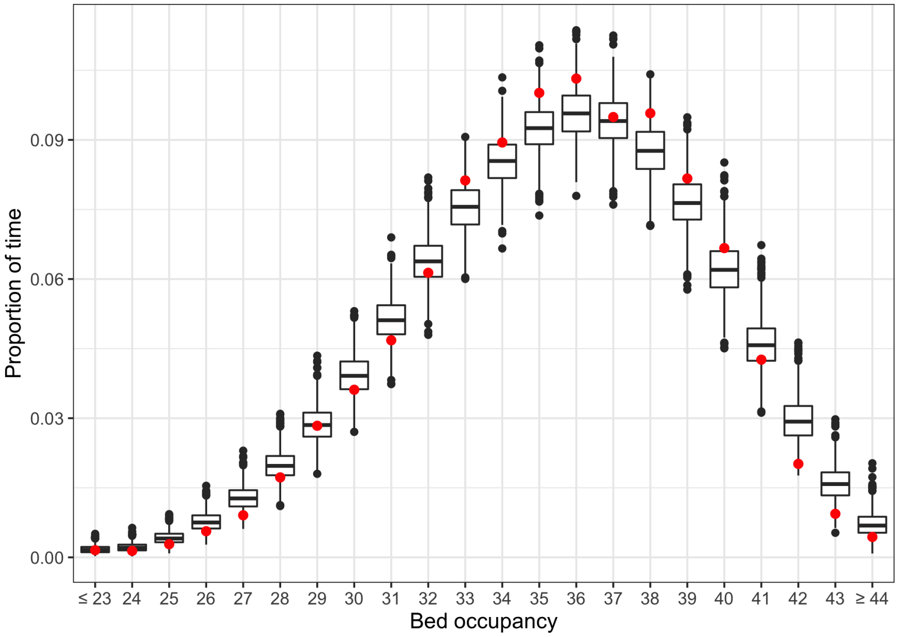


Figure E.2.21: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Figure E.2.22 shows that for the most frequently visited bed occupancies, the transition probabilities of the fitted two-phase structured QBD process are similar to that observed in the RAH ICU data set. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

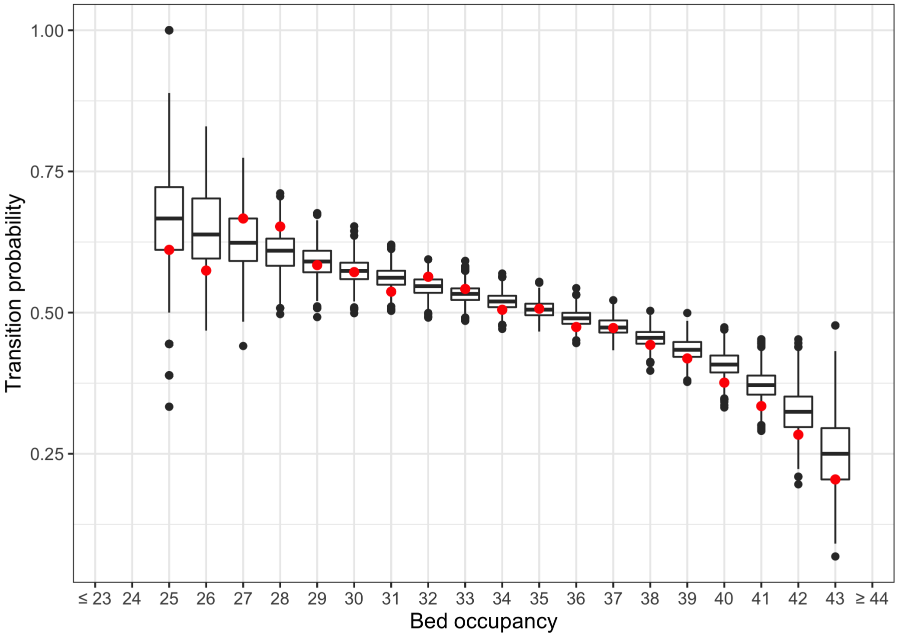
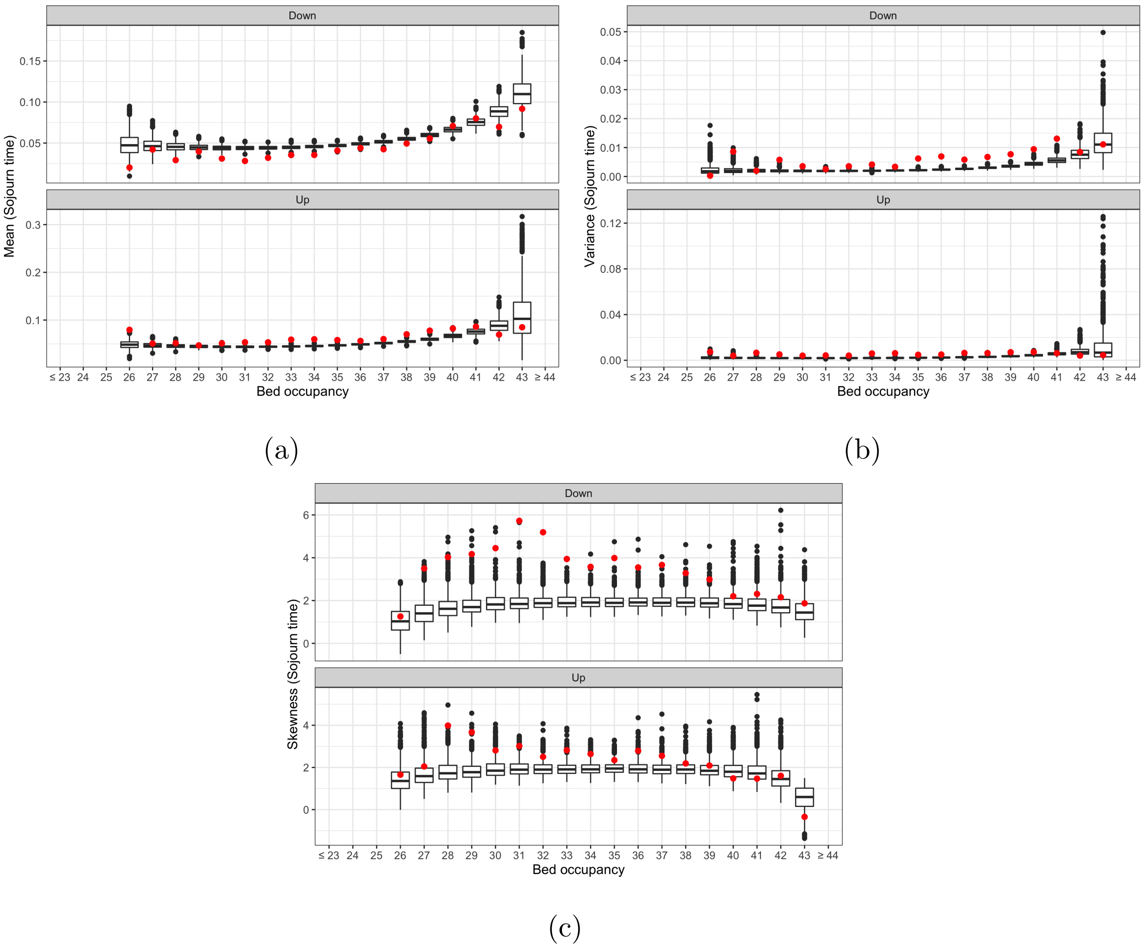


Figure E.2.22: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 25 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite observing positive results for the previous two behavioural components, the observed distributions of conditional sojourn times are incorrectly estimated by the fitted two-phase structured QBD process, as illustrated in Figures E.2.23a, E.2.23b, and E.2.23c. This result is further illustrated in Figures E.2.24 and E.2.25.

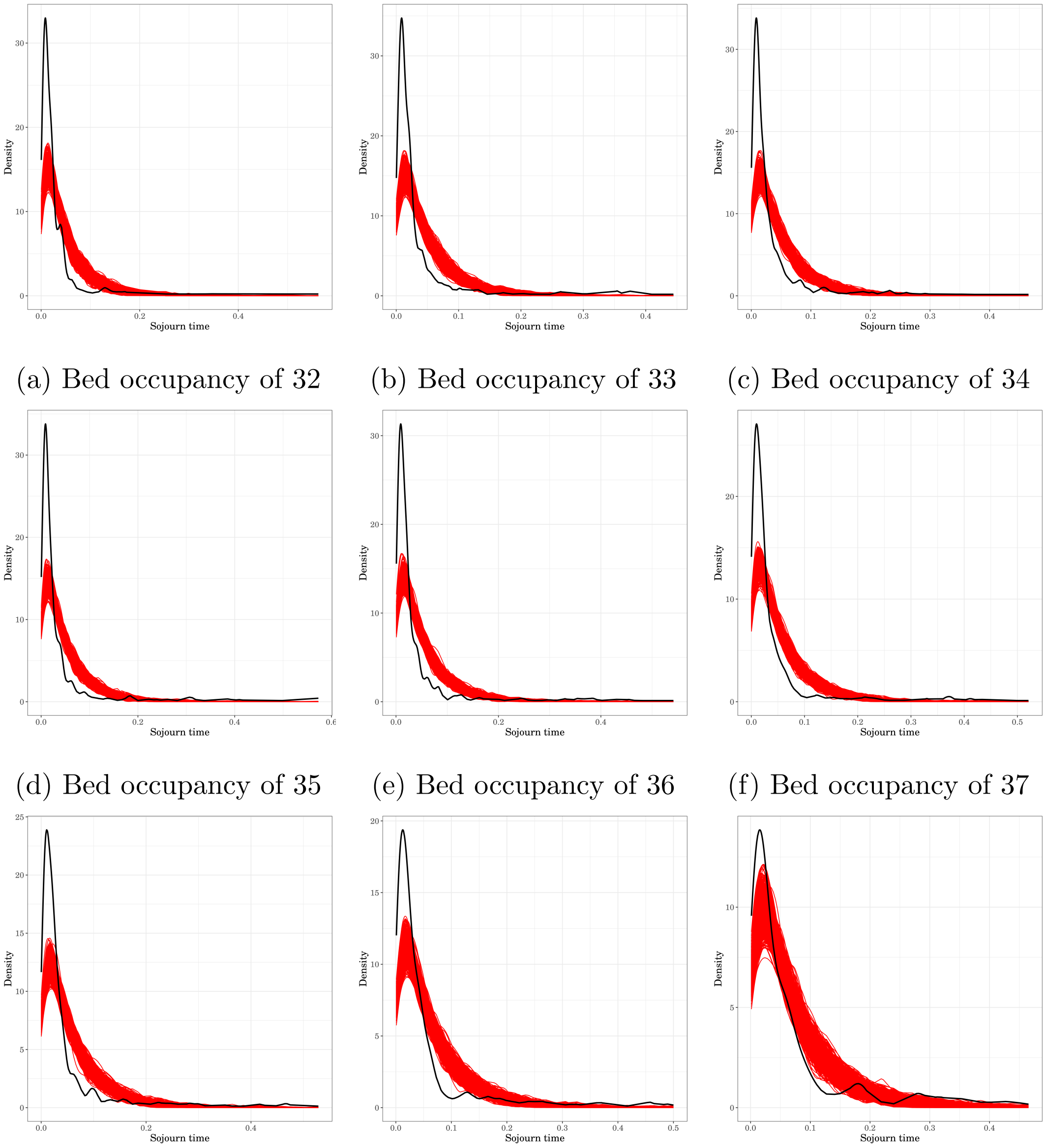


(a)

(b)

(c)

Figure E.2.23: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that the conditional sojourn times for bed occupancies below 26 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

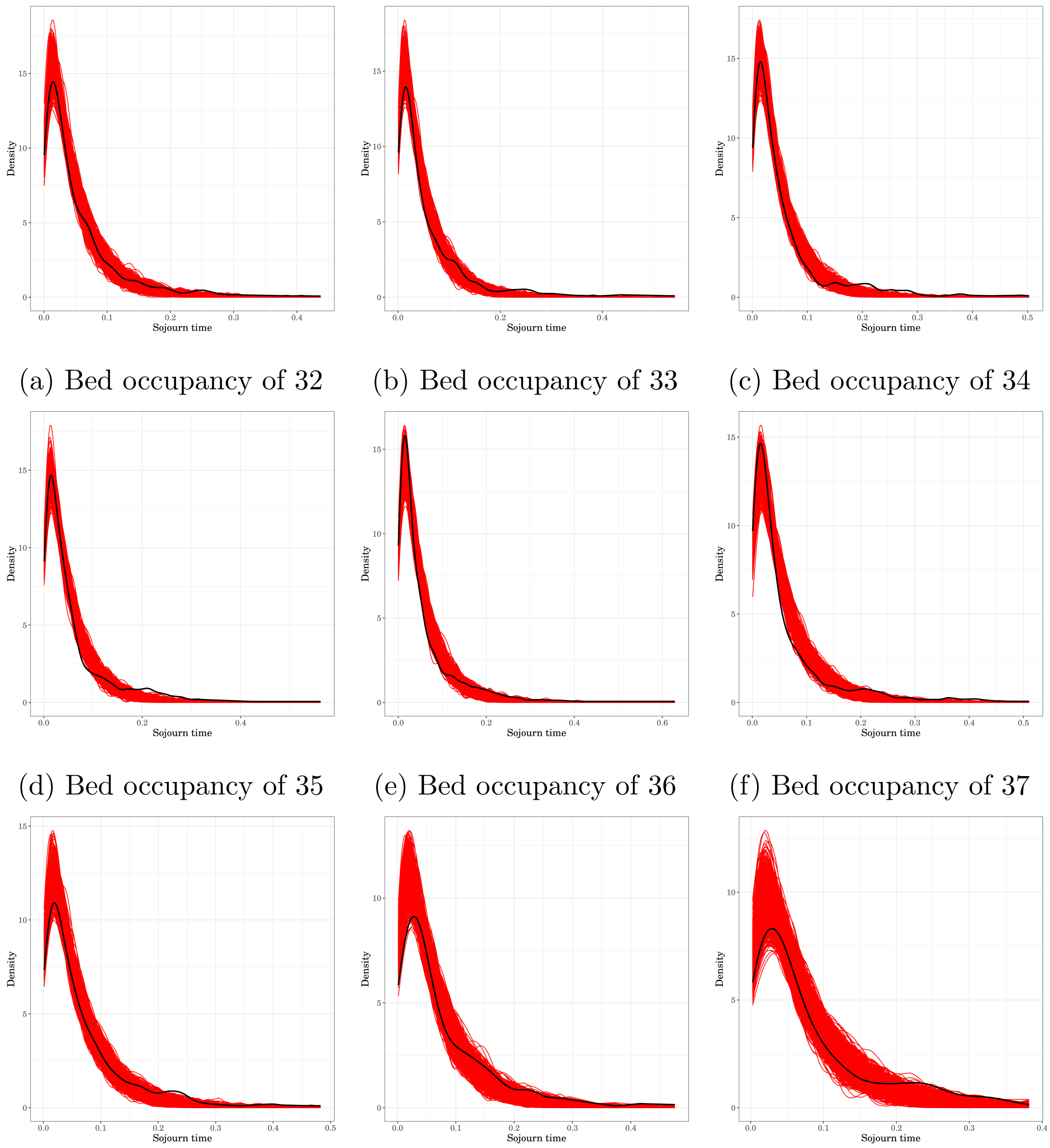


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.24: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level and phase transition dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.25: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted two-phase structured QBD process with level and phase transition dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

E.2.6 Three-phase structured QBD process with level and phase transition dependent polynomial forms

In this section, we consider a three-phase structured QBD process, such that the null and alternative hypotheses are defined as

$$\begin{aligned} H_0 &: g \in \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \\ H_a &: g \notin \{\mathcal{G} : \mathcal{G} = f(\mathcal{Q}), \mathcal{Q} \in \Omega\}, \end{aligned}$$

where Ω is the set of all three-phase structured QBD processes with an infinitesimal generator matrix of the form

$$Q = \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & 0 & \dots & 0 & 0 & 0 \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \dots & 0 & 0 & 0 \\ 0 & A_-^{(2)} & A_0^{(2)} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_0^{(23)} & A_+^{(23)} & 0 \\ 0 & 0 & 0 & \dots & A_-^{(24)} & A_0^{(24)} & A_+^{(24)} \\ 0 & 0 & 0 & \dots & 0 & A_-^{(25)} & A_0^{(25)} \end{bmatrix},$$

where the functional forms for the block matrices describing a decrease in level, for $1 \leq \ell \leq 25$, take the form

$$[H_\ell]_{i,j} = h_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,1), (1,3), (2,1), (3,2), (3,3)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1), & \text{if } (i,j) \in \{(2,3)\}, \\ 1 + \beta_1^{h,i,j}(\ell - 1) + \beta_2^{h,i,j}(\ell - 1)^2, & \text{if } (i,j) \in \{(1,2), (2,2), (3,1)\}, \end{cases}$$

the functional forms for the block matrices describing within level changes, for $0 \leq \ell \leq 25$, take the form

$$[G_\ell]_{i,j} = g_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,3),\,(2,1),\,(2,3),\,(3,2)\}, \\ 1 + \beta_1^{g,i,j}\ell, & \text{if } (i,j) \in \{(1,2)\}, \\ 1 + \beta_1^{g,i,j}\ell + \beta_2^{g,i,j}\ell^2, & \text{if } (i,j) \in \{(3,1)\}, \end{cases}$$

and the functional forms for the block matrices describing an increase in level, for $0 \leq \ell \leq 24$, take the form

$$[F_\ell]_{i,j} = f_{\ell,i,j} = \begin{cases} 1, & \text{if } (i,j) \in \{(1,2),\ (1,3),\ (2,1),\ (3,1), \\ & (3,2)\}, \\ 1 + \beta_1^{f,i,j}\ell, & \text{if } (i,j) \in \{(2,2),(2,3)\}, \\ 1 + \beta_1^{f,i,j}\ell + \beta_2^{f,i,j}\ell^2, & \text{if } (i,j) \in \{(1,1),(3,3)\}. \end{cases}$$

As illustrated in Figure E.2.26, the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted three-phase structured QBD process under-estimates the observed proportion of time spent in each bed occupancy from the RAH ICU data set for the most frequently visited bed occupancies.

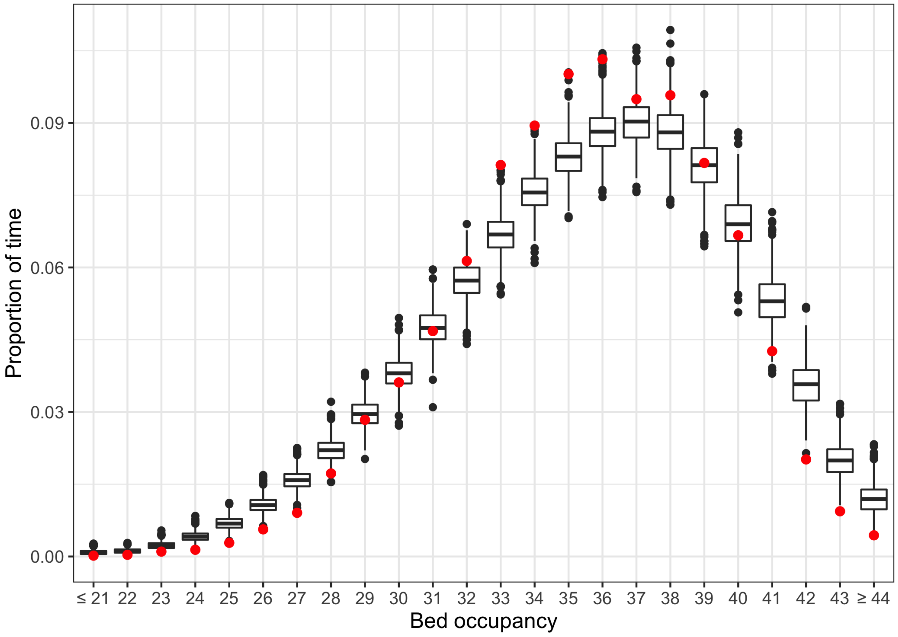


Figure E.2.26: Comparison of the observed proportion of time spent in each bed occupancy (red points) to the observed proportion of time spent in each bed occupancy based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

The transition probabilities of the fitted three-phase structured QBD process for the most frequently visited bed occupancies are similar to that observed in the RAH ICU data set, as illustrated in Figure E.2.27. Note that the variation towards the lower and higher bed occupancies is due to minimal data.

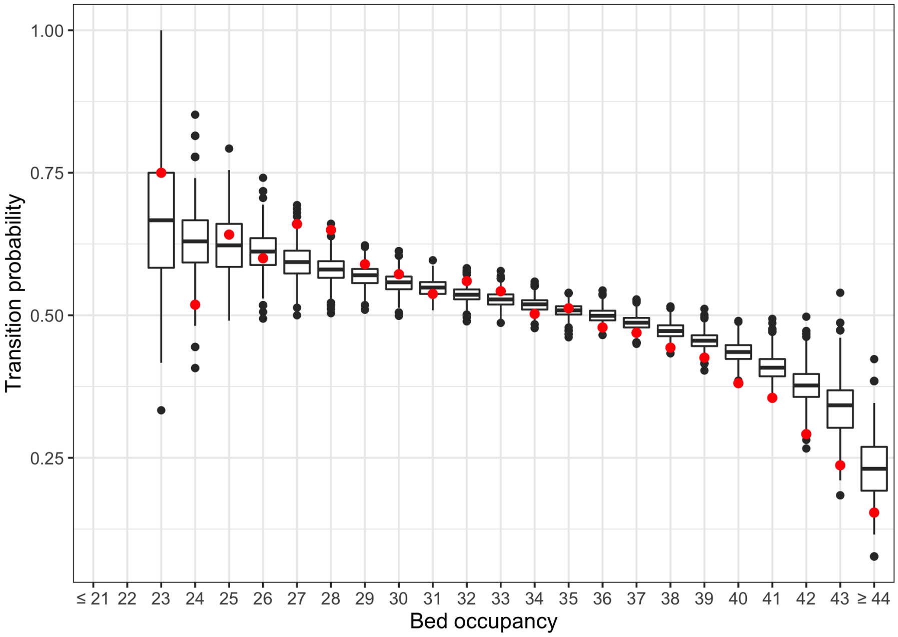
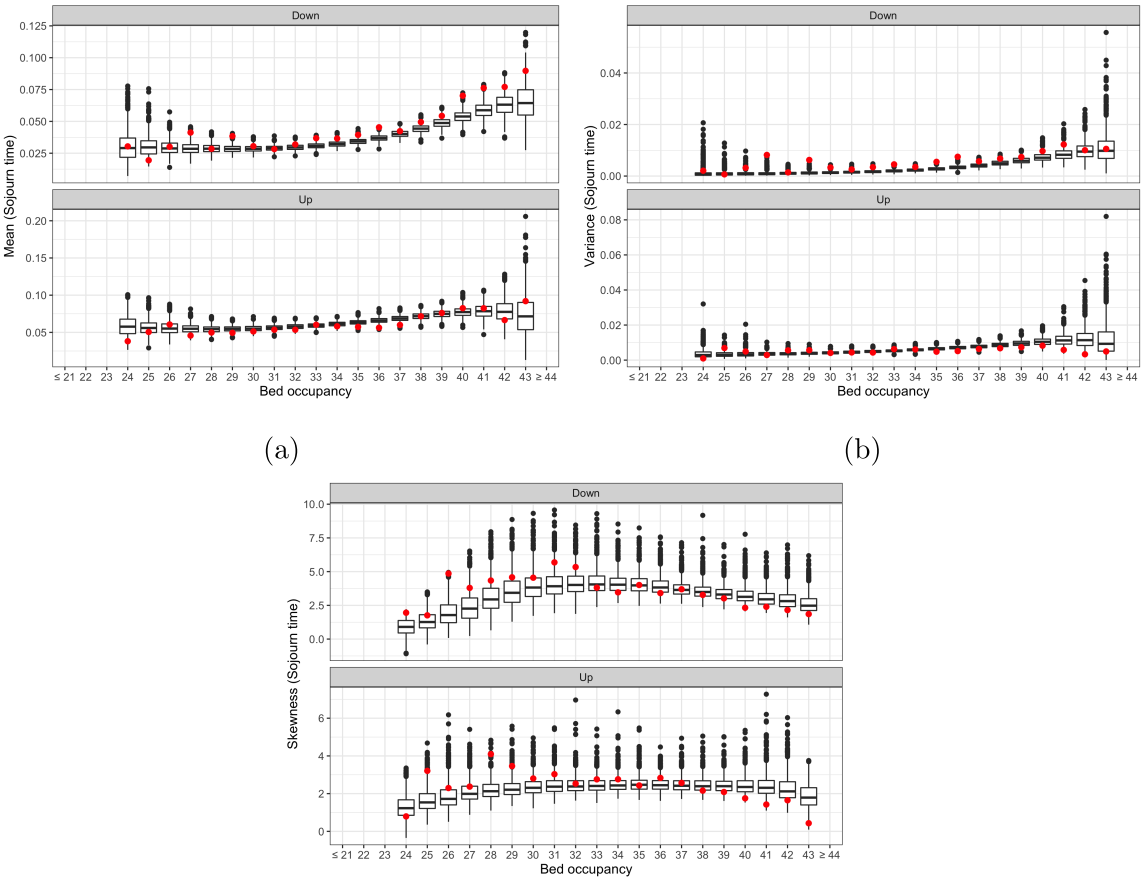


Figure E.2.27: Comparison of the observed transition probabilities between bed occupancies (red points) to the observed transition probabilities between bed occupancies based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that the transition probabilities for bed occupancies below 23 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Despite introducing another phase, the distribution of sojourn times conditioned on moving up a bed occupancy are slightly over-estimated, whereas the distribution of sojourn times conditioned on moving down a bed occupancy are slightly under-estimated, as demonstrated in Figures E.2.28a, E.2.28b, and E.2.28c. Figures E.2.29 and E.2.30 further illustrate the difference in the distribution of conditional sojourn times observed in the RAH ICU data compared to those of the fitted three-phase structured QBD process.

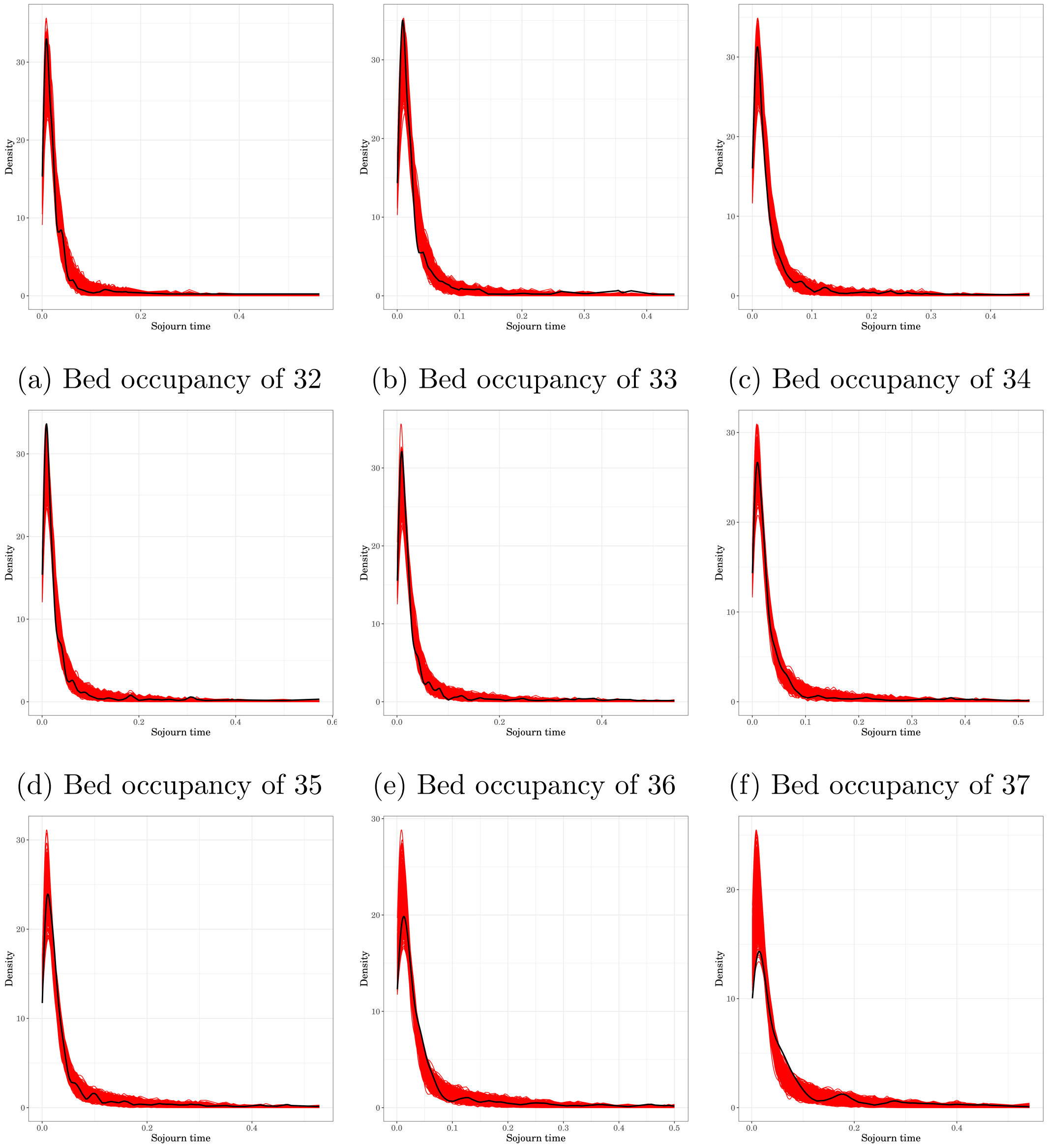


(a)

(b)

(c)

Figure E.2.28: Comparison between the first (a), second (b), and third (c) moments of the distribution of conditional sojourn times for each bed occupancy based on observed data (red points) and the distribution of conditional sojourn times for each bed occupancy based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level and phase transition dependent polynomial forms (box-plots). Note that the conditional sojourn times for bed occupancies below 24 and above 43 are omitted due to minimal data. Also note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

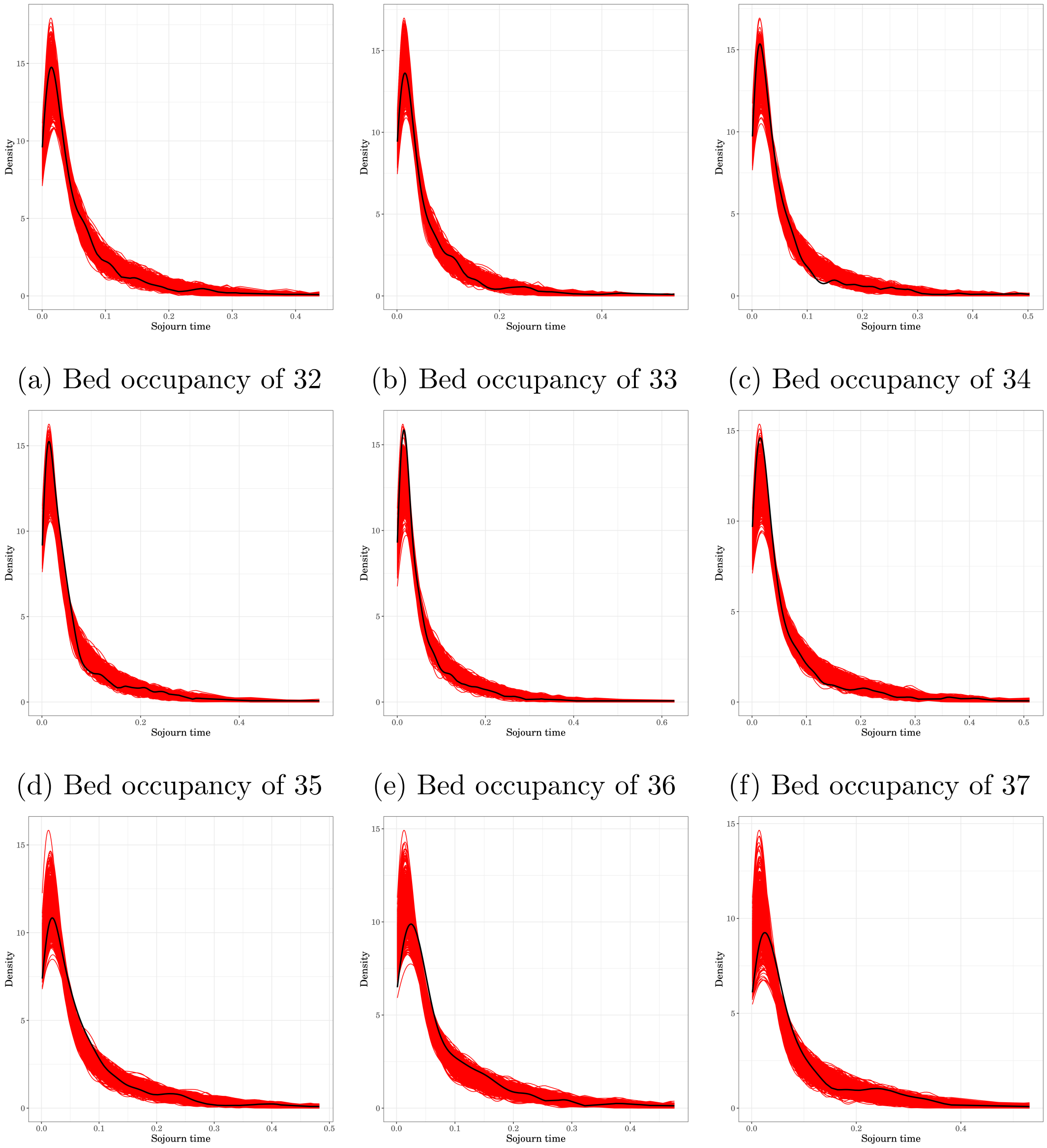


(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.29: Comparison of the observed density of sojourn times conditioned on downward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on downward transitions based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level and phase transition dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.



(a) Bed occupancy of 32 (b) Bed occupancy of 33 (c) Bed occupancy of 34

(d) Bed occupancy of 35 (e) Bed occupancy of 36 (f) Bed occupancy of 37

(g) Bed occupancy of 38 (h) Bed occupancy of 39 (i) Bed occupancy of 40

Figure E.2.30: Comparison of the observed density of sojourn times conditioned on upward transitions for the most frequently visited bed occupancies (black line) to the observed densities of sojourn times conditioned on upward transitions based on 1000 simulated data sets generated from the fitted three-phase structured QBD process with level and phase transition dependent polynomial forms (red lines). Note that each simulated data set contains the same number of changes in bed occupancy as that observed in the RAH ICU data set.

Appendix F

Transient behaviour analysis from Section 10.4

In this section, we complete the comparison of the $M/M/\bullet$ queueing models and the fitted four-phase structured QBD process with level and phase transition dependent polynomial forms in terms of capturing the transient behaviour present in the RAH ICU bed occupancy data set.

The four-phase structured QBD process with level and phase transition dependent polynomial forms is able to capture more of the transient behaviour compared to the $M/M/\bullet$ queueing models, as illustrated in Figures F.0.1, F.0.2, F.0.3, and F.0.4.

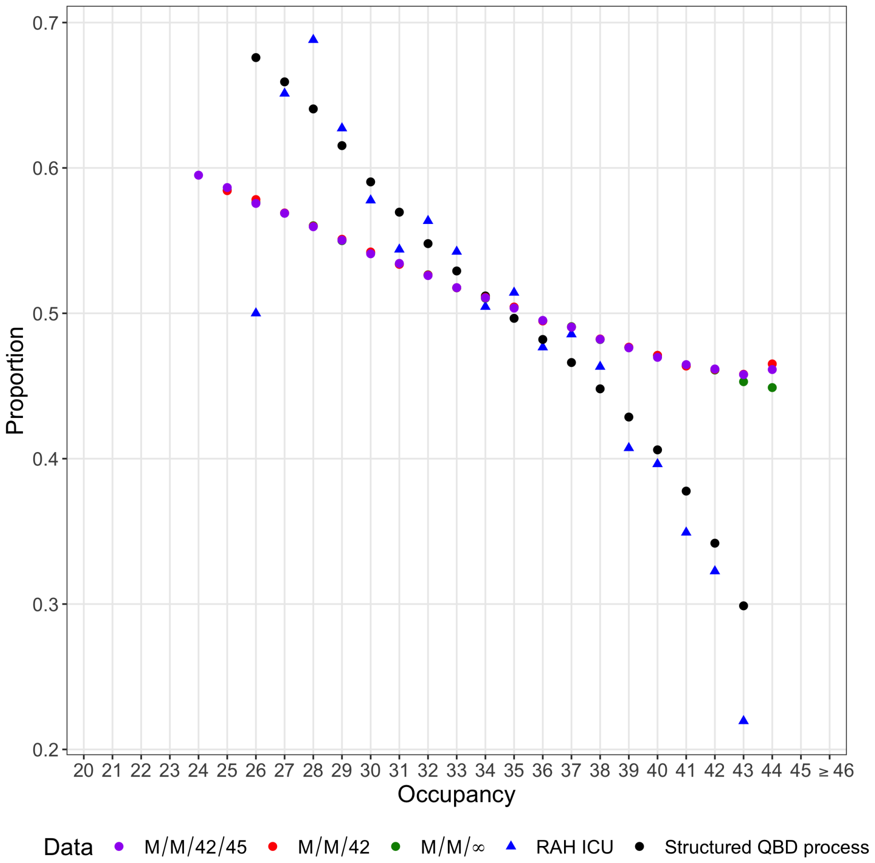


Figure F.0.1: Comparison of the transition probabilities between bed occupancies (blue) versus the transition probabilities between bed occupancies assuming an $M/M/42/45$ queueing model (purple), assuming an $M/M/42$ queueing model (red), assuming an $M/M/\infty$ queueing model (green), and assuming a four-phase structured QBD process with level and phase transition dependent polynomial forms (black). Note that the conditional sojourn times for levels below 24 and above 44 are omitted due to minimal data.

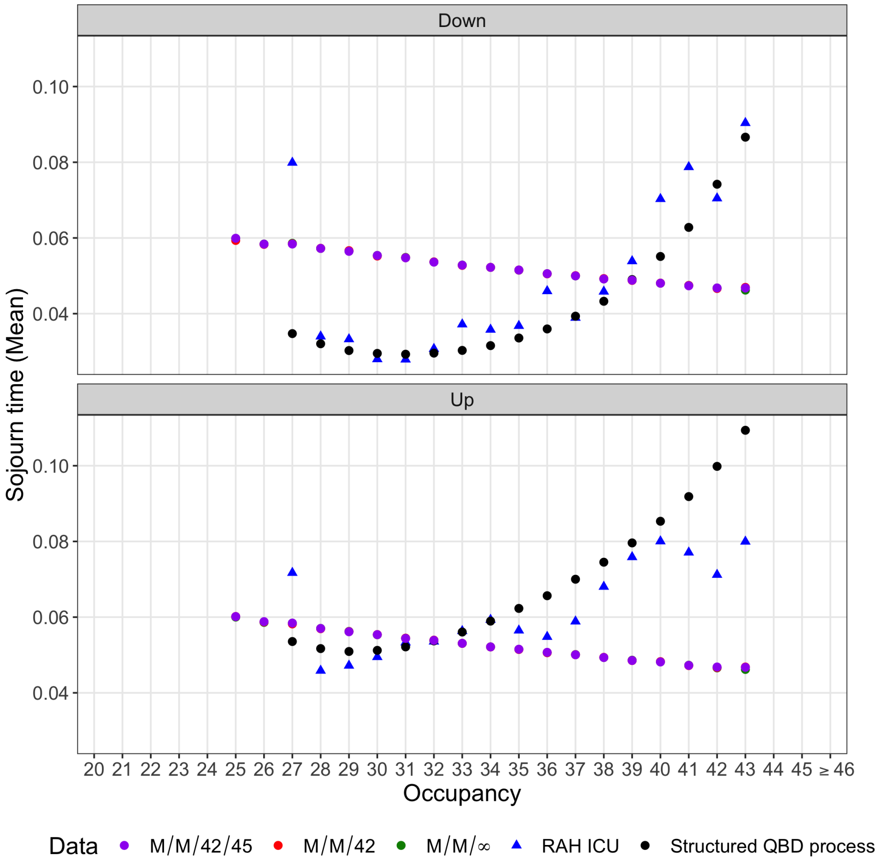


Figure F.0.2: Comparison between the first moment of the distribution of conditional sojourn times for each level based on observed data (blue), the distribution of conditional sojourn times for each level assuming an $M/M/42/45$ queueing model (purple), assuming an $M/M/42$ queueing model (red), assuming an $M/M/\infty$ queueing model (green), and assuming a four-phase structured QBD process with level and phase transition dependent polynomial forms (black). Note that the conditional sojourn times for levels below 25 and above 43 are omitted due to minimal data.

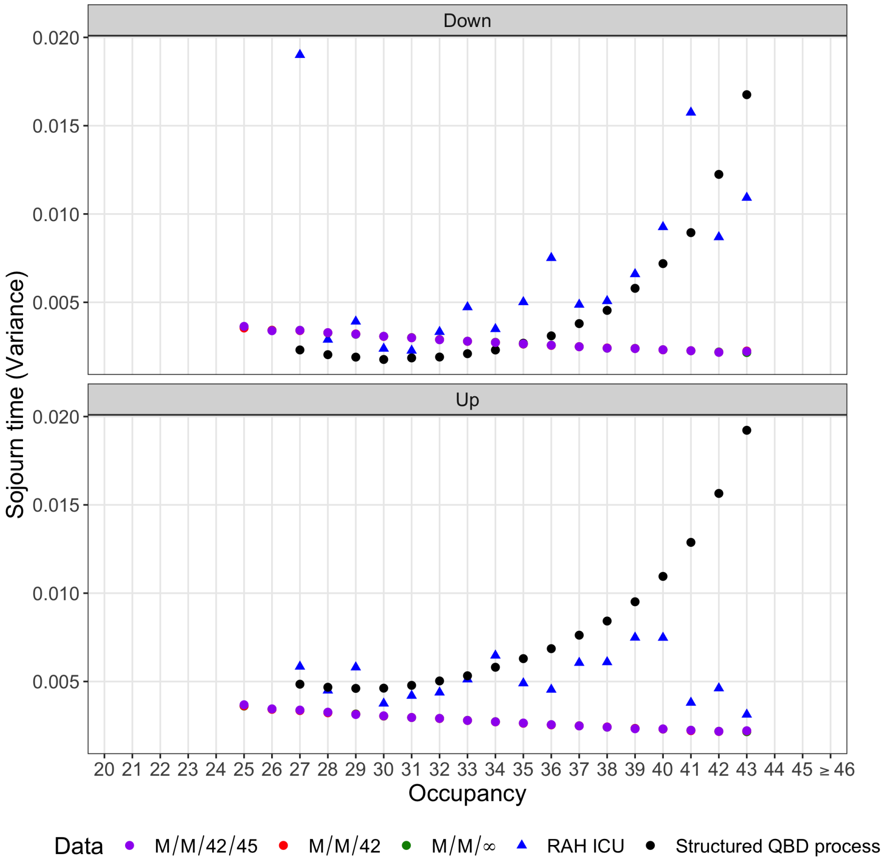


Figure F.0.3: Comparison between the second moment of the distribution of conditional sojourn times for each level based on observed data (blue), the distribution of conditional sojourn times for each level assuming an $M/M/42/45$ queueing model (purple), assuming an $M/M/42$ queueing model (red), assuming an $M/M/\infty$ queueing model (green), and assuming a four-phase structured QBD process with level and phase transition dependent polynomial forms (black). Note that the conditional sojourn times for levels below 25 and above 43 are omitted due to minimal data.

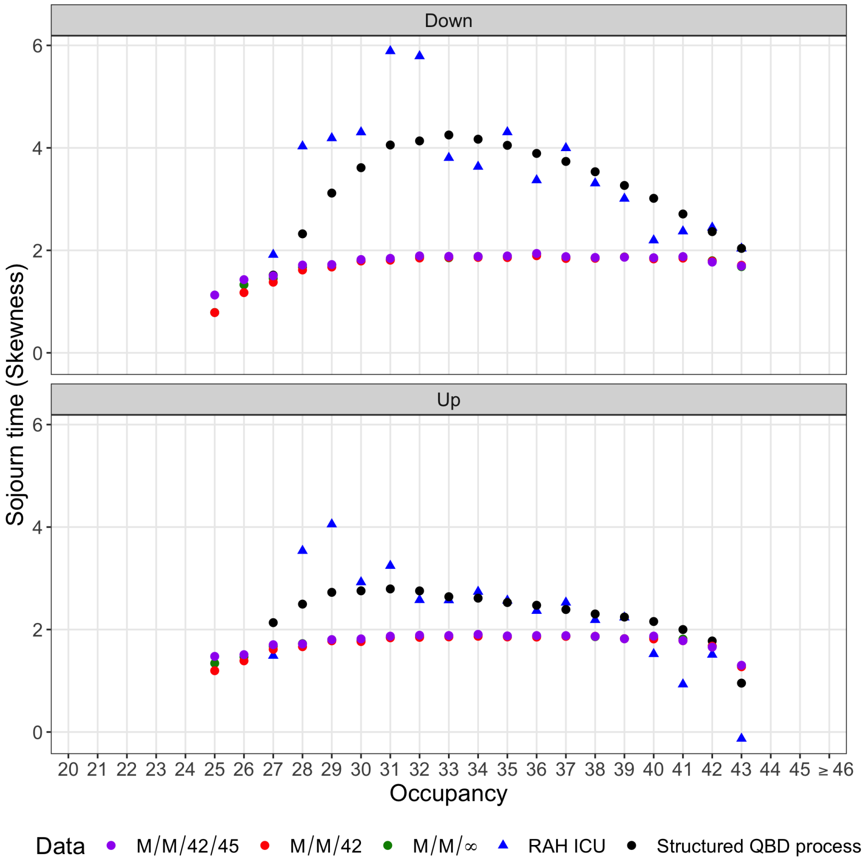


Figure F.0.4: Comparison between the third moment of the distribution of conditional sojourn times for each level based on observed data (blue), the distribution of conditional sojourn times for each level assuming an $M/M/42/45$ queueing model (purple), assuming an $M/M/42$ queueing model (red), assuming an $M/M/\infty$ queueing model (green), and assuming a four-phase structured QBD process with level and phase transition dependent polynomial forms (black). Note that the conditional sojourn times for levels below 25 and above 43 are omitted due to minimal data.

Bibliography

[1] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, pages 419–441, 1996.

[2] J. Bai, A. Fügener, J. Schoenfelder, and J. O. Brunner. Operations research in intensive care unit management: a literature review. *Health Care Management Science*, 21(1):1–24, 2018.

[3] H. Baumann and W. Sandmann. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Computer Science*, 1(1):1561 – 1569, 2010. ICCS 2010.

[4] N. G. Bean, M. Fackrell, and P. Taylor. Characterization of matrix-exponential distributions. *Stochastic Models*, 24(3):339–363, 2008.

[5] A. Ben-Israel and T. N. Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

[6] M. Bladt. *Renewal Theory and Queueing Algorithms for Matric-exponential Distributions*.

[7] M. Bladt, A. Gonzalez, and S. L. Lauritzen. The estimation of Phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal*, 2003(4):280–300, 2003.

[8] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):47–50, 1983.

[9] L. Bright and P. G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11(3):497–525, 1995.

[10] P. Buchholz. An EM-algorithm for MAP fitting from real traffic data. In *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 218–236. Springer, 2003.

[11] M. P. Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.

[12] D. R. Cox. 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, 1955.

[13] D. R. Cox. 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, 1955.

[14] F. W. Crawford, V. N. Minin, and M. A. Suchard. Estimation for general birth-death processes. *Journal of the American Statistical Association*, 109(506):730–747, 2014.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[16] B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[17] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[18] M. Fackrell. Fitting with matrix-exponential distributions. *Stochastic models*, 21(2-3):377–400, 2005.

[19] M. M. Fadiloglu and S. Yeralan. Models of production lines as quasi-birth-death processes. *Mathematical and computer modelling*, 35(7-8):913–930, 2002.

[20] C. Fauconnier and G. Haesbroeck. Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology*, 6(4):363–379, 2009.

[21] D. Gaver, P. Jacobs, and G. Latouche. Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability*, 16(4):715–731, 1984.

[22] H. Ghorbani. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Univ Ser Math Inform*, 34(3):583–95, 2019.

[23] T. P. Gonzenbach, S. P. McGuinness, R. L. Parke, and T. M. Merz. Impact of Nonpharmaceutical Interventions on ICU Admissions During Lockdown for Coronavirus Disease 2019 in New Zealand—A Retrospective Cohort Study. *Critical Care Medicine*, 49(10):1749, 2021.

[24] S. Hautphenne and M. Fackrell. An EM algorithm for the model fitting of Markovian binary trees. *Computational Statistics & Data Analysis*, 70:19–34, 2014.

[25] Q.-M. He and H. Zhang. On matrix exponential distributions. *Advances in Applied Probability*, 39(1):271–292, 2007.

[26] R. A. Johnson, D. W. Wichern, et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

[27] R. Jones. Hospital bed occupancy demystified. *British Journal of Healthcare Management*, 17(6):242–248, 2011.

[28] S. Karlin and J. McGregor. Linear growth, birth and death processes. *Journal of Mathematics and Mechanics*, 7(4):643–662, 1958.

[29] D. Karlis and E. Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577 – 590, 2003. Recent Developments in Mixture Model.

[30] A. D. Keegan. Hospital bed occupancy: more than queuing for a bed. *Medical Journal of Australia*, 193(5):291–293, 2010.

[31] N. Keiding et al. Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2):363–372, 1975.

[32] L. Kleinrock. Queueing systems. Technical report, 1975.

[33] A. Klemm, C. Lindemann, and M. Lohmann. Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation*, 54(2):149–173, 2003.

[34] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*, volume 5. Siam, 1999.

[35] R. Li, C. Rivers, Q. Tan, M. B. Murray, E. Toner, and M. Lipsitch. The demand for inpatient and ICU beds for COVID-19 in the US: lessons from Chinese cities. *MedRxiv*, 2020.

[36] L. Lipsky and V. Ramaswami. A unique minimal representation of Coxian service centres. *Dept. Comput. Sci. Eng., Univ. Nebraska, Lincoln, NE, USA, Tech. Rep*, 1985.

[37] F. J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[38] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[39] C. A. O'Cinneide. On non-uniqueness of representations of phase-type distributions. *Communications in Statistics. Stochastic Models*, 5(2):247–259, 1989.

[40] C. A. O'Cinneide. Characterization of phase-type distributions. *Stochastic Models*, 6(1):1–57, 1990.

[41] H. Okamura and T. Dohi. Faster maximum likelihood estimation algorithms for Markovian arrival processes. In *Quantitative Evaluation of Systems, 2009. QEST'09. Sixth International Conference on the*, pages 73–82. IEEE, 2009.

[42] M. Olsson. Estimation of phase-type distributions from censored data. *Scandinavian journal of statistics*, pages 443–460, 1996.

[43] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[44] B. Ramos-Lacuey, M. H. Aguirre, C. C. Gallego, A. I. L. de Munain, E. G. Esarte, and L. Moreno-Galarraga. ECIEN-2020 study: the effect of COVID-19 on admissions for non-COVID-19 diseases. *World Journal of Pediatrics*, 17(1):85–91, 2021.

[45] W. J. Roberts, Y. Ephraim, and E. Dieguez. On Ryden's EM algorithm for estimating MMPPs. *IEEE Signal Processing Letters*, 13(6):373–376, 2006.

[46] L. Tao and J. Liu. *Healthcare Service Management: A Data-Driven Systems Approach.* Springer, 2019.

[47] J. Varney, N. Bean, and M. Mackay. The self-regulating nature of occupancy in ICUs: stochastic homoeostasis. *Health care management science*, 22(4):615–634, 2019.

[48] Z. Wu and J. M. McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama*, 323(13):1239–1242, 2020.