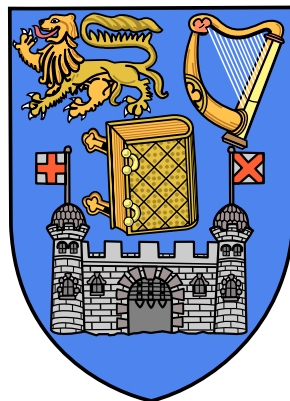


# MCMC for Inference on Phase-type and Masked System Lifetime Models

A thesis submitted to the University of Dublin, Trinity College  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy

Department of Statistics, Trinity College Dublin



December 2012

**Louis J. M. Aslett**

*I dedicate this thesis to my wonderful and supportive parents*

*Heli & Martin Aslett*

## Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

The copyright remains solely with the author, Louis J. M. Aslett.

---

Louis J. M. Aslett

Dated: 18 December 2012

# Abstract

Commonly reliability data consist of lifetimes (or censoring information) on all components and systems under examination. However, masked system lifetime data represents an important class of problems where the information available for statistical analysis is more limited: one only has failure times for the system as a whole, but no data on the component lifetimes directly, or even which components had failed. For example, such data can arise when system autopsy is impractical or cost prohibitive.

The problem of statistical inference with such masked system lifetime data is considered in the situation where component repair is and is not possible.

Continuous-time Markov chains are widely-used models for repairable systems with redundancy, with an absorbing state representing failure. The failure time is then the first passage time to the absorbing state, which is modelled by a Phase-type distribution.

Bayesian inference via MCMC for Phase-type distributions when data consist only of absorption times is considered. Extensions to the existing methodology of Bladt *et al.* (2003) are presented which accommodate censored data, enable a structure to be imposed on the underlying continuous-time Markov process and expand computational tractability to a wider class of situations. This part of the research is also broadly applicable outside the reliability interpretation presented in this thesis.

For non-repairable systems, a novel signature based data augmentation scheme is presented which enables inference for a wide class of component lifetime models for an exchangeable population of systems. It is shown that the approach can be extended to enable topological inference of the underlying system design.

Finally, two R packages ('PhaseType' and 'ReliabilityTheory') are provided which enable reliability practitioners to make use of the theoretical contributions of this research easily.

# Acknowledgements

I would like to express sincere gratitude to my supervisor Simon Wilson first. He has always given superb guidance, whilst never making me feel foolish even if my questions frequently were! Yet he also provided healthy doses of autonomy and space to explore my own ideas and interests, combined with supreme patience and good humour when some avenues of my naïve enthusiasm proved fruitless. He has also encouraged and enabled wonderful opportunities to present at numerous international conferences. I am indebted to Brett Houlding who was so generous with his time as to be akin to a second supervisor, and who I now consider among my closest friends.

Indeed, all of the statistics faculty have been fabulous to engage with over these four years, and the shared journey with my fellow postgraduates has been truly memorable: I thank them all for their camaraderie. In particular, I have shared many superb work and social times with Arnab Bhattacharya. Outside the department, I owe special thanks to Donal O'Donovan without whose encouragement and assistance a return to academia may not have been possible.

On a personal level, I am enduringly grateful to my parents, Heli and Martin, and my brother and his wife, Pépin and Krista, for unfailing encouragement and support. I thank their two wonderful bundles of life, Heidi and Hugo, who burst onto the scene since I began the PhD, for making me a proud and happy uncle. My Irish cousins, the Cusack family (Brian, Mo, Andrew, Julianne & Nicola), have truly been like a second family to me here and welcomed me in to the family home for a long time.

On the occasions we couldn't meet, the patience of my longest and best of friends, Scott Gould and Luke Mundy, has been tremendous. The Staniek family (Magda, Miko, Milosh & Iryda) have been such special friends and so kind to me. Finally, Susan and Ciaran Walsh have always been gracious and caring, never failing to exceed any expectation.

The generous support of a three year IRCSET postgraduate fellowship and a one year SFI funded extension under the STATICA project made it all feasible.

**Louis J. M. Aslett**

*Trinity College Dublin*

*December 2012*

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>v</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Overview of chapters . . . . .	3
1.2 Research contributions . . . . .	4
<b>Chapter 2 Reliability Theory and Phase-type Distributions</b>	<b>6</b>
2.1 An introduction to reliability theory . . . . .	6
2.1.1 Component reliability modelling . . . . .	7
2.1.2 Inferential considerations . . . . .	10
2.1.3 Structural reliability . . . . .	11
2.1.3.1 System structure functions . . . . .	11
2.1.3.2 Path and cut sets . . . . .	13
2.1.3.3 System reliability . . . . .	14
2.1.3.4 Coherent systems . . . . .	15
2.1.3.5 System signature . . . . .	16
2.1.3.6 Summary . . . . .	18
2.1.4 Masked system lifetime inference . . . . .	19
2.1.5 Network reliability . . . . .	19
2.1.6 Repairable systems . . . . .	20
2.2 Stochastic processes . . . . .	20

2.2.1	Continuous-time Markov chains . . . . .	22
2.2.1.1	Formulating a CTMC model . . . . .	24
2.2.1.2	Relationship to reliability modelling . . . . .	25
2.2.1.3	Transition probabilities . . . . .	25
2.2.1.4	Limiting and stationary distributions . . . . .	26
2.2.1.5	Simulation of a continuous-time Markov chain . . . . .	27
2.2.1.6	Absorbing states . . . . .	28
2.2.1.7	Summary . . . . .	28
2.2.2	Phase-type distributions . . . . .	29
2.2.2.1	Random variate generation . . . . .	30
2.2.2.2	Proposed usage . . . . .	31
<b>Chapter 3 Statistical Methodology</b>		<b>32</b>
3.1	Bayesian inference . . . . .	32
3.1.1	Likelihood function . . . . .	34
3.1.2	Prior distribution . . . . .	35
3.1.2.1	Common prior choices . . . . .	35
3.1.3	Posterior analysis . . . . .	37
3.1.4	Hierarchical Bayesian models . . . . .	38
3.1.5	Exchangeability . . . . .	39
3.1.6	Summary . . . . .	40
3.2	Markov chain Monte Carlo . . . . .	41
3.2.1	Standard Monte Carlo . . . . .	42
3.2.1.1	Generality of expectation . . . . .	42
3.2.1.2	Beyond standard Monte Carlo . . . . .	43
3.2.2	(Discrete time) Markov chains . . . . .	43
3.2.3	Markov chain sampling algorithms . . . . .	45
3.2.3.1	The Metropolis-Hastings algorithm . . . . .	45
3.2.3.2	The Gibbs sampler . . . . .	47
3.2.3.3	Data augmentation . . . . .	50
3.2.4	Convergence . . . . .	50
3.2.4.1	Markov chain Central Limit Theorem . . . . .	53
3.2.5	Implementing MCMC . . . . .	54

3.2.5.1	Burn-in . . . . .	55
3.2.5.2	Run length . . . . .	55
3.2.5.3	Diagnostics . . . . .	56
3.2.6	Summary . . . . .	56
3.3	Inference for Phase-type models . . . . .	57
3.3.1	Latent process likelihood . . . . .	57
3.3.2	The approach of Asmussen <i>et al.</i> (1996) . . . . .	58
3.3.3	The approach of Bladt <i>et al.</i> (2003) . . . . .	59
3.3.3.1	The methodology . . . . .	60
3.3.3.2	Simulating from $f_{\Phi \Psi}(\phi_i   \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$ . . . . .	60
3.3.3.3	Summary . . . . .	62
<b>Chapter 4 MCMC for Phase-type Models</b>		<b>63</b>
4.1	A study of the original methodology . . . . .	63
4.1.1	The model set-up . . . . .	64
4.1.1.1	Prohibited transitions . . . . .	64
4.1.1.2	Equivalent transitions . . . . .	64
4.1.1.3	Scope of the issue . . . . .	65
4.1.2	Censored data . . . . .	68
4.1.3	Computational performance . . . . .	68
4.1.3.1	I. Exploration of parameter space . . . . .	69
4.1.3.2	II. Zero constraints for absorbing moves . . . . .	70
4.1.3.3	III. Convergence of process samples to stationarity . . . . .	71
4.1.4	Summary . . . . .	75
4.2	Extensions to the methodology . . . . .	76
4.2.1	Model reformulation . . . . .	76
4.2.2	Latent process simulation . . . . .	78
4.2.2.1	Pilot solutions . . . . .	78
4.2.2.2	Exact Conditional Sampling (ECS) . . . . .	80
4.2.3	Incorporating censored data . . . . .	86
4.3	Examining the extended methodology . . . . .	90
4.3.1	Effect of model reformulation . . . . .	90
4.3.2	Computational performance . . . . .	92



4.3.2.1	Example with no major problems . . . . .	93
4.3.2.2	Example exhibiting acute problems . . . . .	93
4.3.2.3	Tail depth speed improvements . . . . .	94
4.3.2.4	Effect of censoring . . . . .	95
4.3.3	Realistic 5 component bridge system . . . . .	95
4.4	Sensitivity to the prior . . . . .	98
4.5	Exchangeability . . . . .	101
4.6	Generalisations . . . . .	101
4.7	Summary . . . . .	102
<b>Chapter 5 Networks, Simply Connected Systems &amp; Their Signature</b>		<b>103</b>
5.1	Colliding nomenclatures . . . . .	104
5.2	Simply connected systems . . . . .	105
5.3	Coherent networks . . . . .	108
5.4	Representations and their signature . . . . .	109
5.4.1	Simply connected coherent system signatures . . . . .	110
5.4.2	Coherent network signatures . . . . .	111
5.5	Summary . . . . .	111
<b>Chapter 6 Parametric and Topological Inference with Masked Lifetime</b>		
	<b>Data</b>	<b>114</b>
6.1	The problem setting . . . . .	115
6.1.1	Independent systems . . . . .	115
6.1.2	Exchangeable systems . . . . .	116
6.2	Signature based MCMC algorithm . . . . .	117
6.2.1	Inferring component lifetime parameters . . . . .	117
6.2.1.1	Sampling Latent Component Failure Times . . . . .	119
6.2.2	Topological inference . . . . .	121
6.2.3	Summary of the MCMC algorithm . . . . .	124
6.3	Examples . . . . .	125
6.3.1	IID systems with Exponential components . . . . .	126
6.3.1.1	Model set-up . . . . .	126
6.3.1.2	Parametric inference only . . . . .	127

6.3.1.3	Parametric and topological inference . . . . .	127
6.3.2	Exchangeable systems with Exponential components . . . . .	129
6.3.2.1	Model set-up . . . . .	129
6.3.2.2	Parametric inference only . . . . .	133
6.3.2.3	Parametric and topological inference . . . . .	133
6.3.3	IID systems with Phase-type components . . . . .	135
6.3.3.1	Model set-up . . . . .	135
6.3.3.2	Parametric and topological inference for simply con- nected systems . . . . .	135
6.3.4	Exchangeable systems with Phase-type components . . . . .	137
6.3.4.1	Model set-up . . . . .	137
6.3.4.2	Parametric inference . . . . .	140
6.3.4.3	Topological inference . . . . .	141
6.3.5	Sensitivity to the prior . . . . .	142
6.3.6	Summary . . . . .	143
<b>Chapter 7 Statistical Computing</b>		<b>145</b>
7.1	‘PhaseType’ package . . . . .	146
7.1.1	Methodology of Bladt <i>et al.</i> (2003) . . . . .	146
7.1.2	Methodology of Chapter 4 . . . . .	148
7.2	‘ReliabilityTheory’ package . . . . .	150
7.2.1	Systems and networks . . . . .	150
7.2.2	Making PhaseType easier for reliability problems . . . . .	150
7.2.3	Parametric and topological inference . . . . .	152
7.2.3.1	Simple interface . . . . .	152
7.2.3.2	General framework . . . . .	153
7.3	Summary . . . . .	155
<b>Chapter 8 Conclusion</b>		<b>156</b>
8.1	Future work . . . . .	157
<b>Appendix A Signatures</b>		<b>160</b>
A.1	Coherent network signatures . . . . .	160
A.2	Simply connected coherent system signatures . . . . .	163

# List of Tables

2.1	A labelling of all combinations of component states for which system is operational (1–16) and various failure states (17–32) for bridge system (see Figure 2.3). . . . .	21
2.2	A labelling for the elementary two component parallel system. . . . .	22
4.1	Posterior summary statistics for simulations from known ground truth using Bladt <i>et al.</i> (2003). Bracketed figures are MCMC standard errors. . . . .	67
4.2	Example probability mass function of latent process state at time $y_i = 500$ , for $\lambda_f = 1, \lambda_r = 1000$ in the (repairable) two component parallel system. . . . .	71
4.3	Posterior summary statistics for simulations from known ground truth using extended methodology. Bracketed figures are MCMC standard errors. . . . .	91
4.4	Posterior summary statistics for simulations from known ground truth using extended methodology for 5 component bridge system. Bracketed figures are MCMC standard errors. . . . .	97
5.1	All simply connected systems of order 2, together with system signature. Contained in data set <code>sccs02</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	110
5.2	All simply connected systems of order 3, together with system signature. Contained in data set <code>sccs03</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	111
5.3	All simply connected systems of order 4, together with system signature. Contained in data set <code>sccs04</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	112

5.4	All coherent networks of node order 2, together with system signature. Black nodes are start/terminal. Contained in data set <code>cn02</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	113
6.1	Uniqueness of signatures for coherent systems. . . . .	122
A.1	All coherent networks of node order 3, together with system signature. Black nodes are start/terminal. Contained in data set <code>cn03</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	162
A.2	All simply connected systems of order 5, together with system signature. Contained in data set <code>sccs05</code> of <code>ReliabilityTheory</code> package (see §7.2.1, page 150). . . . .	166

# List of Figures

2.1	A series (left) and parallel (right) system structure. . . . .	12
2.2	A simple 4 component system. . . . .	12
2.3	Addition of bridging component to simple 4 component system. . . . .	13
2.4	Example realisation of stochastic process for a two component parallel system. . . . .	22
4.1	Boxplot for posterior simulations from known ground truth using Bladt <i>et al.</i> (2003). Solid vertical lines show ground truth values. . . . .	66
4.2	Total variation distance from the target distribution. . . . .	75
4.3	Boxplot for posterior simulations from known ground truth using extended methodology. Solid vertical lines show ground truth values. . . . .	92
4.4	Speed of original (MH+RS) and new (ECS) sampling methodologies at varying right tail probabilities, with bootstrap resampled confidence intervals from 50 simulations at each $x$ . . . . .	94
4.5	Boxplot for posterior simulations from known ground truth using extended methodology for 25% censoring on data. Solid vertical lines show ground truth values. . . . .	95
4.6	Boxplot for posterior simulations from known ground truth using extended methodology for 5 component bridge system. Solid vertical lines show ground truth values. . . . .	98
4.7	Marginal failure rate posteriors for all combinations of 4 different failure and repair rate priors. Prior parameters shown in the dark grey boxes (failure rate across top, repair rate down the side), with failure rate prior densities depicted by dashed line at the top. Ground truth is dotted vertical line. . . . .	99

4.8	Marginal repair rate posteriors for all combinations of 4 different failure and repair rate priors. Prior parameters shown in the dark grey boxes (repair rate across top, failure rate down the side), with repair rate prior densities depicted by dashed line at the top. Ground truth is dotted vertical line. . . . .	100
6.1	A simple 3 component system. . . . .	116
6.2	Graphical model representing exchangeable systems. . . . .	117
6.3	Boxplot for posterior simulations for each of 11 different simply connected coherent system from known ground truth, $\lambda = 3.14$ . Solid horizontal line shows ground truth. . . . .	128
6.4	Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with $\lambda = 3.14$ . $\mathcal{M}$ consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed. Solid lines show ground truth. . . . .	128
6.5	Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with $\lambda = 3.14$ . $\mathcal{M}$ consisted of all 52 order 2, 3, 4 and 5 simply connected coherent systems. Topologies with posterior probability below 0.04 suppressed. Solid lines show ground truth. . . . .	130
6.6	Diagnostic plots of $\log f_{\Theta Y}(\nu, \zeta   \mathbf{y})$ before (left) and after (right) reparameterisation. . . . .	132
6.7	Ground truth, prior predictive and posterior predictive densities for the exchangeable population density of $\lambda$ . . . . .	133
6.8	Topology posterior for ground truth system 3. $\mathcal{M}$ consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed. . . . .	134
6.9	Topology posterior for ground truth network 7. $\mathcal{M}$ consisted of all 24 node order 3 coherent networks. Zero probability topologies suppressed. . . . .	134

6.10	Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with $\lambda_f = 1.8$ and $\lambda_r = 9.5$ . $\mathcal{M}$ consisted of all 52 order 2, 3, 4 and 5 simply connected coherent systems. Only topologies with posterior probability $> 0.04$ shown. Solid lines show ground truth. . . .	136
6.11	Ground truth, prior predictive and posterior predictive densities for the exchangeable population density of $\lambda_f$ and $\lambda_r$ . . . . .	140
6.12	Topology posterior for ground truth system 3. $\mathcal{M}$ consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed. . . . .	141
6.13	Example parameter and topology posterior for high variance prior. $\mathcal{M}$ consisted of all 11 order 4 simply connected coherent systems. Zero probability topologies suppressed. . . . .	142
6.14	Marginal failure rate posteriors in solid line for different priors in dashed line. Prior parameters shown in the dark grey boxes, with failure rate prior densities depicted by dashed line. Ground truth is dotted vertical line. . . . .	143
6.15	Marginal topology posteriors for different failure rate priors. Prior parameters shown in the dark grey boxes. Ground truth is solid vertical line. . . . .	144

# Chapter 1

## Introduction

The motivating preoccupation of this thesis is statistical inference for reliability models of both repairable and non-repairable coherent systems in the presence of masked system lifetime data.

The availability of a system depends on the operational state of its components. Thus, a simple inferential analysis may traditionally assume that the operational state and failure times (or censoring information) for all constituent components of the system are known. In principle, one may then proceed with a Bayesian analysis in a straightforward manner to learn about the parameters of the component lifetime distribution used in the model.

However, the scenario considered herein assumes that only the overall system failure time is known, and not those of the constituent components of the system. Indeed, the particular operational state of the components comprising the system is also considered unknown at the system failure time. It is data of this kind which are commonly referred to as *masked system lifetime data*.

Such data can arise in a variety of applications. For example, repair of a complex system may involve identifying which major subsystem has failed and replacing it entirely, so that the subsystem lifetime is known, but not the lifetimes of the components within it or even which components had failed, unless an autopsy is performed. Such an autopsy may be impractical or cost prohibitive relative to the repair, such as in consumer computer repair. Another example would be where many systems are sent into the field, but upon failure the system may be discarded, and only failure time of the system as a whole reported.



Two cases are considered: when components cannot be repaired, and also when non-disruptive component repair prior to the first system failure is possible, with repair times having a random element.

The repairable case is handled by using a continuous-time Markov chain model of the process of failure and repair through which a system passes before ultimately failing. With an absorbing state representing failure, the first passage time to that state is then Phase-type distributed. Consequently, focus is on Bayesian inference for Phase-type distributions with generator matrices designed to represent the failure and repair process.

The non-repairable case is tackled by development of a system signature based data augmentation scheme, leading to a highly flexible approach. As a result, the topology of the system may be jointly incorporated into the inference so that learning may take place not only on the parameters of the component lifetime distribution, but also on the system design.

There is potentially broad scope for real world application of this work. Indeed, the advances of Chapter 4 are applicable anywhere that the problem under investigation may be modelled by a Phase-type distribution. Restricted to reliability theory, the focus in the sequel, applications still abound in power, telecommunications and computer systems, among many other engineered systems.

For example, an offshore wind farm may only be subject to very periodic repairs, so that it may be modelled as a non-repairable system, and moreover catastrophic failure can lead to sufficient damage to render autopsy impossible, leading to fully masked lifetime data.

Another example, this time repairable, may be of a business reliant on a telecommunications network for connecting mission critical systems. If not the operator of the network, then the exact process of failure and repair may be unknown but failure times will be observed, again leading to fully masked lifetime data.

Finally, learning of the system design is the most ‘blue skies’ aspect of the research, with fewer immediately obvious applications. One application which has been suggested is an adversarial military or game theoretic setting where the design of the opposing force’s system is unknown but failure data is available to then learn its design.

## 1.1 Overview of chapters

This thesis should provide a perspicuous account of the research assuming only knowledge of the compulsory elements of an undergraduate degree in mathematics, such as the moderatorship programme taught at Trinity College Dublin.

An overview of the remainder of the thesis is as follows:

### **Chapter 2: Reliability Theory and Phase-type Distributions**

The fundamental concepts involved in structural reliability theory are introduced in this chapter, from the perspective of both the structure function and the system signature. The second part of the chapter introduces stochastic processes, specifically continuous-time Markov chains, and Phase-type distributions with an emphasis on their potential application in reliability theory.

### **Chapter 3: Statistical Methodology**

The Bayesian inferential method underlies the work in this thesis and so a brief introduction to the Bayesian perspective on inference is provided for those outside the statistics research community, who may be more familiar with frequentist ideas. The chapter continues with an introduction to the foundational ideas of Markov chain Monte Carlo theory, atop which all of the research in this thesis sits. The chapter concludes by reviewing the two main works in the literature for general inference on Phase-type models.

### **Chapter 4: MCMC for Phase-type Models**

This chapter marks the commencement of the main body of contributions.

There is a detailed study of the MCMC algorithm of Bladt *et al.* (2003), with special emphasis given to understanding its potential application to masked system lifetime inference when components are repairable. In light of certain problems which become apparent, the chapter continues by presenting extensions to the methodology which eliminate or ameliorate them. This includes reformulation of the model, development of a new approach to sampling the latent continuous-time Markov process and accommodation of censored data. The chapter concludes with an examination of the

extended methodology and a realistic size reliability example.

The contributions of Chapter 4 are generally applicable to any Phase-type model, but to maintain a cogent thread through the research the emphasis is always on reliability problems.

### **Chapter 5: Networks, Simply Connected Systems & Their Signature**

This is a short chapter which lays some prosaic but essential groundwork for the remainder of the research. It clarifies terminology used in different ways in various parts of the literature, making clear a separate subclass of coherent system which is useful for some applications. It also catalogues graph representations and system signatures for these systems and additionally applies the system signature to unreliable links rather than unreliable components in such systems.

### **Chapter 6: Parametric and Topological Inference with Masked Lifetime Data**

Chapter 6 considers the non-repairable setting. A signature based data augmentation scheme is developed which is applicable for a broad range of component lifetime distributions, enabling parametric inference on the lifetime distribution parameters. This is immediately extended to enable inference on the topology of the system which generated the masked lifetime data.

The chapter includes extensive examples of the methodology, both linking it with, and extending, the work of Chapter 4.

### **Chapter 7: Statistical Computing**

The final chapter presents a succinct introduction to the two R packages, ‘PhaseType’ and ‘ReliabilityTheory’, which are provided to enable reliability practitioners to immediately and easily start using the theoretical contributions presented throughout.

## **1.2 Research contributions**

Following is an overview of the contributions made by this research:

1. Detailed study of the inferential and computational characteristics of the methodology of Bladt *et al.* (2003), including the statement and proof of Theorem 4.1 and Corollary 4.2.
2. The model reformulation of §4.2.1, enabling structure to be imposed on the latent continuous-time Markov chain generator of the Phase-type model.
3. Development of the Exact Conditional Sampling (ECS) technique for simulation of an absorbing continuous-time Markov process conditional on an exact absorption time. This is Algorithm 4.1, including statement and proof of associated Lemmas 4.3, 4.4 and 4.5.
4. Extension of ECS to right censored observations in Algorithm 4.2, including statement and proof of associated Lemma 4.6 and Corollary 4.7.
5. The development of Algorithm 5.1 and proof of Theorem 5.1, for enumeration of all simply connected coherent systems and their signature. Also the development of Algorithm 5.2 to likewise enumerate all coherent networks and their signature. These lead to the useful results of Tables 5.1, 5.2, 5.3, A.2, 5.4 and A.1.
6. The MCMC approach to parametric and topological inference in Algorithm 6.2, including the signature based data augmentation of Algorithm 6.1 and statement and proof of associated Lemma 6.1.
7. The derivation of the prerequisites for implementation of Algorithm 6.2 for the model formulations in §6.3.1, §6.3.2, §6.3.3 and §6.3.4. §6.3.4 also includes extension of Chapter 4 to the weaker assumption of exchangeable systems.
8. Development and public release of the R packages ‘PhaseType’ (Aslett, 2011) and ‘ReliabilityTheory’ (Aslett, 2012) which provide easy interfaces to use the contributions of this research in real-world applications.

## Chapter 2

# Reliability Theory and Phase-type Distributions

Although some of the contributions of this thesis are generally applicable, there was a motivating application in mind which led to their development. That application relates to questions of inference surrounding physical systems which are subject to repairable (or non-repairable) component failures, but where there may be redundancies built in so that unless multiple failures occur before repair is complete there is no interruption in service.

This chapter introduces the field of reliability theory as studied by mathematicians and probabilists (as distinct from engineers who favour fault-tree style analysis), with an emphasis on structural reliability of systems. It then introduces the study of stochastic processes — in particular Continuous-time Markov Chains (CTMCs) — and relates it to the study of reliability. The chapter concludes with an exposition of Phase-type distributions which is proposed as a model in problems of inference for repairable redundant systems.

### 2.1 An introduction to reliability theory

The origins of modern reliability theory can be traced to World War II and the advent of complex military systems whose failure characteristics had to be well understood. A fascinating historical exposition of some of this pioneering work following its declassification is in Mangel and Samaniego (1984).

Reliability theory as discussed here focuses on quantifying the uncertainty surrounding the correct operation of components and systems of components using the language of probability theory, for example, by seeking to model time to failure. Although the focus in this thesis is on engineered physical systems there is overlap with biological reliability, conventionally termed ‘survival analysis’. The main distinction in research direction stems from interest in structural reliability of engineered systems: that is, reliability of systems composed from many components each with certain failure characteristics. In contrast, survival analysis traditionally treats individuals as discrete units.

Broadly speaking, there are two avenues of study pursued in the literature. The first route is a probabilistic analysis of component lifetimes, system lifetimes, and indeed matters of ‘optimal’ design of those systems — here the model and parameters are usually treated as known. The second route, given data on the failure of components or systems, is a statistical analysis to infer the parameters of the chosen model and check the adequacy of that model to describe the failure data.

This thesis is concerned primarily with the latter, although there is naturally much interplay between the areas. §3.1 provides background on Bayesian inference which is the methodology used in this thesis. A genuinely classic text concerning probabilistic aspects of reliability theory is Barlow and Proschan (1981). As noted in that work, although on the surface reliability theory appears to be merely another generic application of probability theory, there are, in fact, extensive subject-matter considerations which have led to its fruitful development as a sub-discipline in its own right.

Finally, the words ‘component’ and ‘system’ are employed in the most general sense possible. An application area of particular interest to the author is telecommunication networks: any likewise interested reader may profitably think of ‘nodes’ and ‘networks’ respectively where appropriate.

### **2.1.1 Component reliability modelling**

The term ‘component’ refers to any unit which is atomic from the perspective of the analysis, meaning that no constituent parts of the unit are modelled directly, only the unit as a whole. Note that a component in an analysis may in physical reality be a complex system.

The operational state of a component is considered to be binary: correctly functioning, or failed. Let  $Y$  denote the binary random variable for the state of a general component, with  $Y = 1$  representing correct operation and  $Y = 0$  representing failed. Such notation is conventional in the reliability literature and an elementary model is  $Y \sim \text{Bernoulli}(p)$ .

However, it is vital to bear in mind that the value  $Y$  attains will depend on the time at which the component is observed. Unless otherwise stated, in this section it should be taken as implicit that all random variables under consideration are observed at a common fixed time of no particular note. In instances where the time of observation is pertinent, or not the same for all the random variables under consideration, an extended notation is useful:

$$Y^{\{t\}} = \begin{cases} 0 & \text{if } Y \text{ is not functioning at time } t \\ 1 & \text{if } Y \text{ is operational at time } t \end{cases} \quad (2.1)$$

This makes explicit the dependence on time when necessary. By natural extension,  $Y^{[0,t)}$  denotes the state of the component on the half closed interval from time 0 to  $t$ . Indeed, in full generality  $Y^\Omega$  represents the state of the component for all  $t \in \Omega$ . Thus the aforementioned elementary model would be a Bernoulli distribution with parameter  $p$  expressing the probability that  $Y^{\{t\}} = 1$  for some fixed value of  $t$ . Hereinafter the relevant notation ( $Y, Y^{\{t\}}, Y^{[0,t)}$ , etc.) will be used as appropriate in the context.

If repair of a component is not possible then there exists some  $t$  such that  $Y^{[0,t)} = 1$  and  $Y^{[t,\infty)} = 0$ , where  $t$  is termed the *failure time* of the component. This failure time is itself a random variable,  $T$ , and is commonly modelled by some lifetime probability distribution  $F_T(t) = \mathbb{P}(T < t)$  with non-negative support  $\{t : F_T(t) > 0\} \subset [0, \infty)$ . Discretely distributed random failure times are a possibility, but not considered here. Of particular interest in a reliability context is the *survival function*, defined as:

$$\bar{F}_T(t) := \mathbb{P}(T > t) = 1 - F_T(t)$$

Commonly used lifetime probability distributions include the Exponential, Gamma, Weibull, Gumbel, Extreme Value, truncated Normal and log-Normal distributions, among others. There is also a rich literature of non-parametric reliability theory, from the classical Kaplan-Meier estimator (Kaplan and Meier, 1958) to more recent and sophisticated Nonparametric Predictive Inference (NPI) methods (Coolen *et al.*, 2002)

which allow vagueness in modelling by using upper and lower probabilities with few assumptions, though these approaches are not considered further here.

Arguably the foremost probability concept connected to reliability theory (but of only minor interest in general probability theory) is the hazard rate:

**Definition 2.1 (Hazard rate)** *For a lifetime distribution  $F_T(t)$ , the hazard rate is given by*

$$h(t) = \lim_{\delta \rightarrow 0^+} \frac{\mathbb{P}(t < T < t + \delta | T > t)}{\delta} = \frac{1}{\bar{F}_T(t)} \frac{\partial F_T(t)}{\partial t} = \frac{f_T(t)}{\bar{F}_T(t)}$$

The hazard rate is the instantaneous risk of failure and encapsulates the changing failure characteristics with time. Many physical components will have non-decreasing hazard rate: their exposure to risk of failure at any instant is constant or increasing over time due to wear. An interesting counterpoint to this is software reliability, where hazard rates are commonly non-increasing: as programming bugs are discovered and fixed the reliability tends to improve (Singpurwalla and Wilson, 1999). Probability distributions used in reliability theory which satisfy the former are termed IFR (Increasing Failure Rate) and the latter DFR (Decreasing Failure Rate) and such a simple restriction leads to often quite sharp bounds on survival probabilities.

Therefore the hazard rate is an important property of any lifetime probability distribution, encapsulating an intuitive notion of time-varying risk which one would seek to be concordant with reality when choosing a model. In particular, solving a simple differential equation shows that there is an injective relationship between a hazard rate function and a survival function:

$$\bar{F}_T(t) = \exp \left\{ - \int_0^t h(u) du \right\}$$

meaning the reliability model could even be chosen by considering the hazard rate directly and then deriving the probability distribution. There are some caveats to this approach aside from the obvious necessity of assuming absolute continuity of  $F_T$  (see Singpurwalla and Wilson, 1995)

The Exponential distribution (rate  $\lambda$ ) deserves individual mention as a special case since it has constant hazard,  $h(t) = \lambda \forall t$ . This is as might be expected from the characterising ‘memoryless’ property of the Exponential distribution and so models components which are not subject to wear or online improvement, such as electronic



components. It is surprisingly pervasive in reliability: for example, even in complex series systems with independent but not identical components, the inter-failure times converge to an Exponential distribution irrespective of whether the individual failure distributions are Exponential (Barlow and Proschan, 1965).

Since engineers in particular find the hazard an appealing concept, inference may focus simply on the hazard rate directly rather than on the parameters of the model. Indeed, the seminal work of Cox (1972) which is widely adopted in survival analysis developed proportional hazard regression models, obviating any distributional assumptions entirely to infer the multiplicative effect of explanatory variables on an unknown baseline hazard.

## 2.1.2 Inferential considerations

As already alluded to, the interest in this work is primarily on problems of inference. With a standard probability model defined for the failure time of individual components, inference on the model parameters can usually proceed in a relatively straight-forward manner. However, a prominent subject-matter concern is the data which are typical in reliability settings.

One aspect requiring care when applying reliability theory to real world data is the measurement of ‘time’. Reliability theory models depend on ‘time’ in the abstract sense of a measure of ageing, but this need not be wall-clock time. For example, equipment which is not always on may better be measured by the time in operation, and an example lending itself to an alternative measure of age would be vehicles, which may more properly be aged by distance travelled.

Another facet requiring care in application is the definition of failure. Traditionally the reliability literature considers there to be an instant in time before which the unit was operational and after which it was failed. There may be systems which can operate while severely degraded but in an out-of-specification manner, which then requires care in formally elucidating a consistent definition of failure.

A concern which has an impact at the methodological level — and is therefore of more direct interest in this thesis, especially for a contribution in Chapter 4 — is that of censoring of observations. Censored data can appear in different guises and are ubiquitous in most real world reliability datasets: simply ignoring it can lead to

seriously erroneous and biased conclusions. A compelling example can be found in the post-mortem of analysis conducted immediately preceding the launch of NASA's Challenger shuttle: test flights where no failure occurred were omitted (count censoring) leading to catastrophically poor decisions (Dalal *et al.*, 1989). Another common guise is in lifetime data when there may be right censoring of units in a limited duration experiment: it will not be uncommon for some units to survive beyond the end of the trial time, meaning it is only known that the unit would have failed at some time beyond the end of observation. Accounting for this usually takes place in the likelihood (see §3.1.1), by appropriate use of the survival function in place of the density (the details of exactly how depending on model assumptions, see §3.1.5). Similar procedures enable incorporation of left and interval censored data. All censoring in this thesis is considered to be at random (and thus uninformative censoring). Kalbfleisch and Prentice (2002) has further discussion of censoring mechanisms and their impact on the likelihood.

### 2.1.3 Structural reliability

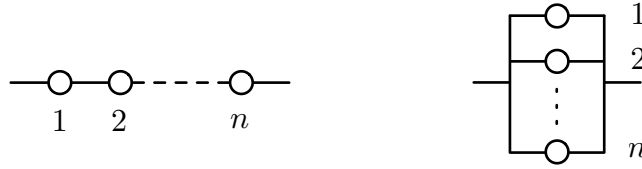
With an appropriate model for components, attention can turn to the study of systems of components, or so-called structural reliability. The pioneering work of Birnbaum *et al.* (1961) set the foundation for the study of structural reliability via the 'structure function', which is briefly covered here. It was later in Samaniego (1985) and Kochar *et al.* (1999) that the 'signature' was developed as another avenue for structural reliability research and is the foundation on which the contributions of Chapters 5 and 6 are built.

#### 2.1.3.1 System structure functions

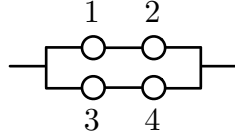
If a system of components  $Y_1, \dots, Y_n$  is formed, denote by  $\phi$  the binary random variable representing its state (or  $\phi^{\{t\}}$  akin to (2.1) if appropriate). The number of components,  $n$ , is termed the *order* of the system.

**Definition 2.2 (Structure function)** *When the state of a system is dependent only on the state of the constituent components, then the binary random variable,  $\phi$ , denoting operation of the system is a functional of the component states.*

$$\phi = \varphi(Y_1, \dots, Y_n)$$



**Figure 2.1:** A series (left) and parallel (right) system structure.



**Figure 2.2:** A simple 4 component system.

The mapping  $\varphi(\cdot) : \{0, 1\}^n \rightarrow \{0, 1\}$  is called the structure function of the system.

The structure function for series and parallel systems (Figure 2.1) is trivial. A *series* structure is one where every component must function for the system to function, whilst a *parallel* structure is one where at least one component must function for the system to function. Letting  $\mathbf{Y} = (Y_1, \dots, Y_n)$  be the state vector of the components of the system, the series system has structure function:

$$\varphi(\mathbf{Y}) = \min(Y_1, \dots, Y_n) = \prod_{i=1}^n Y_i$$

and the parallel system has structure function:

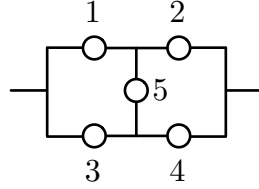
$$\varphi(\mathbf{Y}) = \max(Y_1, \dots, Y_n) = 1 - \prod_{i=1}^n (1 - Y_i)$$

Sometimes one can then derive the structure function of more complex systems by simply combining series and parallel substructures. For example, Figure 2.2 shows a series of components in parallel enabling one to immediately write down the structure function:

$$\begin{aligned} \varphi(\mathbf{Y}) &= \max(\min(Y_1, Y_2), \min(Y_3, Y_4)) \\ &= 1 - (1 - Y_1 Y_2)(1 - Y_3 Y_4) \end{aligned}$$

However, the simple bridging addition in Figure 2.3 results in a system not amenable to such decomposition and leads to a substantially more complicated structure function:

$$\begin{aligned} \varphi(\mathbf{Y}) &= Y_1 Y_2 + Y_3 Y_4 + Y_1 Y_4 Y_5 + Y_2 Y_3 Y_5 - Y_1 Y_2 Y_3 Y_5 - Y_1 Y_3 Y_4 Y_5 - Y_1 Y_2 Y_4 Y_5 \\ &\quad - Y_1 Y_2 Y_3 Y_4 - Y_2 Y_3 Y_4 Y_5 + 2Y_1 Y_2 Y_3 Y_4 Y_5 \end{aligned} \quad (2.2)$$



**Figure 2.3:** Addition of bridging component to simple 4 component system.

Note in particular that any permutation of the indices represents the same system with permuted labels for the components making immediate identification of a system via the structure function challenging.

### 2.1.3.2 Path and cut sets

Derivation of (2.2) may seem mysterious at this juncture. This short subsection defines path and cut sets and their link to the structure function.

**Definition 2.3 ((Minimal) path set)** *A set of components,  $P$ , of a system is said to be a path set if the system functions correctly whenever all the components in  $P$  function correctly.*

*If no proper subset of  $P$  is a path set, then  $P$  is said to be a minimal path set.*

**Lemma 2.1** *Let  $\mathcal{P}$  be the collection of all minimal path sets of a system. If  $X$  is the set of currently operational components, then the system as a whole is operational if and only if  $\exists P_i \in \mathcal{P}$  s.t.  $P_i \subseteq X$*

For example, the path sets of the system in Figure 2.2 are:

$$\{1, 2\} \quad \{3, 4\} \quad \{1, 2, 3\} \quad \{1, 2, 4\} \quad \{1, 3, 4\} \quad \{2, 3, 4\} \quad \{1, 2, 3, 4\}$$

However, only  $\{1, 2\}$  and  $\{3, 4\}$  are minimal path sets. The more complex bridge system of Figure 2.3 has minimal path sets  $\{1, 2\}$ ,  $\{3, 4\}$ ,  $\{1, 4, 5\}$  and  $\{2, 3, 5\}$ .

**Definition 2.4 ((Minimal) cut set)** *A set of components,  $C$ , of a system is said to be a cut set if the system is failed whenever all the components in  $C$  have failed.*

*If no proper subset of  $C$  is a cut set, then  $C$  is said to be a minimal cut set.*

**Lemma 2.2** *Let  $\mathcal{C}$  be the collection of all minimal cut sets of a system. If  $X$  is the set of currently failed components, then the system as a whole is failed if and only if  $\exists C_i \in \mathcal{C}$  s.t.  $C_i \subseteq X$*

For example, the cut sets of the system in Figure 2.2 are:

$$\{1, 3\} \quad \{1, 4\} \quad \{2, 3\} \quad \{2, 4\} \quad \{1, 2, 3\} \quad \{1, 2, 4\} \quad \{1, 3, 4\} \quad \{2, 3, 4\} \quad \{1, 2, 3, 4\}$$

However, only  $\{1, 3\}$ ,  $\{1, 4\}$ ,  $\{2, 3\}$  and  $\{2, 4\}$  are minimal cut sets. The more complex bridge system of Figure 2.3 has minimal cut sets  $\{1, 3\}$ ,  $\{2, 4\}$ ,  $\{1, 4, 5\}$  and  $\{2, 3, 5\}$ .

This leads to the well known result:

**Theorem 2.3 (Barlow and Proschan, 1981, p.12)** *The structure function of a system with collection of all minimal path sets  $P_1, \dots, P_r$  and collection of all minimal cut sets  $C_1, \dots, C_s$  can be expressed in terms of the components of either:*

$$\begin{aligned} \varphi(\mathbf{Y}) &= 1 - \prod_{j=1}^r \left( 1 - \prod_{i \in P_j} Y_i \right) \\ &= \prod_{j=1}^s \left( 1 - \prod_{i \in C_j} (1 - Y_i) \right) \end{aligned}$$

The interested reader can confirm that using the minimal path and cut sets above for Figure 2.3 together with this Theorem does indeed result in equation (2.2). Note that  $Y_i^2 = Y_i \forall i$ .

### 2.1.3.3 System reliability

The structure function enables statements of probability to be made about the system with knowledge of the component lifetime distributions. The extended notation from §2.1.1 is used here.

Let  $T_i$  be the random variable denoting the failure time of component  $Y_i$  and  $\tau$  be the random variable denoting the failure time of the system  $\phi$ . Then:

$$\mathbb{P}(Y_i^{\{t\}} = 1) = \mathbb{E}[Y_i^{\{t\}}] \quad \text{and} \quad \mathbb{P}(Y_i^{\{t\}} = 1) = \mathbb{P}(T_i > t) = \bar{F}_{T_i}(t)$$

for each component. Thus, if components are assumed to be statistically independent then for the system as a whole:

$$\begin{aligned} \mathbb{P}(\phi^{\{t\}} = 1) &= \mathbb{P}(\varphi(\mathbf{Y}^{\{t\}}) = 1) \\ &= \mathbb{E}[\varphi(\mathbf{Y}^{\{t\}})] \\ &= \varphi(\mathbb{E}[Y_1^{\{t\}}], \dots, \mathbb{E}[Y_n^{\{t\}}]) \\ &= \varphi(\bar{F}_{T_1}(t), \dots, \bar{F}_{T_n}(t)) =: h_\varphi(\cdot) \end{aligned}$$

The function  $h_\varphi(\cdot)$  defined above is termed the *reliability function* associated with structure function  $\varphi$ . In particular, this provides the system survival function and hence lifetime distribution of the system.

$$\bar{F}_\tau(t) = h_\varphi(\mathbf{Y}^{\{t\}}) \implies F_\tau(t) = 1 - h_\varphi(\mathbf{Y}^{\{t\}}) \quad (2.3)$$

The lifetime density is then available if  $F_\tau(\cdot)$  is absolutely continuous. For example, with the bridge system in Figure 2.3 and independent and identically distributed (i.i.d.) Exponential( $\lambda$ ) components, (2.2) renders:

$$\begin{aligned} \bar{F}_\tau(t) &= 2e^{-5\lambda t} - 5e^{-4\lambda t} + 2e^{-3\lambda t} + 2e^{-2\lambda t} \\ f_\tau(t) &= 10\lambda e^{-5\lambda t} - 20\lambda e^{-4\lambda t} + 6\lambda e^{-3\lambda t} + 4\lambda e^{-2\lambda t} \end{aligned} \quad (2.4)$$

#### 2.1.3.4 Coherent systems

The class of all systems is commonly restricted to so-called *coherent* systems, though said restriction is not unduly confining since it is in harmony with what one would intuitively consider a reasonable design for a system.

The first requirement is that all components in the system should have some impact on the working of the system. It would be unusual to include components which are entirely extraneous to the correct operation of the system, so this is not limiting.

**Definition 2.5 (Relevant component)** *Consider a system of order  $n$  with state vector of components  $(Y_1, \dots, Y_{i-1}, X, Y_{i+1}, \dots, Y_n)$ . The  $i$ th component  $X$  is said to be irrelevant if:*

$$\varphi(Y_1, \dots, Y_{i-1}, 0, Y_{i+1}, \dots, Y_n) = \varphi(Y_1, \dots, Y_{i-1}, 1, Y_{i+1}, \dots, Y_n)$$

for all possible realisations of  $(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) \in \{0, 1\}^{n-1}$ .

*If a component is not irrelevant, it is defined to be a relevant component.*

The second requirement which is placed on systems commonly studied in reliability theory is that of monotonicity of the structure function. Intuitively, monotonicity excludes the possibility that successfully repairing some component of a system could result in a working system suddenly failing. Again this is compatible with what seems a reasonable design for a system.

**Definition 2.6 (Monotone structure function)** *The structure function  $\varphi(\cdot)$  of an order  $n$  system is said to be monotone if*

$$\mathbf{X} \leq \mathbf{Y} \implies \varphi(\mathbf{X}) \leq \varphi(\mathbf{Y})$$

where  $\mathbf{X}, \mathbf{Y} \in \{0, 1\}^n$  and the inequality on the left is taken element-wise.

In particular, this means that when repair of components is impossible, it must be the case that  $\varphi(\mathbf{Y}^{\{t+\varepsilon\}}) \leq \varphi(\mathbf{Y}^{\{t\}}) \forall t \in [0, \infty), \varepsilon > 0$ . In other words, a coherent system with non-repairable components cannot recover from failure (especially not due to further failure of components) — a logical requirement.

These concepts lead to the formal definition of system coherency.

**Definition 2.7 (Coherent system)** *A system is coherent if and only if the structure function representing the system is monotone and every component is relevant.*

For a given number of components, the requirement for coherency dramatically limits how many structure functions represent admissible systems. For example, when  $n = 3$ , there are  $2^{2^3} = 256$  mappings  $\varphi : \{0, 1\}^3 \rightarrow \{0, 1\}$ , but from Barlow and Proschan (1981) there are only 5 coherent systems of order 3. The following Theorem then allows an entirely set theoretic characterisation of coherent systems.

**Theorem 2.4 (Butterworth, 1972)** *If a coherent system  $\phi$  has collection of all minimal cut sets  $\mathcal{C}$ , then  $\bigcup_{C_i \in \mathcal{C}} C_i$  is the set of all components in  $\phi$ .*

*Conversely, any collection of sets  $\mathcal{C}$  such that  $C_i \not\subset C_j, \forall C_i, C_j \in \mathcal{C}$  defines a collection of minimal cut sets of some coherent system comprising the components  $\bigcup_{C_i \in \mathcal{C}} C_i$ .*

Note that all instances of ‘cut set’ can be replaced by ‘path set’ in Theorem 2.4 without affecting the correctness of the statement.

### 2.1.3.5 System signature

A more recently discovered additional tool for the study of coherent systems with i.i.d. lifetimes is the system signature, providing an elegantly simple representation of a system.

**Definition 2.8 (System signature)** Consider a coherent system of order  $n$ , with independent and identically distributed component lifetimes. The signature of the system is the  $n$ -dimensional probability vector  $\mathbf{s} = (s_1, \dots, s_n)$  with elements:

$$s_i = \mathbb{P}(\tau = T_{i:n})$$

where  $\tau$  is the failure time of the system and  $T_{i:n}$  is the  $i$ th order statistic of the  $n$  component failure times.

Note that the failure time of a coherent system must coincide with the failure of a component of the system. The  $i$ th element of the signature is the probability that it was the  $i$ th component failure which resulted in the system failure.

Computing the signature of a specific system is an exercise in combinatorial mathematics. The series system and parallel systems can be immediately written down as  $\mathbf{s} = (1, 0, \dots, 0)$  and  $\mathbf{s} = (0, \dots, 0, 1)$  respectively: the series system always fails when the first component does and the parallel system always fails when the last component does.

In less trivial cases the signature can be found by considering all permutations of the order in which the  $n$  components may fail. Since the  $T_i$  are considered i.i.d., each of the  $n!$  permutations for order of failure is equally likely. The signature is then the relative frequency with which the  $i$ th ordered failure caused system failure across all permutations.

For example, in the bridge system (Figure 2.3), the permutation  $(3, 4, 2, 1, 5)$  means component 3 failed first, followed by component 4 and so on, so that  $T_3 < T_4 < T_2 < T_1 < T_5$ . Under this ordering, the system would fail at  $\tau = T_2 = T_{3:5}$ . To see this, notice that  $\{3\}$  and  $\{3, 4\}$  do not contain any cut sets, but  $\{3, 4, 2\} \supset \{2, 4\}$ , meaning that the system continued to function until component 2 failed (the third failure). Checking all  $5! = 120$  permutations (which takes less than a second on a computer) and counting the order statistic of failure in each case leads to the signature by a simple tabulation of relative frequencies:

$$\mathbf{s} = \left( \frac{0}{120}, \frac{24}{120}, \frac{72}{120}, \frac{24}{120}, \frac{0}{120} \right) = (0, 0.2, 0.6, 0.2, 0) \quad (2.5)$$

Thus, both structure function and system signature require computation of the cut sets, but the signature thereafter involves simple permutation checks, whilst the structure function involves laborious (albeit elementary) algebra. This means the calculation



of signature is more readily automated on a computer. The author has developed R (R Core Team, 2012) code for calculating the signature of an arbitrary system, available in Aslett (2012), which is a minor contribution of this thesis.

A related problem is computation of all possible signatures for coherent systems of a given order: Navarro and Rubio (2009) is recent work on an algorithmic approach to exhaustively computing this and Chapter 5 of this thesis also makes a related contribution in this area.

With the signature computed, an analogous result to (2.3) enables expression of the system lifetime distribution purely via the lifetime distribution of the components.

**Theorem 2.5 (Samaniego, 1985)** *Let  $T_1, \dots, T_n \sim T$  be the i.i.d. component lifetimes of an order  $n$  coherent system with signature  $\mathbf{s}$ . Let  $\tau$  be the system lifetime. Then,*

$$\bar{F}_\tau(t) := \mathbb{P}(\tau > t) = \sum_{i=1}^n s_i \sum_{j=0}^{i-1} \binom{n}{j} F_T(t)^j \bar{F}_T(t)^{n-j} \quad (2.6)$$

The interested reader can confirm that using the signature in (2.5) together with (2.6) and letting  $T \sim \text{Exponential}(\lambda)$  gives an identical result to that obtained via the structure function (2.4).

**Corollary 2.6 (Samaniego, 2007)** *Let  $T_1, \dots, T_n \sim T$  be the i.i.d. component lifetimes of an order  $n$  coherent system with signature  $\mathbf{s}$ . Let  $\tau$  be the system lifetime. If  $F_T(\cdot)$  is absolutely continuous then,*

$$f_\tau(t) := -(\partial/\partial t)\mathbb{P}(\tau > t) = \sum_{i=1}^n i s_i \binom{n}{i} F_T(t)^{i-1} \bar{F}_T(t)^{n-i} f_T(t) \quad (2.7)$$

### 2.1.3.6 Summary

In summary, the key well-known ideas expounded in this section about structural reliability have built up the results necessary for examining the lifetime distributions of systems of components, based on lifetime distributions for the components themselves.

This is a carefully chosen path for the purposes of this thesis through what is a tremendously rich subject area. Introductions to the extensive other uses and results relating to structure functions can be found in Barlow and Proschan (1981) and similarly for system signatures in Samaniego (2007).

### 2.1.4 Masked system lifetime inference

The most common inferential situation with reliability data is to have lifetimes (or censoring information) on components and systems under examination. However, there is an important class of problems where the information available for statistical analysis is more limited. One such situation — a situation which preoccupies this thesis — is that of so-called *masked system lifetime data*. This is data where one commonly has failure times for the system under examination, but no data on the component lifetimes directly.

Such data can arise in a variety of applications. For example, repair of a complex system may involve identifying which major subsystem has failed and replacing it entirely, so that the subsystem lifetime is known, but not the lifetimes of the components within it or even which components had failed, unless an autopsy is performed. Such an autopsy may be impractical or cost prohibitive relative to the repair, such as in consumer computer repair. Another example would be where many systems are sent into the field, but upon failure the system may be discarded and only failure time of the system as a whole reported.

The inference problem with data of this kind has been considered before. A rich literature (Reiser *et al.*, 1995; Kuo and Yang, 2000, among others) has investigated Bayesian inference in competing risk, or essentially series systems. There has been less work on inference for generally structured systems, the most notable exception being in Ng *et al.* (2012), where frequentist inference is performed on the hazard rate of the components. Gåsemyr and Natvig (2001) provide a Bayesian view by deriving likelihoods and conjugate priors for Exponentially distributed components directly.

The problems addressed in Chapters 4 and 6 relate to data of this kind.

### 2.1.5 Network reliability

In the probability and statistics literature the term network is usually used to indicate a structural reliability setting where the links between components, as opposed to the components themselves, are unreliable. Most of §2.1.3 can be developed from this viewpoint and there is a very natural duality in all the results, so it is not repeated here. Chapter 5 discusses the distinction at greater length.

### 2.1.6 Repairable systems

Up to this juncture, consideration has implicitly only been given to components which are not repairable. In this way, the system always moves toward a more degraded state until ultimately it fails once there is no longer a functioning path through the system.

The next section introduces the tools to enable study of repairable redundant systems, whereby repair can be non-disruptively performed on a system and Chapter 4 focuses entirely on inferential challenges in such repairable model settings.

## 2.2 Stochastic processes

The methods described in the previous section are a common approach in reliability theory. It focuses attention on either the random operational status at a point-in-time  $(Y^{\{t\}}, \phi^{\{t\}})$  or else focuses on the random failure time  $(T, \tau)$ .

When considering systems — in particular systems whose components are repairable — there is an expanded view which makes more explicit the fact that the state of the system is changing over time: that is to acknowledge that the state of the system,  $\phi$ , is a *stochastic process*.

**Definition 2.9 (Stochastic process)** *A stochastic process is a set of random variables  $\{\Phi^{(t)} : t \in T\}$ , where  $T$  denotes the index set. The state space of a stochastic process is the space of realisable values of  $\Phi^{(t)}$ .*

Commonly  $T$  will be a spatial, time or spatio-temporal index. Only time indices are considered herein. Thus in the reliability setting of this chapter, the extended notation,  $\phi^{\{t\}}$  (or  $Y^{\{t\}}$ ), denotes the state of the stochastic process representing the operational condition of a system (or component) at time  $t$ , where time is now allowed to vary. Hereinafter, when modelling reliability this thesis will take the state space of the stochastic process to be discrete and finite (denoted  $\Omega$ ) and the time index space to be  $[0, \infty)$ .

Rather than contemplating only overall failure and operation ( $\{0, 1\}$ ) as the system states, the operational state of the constituent components may lead to a multi-state picture. There are different ways to formulate this. For now, consider the conceptually simplest formulation: take all  $2^n$  combinations of possible component states and add

System State	Component States					System State	Component States				
	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$		$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$
1	1	1	1	1	1	10	1	0	1	1	0
2	1	1	1	1	0	11	1	0	0	1	1
3	1	1	1	0	1	12	0	1	1	1	1
4	1	1	1	0	0	13	0	1	1	1	0
5	1	1	0	1	1	14	0	1	1	0	1
6	1	1	0	1	0	15	0	0	1	1	1
7	1	1	0	0	1	16	0	0	1	1	0
8	1	1	0	0	0	17–32	<i>system failed states</i>				
9	1	0	1	1	1						

**Table 2.1:** A labelling of all combinations of component states for which system is operational (1–16) and various failure states (17–32) for bridge system (see Figure 2.3).

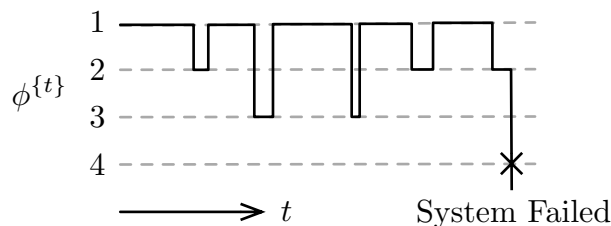
one numbered system state for each combination. For reasons that will become clear in §2.2.2 the combinations where the system is failed are not separately identified but rather pooled together.

For example, a possible labelling for the bridge system of Figure 2.3 is shown in Table 2.1. Now  $\phi^{\{t\}} \in \{1, 2, \dots, 32\} \forall t \in [0, \infty)$  and moves between states are determined by the underlying component failures. Without repair the chosen labelling implies  $\phi^{\{t+\varepsilon\}} \geq \phi^{\{t\}} \forall t \in [0, \infty), \varepsilon > 0$ . When component repair is introduced, the stochastic process can also make ‘backward’ moves to previous states.

To reduce deforestation and simplify exposition, hereinafter the example used to illustrate the concepts will be an elementary two component parallel system with labelling as in Table 2.2, although the reader may hold in mind the bridge system as a more complex scenario. The general evolution of the stochastic process can be visualised as in Figure 2.4. The first state change represents a failure of component  $Y_2$ , and the subsequent return to state 1 represents its repair. In such a system, there would be correct function up until the ultimate move to state 4. This represents the system in its fullest generality, with random failure times and random repair times. From a modelling perspective, this means that in addition to component lifetime distributions, one must now also specify component repair time distributions to represent the random

System State	Component States	
	$Y_1$	$Y_2$
1	1	1
2	1	0
3	0	1
4	0	0

**Table 2.2:** A labelling for the elementary two component parallel system.



**Figure 2.4:** Example realisation of stochastic process for a two component parallel system.

time required to repair or replace a component to correct operation.

An essential point to note is that there is commonly a restriction on which states the stochastic process may move between. In the current example, a move  $2 \rightarrow 3$  or  $3 \rightarrow 2$  would not be possible: it represents simultaneous failure of one component and repair of the other, which would require two continuous random variables (failure time and repair time) to attain identical values, an event of probability measure zero.

### 2.2.1 Continuous-time Markov chains

There are two simplifying assumptions which make stochastic processes amenable to analysis. These assumptions are that the stochastic process satisfies the so-called *Markov property* and *homogeneity*.

**Definition 2.10 (Markov property)** *A stochastic process satisfies the Markov property if*

$$\mathbb{P}(\phi^{\{t_n\}} = i_n \mid \phi^{\{t_1\}} = i_1, \dots, \phi^{\{t_{n-1}\}} = i_{n-1}) = \mathbb{P}(\phi^{\{t_n\}} = i_n \mid \phi^{\{t_{n-1}\}} = i_{n-1})$$

for all  $i_1, \dots, i_{n-1}, i_n \in \Omega$  and any  $t_1 < t_2 < \dots < t_n \in [0, \infty)$ .

**Definition 2.11 (Homogeneous process)** *Let the transition probability be defined as:*

$$p_{ij}(s, t) := \mathbb{P}(\phi^{\{t\}} = j \mid \phi^{\{s\}} = i) \quad \text{where } s \leq t$$

*Then  $\phi$  is a homogeneous process if  $p_{ij}(s, t) = p_{ij}(0, t - s) \forall i, j \in \Omega, s \leq t \in [0, \infty)$ .*

*The transition probability is then written simply  $p_{ij}(t - s)$ .*

The Markov property simplifies affairs greatly, since subsequent moves in the state space depend only upon the most recently observed state and not upon the entire history of the process. In a similar vein, homogeneity means that state move probabilities after a specific sojourn time are independent of the total elapsed time up to that point.

**Definition 2.12 (Continuous-time Markov chain)** *A stochastic process with time indexed by the set  $[0, \infty)$  is called a continuous-time Markov chain (CTMC) if it is homogeneous and satisfies the Markov property.*

There is an important property which flows from this:

**Theorem 2.7 (Brémaud (1999), p.346)** *A continuous-time Markov chain satisfies the strong Markov property. That is, conditional on knowing the state of the chain at some stopping time  $t$ ,  $\phi^{\{t\}} = k$  say, the chain before time  $t$  and the chain after time  $t$  are independent.*

A *stopping time* is defined rigorously in Brémaud (1999). For the purposes of this thesis, the strong Markov property is only invoked for times at which transition to a state occurs (or ‘jump times’) which are in fact stopping times.

Further detailed treatment is omitted for brevity, but it is important to note that the characterisation in the remainder of this subsection is not a prescriptive definition, but rather is one of several equivalent views of continuous-time Markov chains which follow from the Markov and homogeneity assumptions. The interested reader should consult Grimmett and Stirzaker (2001) or Brémaud (1999) for the proof of results which form the narrative elucidated here. The objective is to maintain focus on an intuitive link to the subject-matter of reliability in this chapter.

### 2.2.1.1 Formulating a CTMC model

Continuous-time Markov chains are stochastic processes whose state sojourn times are Exponentially distributed, with fixed state move probabilities. Further, the sojourn time may be viewed as the minimum of competing Exponentially distributed ‘clocks’, one for each permissible move. In other words, given that the current state is  $i$  and given an Exponentially distributed timer for each permissible move  $i \rightarrow j, i \neq j$ , the time until any move is the Exponentially distributed minimum of the collection<sup>1</sup>, the corresponding minimal timer being the state to which the process moves.

Therefore, one strategy for defining a continuous-time Markov chain is to define the (Exponential) rate at which each state moves to each other state, the state sojourn then being automatically determined.

This finds natural expression in the  $m \times m$  square *generator matrix* traditionally used to represent an  $m$  state continuous-time Markov chain:

$$\mathbf{G} = \begin{pmatrix} -\lambda_1 & \lambda_{12} & \lambda_{13} & \cdots & \lambda_{1m} \\ \lambda_{21} & -\lambda_2 & \lambda_{23} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \lambda_{m3} & \cdots & -\lambda_m \end{pmatrix} \quad (2.8)$$

where  $\lambda_{ij}$  is the rate of the Exponentially distributed timer for the state move  $i \rightarrow j$  and  $\lambda_i := \sum_{j \in \Omega, j \neq i} \lambda_{ij}$  is the rate of the Exponentially distributed sojourn in state  $i$ . Note that  $\sum_{j=1}^m G_{ij} = 0 \forall i$ . The special case  $\lambda_{ij} = 0$  implies that a move  $i \rightarrow j$  is impossible.

One may think of rows representing the current state and columns as candidate moves. Conceptually, if the process starts in state 1, then the collection of ‘timers’  $\{X_i \sim \text{Exp}(\lambda_{1i})\}_{i=2}^m$  begin. Then if  $x_j$  is the minimal realisation of the timers, a move  $i \rightarrow j$  is inserted at time  $x_j$ . Again a new set of timers corresponding to state  $j$  are started to determine the next move.

The model formulation is completed by defining the starting distribution of states at  $t = 0$  by a stochastic vector  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)^T$  with  $\sum_{i=1}^m \pi_i = 1$ .

<sup>1</sup>Recall the elementary result:

$$X_1 \sim \text{Exp}(\lambda_1), \dots, X_n \sim \text{Exp}(\lambda_n) \implies \min(X_1, \dots, X_n) \sim \text{Exp}(\sum_{i=1}^n \lambda_i)$$

### 2.2.1.2 Relationship to reliability modelling

A continuous-time Markov chain provides a natural model of a repairable system when the component lifetime and repair time distributions are Exponential. Consider the two component parallel system. As per Table 2.2, there are 4 possible states leading to a  $4 \times 4$  generator matrix. If the lifetime of a component is Exponential( $\lambda_f$ ) distributed and the repair time of a component is Exponential( $\lambda_r$ ) distributed, then it is claimed that a continuous-time Markov chain models this system, with generator matrix:

$$\mathbf{G} = \begin{pmatrix} -2\lambda_f & \lambda_f & \lambda_f & 0 \\ \lambda_r & -\lambda_f - \lambda_r & 0 & \lambda_f \\ \lambda_r & 0 & -\lambda_f - \lambda_r & \lambda_f \\ 0 & \lambda_r & \lambda_r & -2\lambda_r \end{pmatrix} \quad (2.9)$$

If the system always commences operation in fully functioning order, then one may simply specify  $\boldsymbol{\pi} = (1, 0, 0, 0)^T$ . Note in particular that state moves which require more than one simultaneous failure or repair have rate set to 0, since it represents the situation where two Exponentially distributed random variables attain identical values, which has probability measure zero.

The claim may be seen to be true as a consequence of the memoryless property of the Exponential distribution: the individual component failure timers restarting on every state move does not affect the overall component lifetime distribution. More precisely, given that the time has reached  $t > 0$ , the residual lifetime of an Exponentially distributed component is still Exponentially distributed with the same parameter.

Thus in general, a continuous-time Markov chain model can be established by having a chain state for each system state, with Exponential rates between those states which are adjacent in the sense of differing by at most one component state.

### 2.2.1.3 Transition probabilities

The generator matrix provides a means of explicitly computing the transition probabilities. These are represented by a matrix  $\mathbf{P}_t$  in which are entries  $p_{ij}(t) = \mathbb{P}(\phi^{\{s+t\}} = j | \phi^{\{s\}} = i)$  for arbitrary  $s \geq 0$ . Note that  $\phi^{\{s,s+t\}}$  is not necessarily identically  $i$ : multiple state moves may occur in that interval.



**Lemma 2.8** *If a continuous-time Markov chain is represented by a generator  $\mathbf{G}$ , then the matrix of transition probabilities is:*

$$\mathbf{P}_t = e^{t\mathbf{G}}$$

$e^{t\mathbf{G}}$  is the *matrix exponential*, defined via the usual series expansion of the exponential function:

$$e^{t\mathbf{G}} = \sum_{n=0}^{\infty} \frac{t^n \mathbf{G}^n}{n!}$$

Although this constitutes the definition of the matrix exponential, it is not prudent to attempt calculating it directly in this manner, taking powers of matrices. The humorously titled paper by Moler and Van Loan (2003) provides detailed insights into the issues which are critical in implementation.

#### 2.2.1.4 Limiting and stationary distributions

An interesting property (if it exists) of a continuous-time Markov chain is the so-called *stationary distribution*, which represents a chain in a stochastic steady state.

**Definition 2.13 (Stationary distribution)** *A stochastic vector  $\boldsymbol{\beta}$  is the stationary distribution of a continuous-time Markov chain with transition probability matrix  $\mathbf{P}_t$  if  $\boldsymbol{\beta}^T \mathbf{P}_t = \boldsymbol{\beta}^T \forall t > 0$ .*

Thus, if the chain is randomly started according to the stochastic vector  $\boldsymbol{\beta}$ , then the distribution of states is unchanged at all future times.

**Corollary 2.9** *A stochastic vector  $\boldsymbol{\beta}$  is the stationary distribution of a continuous-time Markov chain with generator  $\mathbf{G}$  if  $\boldsymbol{\beta}^T \mathbf{G} = \mathbf{0}$ .*

For example, solving the equation from Corollary 2.9 for  $\boldsymbol{\beta}$  shows that the stationary distribution of the two component parallel system is

$$\left( \frac{\lambda_r^2}{(\lambda_f + \lambda_r)^2}, \frac{\lambda_f \lambda_r}{(\lambda_f + \lambda_r)^2}, \frac{\lambda_f \lambda_r}{(\lambda_f + \lambda_r)^2}, \frac{\lambda_f^2}{(\lambda_f + \lambda_r)^2} \right)$$

The technical requirement for the existence of a stationary distribution is that the chain be *irreducible*.

**Definition 2.14 (Irreducible chain)** *A continuous-time Markov chain is irreducible if and only if every state is accessible from every other state. Equivalently, there must be a  $t > 0$  such that every element of  $\mathbf{P}_t$  is strictly positive.*

Another related natural question arises of the long-run behaviour of chains when not necessarily started according to the stationary distribution ( $\boldsymbol{\pi} \neq \boldsymbol{\beta}$ ). In subject-matter terms: if a system is in the field for an extended duration and undergoes failure and repair as described by a continuous-time Markov chain, what can be said about the proportion of time the system is operational, degraded and failed?

**Lemma 2.10** *Given an irreducible continuous-time Markov chain with stationary distribution  $\boldsymbol{\beta}$ , the stationary distribution is unique and  $\lim_{t \rightarrow \infty} p_i(t) = \boldsymbol{\beta} \forall i$*

Thus, irrespective of starting state, in the limit the distribution of states converges to the (unique) stationary distribution of the chain. In the two component parallel system, this leads to the intuitively obvious result that in the long-run the system will have higher uptime when the rate of repair is substantially larger than the rate of failure. In particular, one can quantify that to achieve on average “five nines” uptime<sup>2</sup> from such a system in the long run, the rate of repair must be at least 316 times the rate of failure, or roughly speaking components with failure rate  $1\text{yr}^{-1}$  would need to be repaired at a rate  $\approx 1\text{day}^{-1}$ .

### 2.2.1.5 Simulation of a continuous-time Markov chain

The characterisation described above makes simulation easy, since it is descriptive of the probabilistic process. However, rather than having to simulate up to  $m - 1$  Exponential random variables for each state move, there is the following algorithm which simplifies matters:

**Algorithm 2.1** *To simulate a continuous-time Markov chain with generator  $\mathbf{G}$  as in (2.8) and starting distribution  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)^T$ :*

1. Let  $t = 0$ , draw the starting state,  $i$ , from the probability vector  $\boldsymbol{\pi}$  and set  $\phi^{\{0\}} = i$ .
2. Draw the sojourn time,  $s$ , in state  $i$  from an Exponential distribution with rate  $\lambda_i$  and set  $\phi^{[t, t+s)} = i$ .

---

<sup>2</sup>A reliability engineering term meaning 99.999% availability over a given period of time. For example, five nines uptime in a year requires a system to be unavailable for no more than a total of 5 minutes and 15 seconds over a one year period.

3. Draw the state move,  $j$ , from the discrete distribution:

$$\mathbb{P}(j) = \frac{\lambda_{ij}}{\lambda_i}, \quad j \neq i$$

4. Set  $t = t + s$  and set  $i = j$ , then loop to step 2.

The algorithm should be looped for as long as necessary and  $\phi$  will then be a realisation of the stochastic process.  $\square$

Generation of Exponentially and discretely distributed random variates is a well understood problem, an excellent reference on random variate generation being Devroye (1986).

#### 2.2.1.6 Absorbing states

If there exists some  $i$  such that  $\lambda_{ij} = 0 \forall j$ , then the state  $i$  is said to be *absorbing*. If the process enters this state at time  $t$ , then necessarily  $\phi^{[t,\infty)} = i$  — that is, once entered an absorbing state can never be left. In particular, note that this means any chain with at least one absorbing state is not irreducible and thus there is not necessarily any stationary or limiting distribution. If there is exactly one absorbing state (which is accessible from at least one of the other states) and the subgenerator of non-absorbing states is irreducible, then the limiting distribution is trivially the absorbing state with probability 1.

Absorbing states can be artificially introduced to aid analysis of a model if a particular state or set of states is of interest. This idea leads to the development of Phase-type distributions so deeper discussion follows in the sequel.

#### 2.2.1.7 Summary

This subsection has provided a high-level description of continuous-time Markov chains and how they can provide a natural model for repairable systems when component lifetimes and repair times are Exponentially distributed. Also discussed were important properties and a means of simulation of such processes.

For a deeper treatment of the theory (without link to the subject matter), see Brémaud (1999) and Grimmett and Stirzaker (2001).

## 2.2.2 Phase-type distributions

As mentioned, absorbing states can be artificially introduced into continuous-time Markov chains to aid analysis of the model when a particular state or set of states is of interest. For example, in a reliability context it may be desirable to have an absorbing state represent all component combinations for which the system is failed. This is artificial in the sense that once a component failure causes system failure, the model implies no repair is ever attempted. However, it then enables one to answer questions such as: what is the probability the component has failed by time  $t$ ?

Taking the (repairable) two component parallel system as an example, one may replace the final row of the generator matrix in (2.9) with all zeros, resulting in a chain with one absorbing state (the state representing system failure), as follows:

$$\mathbf{G} = \begin{pmatrix} -2\lambda_f & \lambda_f & \lambda_f & 0 \\ \lambda_r & -\lambda_f - \lambda_r & 0 & \lambda_f \\ \lambda_r & 0 & -\lambda_f - \lambda_r & \lambda_f \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (2.10)$$

Then, using Lemma 2.8 the probability that the system has experienced a failure by time  $t$  is simply:

$$\begin{aligned} \mathbb{P}(\phi^{\{t\}} = 4) &= \mathbb{P}\left(\bigcup_{i=1}^4 [\phi^{\{0\}} = i \cap \phi^{\{t\}} = 4]\right) \\ &= \sum_{i=1}^4 \mathbb{P}(\phi^{\{0\}} = i) \mathbb{P}(\phi^{\{t\}} = 4 \mid \phi^{\{0\}} = i) \\ &= \sum_{i=1}^4 \pi_i p_{i4}(t) \end{aligned}$$

This is simply the fourth element of the vector formed from the product  $\boldsymbol{\pi}^T e^{t\mathbf{G}}$ , which affords some intuition for the formal definition of a Phase-type distribution.

**Definition 2.15 (Phase-type distribution)** *Take any continuous-time Markov chain on a finite discrete state space of  $n+1$  states, one of which is absorbing. Without loss of generality, the generator of the Markov chain can be expressed as:*

$$\mathbf{G} = \begin{pmatrix} \mathbf{S} & \mathbf{s} \\ \mathbf{0}^T & 0 \end{pmatrix} \quad (2.11)$$

where  $\mathbf{S} = (S_{ij})$  is the matrix of transition rates between non-absorbing states  $i$  and  $j$  for  $i \neq j$  and  $i, j \in \{1, \dots, n\}$ , whilst  $\mathbf{s} = (s_1, \dots, s_n)^\top$  is the vector of transition rates from state  $i$  to the absorbing state and  $\mathbf{0}$  is a vector of  $n$  zeros.

A (continuous) Phase-type distribution (PHT) is defined to be the distribution of the time to entering the absorbing state of a continuous-time Markov chain with generator  $\mathbf{G}$  and vector of initial state probabilities  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$ .

The distribution and density functions are traditionally expressed in terms of the sub-generator of the rates between non-absorbing states and the vector of exit rates:

$$T \sim \text{PHT}(\boldsymbol{\pi}, \mathbf{G}) \implies \begin{cases} F_T(t) = 1 - \boldsymbol{\pi}^\top \exp(t\mathbf{S})\mathbf{e} \\ f_T(t) = \boldsymbol{\pi}^\top \exp(t\mathbf{S})\mathbf{s} \end{cases}$$

where  $\mathbf{e}$  is a vector of 1's of the appropriate dimension;  $t \in [0, \infty)$  is the time to absorption; and  $\exp(t\mathbf{S})$  is the matrix exponential.

Note in particular that because  $\mathbf{G}$  has a final row of zeros, the top left  $n \times n$  block of  $\mathbf{G}^k$  is identically  $\mathbf{S}^k$  for all  $k \geq 0$ . Together with the fact that the row sums of  $e^{t\mathbf{G}}$  are 1, this means  $1 - \boldsymbol{\pi}^\top \exp(t\mathbf{S})\mathbf{e} = \boldsymbol{\pi}^\top \exp(t\mathbf{G})(0, \dots, 0, 1)^\top$ .

Phase-type distributions were first introduced in Neuts (1975). They are closely related to the retrospectively named Coxian Phase-type distribution (Cox, 1955) which is a structured Phase-type distribution with non-zero entries only on the diagonal and upper diagonal. Indeed all acyclic Phase-type distributions (those with upper-triangular generators) have a unique Coxian representation (see Dehon and Latouche, 1982; Cumani, 1982, the latter being a reliability perspective). The second chapter of Neuts (1994) and the paper by Asmussen (2000) contain a general background.

### 2.2.2.1 Random variate generation

Generation of Phase-type random variates can, in principle, use Algorithm 2.1 as a base: one simply uses the full generator  $\mathbf{G}$  in simulation until state  $n + 1$  is entered, the time of such a transition then being a realisation of a  $T \sim \text{PHT}(\boldsymbol{\pi}, \mathbf{G})$  random variate. Some efficiency can be gained by initially ignoring the generation of sojourn times, only recording the path of the implicitly embedded discrete Markov chain: the total time in each non-absorbing state is then Erlang( $k_i, \lambda_i$ ) distributed, where  $k_i$  is the total number of excursions to state  $i$ . Details of this approach are in Neuts and Pagano (1981).

### 2.2.2.2 Proposed usage

It is proposed to use Phase-type distributions along the lines given to motivate them above: as a means of modelling failure time of repairable redundant systems. Specifically, by merging those system states representing failure into one (or more) absorbing states, a model is provided which can be used to infer component lifetime and repair time rates given only data on the overall system failure. Relating it to a more complex example again, the reason for pooling of states in Table 2.1 should now be clear: these represent failure and are absorbing states of a continuous-time Markov chain. As such labelling could cease at 17 and then  $\phi^{\{t\}} \in \{1, 2, \dots, 17\} \forall t \in [0, \infty)$  with  $\phi^{[t_f, \infty)} = 17$  for  $t_f = \inf\{t : \phi^{\{t\}} = 17\}$ .

The high level purpose bears repetition: it is now possible to consider questions of inference for masked system lifetime data in the more complex setting of a repairable system. In this scenario the component lifetimes and indeed the ultimate pre-failure state of the system are unknown.

The use of Markov chains to model repairable systems is not a new idea in itself: as early as Barlow and Proschan (1965) absorbing Markov chains are considered in this way. Phase-type distributions specifically were used in Neuts and Meier (1981) and Neuts *et al.* (2000) for an analysis of repairable systems. However, bringing the chapter full-circle (recall p.7), the overwhelming majority of usage in the literature has been in ...

*“... probabilistic analysis of component lifetimes, system lifetimes, and indeed matters of ‘optimal’ design of those systems — here the model and parameters are usually treated as known”.*

It is the second route, that of inference, which has been less explored with such models. Cano *et al.* (2010) is a recent example of this kind of model, but crucially the data available there are not masked: the component lifetimes are known directly, as is the schedule of repair. This matter will be revisited with a review of inferential procedures for Phase-type distributions at the end of the next chapter which covers statistical methodology.

# Chapter 3

## Statistical Methodology

The work in this thesis is concerned primarily with statistical inference. Given observed data and a family of models for the data generating process, statistical inference seeks to establish details of the model — commonly estimates of the parameter values which index the family and associated uncertainty — from the data. The objective is to infer not just the estimates with mathematical rigour, but also to quantify the degree of uncertainty in those estimates through the language of probability theory.

This chapter provides an overview of the Bayesian approach to inference, contrasting it where appropriate with the frequentist approach which, outside the statistics community, currently remains better known. It continues by introducing Markov chain Monte Carlo (MCMC), the computational framework often required due to the intractable nature of many of the integrals which arise in Bayesian applications. The chapter concludes with a description of the MCMC methodology for Bayesian inference of Phase-type models developed by Bladt *et al.* (2003), the extension of which forms the basis for the novel contributions of Chapter 4.

### 3.1 Bayesian inference

The dominant theory of inference for the majority of the twentieth century was so-called ‘frequentism’. In the frequentist school of thought, probability is taken to represent the hypothetical long term proportion of the time an event of interest occurs and, crucially, parameters of probability models are taken to be fixed but unknown quantities. Pearson worked extensively on frequentist type problems at the turn of the twentieth century,

though many of the foundations of frequentist inference as practiced today are found in the classical work of Fisher (1922).

Bayesian inference technically predates the frequentist approach, being traced back to the thought experiment of throwing balls onto a square table by Bayes (1763) and to the ‘inverse probability’ discovered by Laplace (1774). Both used uniform prior distributions: for Bayes it was a consequence of the ball throwing process, whilst Laplace considered it an intuitive axiom (‘principle of insufficient reason’). However, Bayesian inference largely faded until reappearing in the modern form recognisable today due to, among many others, Jeffreys (1939) and Savage (1954).

Philosophically the Bayesian school encompasses both objective probability (similar in interpretation to frequency probabilities) and subjective probability (viewing probability as expressing a degree of personalistic belief). The pivotal additional factor differentiating Bayesian inference from frequentist theory is the treatment of parameters of probability models as realisations of a random variable, enabling direct statements of probability about them to be formulated.

Consider observations  $\mathbf{y} = \{y_1, \dots, y_n\}$  and a probability model which it is believed generated the data, having density function  $f_Y(y; \psi)$  where the dependence on the parameter(s)  $\psi$  is made explicit. Without loss of generality,  $\psi$  may be a vector of parameters. Bayesian inference considers  $\psi$  to be the unknown realisation of a random variable, so that one can consider the joint density of the random variables  $Y$  and  $\Psi$ :

$$f_{Y,\Psi}(y, \psi) = f_{Y|\Psi}(y|\psi) f_{\Psi}(\psi) \tag{3.1}$$

where simply  $f_{Y|\Psi}(y|\psi) := f_Y(y; \psi)$ .

Thus given data  $\mathbf{y}$ , by application of Bayes’ Theorem one can directly encapsulate the uncertainty in  $\psi$  as:

$$\begin{aligned} f_{\Psi|Y}(\psi|\mathbf{y}) &= \frac{f_Y(\mathbf{y}; \psi) f_{\Psi}(\psi)}{\int_{\Omega} f_Y(\mathbf{y}; \psi) f_{\Psi}(d\psi)} \\ &\propto f_Y(\mathbf{y}; \psi) f_{\Psi}(\psi) \end{aligned} \tag{3.2}$$

where  $\Omega$  is the sample space of  $\Psi$ . Proportionality follows from noticing the denominator is independent of  $\psi$  due to the integral. (3.2) is referred to as the posterior distribution of  $\psi$  and encapsulates all knowledge about the unknown parameters.

In practice the integral in the denominator of (3.2) — the normalising constant — is often analytically intractable. This fact impeded widespread adoption of Bayesian



inference until the advent of the computer and the development of computational techniques like Markov chain Monte Carlo. Thus, the popularity of the Bayesian approach only really erupted in the applied statistics community as recently as the 1990s.

### 3.1.1 Likelihood function

The first term in the numerator of (3.2) is often written  $L(\psi; \mathbf{y}) := f_Y(\mathbf{y}; \psi)$  to emphasise the fact that  $\psi$  is actually variable and the data  $\mathbf{y}$  are fixed in this context. This was termed the likelihood by Fisher (1922), who argued that it was often all that was necessary for inference, whereby one simply maximises  $L(\psi; \mathbf{y})$  with respect to  $\psi$  obtaining the so-called maximum likelihood estimate,  $\hat{\psi}$ . However, the likelihood itself is not a probability distribution for  $\psi$  because in general  $\int_{\Omega} L(\psi; \mathbf{y}) d\psi \neq 1$ , making expression of uncertainty about the point estimate  $\hat{\psi}$  a somewhat more opaque task than the natural expression of posterior probability in Bayesian inference.

Fisher (1930) argued in favour of the ‘fiducial probability’ of  $\psi$  which he defined as  $\frac{\partial}{\partial \psi} F_Y(y; \psi)$  in an attempt to attach a direct probabilistic uncertainty to  $\psi$ , but almost nobody adheres to this idea in the modern literature after extensive criticism (e.g. Lindley, 1958). Thus for non-standard models expression of uncertainty in  $\hat{\psi}$  is limited to stating upper/lower confidence limits, a specified hypothetical long run proportion of which contain the ‘true’  $\psi$ , based on asymptotic analysis.

None-the-less, the likelihood is the only way in which the data enter the posterior distribution and so as the quantity of data observed increases, the likelihood term dominates the posterior, leading in many cases to approximate agreement between frequentist maximum likelihood and highest a posteriori density values (the mode of the posterior distribution). In particular, Bayesian inference also obeys the ‘likelihood principle’, which is to say that any two models with the same likelihood  $L(\psi; \mathbf{y})$  result in the same inferences: the likelihood conveys all the information in the data (and pedantically, the minimal sufficient statistic in the likelihood conveys all the information in the data about  $\psi$  specifically).

### 3.1.2 Prior distribution

The prior distribution is the factor  $f_{\Psi}(\psi)$  from (3.2). In admitting the treatment of the parameters as a random variable, a Bayesian analysis must specify the distribution of that random variable independently of the data due to the decomposition (3.1). It represents the ‘state of knowledge’ of the analyst prior to observing the specific data under examination and is often considered the most controversial aspect of Bayesian inference.

This controversy is because the choice of prior is always a subjective decision, even when that decision is to attempt to model ignorance. However, it can be argued that so long as the prior choice is transparent it increases the flexibility of modelling: one can intentionally choose a ‘sceptical prior’ — one which assigns lower probability to some favourable outcome — meaning probability weight in favour of that outcome is even more compelling. In any event, one should bear in mind an oft overlooked fact that the choice of the model for the data,  $f_Y(\cdot; \psi)$ , is often just as much a subjective choice as the prior.

#### 3.1.2.1 Common prior choices

The prior is most commonly from some parametric family of distributions and the parameters of the prior are referred to as hyperparameters. There are several approaches commonly taken to prior specification. Broadly speaking, these would include the following specifications: by a personal decision maker; to accord with expert opinion; for mathematical/interpretive convenience; or, to attempt to seek objectivity.

Full specification of the prior in accordance with the beliefs of the particular individual performing the analysis is highly appropriate in the subjective probability and decision theoretic usage of Bayesian inference: the result of the inference is then the mathematically rational updating of Your beliefs about  $\psi$  in light of evidence observed in the form of the data under examination.

Specification of a prior to accord with expert opinion is an open area of research in itself, called ‘prior elicitation’. At the simplest level, this might involve pooling the opinions of experts on credible ranges for the parameters of the model and using a Gaussian prior with upper and lower  $\alpha\%$  points fixed. Of course, this arguably raises many more questions than it solves if the experts do not understand probability

theory well. However, unless the parameters of the model are easily interpretable, then even experts in probability theory will struggle to translate their knowledge into a probabilistic prior. Garthwaite *et al.* (2005) and O'Hagan *et al.* (2006) are relatively recent reviews of techniques to address this.

Consequently, the prior is often chosen for reasons of convenience, particularly where conjugacy results exist. A conjugate prior distribution is one which, when multiplied by the likelihood, results in an expression which is algebraically from the same parametric family as the prior distribution, up to a normalising constant. The advantage of this is that the posterior normalising constant can then be immediately written down merely by inspection, without having to carry out the integral in the denominator of (3.2). In particular, if the likelihood is from the exponential family in canonical form:

$$f_Y(y; \psi) := m(y) \exp\{\mathbf{s}(y)^\top \psi - k(\psi)\}$$

where  $\mathbf{s}(y)$  is a sufficient statistic, then taking the prior specification as Exponential family  $f_\Psi(\psi) \propto \exp\{\eta^\top \psi - \nu k(\psi)\}$  leads to the conjugate posterior:

$$f_{\Psi|Y}(\psi | \mathbf{y}) \propto \exp\{(\mathbf{s}(\mathbf{y}) + \eta)^\top \psi - (1 + \nu)k(\psi)\}$$

For example, in canonical form a likelihood of  $n$  independent Exponentially distributed observations has:

$$\text{Exp}(\lambda) \implies m(\mathbf{y}) = 1, s(\mathbf{y}) = -n\bar{y}, \psi = \lambda, k(\psi) = -n \log \lambda$$

Note also that a Gamma distribution in canonical form has:

$$\begin{aligned} \text{Gamma}(\alpha, \beta) \implies m(y) = y^{-1}, \mathbf{s}(y)^\top &= (\log y, -y), \psi^\top = (\alpha, \beta), \\ k(\psi) &= \log \Gamma(\alpha) - \alpha \log \beta \end{aligned}$$

Consequently, the conjugate prior of an Exponential likelihood is in the canonical form  $f_\Lambda(\lambda) \propto \exp\{\eta\lambda + \nu n \log \lambda\}$ , which can be identified as Gamma with  $\alpha = n\nu$  and  $\beta = -\eta$ . This leads to a posterior:

$$\begin{aligned} f_{\Lambda|Y}(\lambda | \mathbf{y}) &\propto \exp\{(-n\bar{y} + \eta)\lambda + (1 + \nu)n \log \lambda\} \\ &\propto \exp\{(-n\bar{y} - \beta)\lambda + (n + \alpha) \log \lambda\} \end{aligned}$$

which is also Gamma with shape  $\alpha + n$  and reciprocal scale  $\beta + n\bar{y}$ . Simply put, a  $\text{Gamma}(\alpha, \beta)$  is the conjugate prior for an Exponential likelihood, with the posterior being  $\text{Gamma}(\alpha + n, \beta + n\bar{y})$ .

Another convenience prior is one which is often purported to express ‘ignorance’ about  $\psi$  a priori, though it is not really an expression of ignorance. This often involves a (possibly conjugate) prior with hyperparameters chosen to result in very high variance. Alternatively, since the prior is in both numerator and denominator of (3.2), any multiple of the prior effectively makes no difference, so some practitioners more controversially use priors which do not integrate to 1 — such improper priors can be dangerous though, because the posterior is no longer guaranteed to be a proper probability distribution. The extreme version of this is a flat improper prior,  $f_{\Psi}(\psi) = c \forall \psi$  ( $c > 0$ ), which effectively results in solutions which correspond to Fisher’s fiducial probabilities and is not an approach advocated in this thesis.

There have been at least two attempts at creating a broadly prescriptive means of specifying the prior in an objective fashion. The first was developed by Jeffreys (1946) to address the concern that Bayesian inference is sensitive to the parameterisation chosen for the problem. The Jeffreys prior is specified as  $f_{\Psi}(\psi) \propto J(\psi)^{1/2}$  where  $J(\psi)$  is the Fisher information for  $\psi$ . This prior ensures invariance of the posterior to reparameterisation of the model. However, this can result in improper priors and the results for multiparameter models are controversial.

The second, developed by Bernardo (1979) defines a ‘reference prior’ which seeks to maximise the contribution of the likelihood to the posterior, by maximizing the expected Kullback-Leibler divergence between prior and posterior with respect to the distribution of a sufficient statistic of the parameter. However, again even in simple models improper priors can result.

### 3.1.3 Posterior analysis

Formally, all questions of interest in a Bayesian inferential setting can then be answered using the posterior distribution. Commonly, this may involve simple density plots or summaries such as expectation and variance; posterior modal value; or highest posterior density intervals. It may often arise that interest is actually in some functional of the parameters which can also be handled relatively straight-forwardly.

Indeed, situations which lead to greatly increased complexity in a frequentist setting, such as some elements of  $\psi$  being nuisance parameters, can be handled elegantly through standard probability theory. For example, if  $\psi = (\theta, \eta)$  where  $\theta$  is of inter-

est and  $\eta$  is a nuisance parameter required for completeness of the model, then in a Bayesian context one can simply marginalise to obtain the posterior for  $\theta$ :

$$f_{\Theta}(\theta | \mathbf{y}) = \int_{\Omega_{\eta}} f_{\Psi|Y}(\theta, \eta | \mathbf{y}) d\eta$$

The fully probabilistic structure of Bayesian inference opens up other possibilities, such as the posterior predictive distribution which provides the predictive distribution for a future observation,  $y^*$ , from the same data generating process with all knowledge and uncertainty in  $\psi$  taken into consideration:

$$f_{Y^*|Y}(y^* | \mathbf{y}) = \int_{\Omega} f_Y(y^*; \psi) f_{\Psi|Y}(\psi | \mathbf{y}) d\psi \quad (3.3)$$

Commonly, in a frequentist setting the predictive distribution would result from simply substituting a point estimate:  $f_{Y^*|Y}(y^* | \mathbf{y}) := f_Y(y^*; \hat{\psi})$ , which fails to account for uncertainty in  $\hat{\psi}$ . Although frequentist prediction intervals and predictive distributions with correctly calibrated frequency probability properties can be found (Lawless and Fredette, 2005), they do not appear to be used extensively by practitioners.

### 3.1.4 Hierarchical Bayesian models

The simple specification of likelihood and prior described thus far is often limiting when trying to model complex real-world applications. A powerful modelling option which sees natural expression in a Bayesian framework for analysis is the ‘hierarchical model’.

Recall that the parameters of the probability model for the data generating process  $f_Y(\cdot; \psi)$  are formally considered to be the realisation,  $\psi$ , of a random variable,  $\Psi$ . The posterior then expresses the distribution of the value of the *particular realisation*  $\psi$  which resulted in the data under examination. Consider instead multiple realisations of the parameters,  $\psi_i$ , leading to different groups in the data (one group  $y_{i1}, \dots, y_{in_i}$  for each realisation of the parameters). The basic Bayesian concept can be extended by placing additional ‘hyper’-priors on the parameters of  $f_{\Psi}(\cdot)$ . Thus, the distribution of the multiple  $\psi_i$  is estimated as part of the inferential procedure, through a hierarchy of conditional probabilities as will now be clarified.

For example, consider data  $\mathbf{y} = \{y_{11}, \dots, y_{mn_m}\}$  where the first index indicates

group membership. In specifying the model, define the following:

$$f_{Y|\Psi}(\cdot | \psi_i) \tag{3.4}$$

$$f_{\Psi|\Lambda}(\cdot | \lambda) \tag{3.5}$$

$$f_{\Lambda}(\cdot) \tag{3.6}$$

(3.4) is the probability model for the data generating process of group  $i$ . Note that the parametric family is the same, but the parameter for each group differs (as with the discussion thus far, each  $\psi_i$  may be a vector of parameters). (3.5) is the prior distribution of the parameters, with the hyperparameters made explicit as  $\lambda$ . Crucially,  $\lambda$  is not specified directly, but has hyper-prior distribution (3.6). This full model involves parameters  $\xi = (\psi_1, \dots, \psi_n, \lambda)$ , so that the posterior is now:

$$\begin{aligned} f_{\Xi|Y}(\xi | \mathbf{y}) &\propto f_{Y|\Xi}(\mathbf{y} | \xi) f_{\Xi}(\xi) \\ &= f_Y(\mathbf{y}; \boldsymbol{\psi}) f_{\Psi}(\boldsymbol{\psi}; \lambda) f_{\Lambda}(\lambda) \end{aligned}$$

The simplification of the first term to the usual likelihood is due to the conditional independence of  $\mathbf{y}$  and  $\lambda$  given  $\boldsymbol{\psi}$ , and the joint prior decomposes naturally due to the model formulation.

Commonly, interest will then be in the posterior predictive distribution of  $\psi$ , since this incorporates all knowledge learned about that parameter from the observed data:

$$f_{\Psi^*|Y}(\psi^* | \mathbf{y}) = \int f_{\Psi}(\psi^*; \lambda) f_{\Lambda|Y}(\lambda | \mathbf{y}) d\lambda \tag{3.7}$$

The above example is the simplest possible hierarchical model, but the principle extends easily to highly complex dependencies, most easily depicted through directed acyclic graphs which make identification of conditional dependencies and independencies easier.

### 3.1.5 Exchangeability

Rather brushed over in the above discussion are what assumptions are made in relation to the observed data. The usual starting point in a frequentist analysis is the assumption that the observations are independent and identically distributed (i.i.d.), leading to the simplification in likelihood:  $f_Y(\mathbf{y}; \psi) = \prod_{i=1}^n f_Y(y_i; \psi)$ . Although it is not unusual to see this strong assumption made in simple Bayesian analysis, de Finetti (1974)

emphasised development of Bayesian techniques relying only on the weaker assumption of ‘exchangeability’.

**Definition 3.1 (Exchangeability)** *A sequence of observations  $y_1, \dots, y_n$  is considered exchangeable if the joint distribution is unaffected by any permutation of the indices:*

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) \equiv F_{Y_1, \dots, Y_n}(y_{\sigma(1)}, \dots, y_{\sigma(n)})$$

Note that i.i.d.  $\implies$  exchangeable, but not vice-versa, meaning that exchangeability covers a wider class of models. An example of an exchangeable sequence which is not i.i.d. is a sequence resulting from simple random sampling without replacement.

In practice, a simple non-hierarchical Bayesian analysis models the observations from an exchangeable distribution as being i.i.d. given the parameter(s)  $\psi$ . The real usefulness of exchangeability is apparent in hierarchical models, where one considers exchangeability at each level of the hierarchy, conditional on the parameters of the higher level. This leads to a mixture of i.i.d. distributions representing the exchangeable distribution. For example, the exchangeable distribution of  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)$  from §3.1.4 is:

$$f_{\Psi}(\boldsymbol{\psi}) = \int_{\Omega_{\lambda}} \left\{ \prod_{i=1}^n f_{\Psi|\Lambda}(\psi_i | \lambda) \right\} f_{\Lambda}(d\lambda)$$

In most applications, such mixture of i.i.d. distributions is entirely adequate. This nicely mirrors the theoretical result in de Finetti’s Theorem which proves every (infinite) sequence of exchangeable Bernoulli random variables can be expressed in mixture i.i.d. form.

### 3.1.6 Summary

This section has introduced some ideas relating to the Bayesian approach to inference. A classic text for much more detailed background is Bernardo and Smith (2007). A more hands-on introduction is offered in Gelman *et al.* (2004). Finally, Cox (2006) compares and contrasts Bayesian theory with frequentist approaches.

## 3.2 Markov chain Monte Carlo

The playfully dubbed ‘Monte Carlo’ idea of simulating random processes to learn about a quantity of interest was first systematically developed in 1944 by researchers at Los Alamos working on the atomic bomb during World War II (Hammersley and Handscomb, 1964). Markov chain Monte Carlo followed quickly, the earliest published work being the classic paper of Metropolis *et al.* (1953). However, it was a long time before the ideas permeated the statistics community: despite Hastings (1970) generalising the Metropolis algorithm in a statistics journal, it was only later after the Gibbs sampler ideas from Tanner and Wong (1987) and Gelfand and Smith (1990) that it took-off and saw wide spread adoption.

Before the advent of Markov chain Monte Carlo (MCMC), application of Bayesian methods relied heavily on conjugacy results, non-informative priors and sophisticated approaches to analytically evaluating posterior distributions. Looking at a respected text such as Box and Tiao (1973) from the decade before MCMC gained traction in statistical research, and then recent papers, reveals the real paradigm shift: the nature and complexity of models which can be comfortably handled continues to grow and the modern prominence of algorithmic over closed form solutions is conspicuous.

As already indicated, the challenges stem from evaluation of the normalising constant in (3.2), which is often a sum or integral of very high dimension. Let

$$p_u(\psi) = f_Y(\mathbf{y}; \psi) f_\Psi(\psi)$$

be the un-normalised posterior density. Then  $p_u(\cdot)$  gives rise to a probability measure,

$$\pi(A) = \frac{\int_A p_u(\psi) d\psi}{\int_\Omega p_u(\psi) d\psi}$$

which hereinafter is referred to as the ‘target distribution’. MCMC provides a means of estimating expectations of functions  $g : \Omega \rightarrow \mathbb{R}$  with respect to this measure:

$$\mathbb{E}_\pi[g(\Psi)] = \int g(\psi)\pi(d\psi)$$

Given a sample from the target distribution, the idea behind expectation estimation mirrors the well known standard Monte Carlo. The ‘Markov chain’ part of the nomenclature addresses producing the samples in difficult situations like Bayesian posterior analysis.



First, a quick recap of standard Monte Carlo provides motivation. The theory in this section can be applied in many areas aside from posterior exploration, so this is stressed by the presentation in terms of arbitrary random variables,  $X$ , and general unnormalised density functions,  $p_u(x)$ .

### 3.2.1 Standard Monte Carlo

In standard Monte Carlo a sequence of random variables  $X_1, X_2, \dots$  which are independent and identically distributed according to the target distribution are used to compute estimates of  $\mu = \mathbb{E}_\pi[g(X)]$ . Let  $\sigma^2 = \text{Var}_\pi(g(X))$ . Then the strong law of large numbers implies that the estimate:

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n g(X_i)$$

converges almost surely to  $\mu$ . Moreover, if  $g \in L^2(\pi)$  then the estimate is approximately Normally distributed  $\hat{\mu}_n \approx N(\mu, n^{-1}\sigma^2)$ , by the Central Limit Theorem, and the estimator variance can be approximated well by the regular naïve estimate:

$$n^{-1}\hat{\sigma}_n^2 = \frac{1}{n^2} \sum_{i=1}^n (g(X_i) - \hat{\mu}_n)^2$$

Thus, standard Monte Carlo provides an extremely simple framework for computing estimates complete with some quantification of the uncertainty in the estimated value.

#### 3.2.1.1 Generality of expectation

At this point it is apposite to recall the salient fact that once estimates of  $\mu = \mathbb{E}_X[g(X)]$  can be computed, it is then possible to compute an estimate of almost any quantity of interest, due to the generality of expectations. For example:

$$F_X(x) = \mathbb{P}(X < x) = \int_{-\infty}^x f_X(du) = \int_{-\infty}^{\infty} \mathbb{I}_{(-\infty, x)}(u) f_X(du) = \mathbb{E}_X[\mathbb{I}_{(-\infty, x)}(X)]$$

where  $\mathbb{I}_{(-\infty, x)}(\cdot)$  is the indicator function:

$$\mathbb{I}_A(u) = \begin{cases} 0 & \text{if } u \notin A \\ 1 & \text{if } u \in A \end{cases}$$

Thus, entirely unsurprisingly,  $n^{-1} \sum_{i=1}^n \mathbb{I}_{(-\infty, x)}(X_i)$  is an estimate of the probability that the random variable  $X$  is less than  $x$ . More importantly, standard Monte Carlo

theory additionally provides insight into the uncertainty in the estimate: the distribution of the estimated probability is asymptotically Normal with approximate variance  $n^{-2} \sum_{i=1}^n (\mathbb{I}_{(-\infty, x)}(X_i) - \hat{\mu}_n)^2$ .

Many quantities of interest, complete with uncertainty estimate, follow naturally from this. For example, marginalised densities, predictive densities (see (3.3)), posterior moments, quantiles and indeed almost any statement of probability can be expressed as an integral in the form of an expectation.

### 3.2.1.2 Beyond standard Monte Carlo

Standard Monte Carlo theory only applies to independent draws from the target distribution. However, it is possible to extend the methodology to non-independent draws: the objective is to create a sequence (a Markov chain) of random variables  $X_1, X_2, \dots$  whose distribution is the target distribution of interest, without making independent and identically distributed draws directly from that target. Furthermore, equivalent uncertainty estimates to those in standard Monte Carlo are required for these non-independent samples.

## 3.2.2 (Discrete time) Markov chains

This subsection addresses the underlying theory of the ‘Markov chain’ part of MCMC: what set-up enables one to draw (non-independent) samples from  $\pi(\cdot)$  without sampling it directly?

In §2.2.1 continuous-time Markov chains on a discrete state space were described, albeit the bulk of the theory was omitted to maintain a clear connection to the subject matter of reliability. The fundamental set-up here is similar, but there is no direct subject matter link as such: the purpose of constructing the stochastic process at this point is purely computational. In contrast to §2.2, the stochastic process is discretely indexed (by  $\mathbb{N}^+$ ) on the state space of the target distribution,  $\Omega$ , which may be continuous and/or infinite.

The evolution of the discrete-time stochastic process from its initial state  $X_1$  and through all subsequent steps is described by a Markov kernel:

$$\mathbb{P}(X_{n+1} \in A \mid X_n = x) = K(x, A) := \int_A K(x, dy)$$

for all  $K$ -measurable subsets  $A \subset \Omega$ , where  $K(x, \cdot)$  is a probability measure on the space  $\Omega$ . If  $K(x, \cdot)$  is independent of the discrete time,  $n$ , it will result in a homogeneous (Definition 2.11) and Markov (Definition 2.10) stochastic process called a *Markov chain*.

Ultimately then, the object of the exercise is to construct a probability measure  $K(x, \cdot)$  from which it is easy to sample for all  $x$ , but such that the resulting sequence  $X_1, X_2, \dots$  is equal in distribution to the target distribution. Achieving this hinges primarily on two concepts: *stationarity* and *reversibility*.

**Definition 3.2 (Stationary distribution)** *A probability measure  $\pi(\cdot)$  on  $\Omega$  is said to be the stationary distribution of a Markov chain with Markov kernel  $K(x, \cdot)$  if and only if:*

$$\int K(x, A)\pi(dx) = \pi(A)$$

for all  $\pi$ -measurable subsets  $A \subset \Omega$ .

Intuitively, if a Markov chain has stationary measure  $\pi(\cdot)$  and  $X_1$  is distributed according to  $\pi(\cdot)$ , then  $X_2$  (drawn from  $K(X_1, \cdot)$ ) will also be distributed as  $\pi(\cdot)$ . Recursively, so will  $X_3, X_4, \dots$ . This is a very strong stability condition.

**Definition 3.3 (Reversibility)** *A Markov chain with Markov kernel  $K(x, \cdot)$  is reversible with respect to a probability measure  $\pi(\cdot)$  if and only if:*

$$\begin{aligned} \langle Kf, g \rangle &= \langle f, Kg \rangle \\ \iint f(y)g(x)K(x, dy)\pi(dx) &= \iint g(y)f(x)K(x, dy)\pi(dx) \end{aligned}$$

for all  $f, g \in L^2(\pi)$ .

Given an appropriate  $\sigma$ -finite reference measure  $\mu(dx)$  on  $\Omega$  (e.g. Lebesgue measure for continuous target densities and counting measure for discrete), the Markov kernel  $K$  can usually be written as a transition kernel  $K(x, dy) = k(x, y)\mu(dy)$ . Likewise  $\pi(dx)$  has density function  $\pi(dx) = p(x)\mu(dx)$ . This allows a simpler definition which it will be seen is a sufficient condition for reversibility (note that it is not necessary, so cannot entirely supersede the definition of reversibility above):

**Definition 3.4 (Detailed balance)** *A transition kernel  $k(x, \cdot)$  and probability density  $p(\cdot)$  are said to satisfy detailed balance if the following equality holds:*

$$p(x)k(x, y) = p(y)k(y, x) \quad \forall x, y \in \Omega$$

Detailed balance, reversibility and stationarity are then all linked by the following crucial result.

**Theorem 3.1** *For some given measure, if a Markov chain with transition kernel  $k(x, \cdot)$  and the probability density  $p(\cdot)$  satisfy detailed balance, then:*

1. *the Markov chain is reversible with respect to  $\pi(\cdot)$ ; and*
2.  *$\pi(\cdot)$  is the stationary distribution of the Markov chain.*

This important Theorem forms the heart of the majority of modern Markov chain sampling strategies used in Bayesian inference. In summary, it states that finding a Markov kernel which is in detailed balance with the target distribution is sufficient to ensure the Markov chain is reversible and, consequently, the target distribution is the stationary distribution. Therefore, each member of the sequence  $X_2, X_3, \dots$  is distributed according to the target distribution whenever  $X_1$  is.

The last three words of that paragraph, ‘*whenever  $X_1$  is*’, can often be the crux of using Markov chain sampling in practice and will be revisited in §3.2.4. For now, blissful ignorance of this obstacle is assumed.

### 3.2.3 Markov chain sampling algorithms

The previous subsection has indicated the general course of action to finding a Markov chain which may generate the necessary samples. The remarkable work of Metropolis *et al.* (1953) (symmetric proposals) and later generalisation by Hastings (1970) provides a prescriptive means to constructing the necessary transition kernel. Chib and Greenberg (1995) provide an accessible tutorial paper.

#### 3.2.3.1 The Metropolis-Hastings algorithm

The objective is to sample from the probability measure  $\pi(\cdot)$  which has un-normalised density  $p_u(\cdot)$  (with respect to the appropriate  $\sigma$ -finite reference measure). The user of the algorithm must essentially choose some proposal transition kernel  $q(x, \cdot)$ , which clearly will not necessarily be the requisite kernel to be reversible with  $\pi(\cdot)$ . A very common choice for  $q(x, \cdot)$  in continuous settings of dimension  $\geq 2$  is the multivariate Normal distribution.

The Metropolis-Hastings kernel is then defined according to the following algorithm:

**Algorithm 3.1 (Metropolis-Hastings kernel)**

1. If the current state is  $x$ , then propose a new state  $y$  from  $q(x, \cdot)$ .
2. Compute the probability (termed the ‘acceptance ratio’):

$$\alpha(x, y) = \min \left\{ 1, \frac{p_u(y)q(y, x)}{p_u(x)q(x, y)} \right\}$$

3. With probability  $\alpha(x, y)$  the next state is  $y$  and with probability  $1 - \alpha(x, y)$  it is  $x$  again. □

Thus, the transition kernel is:

$$k(x, y) = q(x, y)\alpha(x, y) = q(x, y) \min \left\{ 1, \frac{p_u(y)q(y, x)}{p_u(x)q(x, y)} \right\}$$

Remarkably, this transition kernel defined above via the proposal kernel is then adequate:

**Theorem 3.2 (Metropolis-Hastings)** *Given a probability measure  $\pi(\cdot)$ , having unnormalised probability density  $p_u(\cdot)$ , and a proposal kernel  $q(x, \cdot)$ , the transition kernel defined by Algorithm 3.1 results in a Markov chain whose stationary distribution is  $\pi(\cdot)$ .*

There are different flavours of Metropolis-Hastings, including:

**Metropolis:** where  $q(x, y) = q(y, x)$ , leading to a simplified acceptance ratio.

**Independence Metropolis-Hastings:** where  $q(x, y) = q(y)$ , so that the proposal distribution is the same no matter the current state of the chain.

**Random-walk Metropolis-Hastings:** where  $q(x, y)$  is of the form  $x + e$  where  $e$  is random and independent of the current state (implemented in Geyer, 2010).

Thus, apart from the independence Metropolis-Hastings sampler, moves are made locally with probability determined by the acceptance ratio. There is a trade-off involved here when selecting a proposal distribution: if the resulting acceptance ratio

tends to be high then moves are likely local and highly dependent, consequently exploration of the posterior may be slow. On the other hand, if the acceptance ratio tends to be low, then the chain may become ‘stuck’ and again exploration may be slow. For example, Roberts *et al.* (1997) showed that the asymptotically optimal acceptance ratio is 0.234 for a random-walk Metropolis algorithm, under quite general conditions.

The Metropolis-Hastings algorithm was the most general form of Markov chain sampling technique up until Green (1995) who fully generalised it through a measure theoretic treatment. Section 1.17 of Geyer (2011) is a nice review of this in its full generality. The best known consequence is the ability to do so-called *reversible jump* between models, thus incorporating model inference into the MCMC scheme.

### 3.2.3.2 The Gibbs sampler

The Gibbs sampler has origins in statistical physics and its first appearance in a more traditional statistical application was Geman and Geman (1984). Tanner and Wong (1987) introduced the idea for missing data problems, calling it ‘data augmentation’. Gelfand and Smith (1990) marked the start of mainstream use of Gibbs sampling in general Bayesian problems and shortly after Casella and George (1992) wrote a very accessible tutorial paper.

The Gibbs sampler is actually a special case of the Metropolis-Hastings algorithm, but there is a convention of explicitly describing it in the literature, because it involves a particularly powerful simplification. Gibbs sampling makes use of the collection of so-called ‘full conditional’ distributions of the target distribution. Given a target  $p(x_1, \dots, x_n)$ , the full conditionals are:

$$\begin{aligned} p(x_1 | \mathbf{x}_{(-1)}) &:= p(x_1 | x_2, \dots, x_n) \\ p(x_2 | \mathbf{x}_{(-2)}) &:= p(x_2 | x_1, x_3, \dots, x_n) \\ &\vdots \\ p(x_n | \mathbf{x}_{(-n)}) &:= p(x_n | x_1, \dots, x_{n-1}) \end{aligned}$$

Let  $\mathbf{X}^{(i)} = (X_1^{(i)}, \dots, X_n^{(i)})$  denote the  $i$ th Markov chain random variate. Then, the two main variants of Gibbs sampling are the systematic-scan Gibbs sampler and the random-scan Gibbs sampler.

**Algorithm 3.2 (Systematic-scan Gibbs sampler)** *Initialise the chain with a fixed  $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$  and for  $i = 1, 2, \dots$  iterate:*

1. Draw a sample  $x_1^{(i)}$  from  $p(x_1 | x_2^{(i-1)}, \dots, x_n^{(i-1)})$
- $\vdots$
- $j$ . Draw a sample  $x_j^{(i)}$  from  $p(x_j | x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \dots, x_n^{(i-1)})$
- $\vdots$
- $n$ . Draw a sample  $x_n^{(i)}$  from  $p(x_n | x_1^{(i)}, \dots, x_{n-1}^{(i)})$  □

**Algorithm 3.3 (Random-scan Gibbs sampler)** *Again, initialise the chain with a fixed  $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$  and for  $i = 1, 2, \dots$  iterate:*

1. Choose a component  $j$  to update uniformly from  $\{1, \dots, n\}$
2. Draw a sample  $x_j^{(i)}$  from  $p(x_j | x_1^{(i-1)}, \dots, x_{j-1}^{(i-1)}, x_{j+1}^{(i-1)}, \dots, x_n^{(i-1)})$ , and the remaining components are unchanged:  $x_k^{(i)} = x_k^{(i-1)} \forall k \neq j$  □

Algorithms 3.2 and 3.3 each define a transition kernel for a Markov chain, but it is not immediately obvious that the stationary distribution for these chains is the target distribution  $p(\mathbf{x})$ . The significant Theorem which follows the next definition provides the striking result that the full conditionals of a distribution completely define it and gives the relevant decomposition.

**Definition 3.5 (Positivity condition)** *Let  $(X_1, \dots, X_n)$  be distributed according to  $p(\mathbf{X})$  and let  $p_i(x_i) = \int \dots \int p(x_1, \dots, x_n) d\mathbf{x}_{(-i)}$  be the marginal distribution of the  $i$ th component. If for all  $\mathbf{x} = (x_1, \dots, x_n)$ ,*

$$p_i(x_i) > 0 \quad \forall i \in \{1, \dots, n\} \quad \implies \quad p(x_1, \dots, x_n) > 0$$

*holds, then  $p(\cdot)$  is said to satisfy the positivity condition.*

This positivity condition is somewhat restrictive in that it implies that the support of the joint distribution is the Cartesian product of the supports of all the marginals.

**Theorem 3.3 (Hammersley-Clifford (Besag, 1974))** *Any joint distribution,  $p(\cdot)$ , which satisfies the positivity condition can be decomposed as:*

$$p(x_1, \dots, x_n) \propto \prod_{i=1}^n \frac{p(x_{\sigma(i)} | x_{\sigma(1)}, \dots, x_{\sigma(i-1)}, x'_{\sigma(i+1)}, \dots, x'_{\sigma(n)})}{p(x'_{\sigma(i)} | x_{\sigma(1)}, \dots, x_{\sigma(i-1)}, x'_{\sigma(i+1)}, \dots, x'_{\sigma(n)})}$$

for all permutations  $\sigma$  on  $\{1, \dots, n\}$  and every  $\mathbf{x}' \in \Omega$ , where the constant of proportionality is  $p(x'_1, \dots, x'_n)$ .

The Hammersley-Clifford Theorem makes it easy to see that a single component update of the Gibbs sampler is in detailed balance with the target distribution when it is written using this decomposition into full conditionals (with  $\sigma$  and  $\mathbf{x}'$  chosen appropriately). In fact, the random-scan Gibbs sampler is in detailed balance with the target distribution for multiple updates. However, the systematic-scan is not guaranteed to be in detailed balance, except in the 2 component case where there is component-wise reversibility — joint reversibility is possible with an additional latent sampling step (Robert and Casella, 2004, p.351).

The power of the Gibbs sampler is in its simplicity. In particular, note that the target distribution need not be known in closed form: thus a directed acyclic graph can be used to specify a model and associated dependencies and then, due to the Hammersley-Clifford Theorem, only having the full conditional distributions from each node is adequate to proceed with an MCMC sampler without ever specifying (or knowing) the full target distribution explicitly. The only proviso is that one needs to know that the joint distribution exists: the Hammersley-Clifford Theorem provides the decomposition but does not prove existence of a valid joint distribution.

The Gibbs sampler is, none-the-less, technically a special case of the Metropolis-Hastings algorithm: the full conditionals are the proposal kernel for a single site update leading to acceptance probability always equal to 1. However, it is still hugely important because in principle the insights of the Hammersley-Clifford Theorem enable sampling from a multivariate density by performing only univariate sampling. Indeed, in something of a misnomer, ‘Metropolis-Hastings within Gibbs’ is a common strategy whereby a Gibbs scheme is used but the full conditionals are sampled via a Metropolis-Hastings update. Also common is a so-called ‘blocked Gibbs’ whereby many components are simultaneously updated conditional on the others.



### 3.2.3.3 Data augmentation

The data augmentation algorithm (Tanner and Wong, 1987) is effectively just Gibbs sampling, though it is given a separate name to emphasise the usefulness of Gibbs samplers in missing data problems.

It may be that some target density of interest,  $p : \Omega \rightarrow [0, \infty)$ , is either prohibitively complex or possibly cannot even be written in closed form. However, if there is a function  $h : \Omega \times \Omega' \rightarrow [0, \infty)$  such that:

$$\int_{\Omega'} h(x, y) dy = p(x)$$

and where simulation from both  $h(x | y)$  and  $h(y | x)$  is easy, then Gibbs sampling can be employed using these two full conditionals. That is, the space  $\Omega$  has been augmented with  $\Omega'$  to facilitate sampling. Typically,  $\Omega'$  is some missing data whose reintroduction through simulation causes an unmanageably complex problem to be simplified due to the fact that Gibbs sampling ensures the resultant chain of  $x$  values is from the marginal.

The main subtlety with regard to data augmentation versus general Gibbs sampling is that although two components ( $X$  and  $Y$ ) are being sampled, interest is often only in one. Thus, the transition kernel of interest is really

$$\begin{aligned} k(x, x') &= k_{Y|X}((x, y), (x, y')) k_{X|Y}((x, y'), (x', y')) \\ &= h(y' | x) h(x' | y') \end{aligned}$$

rather than the joint update. This can have an impact on convergence results in §3.2.4 since focus on the  $X$  chain may prove convergence more easily than the joint chain.

Hobert (2011) is a detailed and accessible introduction to MCMC specifically from the data augmentation viewpoint and contains the most recent research results from, for example, parameter expanded data augmentation.

## 3.2.4 Convergence

As hinted at on page 45, although the above seems a complete and rosy picture at first glance, it is only half the story. There are two key issues:

1. If it was possible to draw the first sample directly from the target distribution,

presumably one would set about continuing to draw i.i.d. samples in that way and perform standard Monte Carlo analysis!

2. The definition of stationarity (Definition 3.2) is slightly weaker than it might first appear. Stationarity alone is not enough to ensure that the Markov chain will produce *successive* draws from the target distribution when started in a *single* fixed place. If this seems surprising, it may be clarified by the following neat example, reproduced from Roberts and Rosenthal (2004) with notation modified.

Let  $\Omega = \{1, 2, 3\}$  and take target mass function  $p(1) = p(2) = p(3) = \frac{1}{3}$ . Consider the transition kernel  $k(1, 1) = k(1, 2) = k(2, 1) = k(2, 2) = \frac{1}{2}$  and  $k(3, 3) = 1$ . Then it is trivial to verify  $k(x, \cdot)$  is stationary with respect to  $p(\cdot)$ . However, if  $X_1 = 1$ , then  $X_n \in \{1, 2\}$  for all  $n$ , so  $\mathbb{P}(X_n = 3) = 0$  for all  $n$ .

To make clear what it is that stationarity does guarantee, consider a collection of chains. Let  $X_j^i$  denote the  $j$ th draw from the  $i$ th chain. If  $\{X_1^i : i = 1, \dots, m\}$  are drawn from  $p(\cdot)$ , then stationarity guarantees that after  $j$  steps the set of random variables  $\{X_j^i : i = 1, \dots, m\}$  is distributed as  $p(\cdot), \forall j \geq 2$ . Thus, stationarity preserves the distribution of an initial collection of random variables as they are evolved forward as distinct from ensuring that each chain, serially, is from the target distribution.

These issues are normally addressed by an appeal to convergence. That is, if  $K^n(x, \cdot)$  is the  $n$ -step Markov kernel:

$$\begin{aligned} K^n(x, A) &:= \mathbb{P}(X_{n+1} \in A | X_1 = x) \\ &= \int_A \int \cdots \int K(x, dy_1) K(y_1, dy_2) \cdots K(y_{n-2}, dy_{n-1}) K(y_{n-1}, d\xi) \end{aligned}$$

then the hope is that  $K^n(x, A) \xrightarrow{a.s.} \pi(A)$  as  $n \rightarrow \infty$  for arbitrary choice of starting position  $x \in \Omega$ .

There is a prodigious amount of literature on convergence of Markov chains, much of it quite technical. The following is necessarily just a passing introduction.

**Definition 3.6 (Total variation distance)** Given two probability measures on the same space,  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$ , the total variation distance between them is:

$$\|\pi_1(\cdot) - \pi_2(\cdot)\| = \sup_A |\pi_1(A) - \pi_2(A)|$$

where the supremum is over all  $\pi$ -measurable sets  $A$ .

Thus the objective is to ensure  $\lim_{n \rightarrow \infty} \|K^n(x, \cdot) - \pi(\cdot)\| = 0$  for all  $x \in \Omega$ . This is stronger than simply convergence to stationarity since it means that for all starting points the chain would serially produce draws from  $\pi(\cdot)$ .

**Definition 3.7 ( $\phi$ -irreducibility)** A Markov chain is  $\phi$ -irreducible if there exists a non-zero  $\sigma$ -finite reference measure  $\phi$  on  $\Omega$  such that  $\exists n, K^n(x, A) > 0$  for all  $A \subseteq \Omega$  where  $\phi(A) > 0$ .

This irreducibility requirement is what ensures pathological examples such as the one above are excluded, so that the Markov chain is never partitioned in to one part of the state space. One further requirement is needed

**Definition 3.8 (Aperiodicity)** A Markov chain is periodic if there exists a collection of disjoint subsets  $\Omega_1, \dots, \Omega_d \subseteq \Omega$ , ( $d \geq 2$ ) such that  $K(x, \Omega_{i+1}) = 1 \forall x \in \Omega_i$  ( $1 \leq i < d$ ), and  $K(x, \Omega_1) = 1 \forall x \in \Omega_d$ .

If no such collection exists, the Markov chain is said to be aperiodic.

In the setting here, aperiodicity is still considered to hold if the only collections are such that  $\pi(\Omega_i) = 0 \forall i$ .

If a Markov chain is both  $\phi$ -irreducible and aperiodic then it is said to be *ergodic*. Ergodicity ensures that for  $\pi$ -almost all starting points  $x \in \Omega$ ,  $\lim_{n \rightarrow \infty} \|K^n(x, \cdot) - \pi(\cdot)\| = 0$ . To ensure it for *all* starting places requires a stronger form of irreducibility:

**Definition 3.9 (Harris recurrence)** Let  $\eta_A$  be the number of passages a Markov chain makes to the set  $A \subseteq \Omega$ . Then, the Markov chain is Harris recurrent if it is  $\phi$ -irreducible and moreover  $\mathbb{P}(\eta_A = \infty) = 1$  for all  $A \subseteq \Omega$  where  $\phi(A) > 0$ .

If a Markov chain is  $\phi$ -irreducible, aperiodic and Harris recurrent then it is said to be *Harris ergodic*. Verifying Harris ergodicity can be awkward in many cases, but the following Lemma provides a commonly satisfied sufficient condition:

**Lemma 3.4** *Given a transition kernel  $k$  of some Markov kernel, the resulting Markov chain is Harris ergodic if  $k(x, y) > 0 \forall x, y \in \Omega$ .*

Once Harris ergodicity is confirmed, then the above statement is strengthened:  $\lim_{n \rightarrow \infty} \|K^n(x, \cdot) - \pi(\cdot)\| = 0, \forall x \in \Omega$ . Indeed, the following Theorem which is a Markov chain analogue of the strong law of large numbers then holds.

**Theorem 3.5 (Ergodic Theorem)** *If a Markov chain on a state space with countably generated<sup>1</sup>  $\sigma$ -algebra is Harris ergodic and  $\pi(\cdot)$  is stationary with respect to the transition kernel  $k(x, \cdot)$ , then for all  $g \in L^1(\pi)$ :*

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n g(X_i) \xrightarrow{a.s.} \mathbb{E}_\pi[g(X)] \text{ as } n \rightarrow \infty$$

This Theorem justifies using samples from a Markov chain to estimate expectations with respect to the stationary measure when the chain is Harris ergodic.

The question this Theorem leaves unanswered is at what rate the convergence occurs. The definition which must be satisfied to assure rapid convergence is as follows:

**Definition 3.10 (Geometric ergodicity)** *A Markov chain is geometrically ergodic if  $\exists M : \Omega \rightarrow [0, \infty)$  and  $\rho \in [0, 1)$  such that:*

$$\|K^n(x, \cdot) - \pi(\cdot)\| \leq M(x)\rho^n$$

for all  $x \in \Omega, n \in \mathbb{N}^+$ .

The literature on rapid convergence is expansive and so a detailed treatment of geometric ergodicity and associated deeper results is well beyond the scope of this overview. The book by Robert and Casella (2004) covers matters of convergence in detail. Two particularly nice papers for orienting oneself in the literature are Jones and Hobert (2001) and Roberts and Rosenthal (2004).

### 3.2.4.1 Markov chain Central Limit Theorem

The final element required to make MCMC a fully meritorious successor to the standard Monte Carlo method for problems involving intractable direct simulation is quantification of uncertainty in the estimates produced. Thus, just as Theorem 3.5 is the

---

<sup>1</sup>Note any finite or countable state space will satisfy this, as will the standard Borel  $\sigma$ -algebra on  $\mathbb{R}^n$ .

analogue of the strong law of large numbers, there is also a Markov chain analogue of the Central Limit Theorem. There are many versions of it, differing only in the sufficient conditions they use, the first Theorem in Jones *et al.* (2006) providing a referenced list. One of the versions which can be stated with the definitions encountered hereinbefore is:

**Theorem 3.6** *Take a geometrically Harris ergodic Markov chain with stationary distribution  $\pi(\cdot)$  and let  $g : \Omega \rightarrow \mathbb{R}$ , with  $g \in L^2(\pi)$ . Then, for any initial starting state,*

$$\sqrt{n}(\hat{\mu}_n - \mathbb{E}_\pi[g(X)]) \xrightarrow{d} \mathcal{N}(0, \sigma_g^2)$$

as  $n \rightarrow \infty$ , where

$$\sigma_g^2 := \text{Var}_\pi(g(X_1)) + 2 \sum_{i=2}^{\infty} \text{Cov}_\pi(g(X_1), g(X_i))$$

There is a notable jump in complexity compared to standard Monte Carlo: the variance can no longer be computed by an easy naïve estimate. The work in Jones *et al.* (2006) and accessible review in Flegal *et al.* (2008) provide details on estimating  $\sigma_g^2$  using batch means, which is implemented in Flegal and Hughes (2012). The ‘gold-standard’ method of estimating the variance is via so-called regeneration techniques, but these are often hard to find in practice, so the generally applicable batch means approach is used in the sequel.

### 3.2.5 Implementing MCMC

Assuming that a (geometrically) Harris ergodic Markov chain can be chosen with the target distribution as the stationary distribution, there remain some issues to address in implementation. These include:

1. Having started the chain at some  $x \in \Omega$ , how many iterations should be performed before assuming convergence to the target has occurred?
2. How many iterations should be performed before calculating the estimated quantities of interest?

### 3.2.5.1 Burn-in

The answer to the first question is commonly called the *burn-in*. The idea is that one runs the chain for some specified duration and then discards these once the chain (subjectively or diagnostically) appears to be stochastically stationary — the final value is then used as the starting point for the Markov chain samples which will actually be used in the analysis.

It is not entirely uncontentious. Geyer (2011, p.20) observes: “Any point you don’t mind having in a sample is a good starting point.” That is, any reasonable choice of starting point which is not too far from stationarity obviates the need for extensive burn-in. Thus, in a Bayesian inferential context, if one believes the prior choice to be sensible then a random sample from the prior would be an acceptable starting point without burn-in. Likewise, if the target distribution mode can be found before starting an MCMC run using a numerical optimiser, then this would be a reasonable starting point obviating any need for burn-in.

However, caution is certainly required. If the parameter space is high dimensional, such choices may be difficult to make and at a minimum trace plots of parameter samples should always be examined to ensure there is no discernible trend.

### 3.2.5.2 Run length

In the 1990s, the amount of *thinning* performed was intimately connected with answering the second question. Thinning is the practice of retaining only every  $m$ th iteration, with  $m$  chosen to reduce the serial correlation (or autocorrelation) to near zero. This was done in an effort to create the illusion of i.i.d. samples so that an appeal to the standard Monte Carlo Central Limit Theorem could be made: the number of iterations could then be chosen by deciding the desired standard error accuracy required and computing the requisite  $n$ , then drawing  $nm$  MCMC samples and thinning.

It is mentioned because the practice is still widespread despite the advent of the Markov chain Monte Carlo Central Limit Theorem (Theorem 3.6) toward the end of the millennium, which obviates the need to do so. Geyer (2011, p.27) quotes Elizabeth Thompson: “You don’t get a better answer by throwing away data.” Thus, an estimate of the number of iterations required can be computed by appeal to Theorem 3.6. A thinning style approach may be justified in the case that autocorrelation is so severe

— and inflates the variance so much — that storing all the required iterations is computationally infeasible.

### 3.2.5.3 Diagnostics

Hence, whilst burn-in may be necessary, thinning should be required less often. This still leaves open the question of how to determine whether convergence has occurred. Regrettably, there is no definitive means of answering this question because an MCMC chain may make a sudden change 1 iteration beyond wherever one ceased. However, there are certain diagnostic tests which, assuming any trend that will occur has started to do so, will provide guidance as to whether the chain has converged.

There is a soupçon of irony that these tests tend to be based on frequentist methods, given the heavy use of MCMC for Bayesian inference. Gelman and Rubin (1992) developed the Gelman-Rubin shrink factor (a form of ANOVA between multiple chains), Heidelberger and Welch (1983) introduced Heidelberger and Welch’s convergence diagnostic (based on the Cramér-von-Mises test statistic), and Geweke (1992) presented the Geweke statistic (performing a difference of means test on non-overlapping parts of the chain). All these above tests can be applied with progressively increasing burn-in sizes to gain insight into convergence of the chain.

### 3.2.6 Summary

This section has introduced Markov chain methods of sampling from intractable target probability distributions. It also has briefly mentioned the convergence and asymptotic results which enable a Monte Carlo usage of sequential draws of a Markov chain to estimate expectations of functionals of that target distribution and their associated uncertainty. There was a short explanation of the two most common implementation concerns surrounding convergence and stopping rules.

The interested reader looking for more depth can follow some of the cited tutorial and overview papers or should consult the seminal book Meyn and Tweedie (1993) (recently updated in Meyn and Tweedie, 2009), or Robert and Casella (2004). Diaconis (2009) is a nice review paper which includes interesting applications, and an excellent reference is the recent collection of papers in Brooks *et al.* (2011).

### 3.3 Inference for Phase-type models

Phase-type distributions were introduced in §2.2.2, page 29. In their most general form they comprise  $n^2 + n$  parameters, where  $n$  is the number of non-absorbing states. Statistically, interest may lie in inferring these parameters given some data. As alluded to on page 31, the data associated with Phase-type distributions may vary: it could consist of details of the latent stochastic process, or it may consist only of the absorption times (which correspond to masked system lifetime data in the reliability setting). Concern in this thesis is with the latter scenario, so that observations  $y_1, \dots, y_N$  of random variables  $Y_1, \dots, Y_N$  are modelled as  $Y_i \stackrel{\text{i.i.d.}}{\sim} \text{PHT}(\boldsymbol{\pi}, \mathbf{G})$ , with  $\mathbf{G}$  comprised of  $\mathbf{S}$  and  $\mathbf{s}$  as in (2.11), page 29. The exchangeable rather than i.i.d. setting is considered later in Chapter 6.

There has already been fruitful work in the literature to perform inference on such a model and data, including a maximum likelihood approach developed by Asmussen *et al.* (1996) using the EM-algorithm (Dempster *et al.*, 1977) and a fully Bayesian approach developed by Bladt *et al.* (2003) using Markov chain Monte Carlo. Both exploit the ‘missing data’ connection to the unobserved latent process: Asmussen *et al.* (1996) performs expectation-maximisation on the sufficient statistics of the latent process likelihood, and Bladt *et al.* (2003) simulate latent processes concordant with the data to generate random samples of the same sufficient statistics.

#### 3.3.1 Latent process likelihood

Consider the full latent process data, consisting of a separate process,  $\phi_i$ , for each observed absorption time,  $y_i$ . For  $m$  observations there are thus  $m$  sets of latent processes, each comprising: the starting state; sojourns in each state; and state moves through to absorption. Then the likelihood function can be written:

$$f_{\Phi}(\boldsymbol{\phi}; \boldsymbol{\pi}, \mathbf{S}) = \prod_{i=1}^n \pi_i^{b_i} \prod_{i=1}^n \exp\{S_{ii}z_i\} \prod_{i=1}^n \prod_{\substack{j=0 \\ j \neq i}}^n S_{ij}^{N_{ij}} \quad (3.8)$$

where  $\mathbf{b}$ ,  $\mathbf{z}$  and  $\mathbf{N}$  are the sufficient statistics of  $\boldsymbol{\phi}$  with:

$S_{i0} := s_i$  (recall Definition 2.15, page 29)

$b_i :=$  the number of processes with starting state equal to  $i$



$z_i :=$  the total time spent in state  $i$  across all processes

$N_{ij} :=$  the total number of state moves  $i \rightarrow j$  across all processes

$N_{i0} :=$  the total number of absorbing moves  $i \rightarrow n + 1$  across all processes

In summary: Asmussen *et al.* (1996) calculate  $\mathbf{b}$ ,  $\mathbf{z}$  and  $\mathbf{N}$  on the ‘E’ step (numerically, by Runge-Kutta methods) and update  $\pi_i$  and  $S_{ij}$  on the ‘M’ step; whereas Bladt *et al.* (2003) use the current  $\pi_i$  and  $S_{ij}$  to simulate a latent process concordant with  $y_i$ , compute  $\mathbf{b}$ ,  $\mathbf{z}$  and  $\mathbf{N}$  and then use conjugacy results to update  $\pi_i$  and  $S_{ij}$ .

These are now discussed in detail for the remainder of this chapter.

### 3.3.2 The approach of Asmussen *et al.* (1996)

There are certain theoretical limitations of this approach. The first is that the maximum likelihood parameter estimates are point estimates, with no well defined estimate of uncertainty. Indeed this is still a topic of active research, most recently addressed by Bladt *et al.* (2011), who derived the Fisher information matrix to provide asymptotic variances and covariances of the parameter estimates.

The second limitation is the fact that Phase-type distributions are heavily over parameterised. Specifically, a general Phase-type with  $n^2 + n$  parameters has an equivalent parameterisation of dimension  $2n - 1$  (Asmussen *et al.*, 1996), resulting in severe problems of identifiability. The limitation is acknowledged in that work, but is justifiably dismissed because the emphasis is not on the underlying process, rather on attendant estimable quantities like the distribution, density and hazard rate. However, in light of the subject-matter connection of Chapter 2, it is something of major relevance in this work.

Numerous elementary examples were explored by the author to determine the gravity of the problem and the following highlights the issue. 100 observations were simulated from the Phase-type distribution with parameters:

$$\boldsymbol{\pi} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} -3.6 & 1.8 & 1.8 \\ 9.5 & -11.3 & 0 \\ 9.5 & 0 & -11.3 \end{pmatrix}$$

This represents the (repairable) two component parallel system example in (2.10), page 29, with  $\lambda_f = 1.8$  and  $\lambda_r = 9.5$ . The EM-algorithm was executed using the EMPHT

software (Hägström *et al.*, 1992). The first run resulted in maximum likelihood fit:

$$\hat{\mathbf{S}} = \begin{pmatrix} -0.7 & 0.3 & 0.4 \\ 0.1 & -0.7 & 0 \\ 2.9 & 0 & -17.3 \end{pmatrix} \quad \text{with} \quad l(\hat{\mathbf{S}}; \mathbf{x}) = -189.4$$

Repeating the fit without any changes produced maximum likelihood fit:

$$\hat{\mathbf{S}} = \begin{pmatrix} -16.7 & 8.3 & 8.4 \\ 0.1 & -0.7 & 0 \\ 1.0 & 0 & -1.1 \end{pmatrix} \quad \text{with} \quad l(\hat{\mathbf{S}}; \mathbf{x}) = -189.4$$

These two runs of the EM-algorithm resulted in the same log-likelihood values to 4 significant figures, but markedly different estimates. The combination of identifiability issues and the well-known problems the EM-algorithm has with local maxima creates significant difficulties.

These limitations lead one to conclude that in the use-case where the latent process — and thus parameter estimates — are of direct interest, this may not be an appropriate avenue to pursue.

### 3.3.3 The approach of Bladt *et al.* (2003)

The Bayesian method holds the prospect of addressing both of these limitations. Firstly, Bayesian inference provides natural uncertainty estimates because the posterior distribution articulates all knowledge of the parameters in light of the data (see §3.1).

Furthermore, the ability to specify a prior distribution should assist with regularising the inference, dampening modes of the likelihood in parts of the parameter space which would certainly be known a priori to be implausible. For example, the second maximum likelihood estimate above can immediately be seen to be implausible, because the failure rate is an order of magnitude larger than the repair rate. Even relatively dispersed priors for failure and repair rates with small overlapping measure would prevent the above mode from dominating.

Consideration of the above points — and, candidly, the Bayesian predilection of the author — motivated this route early in the research programme. Consequently, the full detail of only this method is reviewed here.

### 3.3.3.1 The methodology

The ‘full data’ likelihood representation in (3.8) is natural, because it is a multiparameter curved Exponential family. However, it can be re-written as:

$$f_{\Phi}(\boldsymbol{\phi}; \boldsymbol{\pi}, \mathbf{G}) = \prod_{i=1}^n \pi_i^{b_i} \prod_{i=1}^n s_i^{N_{i0}} \exp\{-s_i z_i\} \prod_{i=1}^n \prod_{\substack{j=1 \\ j \neq i}}^n S_{ij}^{N_{ij}} \exp\{-S_{ij} z_i\} \quad (3.9)$$

since  $G_{ii} = -\sum_{j \neq i} G_{ij}$ . This form makes explicit a product of  $n^2$  unnormalised Gamma distributions and an unnormalised Dirichlet distribution. Consequently, if independent priors for each parameter are used then in product with this likelihood the following conjugacy results hold:

$$\left. \begin{array}{l} s_i \sim \text{Gamma}(\nu_{i0}, 1/\zeta_i) \\ S_{ij} \sim \text{Gamma}(\nu_{ij}, 1/\zeta_i) \\ \boldsymbol{\pi} \sim \text{Dir}(\beta_1, \dots, \beta_n) \end{array} \right\} \implies \left\{ \begin{array}{l} s_i | \boldsymbol{\phi} \sim \text{Gamma}(N_{i0} + \nu_{i0}, 1/(\zeta_i + z_i)) \\ S_{ij} | \boldsymbol{\phi} \sim \text{Gamma}(N_{ij} + \nu_{ij}, 1/(\zeta_i + z_i)) \\ \boldsymbol{\pi} | \boldsymbol{\phi} \sim \text{Dir}(\boldsymbol{\beta} + \mathbf{b}) \end{array} \right. \quad (3.10)$$

where  $\nu_{ij}$  are shape,  $\zeta_i$  are reciprocal scale and  $\beta_i$  are concentration hyperparameters, with  $\mathbf{N}, \mathbf{z}$  and  $\mathbf{b}$  being sufficient statistics of  $\boldsymbol{\phi}$ .

However, recall that the latent processes are in fact unobserved. Thus, a data augmentation scheme (§3.2.3.3) is used to sample the latent processes conditional on the parameters and vice versa. If  $\Psi$  is the random variable representing the parameters of interest and  $\Phi$  is the random variable for the collection of latent processes, then the full conditionals are  $f_{\Psi|\Phi}(\boldsymbol{\pi}, \mathbf{G} | \boldsymbol{\phi}, \mathbf{y})$  and  $f_{\Phi|\Psi}(\boldsymbol{\phi} | \boldsymbol{\pi}, \mathbf{G}, \mathbf{y})$ . Due to the Hammersley-Clifford Theorem, these fully define the joint posterior so that data augmented Gibbs sampling will result in samples from  $f_{\Psi, \Phi}(\boldsymbol{\pi}, \mathbf{G}, \boldsymbol{\phi} | \mathbf{y})$ .

Sampling  $f_{\Psi|\Phi}(\boldsymbol{\pi}, \mathbf{G} | \boldsymbol{\phi}, \mathbf{y})$  is straightforward due to the conjugacy results elucidated above. Sampling  $f_{\Phi|\Psi}(\boldsymbol{\phi} | \boldsymbol{\pi}, \mathbf{G}, \mathbf{y})$  is rather more involved: this requires sampling a latent process  $\phi_i$  such that it absorbs at precisely time  $y_i$  when started in state  $j$  with probability  $\pi_j$  and evolved in time according to the continuous-time Markov chain with generator  $\mathbf{G}$ .

### 3.3.3.2 Simulating from $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$

Simulation of an individual absorbing continuous-time Markov chain for given  $\boldsymbol{\pi}, \mathbf{G}$  is a straightforward exercise: Algorithm 2.1 (page 27) can be used, with the minor

refinement that simulation plainly only terminates once the process enters state  $n + 1$  (where it remains in perpetuity).

In order to sample conditional upon an absorption time, Bladt *et al.* (2003) propose a Metropolis-Hastings algorithm (§3.2.3.1). First, rejection sampling is used to generate a realisation of the latent process  $\phi_i$  from  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i \geq y_i)$  — that is, repeatedly simulate using the refined Algorithm 2.1 until a simulation is obtained whose absorbing move occurs beyond time  $y_i$ . This sample latent process is then modified by truncating at time  $y_i$  with a move to absorption. More precisely, the process is kept intact up to  $y_i$  ( $\phi_i^{[0, y_i]}$  is unchanged), and then the forced change  $\phi_i^{[y_i, \infty)} = n + 1$  is made.

The procedure in the preceding paragraph acts as a proposal distribution,  $q(x, \cdot)$ , in the Metropolis-Hastings algorithm. Note that in fact this is an independence Metropolis-Hastings algorithm (page 46).

A crucial simplification comes from noting that, due to the strong Markov property (Theorem 2.7) of the continuous-time Markov chain  $\phi_i$ , the following equality holds:

$$f_{\Phi|\Psi}(\phi_i^{[0, y_i]} | \boldsymbol{\pi}, \mathbf{G}, Y_i \geq y_i, \phi_i^{\{y_i^-\}}) = f_{\Phi|\Psi}(\phi_i^{[0, y_i]} | \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i, \phi_i^{\{y_i^-\}})$$

where  $y_i^-$  indicates the instant before time  $y_i$ . That is, the distribution of the process up to time  $y_i$  is conditionally independent of the absorption time, given knowledge of the state immediately prior to absorption. The consequence of this is a dramatic simplification of the acceptance ratio. If  $\phi_i$  is the current process and  $\phi'_i$  is the proposal process, and suppressing conditioning on  $\boldsymbol{\pi}$  and  $\mathbf{G}$  (present in all densities) for legibility:

$$\begin{aligned} \alpha(\phi_i, \phi'_i) &= \frac{f_{\Phi|\Psi}(\phi'_i | Y_i = y_i) f_{\Phi|\Psi}(\phi_i | Y_i \geq y_i)}{f_{\Phi|\Psi}(\phi_i | Y_i = y_i) f_{\Phi|\Psi}(\phi'_i | Y_i \geq y_i)} \\ &= \frac{f_{\Phi|\Psi}(\phi'_i | Y_i = y_i, \phi_i^{\{y_i^-\}}) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i = y_i)}{f_{\Phi|\Psi}(\phi_i | Y_i = y_i, \phi_i^{\{y_i^-\}}) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i = y_i)} \\ &\quad \times \frac{f_{\Phi|\Psi}(\phi_i | Y_i \geq y_i, \phi_i^{\{y_i^-\}}) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i \geq y_i)}{f_{\Phi|\Psi}(\phi'_i | Y_i \geq y_i, \phi_i^{\{y_i^-\}}) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i \geq y_i)} \\ &= \frac{f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i = y_i) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i \geq y_i)}{f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i = y_i) f_{\Phi|\Psi}(\phi_i^{\{y_i^-\}} | Y_i \geq y_i)} \end{aligned}$$

Hence, the acceptance ratio involves only the state of the current and proposal processes at immediately pre-absorption time and does not involve the entire path of

the process. After some routine probability and algebra, Bladt *et al.* (2003) show the acceptance ratio simplifies to:

$$\alpha(\phi_i, \phi'_i) = \frac{s_k}{s_l} \quad \text{where} \quad k = \phi_i^{\{y_i^-\}} \quad \text{and} \quad l = \phi'_i^{\{y_i^-\}}$$

In other words, the acceptance ratio is simply the ratio of absorbing move rates from the pre-absorption states of the proposal process to the current process.

After many steps of this independence Metropolis-Hastings algorithm, the final process may be taken as a sample from the target distribution  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$ . The chain itself is not needed, simply the sufficient statistics  $\mathbf{N}$ ,  $\mathbf{z}$  and  $\mathbf{b}$ , which are then aggregated over all simulated processes  $\phi_i$  and used in sampling updated parameter values.

### 3.3.3.3 Summary

The methodology of Bladt *et al.* (2003) provides a scheme for performing Bayesian inference on absorption time data in Phase-type models, using Gamma and Dirichlet prior distributions. The method is a data augmentation Gibbs sampler:

$$\begin{array}{c} \curvearrowleft f_{\Psi|\Phi}(\boldsymbol{\pi}, \mathbf{G} | \phi, \mathbf{y}) \\ f_{\Phi|\Psi}(\phi | \boldsymbol{\pi}, \mathbf{G}, \mathbf{y}) \curvearrowright \end{array}$$

where the top full conditional relies on conjugacy results and the bottom full conditional is sampled via an independence Metropolis-Hastings algorithm with rejection sampled proposal.

The next chapter engages in a detailed study of the method and as a consequence presents several contributions which extend it, both in model reformulation and computationally.

# Chapter 4

## MCMC for Phase-type Models

The Markov chain Monte Carlo algorithm provided in Bladt *et al.* (2003) enables Bayesian inference to be performed on Phase-type models with dense generator matrices. There are certain difficulties which may not be immediately apparent and the assumption of fully dense generator matrices is a limitation which affects the efficacy of the parameter inference in the subject matter context of Chapter 2.

This chapter commences with the minor contribution of a detailed study of the algorithm, including exact conditions for convergence of the latent process simulation for a certain class of model. It continues with three of the major contributions of the thesis: model reformulation to enable structured generator matrices; computational improvements to the latent process simulation; and incorporation of censored data including computational aspects. The chapter concludes with a comparison of the original and extended methodologies, and draws the theory together into a realistic simulated reliability application performing inference on a repairable five component system with masked system lifetime data.

### 4.1 A study of the original methodology

It is important to note that in their work, Bladt *et al.* (2003) are focused on estimable quantities like the distribution, density and hazard rate which may have generated the observed data, making the Phase-type distribution a germane choice: it is theoretically dense in the space of all distributions on  $[0, \infty)$  (Asmussen, 2000). In that setting, the algorithm they proposed works very well. Thus, the critique here is specifically in

relation to using that work in a way that was not originally envisioned, which leads to some of the contributions of this thesis. Any reader without a direct interest in the latent process and associated parameters may find only the computational advances of particular interest, since issues such as identifiability do not necessarily affect some other inferences of interest, such as prediction.

### 4.1.1 The model set-up

A key feature of the model is the assumption of a dense generator matrix. That is, a generator matrix  $\mathbf{G}$  whose sub-generator  $\mathbf{S}$  and vector of exit rates  $\mathbf{s}$  (see (2.11), page 29) have distinct non-zero parameters in every (off-diagonal) element. This is a feature which may not be desirable when using the Phase-type representation for scientific modelling of a specific process about which there is some domain knowledge. There are at least two kinds of structure it may be desirable to impose on the generator: prohibited and equivalent state moves.

#### 4.1.1.1 Prohibited transitions

The simplest example of the inflexibility of an assumed dense generator for scientific modelling is that of impossible state transitions. There may be a subject-matter reason to know a priori that the process under investigation could not possibly make direct moves between two states without doing so via another state.

Taking the example of reliability modelling which is of interest in this thesis, this would represent moves between varying levels of operational state which may be theoretically impossible. For example, as noted on page 22, in the (repairable) two component parallel system example in (2.10), page 29, a state transition  $2 \rightarrow 3$  is impossible since it represents repair of one component at precisely the same instant as failure of the other. This is an event of probability measure zero in continuous time<sup>1</sup>.

#### 4.1.1.2 Equivalent transitions

The other example of inflexibility in the model assumptions relates to entirely distinct parameters in the dense generator. Once more, there may be subject-matter knowledge

---

<sup>1</sup>There is the concept of ‘uncovered failure’ in reliability theory, whereby the redundancy or fail-over mechanism itself fails, which would manifest as a simultaneous occurrence of continuous time events. This is not a situation considered here.

about the process under investigation, in this case indicating that certain state moves should be treated as (distributionally) the same, up to a constant multiple. Failure to express this in the model leads to a superfluous inflation in the dimension of the parameter space.

Turning specifically to reliability theory, this may arise through the common modelling assumption of independent and identically distributed components. Again, taking the (repairable) two component parallel system example in (2.10), page 29, certain state transitions should (in an idealised modelling sense) have identical parameters. In particular, state transitions  $2 \rightarrow 1$  and  $3 \rightarrow 1$  should be modelled with the the same rate since there is no reason to believe a distributional difference in the rate of repair for each component. Moreover, with only ultimate system failure time available, often any such difference is in fact unidentifiable. This desirable modelling assumption would allow a significant reduction in the dimension of the parameter space.

#### 4.1.1.3 Scope of the issue

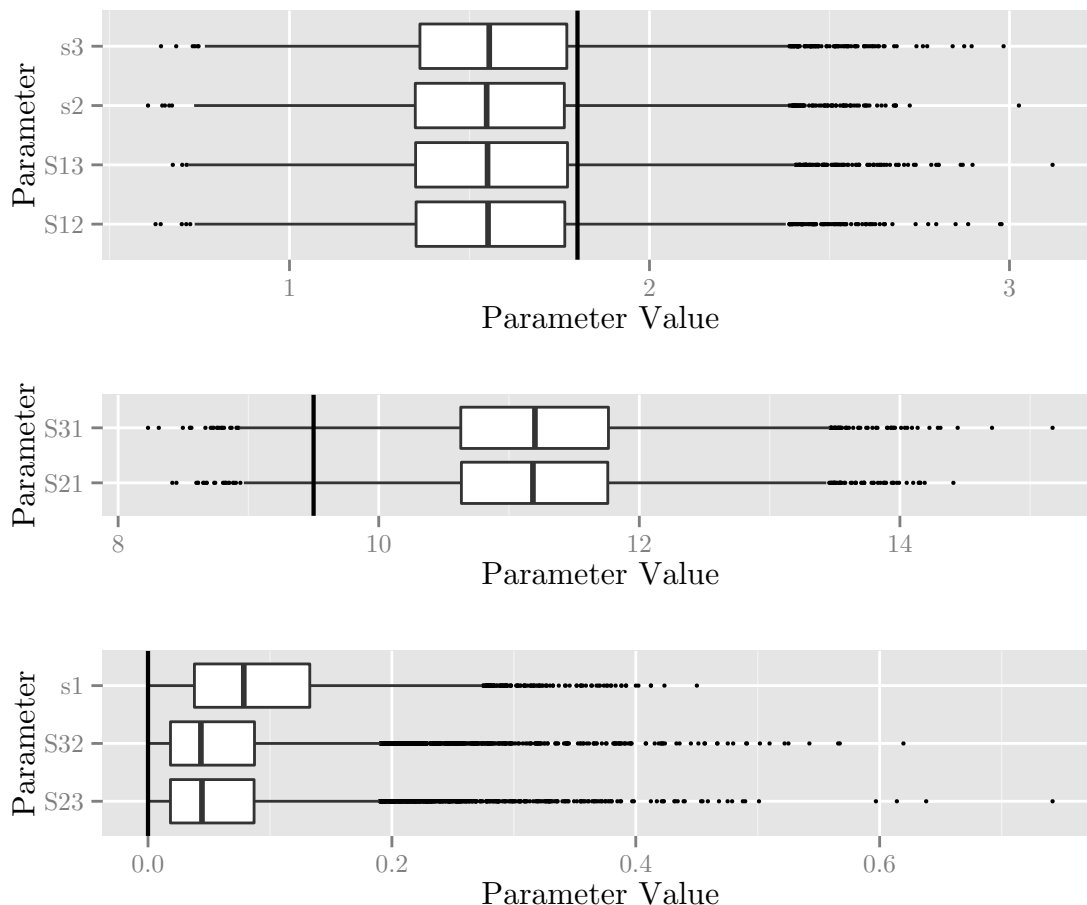
Whilst there seems to be a logical cause for concern about the assumptions made in relation to the generator matrix, it is prudent to at least briefly examine whether these potential drawbacks actually manifest in a way which is detrimental to the inferential conclusions, before perhaps prematurely pursuing a resolution. One informal way to do this is to infer back to some known ground truth which was used to produce simulated data.

Taking as ground truth the generator which represents the (repairable) two component parallel system example in (2.10), page 29, with  $\lambda_f = 1.8$  and  $\lambda_r = 9.5$ :

$$\mathbf{G} = \begin{pmatrix} -3.6 & 1.8 & 1.8 & 0 \\ 9.5 & -11.3 & 0 & 1.8 \\ 9.5 & 0 & -11.3 & 1.8 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

25 absorption times were simulated. Then, the algorithm of Bladt *et al.* (2003) was run for 10,000 iterations, chosen to bring the 95% quantile standard errors down sufficiently to give two to three significant digits. The priors used were as per the formulation in





**Figure 4.1:** Boxplot for posterior simulations from known ground truth using Bladt *et al.* (2003). Solid vertical lines show ground truth values.

(3.10), page 60, with:

$$\begin{aligned}
 \nu_{ij} &= 24 \quad \text{for } (i, j) \in \{(1, 2), (1, 3), (2, 0), (3, 0)\} \\
 \nu_{ij} &= 180 \quad \text{for } (i, j) \in \{(2, 1), (3, 1)\} \\
 \nu_{ij} &= 1 \quad \text{for } (i, j) \in \{(2, 3), (3, 2), (1, 0)\} \\
 \zeta_i &= 16 \quad \forall i
 \end{aligned}$$

Table 4.1 shows the resultant summary estimates, with MCMC standard errors calculated as discussed at the end of §3.2.4. Posterior sample box plots are shown in Figure 4.1.

The most striking positive note from these results is that almost all the posterior estimates of generator entries which should be equal are equal within sampling error. However, even in this toy example the posterior for  $\lambda_r$  does not include the ground

Ground Truth	Parameter	Expectation	Lower 95% quantile	Upper 95% quantile
$\lambda_f = 1.8$	$S_{12}$	1.57 ( 0.00761 )	1.02 ( 0.0107 )	2.21 ( 0.015 )
	$S_{13}$	1.57 ( 0.00894 )	1.02 ( 0.00956 )	2.26 ( 0.0184 )
	$s_2$	1.57 ( 0.00386 )	1.02 ( 0.00658 )	2.21 ( 0.0111 )
	$s_3$	1.57 ( 0.00387 )	1.03 ( 0.0066 )	2.22 ( 0.00969 )
$\lambda_r = 9.5$	$S_{21}$	11.2 ( 0.0107 )	9.65 ( 0.0199 )	12.9 ( 0.0292 )
	$S_{31}$	11.2 ( 0.0113 )	9.63 ( 0.0192 )	12.9 ( 0.0266 )
prohibited, 0	$S_{23}$	0.0632 ( 0.000816 )	0.00163 ( 0.000185 )	0.229 ( 0.00476 )
	$S_{32}$	0.0636 ( 0.000922 )	0.00149 ( 0.000195 )	0.24 ( 0.00531 )
	$s_1$	0.0919 ( 0.00182 )	0.00381 ( 0.000509 )	0.251 ( 0.00361 )

**Table 4.1:** Posterior summary statistics for simulations from known ground truth using Bladt *et al.* (2003). Bracketed figures are MCMC standard errors.

truth in the central 95% quantile. Additionally, there is a relatively extended right tail for those entries which should in fact be zero, despite quite a strong prior weight toward zero on those parameters — this was done because this scenario assumes knowledge of the structure and in the absence of model reformulation a practitioner would place strong prior weight to zero.

The conjecture toward the end of the last chapter that this Bayesian approach would render substantially better results than the frequentist EM-algorithm seems to hold here. However, these results, inferring against a known ground truth, appear to leave the door open to possible further improvement if the model can be reformulated to make use of knowledge about the mechanism which generates the process.

### **4.1.2 Censored data**

It is not uncommon for the kind of data which are available in a reliability analysis to include substantial quantities of right-censored observations. There was no discussion in Bladt *et al.* (2003) of censored observations and, as such, no means of accommodating data sets which include them. As already reviewed in §2.1.2, simply ignoring such observations is not an acceptable solution and so adapting the methodology to enable their incorporation is an important contribution.

### **4.1.3 Computational performance**

Nothing in the original Metropolis-Hastings algorithm for sampling continuous-time Markov chains conditional on an exact absorption time was contingent on a dense generator. Therefore, any model reformulation which facilitated structured generator matrices would, in principle, require no change to simulation of the latent process. However, there are some existing performance problems and, if structured generators are to be accommodated, then in certain circumstances this leads to additional severe degradation in the computational tractability. This is now discussed in problems I, II and III below.

### 4.1.3.1 I. Exploration of parameter space

By nature, an MCMC sampling scheme should explore all parts of the parameter space with non-negligible posterior probability. Thus, the Gibbs sampler will explore the space of  $(\boldsymbol{\pi}, \mathbf{G})$ , sometimes making moves toward areas of lower posterior density, for which the Metropolis-Hastings step must then simulate a latent process. However, the rejection sampling approach can encounter problems for some regions of these low posterior density samples.

Consider the acceptance probability for the rejection sampler. The rejection sampler is merely forward simulating an absorbing continuous-time Markov chain, rejecting sample processes that do not reach the observation time,  $y_i$ , before absorbing. Hence, the acceptance probability is simply  $\mathbb{P}(Y_i > y_i | \boldsymbol{\pi}, \mathbf{G})$ . Therefore, the distribution of the number of trials required to draw a suitably long simulated chain is Geometric, with success probability parameter  $p = \boldsymbol{\pi}^T \exp\{y_i \mathbf{S}\} \mathbf{e}$

Consequently, if  $(\boldsymbol{\pi}, \mathbf{G})$  is in a region of  $\Omega$  where  $p$  is small for some given  $y_i$ , then the rejection sampler can completely stall because it is so unlikely that any chain will be produced which absorbs so far into the tail. For example, if  $K$  is the number of rejection samples required to produce a proposal sample process, then:

$$\mathbb{P}(Y_i > y_i | \boldsymbol{\pi}, \mathbf{G}) = 0.001 \implies \mathbb{E}(K) = 1000, \quad 95\% \text{ CI} = [25, 3687]$$

$$\mathbb{P}(Y_i > y_i | \boldsymbol{\pi}, \mathbf{G}) = 10^{-6} \implies \mathbb{E}(K) = 1000000, \quad 95\% \text{ CI} = [25317, 3688877]$$

Hence, chains are wastefully sampled, potentially millions of times, to find one chain absorbing beyond  $y_i$ . One should bear in mind that:

- i. The Metropolis-Hastings algorithm is run on every iteration of the data augmented Gibbs algorithm.
- ii. Each time, it must generate a latent process for every observation,  $y_i$ , in the data.
- iii. Each such Metropolis-Hastings run may involve many iterations before converging to the stationary target distribution (that is, a process absorbing at precisely  $y_i$ ).
- iv. Each such iteration involves producing a rejection sampled proposal.

It becomes quickly apparent that even modestly small  $p$  can lead to reduced computational tractability due to the amplifying effects listed above. Diffuse priors can further exacerbate this.

This first problem exists whether any model reformulation away from dense generators is performed or not.

#### 4.1.3.2 II. Zero constraints for absorbing moves

This problem would only occur under a model reformulation which facilitated structured generators as discussed in §4.1.1, so does not affect the original methodology.

The acceptance probability in the Metropolis-Hastings algorithm enables truncated samples from  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i > y_i)$  to be used to generate samples which are concordant with the observed absorption times, from  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$ . The truncation of proposal processes needs to be handled with care once there is the possibility that some of the rates of moves to absorption,  $s_i$ , may be zero. There is the possibility that truncating a rejection sampled chain will produce an impossible move: if  $s_j = 0$  is a constraint, for some  $j$ , and the rejection sample is such that  $\phi_i^{\{y_i\}} = j$ , then truncating and inserting a move  $j \rightarrow n + 1$  — that is, forcing  $\phi_i^{\{y_i, \infty\}} = n + 1$  — will result in a sample of the latent process  $\phi_i$  which is invalid and must be discarded.

The severity of this problem can be acute if the states most commonly occupied by the process are those from which absorbing moves are disallowed. Take the (repairable) two component parallel system example in (2.10), page 29: in such a system comprising of reliable components, full operation is the state in which the process spends almost all the time (i.e. state 1). In an idealised modelling sense this is the state from which a direct absorbing move (failure) is impossible. Indeed, if  $\lambda_f$  is several orders of magnitude smaller than  $\lambda_r$  — as would be expected in such a system — the intractability becomes grave. The probability mass function of the state the process will occupy where truncation occurs can be computed directly:

$$\mathbb{P}\left(\phi_i^{\{y_i\}} = j \mid Y_i > y_i\right) = \frac{\boldsymbol{\pi}^T \exp\{y_i \mathbf{S}\} \mathbf{e}_j}{\boldsymbol{\pi}^T \exp\{y_i \mathbf{S}\} \mathbf{e}}$$

An example of what this means for the (repairable) two component parallel system example can be seen in Table 4.2.

Clearly, when dealing with simulating the latent process for highly reliable systems there will be a large number of unusable simulations produced by rejection sampling. Again, the number of simulations required will be Geometrically distributed. The

State, $j$	Meaning	$\mathbb{P}(\phi_i^{\{y_i\}} = j   Y_i > y_i)$
1	both working	0.9980
2	1 failed, 2 working	0.0010
3	1 working, 2 failed	0.0010

**Table 4.2:** Example probability mass function of latent process state at time  $y_i = 500$ , for  $\lambda_f = 1, \lambda_r = 1000$  in the (repairable) two component parallel system.

example in Table 4.2 will have  $p = 1 - 0.9980 = 0.002$  meaning:

$$\mathbb{E}(\# \text{ unusable rej. samp.}) = 500 \quad \text{and} \quad 95\% \text{ CI} = [12, 1872]$$

This second problem can compound with and amplify the first even further. In testing, the algorithm has been observed effectively stalling for days of computation time on a single MCMC iteration in only mildly pathological examples.

#### 4.1.3.3 III. Convergence of process samples to stationarity

The Metropolis-Hastings algorithm can take time to converge to stationarity. That is, the sequence of truncated sample processes proposed takes time to reach the target of processes absorbed at an exact time, as discussed in §3.2.4. Interestingly, in this situation it is more straightforward than usual when implementing Metropolis-Hastings to analyse the time taken to reach stationarity analytically. The novel Theorem proved here seems to proffer good news on at least one front and in particular indicates a possible course of action.

To be specific, an easy case to examine (although the ideas here can be correspondingly used numerically in any case) is when there are only two states from which absorption can occur. This could be the (repairable) two component parallel system example, or indeed for any  $n + 1$ -state latent space where only two  $s_i$  are non-zero. Both the statement and proof of the following Theorem are contributions of this thesis.

**Theorem 4.1** *Let  $\mathbf{G}$  be the generator of a continuous-time Markov chain with an absorbing state, taken without loss of generality to be in the form (2.11), page 29. Let  $\boldsymbol{\eta}$  be the discrete probability mass vector of states occupied by valid proposal processes at the truncation time  $y$ ,*

$$\eta_i = \mathbb{P}(\phi^{\{y\}} = i | \boldsymbol{\pi}, \mathbf{G}, Y \geq y, s_i \neq 0)$$

If  $s_j \neq 0$  for precisely two  $j$ , which may be taken as  $s_1$  and  $s_2$  without loss of generality, then the Metropolis-Hastings algorithm for sampling the latent process conditional on an absorbing move occurring at time  $Y = y$  in Bladt et al. (2003) is geometrically ergodic with rate of convergence:

$$1 - \eta_1 \left( \min \left\{ 1, \frac{s_1}{s_2} \right\} \right) - \eta_2 \left( \min \left\{ 1, \frac{s_2}{s_1} \right\} \right)$$

*Proof:* Recall that here, the Metropolis-Hastings algorithm takes proposal processes which absorb beyond an observation time  $y$  and through the acceptance ratio turns these into processes conditional on absorption at precisely time  $y$  by ensuring the correct proportion of immediate pre-absorption states. Due to the strong Markov property (Theorem 2.7, page 23) this can be properly thought of as a Metropolis-Hastings algorithm on the discrete space of immediate pre-absorption states. Furthermore, that distribution of states immediately before absorption (at  $y$ ) can be explicitly derived for both proposal and target distributions.

Since the generator under consideration has  $s_i = 0 \forall i \notin \{1, 2\}$ , all valid rejection sampled proposal processes  $\phi$  will be in state 1 or 2 at time  $y$ . Denote by  $\rho_i$  ( $i = 1, 2$ ) the target distribution of these two states immediately before absorption:

$$\rho_i = \mathbb{P}(\phi^{\{y^-\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y = y) \propto \boldsymbol{\pi}^T \exp\{y\mathbf{S}\} \mathbf{e}_i s_i$$

where  $y^-$  indicates the instant before time  $y$ . As per the Theorem statement, let  $\eta_i$  denote the probability that a valid (in the sense that truncation at  $y$  does not create an invalid process) rejection sample has pre-absorbing state  $i$ :

$$\eta_i = \mathbb{P}(\phi^{\{y^-\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y \geq y, s_i \neq 0) \propto \boldsymbol{\pi}^T \exp\{y\mathbf{S}\} \mathbf{e}_i$$

Note that  $\eta_i$  is non-zero only for  $i = 1$  and  $i = 2$ . Then, the Metropolis-Hastings algorithm, viewed through the lens of the immediately pre-absorption state, induces a two state discrete time Markov chain, with initial probability vector  $(\eta_1, \eta_2)$  and irreducible stochastic transition matrix:

$$\mathbf{P} = \begin{pmatrix} \eta_1 + \eta_2(1 - \min\{1, \frac{s_2}{s_1}\}) & \eta_2(\min\{1, \frac{s_2}{s_1}\}) \\ \eta_1(\min\{1, \frac{s_1}{s_2}\}) & \eta_2 + \eta_1(1 - \min\{1, \frac{s_1}{s_2}\}) \end{pmatrix}$$

It is not hard to see that this induced Markov chain is  $\phi$ -irreducible for the counting measure and aperiodic. Therefore, since the state space is finite, the induced Markov chain will be ergodic (in fact, geometrically ergodic).

As noted in Diaconis (1996), for discrete probability masses the total variation distance has an easy form for computation:

$$\|\pi_1 - \pi_2\| = \sup_{A \subseteq \Omega} \{|\pi_1(A) - \pi_2(A)|\} = \frac{1}{2} \sum_{x \in \Omega} |\pi_1(x) - \pi_2(x)|$$

This leads to the total variation distance from stationarity after  $k$  steps of the Metropolis-Hastings algorithm:

$$\frac{1}{2} |\boldsymbol{\eta}^T \mathbf{P}^k - \boldsymbol{\rho}^T| \mathbf{e}$$

where  $\boldsymbol{\eta}^T \mathbf{P}^k \xrightarrow{k \rightarrow \infty} \boldsymbol{\rho}^T$ . Clearly, the second eigenvalue of  $\mathbf{P}$  determines the rate of convergence (since the first eigenvalue is  $\lambda_1 = 1$  by the Perron-Frobenius Theorem). The second eigenvalue of  $\mathbf{P}$  is  $\lambda_2 = 1 - \eta_1(\min\{1, \frac{s_1}{s_2}\}) - \eta_2(\min\{1, \frac{s_2}{s_1}\})$ , which may be seen by solving the characteristic equation of  $\mathbf{P}$ , that is:  $\det(\mathbf{P} - \lambda \mathbf{I}) = 0$  (Roman, 2008, p.186). Consequently, the total variation distance from chain to target is  $o(\lambda_2^k)$ , meaning the chain is geometrically ergodic with rate of convergence  $\lambda_2$ .  $\square$

If the objective was to explore the sample space of each latent process (say to learn about some functional of the process), then the geometrically ergodic convergence would be enough. However, the actual usage of the algorithm as part of a data augmentation scheme means that interest is only in producing a single sample. Consequently, although geometric ergodicity reassures one that convergence can be achieved quickly, this may not be adequate: if each iteration involves much work then anything more than a few iterations to produce just one sample process may be prohibitive, especially in light of the amplifying affect of problems I and II in producing rejection sampled proposals.

Since the rate of convergence is known explicitly, the circumstances in which this will occur can be easily seen, as follows.

**Corollary 4.2** *Under the conditions of Theorem 4.1, convergence of the Metropolis-Hastings algorithm to simulate the latent process in Bladt et al. (2003) will require many steps if both  $\eta_1(\min\{1, \frac{s_1}{s_2}\})$  and  $\eta_2(\min\{1, \frac{s_2}{s_1}\})$  are small. This occurs if  $s_1$  and  $s_2$  are orders of magnitude different and  $\eta_i$  is small for the larger  $s_i$ .*



Even with small generator matrices, this has been observed causing many iterations to be required to reduce the total variation distance to a negligible size.

It is interesting to note that the proof of the Theorem does point to a possible route forward: the fact the Metropolis-Hastings algorithm induces a two state chain whose target distribution can be computed means that one alternative would be to rejection sample the chains from this distribution. Of course, the thought of a rejection sampler within a rejection sampler does induce a feeling of foreboding. However, the realisation proved useful in a different guise when the explicit use of this target distribution provided the penultimate solution proposed to this computational problem (see §4.2.2, pilot solutions).

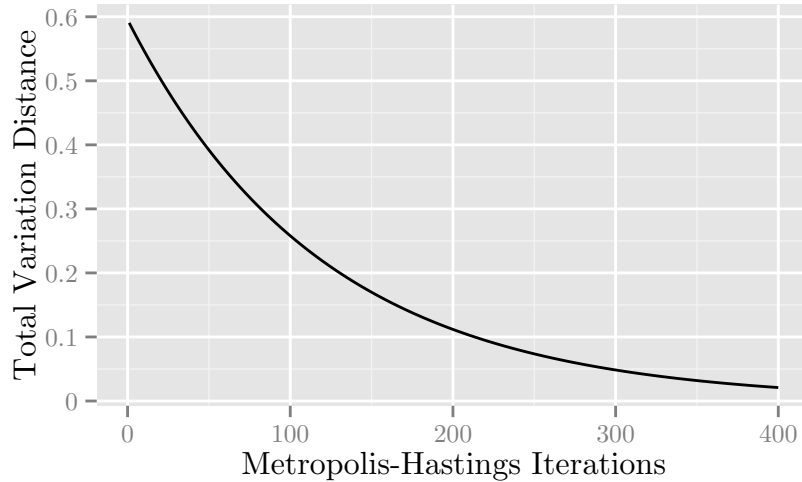
By way of an example, consider that the following generator has been produced by the data augmented Gibbs step and that the Metropolis-Hastings algorithm is being called upon to generate a latent process sample which absorbs at  $y = 600$ , with  $\boldsymbol{\pi}^T = (1, 0, 0)$ .

$$\mathbf{G} = \begin{pmatrix} -2 & 0.01 & 1.99 & 0 \\ 1 & -300 & 0 & 299 \\ 299 & 0 & -300 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (4.1)$$

This satisfies the conditions of Corollary 4.2 and indeed not at all pathologically so. The ratio of exit rates is  $s_2 = 299 : 1 = s_3$  — so around two orders of magnitude difference — and in this case:

$$\begin{aligned} \eta_2 &\propto \boldsymbol{\pi} \exp\{400\mathbf{S}\}\mathbf{e}_2 &\implies &\eta_2 = 0.6004 \\ \eta_3 &\propto \boldsymbol{\pi} \exp\{400\mathbf{S}\}\mathbf{e}_3 &\implies &\eta_3 = 0.3996 \end{aligned}$$

so that  $\eta_2$  could hardly be described as very small. Yet, even in this non-pathological case the total variation distance after  $k$  iterations can be plotted as in Figure 4.2. The total variation distance reaches around 5% at approximately 300 iterations. This is a non-pathological example and 300 Metropolis-Hastings iterations seems a modest requirement. However, tracing this example back through problems I and II highlights the compounding which can occur. Let  $p$  be the probability of a successful rejection



**Figure 4.2:** Total variation distance from the target distribution.

sample for the process above. Then,

$$\begin{aligned}
 p &= \mathbb{P}(Y > 600 \cap \phi^{\{600\}} \in \{2, 3\} \mid \boldsymbol{\pi}, \mathbf{G}) \\
 &= \mathbb{P}(\phi^{\{600\}} \in \{2, 3\} \mid \boldsymbol{\pi}, \mathbf{G}, Y > 600) \mathbb{P}(Y > 600 \mid \boldsymbol{\pi}, \mathbf{G}) \\
 &= 3.34 \times 10^{-7}
 \end{aligned}$$

If one decides a discrepancy from the true distribution of 5% in total variation distance is acceptable, then 300 rejection sampled processes will be required. The total number of samples,  $K$ , required in order to produce a single sample from  $f_{\Phi \mid \Psi}(\phi_i \mid \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$  is then Negative Binomially distributed with  $r = 300, p = 3.34 \times 10^{-7}$ , so that

$$\mathbb{E}[K] = 897,867,259 \quad \text{and} \quad 95\% \text{ CI} = [799,129,604, 1,002,273,104]$$

It is worth pausing at the end of this subsection to reinforce the fact that these figures are to produce one sample process, concordant with one observation, at one realisation of the parameters: this may be roughly multiplied up by the size of the dataset,  $\mathbf{y}$ , and then multiplied up again for the number of data augmented Gibbs steps required to reduce the Markov chain Monte Carlo standard error for the expectation of the functional of interest to an acceptable level.

#### 4.1.4 Summary

This section has provided an analysis of the methodology of Bladt *et al.* (2003) in terms of both model specification and computational performance. In the process,

aspects which either impede effective inference or simply make computing the answer impractical in the context of reliability modelling have been elucidated.

The next section provides several major contributions of this thesis, with extensions to the methodology of Bladt *et al.* (2003) presented.

## 4.2 Extensions to the methodology

The extensions to the original work fall into three areas: model reformulation, computational improvements and censored data.

### 4.2.1 Model reformulation

As discussed in §4.1.1, the algorithm assumes a rate matrix for the Phase-type distribution which is both completely dense and where every rate can vary independently. However, when the objective is scientific modelling of a process about which there is some known structure or theory this may not be a desirable property.

Firstly, constraining a parameter to zero is a straight-forward procedure as was the case for the EM-algorithm of Asmussen *et al.* (1996). The parameter is simply fixed at zero when simulating the latent process and no new values are drawn in the Gibbs step: this solves the problem because this entry of  $\mathbf{G}$  is no longer treated as an unknown parameter, hence excluding it from all inferential procedures.

The second matter — that of constrained equality for certain parameters — requires examination of the posterior density. In the original methodology, parameters on different rows of the Markov chain generator shared a common scale hyperparameter in their Gamma priors with freely varying shape, but when adding special structures this can lead to every parameter in the model sharing a common scale hyperparameter if there is a single parameter which appears in every row (or indeed, if there is just no empty intersection between the rows in which all parameters appear). Thus, we commence by adjusting the prior structure with freely varying scale hyperparameters and thereafter show the model can be reformulated with conjugacy preserved.

Take  $\lambda_k, k = 1, \dots, m$ , as the parameters appearing in the generator of the Phase-type distribution. Then each  $S_{ij} = 0$  or  $c_{ij}\lambda_k$  and each  $s_i = 0$  or  $c_{i0}\lambda_k$  for some  $k$  and

$c_{ij} \in \mathbb{R}^+$ . Let

$$\Lambda_k = \{(i, j) : S_{ij} = c_{ij}\lambda_k\} \cup \{(i, 0) : s_i = c_{i0}\lambda_k\}$$

represent the set of all entries where  $\lambda_k$  appears, for each  $k$ .

Specify independent Gamma priors on each  $\lambda_k$  and retain the Dirichlet prior on  $\boldsymbol{\pi}$ :

$$\lambda_k \sim \text{Gamma}(\text{shape} = \nu_k, \text{scale} = 1/\zeta_k)$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^\top)$$

so that the full prior specification is:

$$f_\Psi(\boldsymbol{\pi}, \mathbf{G}) = \prod_{i=1}^n \pi_i^{\beta_i-1} \prod_{i=1}^m \lambda_i^{\nu_i-1} e^{-\lambda_i \zeta_i}$$

The likelihood function for a Phase-type density can be written as in (3.9), page 60:

$$f_\Phi(\boldsymbol{\phi}; \boldsymbol{\pi}, \mathbf{G}) = \prod_{i=1}^n \pi_i^{b_i} \prod_{i=1}^n s_i^{N_{i0}} \exp\{-s_i z_i\} \prod_{i=1}^n \prod_{\substack{j=1 \\ j \neq i}}^n S_{ij}^{N_{ij}} \exp\{-S_{ij} z_i\}$$

where  $\mathbf{b}, \mathbf{N}$  and  $\mathbf{z}$  are the sufficient statistics of the chain resulting from simulation of the latent process. Recall, these quantities are: number of transitions  $i \rightarrow j$ ,  $N_{ij}$ ; number of absorbing moves  $i \rightarrow n+1$ ,  $N_{i0}$ ; total time in state  $i$ ,  $z_i$ ; and number of chains starting in state  $i$ ,  $b_i$ .

Letting

$$N_k^* = \sum_{(i,j) \in \Lambda_k} N_{ij}$$

$$z_k^* = \sum_{(i,j) \in \Lambda_k} c_{ij} z_i$$

enables the likelihood to be directly expressed in terms of the  $m$  parameters:

$$f_\Phi(\boldsymbol{\phi}; \boldsymbol{\pi}, \mathbf{G}) = \prod_{i=1}^n \pi_i^{b_i} \prod_{i=1}^m \lambda_i^{N_i^*} \exp\{-\lambda_i z_i^*\}$$

The posterior can then be written as:

$$\begin{aligned} f_{\Psi|\Phi}(\boldsymbol{\pi}, \mathbf{G} | \boldsymbol{\phi}) &\propto f_\Psi(\boldsymbol{\pi}, \mathbf{G}) f_\Phi(\boldsymbol{\phi}; \boldsymbol{\pi}, \mathbf{G}) \\ &= \left( \prod_{i=1}^n \pi_i^{\beta_i-1} \prod_{i=1}^m \lambda_i^{\nu_i-1} \exp\{-\lambda_i \zeta_i\} \right) \left( \prod_{i=1}^n \pi_i^{b_i} \prod_{i=1}^m \lambda_i^{N_i^*} \exp\{-\lambda_i z_i^*\} \right) \\ &= \prod_{i=1}^n \pi_i^{\beta_i+b_i-1} \prod_{i=1}^m \lambda_i^{\nu_i+N_i^*-1} e^{-\lambda_i(\zeta_i+z_i^*)} \end{aligned}$$

Thus, the full conditional posteriors for each parameter are now:

$$\begin{aligned}\lambda_k | \phi &\sim \text{Gamma}(\text{shape} = \nu_k + N_k^*, \text{scale} = 1/(\zeta_k + z_k^*)) \\ \boldsymbol{\pi} | \phi &\sim \text{Dirichlet}(\boldsymbol{\beta} = (\beta_1 + b_1, \dots, \beta_n + b_n))\end{aligned}$$

This shows it is possible to impose structure in the form of zeros and equal entries in the generator,  $\mathbf{G}$ , whilst maintaining the conjugacy properties of a Gamma prior, as derived on page 36. This retains the simple form of the data augmented Gibbs sampling procedure, with an elementary change to the parameter updating through  $\mathbf{N}^*$  and  $\mathbf{z}^*$ .

The next section considers improvements to the latent process sampling approach.

## 4.2.2 Latent process simulation

To address the problems detailed in §4.1.3, we develop an alternative to the Metropolis-Hastings approach to simulating the latent process.

### 4.2.2.1 Pilot solutions

In the course of this research, the author progressed through numerous ideas of varying efficacy before arriving at the solution presented in detail in the sequel. Those early ideas are presented in highly abbreviated form here to provide a brief narrative behind the development.

Very early work included a scheme to simulate directly from  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i \geq y_i)$  to eliminate the need for rejection sampling the proposals. However, this only addresses problem I, and not problems II or III. This earliest concept remains in some form as will be seen later when handling censored data.

A second evolution of the idea transpired to be overly tailored to reliability problems. It resolved problem II in situations where there was a fixed starting state which coincided with the most commonly inhabited state of the process, but from which absorption was impossible. For example, state 1 in the (repairable) two component parallel system example in (2.10), page 29, fulfils this criterion. The solution involved simulating the chain in reverse: that is, select the starting state from  $\mathbb{P}(\phi^{\{y\}} = i | \boldsymbol{\pi}, \mathbf{G}, Y = y)$  and then simulate in reverse until time zero, ergo 99.8% of sampled processes are valid (for the example values in Table 4.2). However, the

problem-specific nature of the solution is troublesome and after lengthy efforts there was limited progress toward mathematically proving the resultant samples were from the required target distribution. This avenue was therefore set aside.

The pursuit of a resolution for problem II led to considering the possibility of simulating the required sufficient statistics directly, rather than simulating the latent process. Of course Asmussen *et al.* (1996) provide a means of finding the expectation of the sufficient statistics for any given parameters, but the distributional form (even approximately) of these remained elusive. This route was set aside, but not forgotten. It is interesting that recently, and well after having decided to move on from this avenue, Hobolth and Jensen (2011) examined continuous-time Markov process summary statistics under somewhat similar endpoint conditioning, which it may be possible to adapt. This falls under an exciting avenue of future research.

The first generally applicable breakthrough in addressing problem II was adapting a solution to a different problem, presented in Hobolth and Stone (2009). They consider simulating an irreducible continuous-time Markov process conditional on the process inhabiting a given state at a certain time. In contrast, the processes under consideration here are reducible and conditioned on a particular state move occurring at an exact time (i.e. a move to the closed absorbing state at an exact time, not merely being in that state at a given time). This was adapted in two ways:

The first version adjusted their ‘direct simulation’ method to handle reducibility and additionally so that the end-point conditioning was on a set of states rather than a single one (here, those states  $i$  where  $s_i \neq 0$ ). This was then used to produce the proposal processes in place of rejection sampling, because the resultant samples would be guaranteed to reach  $y$  and to inhabit a valid state for truncation.

The second version followed after the development of Theorem 4.1. In this version, the end-point state is first drawn from the discrete distribution  $\mathbb{P}(\phi^{\{y\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y = y)$  and then the above method used to sample a process which will be in that state at time  $y$ . This obviates any need for a Metropolis-Hastings update at all: the first such process sampled will be from the target distribution upon truncation, by the strong Markov property.

However, the first assumption of Hobolth and Stone (2009) in the derivation is that an eigendecomposition of the irreducible continuous-time Markov chain generator is

available for computation of the matrix exponential. As indicated in Moler and Van Loan (2003), this is not always an ideal choice for computing the matrix exponential, although it does possess very attractive computational advantages. This leads to the solution which is to be presented in detail in this thesis. In a similar spirit to that work, an algorithm was therefore developed which: is tailored specifically to the exact state move problem at hand; admits computation in the event that eigendecomposition is inadequate or unavailable; and yet allows the computational advantage of eigendecomposition when it is available and accurate. This new algorithm has been named ‘Exact Conditional Sampling’ (ECS).

#### 4.2.2.2 Exact Conditional Sampling (ECS)

The approach is to adapt the standard algorithm for unconditional simulation of an absorbing continuous-time Markov chain (Algorithm 2.1) by explicitly conditioning on an absorption time  $Y = y$  in every part of that algorithm. To do so, new expressions for the starting state, sojourn times and state moves must be derived. The discrete nature of the starting state and state move probabilities make them, in principle at least, routine. The statement and proof of the following Lemma provides them and is a contribution of this thesis.

**Lemma 4.3 (ECS starting state and transitions)** *Let  $\mathbf{G}$  be the generator of an  $n + 1$  state absorbing continuous-time Markov process in the form (2.11), and let  $\boldsymbol{\pi}$  be the probability vector for the distribution of the starting state. Conditional on the fact that the process enters state  $n + 1$  at exactly time  $Y = y$ ,*

1. *the distribution of the starting state is:*

$$\mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y = y) = \frac{\mathbf{e}_i^T \exp\{\mathbf{S}y\} \mathbf{s} \pi_i}{\boldsymbol{\pi}^T \exp\{\mathbf{S}y\} \mathbf{s}} \quad (4.2)$$

2. *if the process has reached time  $t$  and it has already been determined that a non-absorbing jump occurs after a further time  $d$  — so that  $t + d < y$  — then the embedded discrete chain state move probabilities become:*

$$\begin{aligned} \mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, Y = y, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \in \{1, \dots, n\} \setminus i) \\ \propto S_{ij} \mathbf{e}_j^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{s} \end{aligned} \quad (4.3)$$

*Proof:* The unconditional selection of starting state is simply a draw from the discrete probability mass function defined by  $\boldsymbol{\pi}$

$$\mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{S}) = \pi_i$$

The distribution of the random variable  $Y$  is Phase-type (Definition 2.15). Therefore, for the first part of the Lemma, the conditional version of the starting state introduces a Phase-type adjustment:

$$\begin{aligned} \mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{S}, Y = y) &= \frac{f_Y(Y = y \mid \boldsymbol{\pi}, \mathbf{S}, \phi^{\{0\}} = i) \mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{S})}{f_Y(Y = y \mid \boldsymbol{\pi}, \mathbf{S})} \\ &= \frac{f_Y(Y = y \mid \mathbf{S}, \phi^{\{0\}} = i) \mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{S})}{f_Y(Y = y \mid \boldsymbol{\pi}, \mathbf{S})} \\ &= \frac{\mathbf{e}_i^T \exp\{\mathbf{S}y\} \mathbf{s} \pi_i}{\boldsymbol{\pi}^T \exp\{\mathbf{S}y\} \mathbf{s}} \end{aligned}$$

as required, where  $\mathbf{e}_i$  is a vector of zeros with a single one in the  $i^{\text{th}}$  position.

Unconditional on any absorption information, selection of a state move, after the sojourn  $d$  has been determined, is a draw from the discrete mass defined by:

$$\mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \in \{1, \dots, n+1\} \setminus i) = -G_{ij}/G_{ii} \quad \forall t \geq 0, d > 0$$

For the second part of the Lemma, the conditional version again introduces a Phase-type adjustment:

$$\begin{aligned} \mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, Y = y, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \in \{1, \dots, n\} \setminus i) \\ \propto f_Y(Y = y \mid \mathbf{G}, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} = j \in \{1, \dots, n\} \setminus i) \\ \times \mathbb{P}(\phi^{\{t+d\}} = j \mid \mathbf{G}, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \in \{1, \dots, n\} \setminus i) \\ = S_{ij} \mathbf{e}_j^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{s} \end{aligned}$$

as required. □

The sojourn time distribution is slightly more complicated. At a time  $t$ , the cumulative distribution function of the sojourn is continuous on  $[0, y - t)$ , with a jump discontinuity at  $\{y - t\}$  to 1 — a generalised probability density function: there is no probability density function with respect to the usual Lebesgue measure. Although the single time  $y - t$  has Lebesgue measure zero, there is a discrete mass of probability because ultimately some jump must be made to absorption at exactly time  $y$ .



Thus, sampling would best be handled by splitting the sojourn time in two: first determine whether the sojourn is exactly  $y - t$  (i.e. absorption) or not; if not the jump time is then from a continuous distribution over a finite half-closed interval (whose probability density function can be written down). The statement and proof of the following Lemma shows how to achieve this and is a contribution of the thesis.

**Lemma 4.4 (ECS sojourn time)** *Let  $\mathbf{G}$  be the generator of an  $n+1$  state absorbing continuous-time Markov process in the form (2.11), and let  $\boldsymbol{\pi}$  be the probability vector for the distribution of the starting state. If the process is currently in state  $i$  at time  $t$ , and conditional on the fact that the process enters state  $n+1$  at exactly time  $Y = y$ ,*

1. *then the probability that the next move is to absorption after exactly  $y - t$  longer is:*

$$\mathbb{P}(\phi^{[t,y]} = i \cap \phi^{\{y\}} = n+1 \mid \mathbf{G}, Y = y, \phi^{\{t\}} = i) = \frac{\exp\{S_{ii}(y-t)\} s_i}{\mathbf{e}_i^T \exp\{\mathbf{S}(y-t)\} \mathbf{s}} \quad (4.4)$$

2. *then the density of the sojourn until the next move, given that the move is not absorbing, is:*

$$\begin{aligned} f_{\Delta|Y}(\delta = d \mid \mathbf{G}, Y = y, \phi^{[t,t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i) \\ = \frac{\mathbf{p}_i^T \exp\{\mathbf{S}(y-t-d)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}d)}{\int_0^{y-t} \mathbf{p}_i^T \exp\{\mathbf{S}(y-t-\delta)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}\delta) d\delta} \end{aligned} \quad (4.5)$$

*Proof:*

$$\begin{aligned} \mathbb{P}(\phi^{[t,y]} = i \cap \phi^{\{y\}} = n+1 \mid \mathbf{G}, Y = y, \phi^{\{t\}} = i) \\ = \frac{\mathbb{P}(Y = y \mid \mathbf{G}, \phi^{[t,y]} = i, \phi^{\{y\}} = n+1) \mathbb{P}(\phi^{[t,y]} = i \cap \phi^{\{y\}} = n+1 \mid \mathbf{G}, \phi^{\{t\}} = i)}{\mathbb{P}(Y = y \mid \mathbf{G}, \phi^{\{t\}} = i)} \\ = \frac{\mathbb{P}(\phi^{[t,y]} = i \mid \mathbf{G}, \phi^{\{t\}} = i) \mathbb{P}(\phi^{\{y\}} = n+1 \mid \mathbf{G}, \phi^{\{t\}} = i, \phi^{[t,y]} = i)}{\mathbb{P}(Y = y \mid \mathbf{G}, \phi^{\{t\}} = i)} \\ = \frac{\exp\{S_{ii}(y-t)\} s_i}{\mathbf{e}_i^T \exp\{\mathbf{S}(y-t)\} \mathbf{s}} \end{aligned}$$

as required.

The sojourn time is now free to be calculated without the concern of discrete probability mass at  $y - t$ . Thus, we condition on the known absorption time and given that

the next move will not be an absorbing one:

$$\begin{aligned}
f_{\Delta|Y}(\delta = d | \mathbf{G}, Y = y, \phi^{[t, t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i) \\
&= f_{Y|\Delta}(Y = y | \mathbf{G}, \phi^{[t, t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i, \delta = d) \\
&\quad \times f_{\Delta}(\delta = d | \mathbf{G}, \phi^{[t, t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i) \\
&\quad \div f_Y(Y = y | \mathbf{G}, \phi^{[t, t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i) \\
&= \frac{\mathbf{p}_i^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}d)}{\int_0^{y-t} \mathbf{p}_i^T \exp\{\mathbf{S}(y - t - \delta)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}\delta) d\delta} \quad \text{for } d \in [0, y - t)
\end{aligned}$$

as required, where  $\mathbf{p}_i = (p_{ij})$  is the vector of *unconditional* jump probabilities  $i \rightarrow j$ , excluding absorption.  $\square$

Equations (4.2) and (4.3) are straight-forward to sample from, being discrete distributions with only  $n$  and  $n - 1$  possible outcomes respectively. Evaluation of (4.4) is also straight-forward and a decision simply involves comparison to a Uniform(0,1) random number. However, (4.5) is a continuous distribution on  $[0, y - t)$  which is rather more complex. The method proposed here is a form of inversion by numerical solution (Devroye, 1986). The statement and proof of the following Lemma describes how and is a contribution of this thesis.

**Lemma 4.5 (Sampling of density (4.5))** *Random samples can be generated from the density in equation (4.5) by drawing  $z$  where  $Z \sim \text{Uniform}(0, 1)$  and numerically solving for  $d \in [0, y - t)$  in:*

$$z = \frac{\mathbf{p}_i^T \exp\{\mathbf{S}(y - t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s}}{\mathbf{p}_i^T \exp\{\mathbf{S}(y - t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-(y - t)(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s}} \quad (4.6)$$

Or, if  $\mathbf{S}$  admits an acceptable eigendecomposition,  $\mathbf{S} = \mathbf{Q}\mathbf{L}\mathbf{Q}^{-1}$ , by which it is meant where  $\|\mathbf{Q}\| \|\mathbf{Q}^{-1}\|$  is not large (see Moler and Van Loan, 2003), then random samples can be generated from the density in equation (4.5) by drawing  $z$  where  $Z \sim \text{Uniform}(0, 1)$  and numerically solving for  $d \in [0, y - t)$  in:

$$\mathbf{p}_i^T \mathbf{Q} (\mathbf{L} - \mathbf{I}S_{ii})^{-1} (\mathbf{V}_z - \mathbf{U}_d) \mathbf{Q}^{-1} \mathbf{s} = 0 \quad (4.7)$$

where

$$\begin{aligned}
\mathbf{V}_z &= (z - 1) \exp\{\mathbf{L}(y - t)\} - z \exp\{(y - t)S_{ii}\} \\
\mathbf{U}_d &= \exp\{\mathbf{L}(y - t) - d(\mathbf{L} - \mathbf{I}S_{ii})\}
\end{aligned}$$

*Proof:* Integrating the density (4.5) analytically to give the conditional jump distribution is possible, since:

$$\begin{aligned}
& \int_0^d \mathbf{p}_j^T \exp\{\mathbf{S}(y-t-x)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}x) dx \\
&= (-S_{ii}) \mathbf{p}_j^T \exp\{\mathbf{S}(y-t)\} \left[ -(\mathbf{S} - \mathbf{I}S_{ii})^{-1} \exp\{-x(\mathbf{S} - \mathbf{I}S_{ii})\} \right]_0^d \mathbf{s} \\
&= (-S_{ii}) \mathbf{p}_j^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s}
\end{aligned}$$

Thus, the cumulative distribution function for the conditional jump time is:

$$F_{\Delta|Y}(d) = \frac{\mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s}}{\mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-(y-t)(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s}}$$

For inverse sampling from the distribution (Devroye, 1986), this means generating a random  $Z \sim \text{Uniform}(0, 1)$  and solving for  $d \in [0, y-t]$  in (4.6), as required.

Note that  $\mathbf{p}_i \mathbf{p}_i^T$  is singular, so this cannot be directly solved. Instead, it is possible to exploit the fact that numerator and denominator are effectively scalar products in order to slightly simplify the problem for a numerical root finder:

$$\begin{aligned}
\implies 0 &= \mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I}z - z \exp\{-(y-t)(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s} \\
&\quad - \mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1} [\mathbf{I} - \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s} \\
&= \langle [\mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1}]^T, [\mathbf{I}z - z \exp\{-(y-t)(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s} \rangle \\
&\quad - \langle [\mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1}]^T, [\mathbf{I} - \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s} \rangle \\
&= \langle [\mathbf{p}_i^T \exp\{\mathbf{S}(y-t)\} (\mathbf{S} - \mathbf{I}S_{ii})^{-1}]^T, \\
&\quad [\mathbf{I}(z-1) - z \exp\{-(y-t)(\mathbf{S} - \mathbf{I}S_{ii})\} + \exp\{-d(\mathbf{S} - \mathbf{I}S_{ii})\}] \mathbf{s} \rangle
\end{aligned}$$

where  $\langle \cdot, \cdot \rangle$  is the scalar product and the matrices operating on  $\mathbf{s}$  are viewed as linear operators. This achieves little immediately, but if  $\mathbf{S}$  admits an acceptable eigendecomposition,  $\mathbf{S} = \mathbf{Q}\mathbf{L}\mathbf{Q}^{-1}$ , then this simplifies to:

$$\mathbf{p}_i^T \mathbf{Q} (\mathbf{L} - \mathbf{I}S_{ii})^{-1} (\mathbf{V}_z - \mathbf{U}_d) \mathbf{Q}^{-1} \mathbf{s} = 0$$

where

$$\begin{aligned}
\mathbf{V}_z &= (z-1) \exp\{\mathbf{L}(y-t)\} - z \exp\{(y-t)S_{ii}\} \\
\mathbf{U}_d &= \exp\{\mathbf{L}(y-t) - d(\mathbf{L} - \mathbf{I}S_{ii})\}
\end{aligned}$$

due to the commutativity of the diagonal  $\mathbf{L}$ . □

Accordingly, this Theorem leads to a choice: in the event that  $\mathbf{S}$  admits an eigendecomposition which computes the matrix exponential stably, one can proceed using (4.7). However, if it does not, then at the cost of some additional computation any matrix exponential algorithm can be used together with equation (4.6), such as the fairly recent and (“uniformly”) superior method due to Higham (2008).

Note that the form of (4.7) is not contrived: it provides a form which enables considerable computational efficiency. In particular,

- The eigendecomposition  $\mathbf{S} = \mathbf{Q}\mathbf{L}\mathbf{Q}^{-1}$ , and the  $1 \times n$  and  $n \times 1$  terms  $\mathbf{p}_i^T \mathbf{Q}(\mathbf{L} - \mathbf{L}S_{ii})^{-1}$  and  $\mathbf{Q}^{-1}\mathbf{s}$  need only be computed once per outer Gibbs iteration. Thereafter, every step within every latent path simulated for every observation in the dataset uses the same decomposition and the two terms.
- $\mathbf{V}_z$  is a diagonal  $n$  dimensional matrix, only needing to be computed once for each sample required from  $F_{\Delta|Y}(d)$ .
- $\mathbf{U}_d$  is a diagonal  $n$  dimensional matrix, and is the only term which needs to be computed on every step of the root finder.

Thus, coded for maximum efficiency, the root finder only involves evaluation of  $n$  exponentials,  $3n$  multiplications and  $3n$  additions for each iteration. Some care is needed if (4.7) approaches machine precision over the whole interval  $d \in [0, y - t)$ . In those edge cases, sacrificing some speed on that sojourn sample and using Adaptive Rejection Metropolis Sampling (Gilks *et al.*, 1995) with the unnormalised density (4.5) has always proved adequate in testing.

Bringing all these Lemmas together leads to the Exact Conditional Sampling algorithm which may now be stated in full.

**Algorithm 4.1 (Exact Conditional Sampling)** *Produce a sample process directly from  $f_{\Phi|\Psi}(\phi | \boldsymbol{\pi}, \mathbf{G}, Y = y)$  as follows:*

- i) Sample a starting state,  $i$ , from the probability mass function (Lemma 4.3, (4.2)):*

$$\mathbb{P}(\phi^{\{0\}} = i | \boldsymbol{\pi}, \mathbf{G}, Y = y) = \frac{\mathbf{e}_i^T \exp\{\mathbf{S}y\}\mathbf{s} \pi_i}{\boldsymbol{\pi}^T \exp\{\mathbf{S}y\}\mathbf{s}}$$

*and set  $t = 0$*

ii) With probability (Lemma 4.4, (4.4)):

$$\mathbb{P}(\phi^{[t,y]} = i \cap \phi^{\{y\}} = n + 1 \mid \mathbf{G}, Y = y, \phi^{\{t\}} = i) = \frac{\exp\{S_{ii}(y - t)\} s_i}{\mathbf{e}_i^T \exp\{\mathbf{S}(y - t)\} \mathbf{s}}$$

set  $\phi^{[t,y]} = i$  and  $\phi^{[y,\infty)} = n + 1$  and end the algorithm; else continue to (iii)

iii) Sample the sojourn time,  $d$ , in the current state,  $i$ , before a non-absorbing move from (Lemma 4.4, (4.5)):

$$\begin{aligned} f_{\Delta|Y}(\delta = d \mid \mathbf{G}, Y = y, \phi^{[t,t+\delta]} = i, \phi^{\{t+\delta\}} \in \{1, \dots, n\} \setminus i) \\ = \frac{\mathbf{p}_i^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}d)}{\int_0^{y-t} \mathbf{p}_i^T \exp\{\mathbf{S}(y - t - \delta)\} \mathbf{s} (-S_{ii}) \exp(S_{ii}\delta) d\delta} \end{aligned} \quad (4.8)$$

using the methods of Lemma 4.5 and set  $\phi^{[t,t+d]} = i$

iv) Sample a state move,  $i \rightarrow j$ , from (Lemma 4.3, (4.3)):

$$\begin{aligned} \mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, Y = y, \phi^{[t,t+d]} = i, \phi^{\{t+d\}} \in \{1, \dots, n\} \setminus i) \\ \propto S_{ij} \mathbf{e}_j^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{s} \end{aligned}$$

and set  $\phi^{\{t+d\}} = j$

v) Update  $t = t + d$  and  $i = j$ , then loop to (ii) □

This algorithm is intended to completely replace the Metropolis-Hastings part of Bladt *et al.* (2003). A full comparison will follow toward the end of this chapter.

### 4.2.3 Incorporating censored data

Ignoring the advances of §4.2.2 momentarily, right-censored observations can be dealt with quite elegantly within the original algorithm. The data will now consist of some actual and some censored observations,  $\mathbf{y} = \{y_1, \dots, y_m, y_{m+1}^c, \dots, y_n^c\}$ , where a superscript ‘c’ denotes right-censoring. Now when the Metropolis-Hastings algorithm is invoked to sample the latent process from  $f_{\Phi|\Psi}(\phi_i \mid \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$ , it should be used as normal for  $\{y_1, \dots, y_m\}$ , but clearly for  $\{y_{m+1}^c, \dots, y_n^c\}$  sampling should be from  $f_{\Phi|\Psi}(\phi_i \mid \boldsymbol{\pi}, \mathbf{G}, Y_i > y_i)$  — that is, the first rejection sampled proposal process can be immediately returned. It is important to note though that the simulation must proceed through to absorption for these observations, where previously it was possible to

stop short once the observation time was reached since truncation would occur. This simple modification allows incorporation of the censored observations in the likelihood through the latent process simulation.

Consequently, censored observations will not suffer computational problems II or III described in §4.1.3, though they will still suffer from problem I. Thus, a similar approach to §4.2.2 is employed (making use of the early work described in that section).

**Lemma 4.6 (ECS for censored absorption times)** *Let  $\mathbf{G}$  be the generator of an  $n + 1$  state absorbing continuous-time Markov process in the form (2.11), and let  $\boldsymbol{\pi}$  be the probability vector for the distribution of the starting state. Then, conditional on the fact that the process enters state  $n + 1$  at some time  $Y > y$ ,*

1. *the distribution of the starting state is:*

$$\mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y > y) = \frac{\mathbf{e}_i^T \exp\{\mathbf{S}y\} \mathbf{e} \pi_i}{\boldsymbol{\pi}^T \exp\{\mathbf{S}y\} \mathbf{e}} \quad (4.9)$$

2. *if the process has reached time  $t < y$  then the density of the sojourn until the next move, is:*

$$f_{\Delta|Y}(\delta = d \mid \mathbf{G}, Y > y, \phi^{\{t\}} = i) \propto \begin{cases} \mathbf{p}_i^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} (-S_{ii}) \exp(S_{ii}d) & \text{for } d \in [0, y - t) \\ -S_{ii} \exp(S_{ii}d) & \text{for } d > y - t \end{cases} \quad (4.10)$$

3. *if the process has reached time  $t < y$  and it has already been determined that a jump occur after a further time  $d$ , then the embedded discrete chain state move probabilities become:*

$$\mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, Y > y, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \neq i) \propto \begin{cases} S_{ij} \mathbf{e}_j^T \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} & \text{if } d < y - t \\ S_{ij} & \text{if } d \geq y - t \end{cases} \quad (4.11)$$

4. *if the process has reached time  $t \geq y$  then sojourn and state move probabilities are as in Algorithm 2.1, page 27.*

*Proof:*

$$\begin{aligned}\mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{G}, Y > y) &= \frac{f_Y(Y > y \mid \boldsymbol{\pi}, \mathbf{G}, \phi^{\{0\}} = i) \mathbb{P}(\phi^{\{0\}} = i \mid \boldsymbol{\pi}, \mathbf{S})}{f_Y(Y > y \mid \boldsymbol{\pi}, \mathbf{S})} \\ &= \frac{\mathbf{e}_i^{\top} \exp\{\mathbf{S}y\} \mathbf{e} \pi_i}{\boldsymbol{\pi}^{\top} \exp\{\mathbf{S}y\} \mathbf{e}}\end{aligned}$$

as required, where  $\mathbf{e}_i$  is a vector with one in the  $i^{\text{th}}$  position and zero elsewhere; and  $\mathbf{e}$  is a vector of all ones.

$$\begin{aligned}f_{\Delta \mid Y}(\delta = d \mid \mathbf{G}, Y > y, \phi^{\{t\}} = i) \\ \propto f_Y(Y > y \mid \mathbf{G}, \phi^{\{t\}} = i, \delta = d) f_{\Delta}(\delta = d \mid \mathbf{G}, \phi^{\{t\}} = i) \\ = \begin{cases} \mathbf{p}_i^{\top} \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} (-S_{ii}) \exp(S_{ii}d) & \text{for } d \in [0, y - t) \\ -S_{ii} \exp(S_{ii}d) & \text{for } d > y - t \end{cases}\end{aligned}$$

as required, where  $\mathbf{p}_i = (p_{ij})$  is the  $n$ -dimensional vector of unconditional jump probabilities  $i \rightarrow j$ . Note that this is an Exponential density beyond  $y - t$ .

$$\begin{aligned}\mathbb{P}(\phi^{\{t+d\}} = j \mid \boldsymbol{\pi}, \mathbf{G}, Y > y, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \neq i) \\ \propto \mathbb{P}(Y > y \mid \mathbf{G}, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} = j \neq i) \\ \quad \times \mathbb{P}(\phi^{\{t+d\}} = j \mid \mathbf{G}, \phi^{\{t,t+d\}} = i, \phi^{\{t+d\}} \neq i) \\ = \begin{cases} S_{ij} \mathbf{e}_j^{\top} \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} & \text{if } d < y - t \\ S_{ij} & \text{if } d \geq y - t \end{cases}\end{aligned}$$

as required.

The final point to note is that once the process has passed time  $y$ , then conditioning on an absorption time  $Y > y$  has no effect — when using Bayes' Theorem to change the conditioning the first term will always equal 1. The final part of the Lemma then follows.  $\square$

At first glance, using (4.10) in Lemma 4.6 could seem problematic: one must know the size of the sampled sojourn before the density can be established. However, the solution is elementary:

**Corollary 4.7** *In order to sample from the density in (4.10) when  $t < y$ :*

With probability  $e^{S_{ii}(y-t)}/\mathbf{e}_i \exp\{\mathbf{S}(y-t)\}\mathbf{e}$  set  $d = y - t + T$  where  $T \sim \text{Exp}(-S_{ii})$ ; else sample  $d$  from the density with support  $[0, y - t)$ :

$$f_{\Delta|Y}(\delta = d | \mathbf{G}, Y > y, \phi^{\{t\}} = i) \propto \mathbf{p}_i^T \exp\{\mathbf{S}(y-t-d)\}\mathbf{e} (-S_{ii}) \exp(S_{ii}d)$$

*Proof:* The first step is to determine whether  $d \in [0, y - t)$  or  $d \in [y - t, \infty)$ . This can be determined as:

$$\begin{aligned} \mathbb{P}(d \in [y - t, \infty)) &= f_{\Delta|Y}(\delta > y - t | \mathbf{G}, Y > y, \phi^{\{t\}} = i) \\ &= \frac{f_{Y|\Delta}(Y > y | \mathbf{G}, \phi^{\{t\}} = i, \delta > y - t) f_{\Delta}(\delta > y - t | \mathbf{G}, \phi^{\{t\}} = i)}{f_Y(Y > y | \mathbf{G}, \phi^{\{t\}} = i)} \\ &= \frac{1 e^{S_{ii}(y-t)}}{\mathbf{e}_i \exp\{\mathbf{S}(y-t)\}\mathbf{e}} \end{aligned}$$

Consequently, with this probability the sample  $d$  exceeds  $y - t$ , where the density is proportional to an Exponential distribution. Thus,  $\delta | \delta > y - t$  is Exponential and so  $d$  can be sampled as  $d = y - t + T$  where  $T \sim \text{Exp}(-S_{ii})$ .

Otherwise,  $d$  is sampled from the part of  $f_{\Delta|Y}$  where  $d \in [0, y - t)$ , which is just a renormalised version of the given density.  $\square$

Actually sampling from  $f_{\Delta|Y}$  on the interval  $[0, y - t)$  is then the objective. Experience has shown that, unlike (4.5), this is often approximately log-linear. This fact, together with a genuinely finite support, means it is particularly well suited to application of Adaptive Rejection Metropolis Sampling (Gilks *et al.*, 1995).

**Algorithm 4.2 (ECS for Censored Observations)** *Produce a sample process directly from  $f_{\Phi|\Psi}(\phi | \boldsymbol{\pi}, \mathbf{G}, Y > y)$  as follows, using the results of Lemma 4.6:*

1. *Sample a starting state,  $i$ , from the probability mass function (4.9):*

$$\mathbb{P}(\phi^{\{0\}} = i | \boldsymbol{\pi}, \mathbf{G}, Y > y) = \frac{\mathbf{e}_i^T \exp\{\mathbf{S}y\}\mathbf{e} \pi_i}{\boldsymbol{\pi}^T \exp\{\mathbf{S}y\}\mathbf{e}}$$

*and set  $t = 0$*

2. *If  $t > y$ , then switch to Algorithm 2.1 (page 27) and pick up from step 2 there, otherwise continue.*
3. *Sample the sojourn time in the current state,  $\delta$ . With probability*

$$e^{S_{ii}(y-t)}/\mathbf{e}_i \exp\{\mathbf{S}(y-t)\}\mathbf{e}$$



the sojourn time is  $d = y - t + T$  where  $T \sim \text{Exp}(-S_{ii})$ . Else, the sojourn time is a sample from the following finitely supported density on  $[0, y - t)$

$$f_{\Delta|Y}(\delta = d | \mathbf{G}, Y > y, \phi^{\{t\}} = i) \propto \mathbf{p}_i^{\text{T}} \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} (-S_{ii}) \exp(S_{ii}d)$$

Set  $\phi^{[t, t+d)} = i$

4. Sample a state move,  $i \rightarrow j$  ( $i \neq j$ ), from

$$\begin{aligned} & \mathbb{P}(\phi^{\{t+d\}} = j | \boldsymbol{\pi}, \mathbf{G}, Y > y, \phi^{[t, t+d)} = i, \phi^{\{t+d\}} \neq i) \\ & \propto \begin{cases} S_{ij} \mathbf{e}_j^{\text{T}} \exp\{\mathbf{S}(y - t - d)\} \mathbf{e} & \text{if } d < y - t \\ S_{ij} & \text{if } d \geq y - t \end{cases} \end{aligned}$$

and set  $\phi^{\{t+d\}} = j$

5. If  $j = n + 1$ , exit loop. Else, update  $t = t + d$  and  $i = j$ , then loop to (ii)  $\square$

This algorithm is intended to complement Algorithm 4.1, being used for those observations which constitute censored observation times.

## 4.3 Examining the extended methodology

Everything in this section was computed using Aslett (2011), which is a major contribution of this thesis, discussed further in Chapter 7.

The novel extensions proposed in the preceding section are now examined in three ways. First, the benefits of the model reformulation in §4.2.1 are explored, when the objective of inference is learning about interpretable parameters. Second, is detailed examination of the computational improvements, from §4.2.2, through simulation of the latent stochastic process in isolation. This shows good performance both in the presence and absence of the problems which motivated its development. Finally, the first example is rerun with some observations randomly censored to show inference remains tenable.

### 4.3.1 Effect of model reformulation

Exactly the same simulated data as in the example from §4.1.1 are used to enable direct comparison. In terms of the new model formulation, the following structure is

Ground Truth	Expectation	Lower 95% quantile	Upper 95% quantile
$\lambda_f = 1.8$	1.79 ( 0.00344 )	1.44 ( 0.00543 )	2.17 ( 0.00676 )
$\lambda_r = 9.5$	11.1 ( 0.0132 )	9.56 ( 0.0253 )	12.8 ( 0.0256 )

**Table 4.3:** Posterior summary statistics for simulations from known ground truth using extended methodology. Bracketed figures are MCMC standard errors.

imposed:

$$\Lambda_f = \{(1, 2), (1, 3), (2, 0), (3, 0)\}$$

$$\Lambda_r = \{(2, 1), (3, 1)\}$$

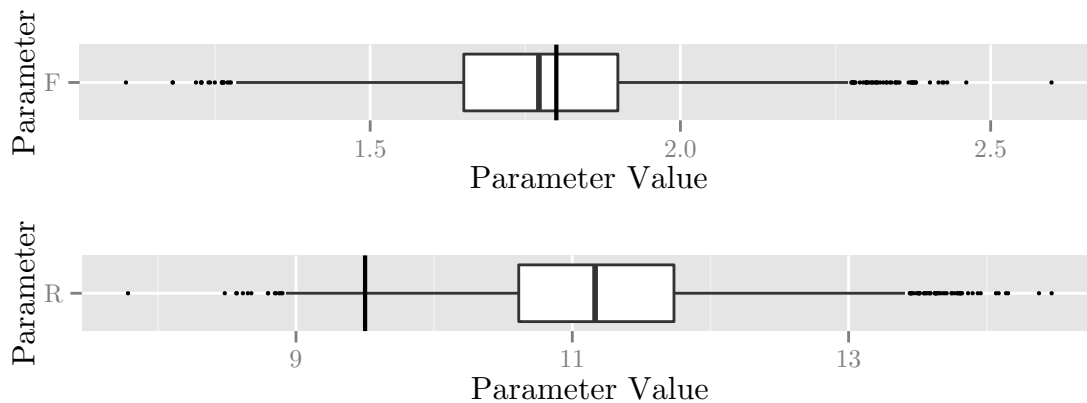
with zero constraints on  $\{(2, 3), (3, 2), (1, 0)\}$ . In this case,  $c_{ij} = 1 \forall (i, j) \in \Lambda_f \cup \Lambda_r$ . Prior specification in the extended methodology is taken as:

$$\nu_{\lambda_f} = 24 \quad \text{and} \quad \zeta_{\lambda_f} = 16$$

$$\nu_{\lambda_r} = 180 \quad \text{and} \quad \zeta_{\lambda_r} = 16$$

which ensures that the prior matches the original example, even though the extended methodology allows for a more flexible prior specification. Again, 10,000 iterations were run and Table 4.3 shows the resultant summary estimates, with MCMC standard errors calculated as discussed at the end of §3.2.4. Posterior sample box plots are shown in Figure 4.3.

The reformulation of the model has, as would be expected, improved estimates of the underlying two parameters and there was no need to estimate the prohibited transition rates. The estimate of  $\lambda_f$  has notably improved accuracy, but the estimate of  $\lambda_r$  is only just still outside the 95% interval and the posterior is effectively a slightly location shifted version of the prior. However, apropos of repair rates, Daneshkhah and Bedford (2008) show system lifetime is considerably less sensitive to rate of repair than rate of failure, which would affect how well such a parameter could be inferred without a substantial number of observations. As a result, with the known correct model here, it may be taken that this is as much as one could hope to learn about the repair rate.



**Figure 4.3:** Boxplot for posterior simulations from known ground truth using extended methodology. Solid vertical lines show ground truth values.

### 4.3.2 Computational performance

Consider simulating a continuous-time Markov process which absorbs at a given time, as required for the data augmented Phase-type MCMC algorithm. The comparison of interest is between the Metropolis-Hastings with rejection sampling of Bladt *et al.* (2003) and the Exact Conditional Sampling of §4.2.2. This is done in two example situations: one in the presence of, and the other in the absence of, the three major problems elucidated in §4.1.3. This is to demonstrate the new approach is acceptably fast in ‘easy’ situations and to highlight the potential gains where the computational problems become acute.

Finally, a comparison of the performance of simulating the latent process at varying depths into the tail of the absorption time distribution is given.

The timings in this section are based on the highly optimised C code publicly available in the `PhaseType` package (Aslett, 2011), running on a MacBook Pro (mid-2007, ‘MacBookPro3,1’) with an Intel Core 2 Duo 2.4GHz CPU.

### 4.3.2.1 Example with no major problems

Take a continuous-time Markov chain which always starts in state 1, whose generator is:

$$\mathbf{G} = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Simulation of this chain, conditional on absorbing at time  $y = 0.7$ , does not exhibit problems I, II or III of §4.1.3. In particular:

- I. The probability of reaching  $y$  for a rejection sample is approximately 0.5.
- II. There are no zero constraints on absorbing moves.
- III. The total variation distance is zero from the first Metropolis-Hastings iteration, so it starts in stationarity.

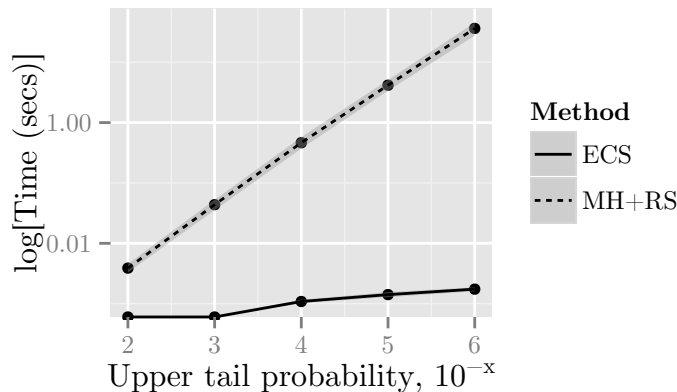
Under such ‘ideal’ conditions (for the original algorithm), the mean processing time for the Metropolis-Hastings algorithm is on the order of 1.6 micro-seconds (standard deviation 104 micro-seconds). The novel Exact Conditional Sampling algorithm has mean processing time of 7.2 micro-seconds (standard deviation 19 micro-seconds). As expected, on average the original algorithm is faster under these conditions, but the new approach could not be characterised as ‘slow’ and has considerably less variable run-time.

### 4.3.2.2 Example exhibiting acute problems

Conversely, consider the continuous-time Markov process of (4.1) which, as already discussed, exhibits all of the problems I–III. Recall that simulation of this process conditional on absorbing at time  $y = 600$  implies that the total number of samples,  $K$ , required in order to produce a single sample from  $f_{\Phi|\Psi}(\phi_i | \boldsymbol{\pi}, \mathbf{G}, Y_i = y_i)$  has:

$$\mathbb{E}[K] = 897,867,259 \quad \text{and} \quad 95\% \text{ CI} = [799,129,604, 1,002,273,104]$$

Further, recall that drawing such samples must be repeated for each observation  $y_i$  in the data set on every Gibbs iteration. For this example, sampling a *single* path



**Figure 4.4:** Speed of original (MH+RS) and new (ECS) sampling methodologies at varying right tail probabilities, with bootstrap resampled confidence intervals from 50 simulations at each  $x$ .

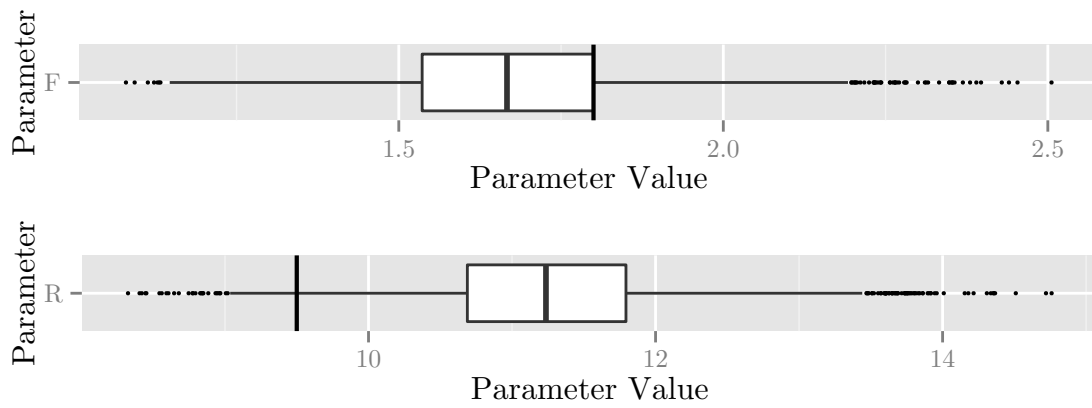
conditional on an exact absorption time took 18.5 hours for the Metropolis-Hastings algorithm, and a mere 0.0087 seconds for the novel Exact Conditional Sampling methodology. This demonstrates that the new algorithm is substantially faster in instances where the problems described earlier occur. In particular, the new algorithm would not cause the MCMC procedure to stall in such scenarios.

#### 4.3.2.3 Tail depth speed improvements

Finally, a direct comparison of the impact of right-tail observations on the simulation time is considered. As described in problem I, the exploration of the parameter space may lead to the right-tail Phase-type probability beyond some observations being very small and this is something which still conceivably affects ECS: since the process is being simulated, larger observations imply more process state changes and longer simulations. Accordingly, it is relevant to examine this problem in isolation.

The impact of decreasing right-tail probability can be illustrated by taking fixed parameters and choosing a  $y$  such that  $\mathbb{P}(Y > y | \boldsymbol{\pi}, \mathbf{G}) = 10^{-x}$  for increasing values  $x$ . Using the two component parallel system example in (2.10), page 29, with  $\lambda_f = 1/478$  and  $\lambda_r = 1/39$ , the time to simulate a chain when conditioning on such  $y$  under old and new methodology is shown in Figure 4.4.

Note in particular that logarithmic time is plotted: consequently, the rate of slow-down in the original methodology is sharp and the confidence bounds fan out extensively. In contrast, ECS has modest increases over the full spectrum of tail probabilities



**Figure 4.5:** Boxplot for posterior simulations from known ground truth using extended methodology for 25% censoring on data. Solid vertical lines show ground truth values. shown.

#### 4.3.2.4 Effect of censoring

By way of comparison, the starting point was exactly the same model and data as used in both §4.1.1 and §4.3.1. The 3rd quartile time was then used as a censoring time to emulate an experiment in which the longest surviving 25% of units were censored due to early termination of the study. The resulting parameter estimates are shown in Figure 4.5 from an MCMC run of 10,000 iterations which again reduced the standard error enough to provide 2–3 significant figures in the 95% quantiles. While the posterior mean is further from the ground truth, the estimate is not too poor given that there are now only 18 exact observations.

The results were not substantively different when performing random censoring, whereby 25% of the data was chosen at random and each replaced with a uniform random draw between zero and the observation time.

### 4.3.3 Realistic 5 component bridge system

As mentioned earlier, the focus thus far has been on a relatively simple example — the dual repairable redundant system — to aid clarity of exposition. However, it is important to note that the technique is not restricted to such elementary models. Consequently, the final example of this chapter is to perform inference in the case of the repairable 5 component bridge system (Figure 2.3) with masked lifetime data.

This also provides an example where some of the parameters are subject to a constant multiple in certain elements of the generator.

This leads to a more sizable generator matrix. Row  $i$  corresponds to the system state column in Table 2.1, and the diagonal is suppressed for readability (it simply being the negative of the row sums):

$$\mathbf{G} = \begin{pmatrix} \cdot & \lambda_f & \lambda_f & 0 & \lambda_f & 0 & 0 & 0 & \lambda_f & 0 & 0 & \lambda_f & 0 & 0 & 0 & 0 & 0 \\ \lambda_r & \cdot & 0 & \lambda_f & 0 & \lambda_f & 0 & 0 & 0 & \lambda_f & 0 & 0 & \lambda_f & 0 & 0 & 0 & 0 \\ \lambda_r & 0 & \cdot & \lambda_f & 0 & 0 & \lambda_f & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_f & 0 & 0 & \lambda_f \\ 0 & \lambda_r & \lambda_r & \cdot & 0 & 0 & 0 & \lambda_f & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\lambda_f \\ \lambda_r & 0 & 0 & 0 & \cdot & \lambda_f & \lambda_f & 0 & 0 & 0 & \lambda_f & 0 & 0 & 0 & 0 & 0 & \lambda_f \\ 0 & \lambda_r & 0 & 0 & \lambda_r & \cdot & 0 & \lambda_f & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\lambda_f \\ 0 & 0 & \lambda_r & 0 & \lambda_r & 0 & \cdot & \lambda_f & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\lambda_f \\ 0 & 0 & 0 & \lambda_r & 0 & \lambda_r & \lambda_r & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\lambda_f \\ \lambda_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \lambda_f & \lambda_f & 0 & 0 & 0 & \lambda_f & 0 & \lambda_f \\ 0 & \lambda_r & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_f & 2\lambda_f \\ 0 & 0 & 0 & 0 & \lambda_r & 0 & 0 & 0 & \lambda_r & 0 & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 3\lambda_f \\ \lambda_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \lambda_f & \lambda_f & \lambda_f & 0 & \lambda_f \\ 0 & \lambda_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r & \cdot & 0 & 0 & \lambda_f & 2\lambda_f \\ 0 & 0 & \lambda_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r & 0 & \cdot & 0 & 0 & 3\lambda_f \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r & 0 & 0 & \lambda_r & 0 & 0 & \cdot & \lambda_f & 2\lambda_f \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r & 0 & 0 & \lambda_r & 0 & \lambda_r & \cdot & 2\lambda_f \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot \end{pmatrix} \quad (4.12)$$

The need to construct such generator matrices may initially seem to be an impediment to adoption of this approach by reliability practitioners, but the reader is referred to Chapter 7 where statistical software contributions of this thesis actually make the task trivial (§7.2.2, page 150).

The generator in (4.12) was used with  $\lambda_f = 1 \text{ yr}^{-1}$  and  $\lambda_r = 365 \text{ yr}^{-1}$  to simulate a data set of 25 observations. These values also represent more realistic parameter values, with an expected component lifetime of 1 year and expected repair time of 1 day.  $\Lambda_f$ ,  $\Lambda_r$  and  $c_{ij}$  were constructed in accordance with the generator and used to

<b>Ground Truth</b>	<b>Expectation</b>	<b>Lower 95% quantile</b>	<b>Upper 95% quantile</b>
$\lambda_f = 1$	0.929 ( 0.0089 )	0.602 ( 0.0175 )	1.29 ( 0.0218 )
$\lambda_r = 365$	390 ( 6.76 )	179 ( 9.27 )	667 ( 36.9 )

**Table 4.4:** Posterior summary statistics for simulations from known ground truth using extended methodology for 5 component bridge system. Bracketed figures are MCMC standard errors.

infer the parameters in conjunction with prior specification:

$$\begin{aligned} \nu_{\lambda_f} &= 2.2 \quad \text{and} \quad \zeta_{\lambda_f} = 1 \\ \nu_{\lambda_r} &= 6 \quad \text{and} \quad \zeta_{\lambda_r} = 0.015 \end{aligned}$$

This provides a 95% prior belief interval for the expected component lifetime of roughly 0.4 to 5.1 years, and for expected repair time of roughly 0.5 to 2 days.

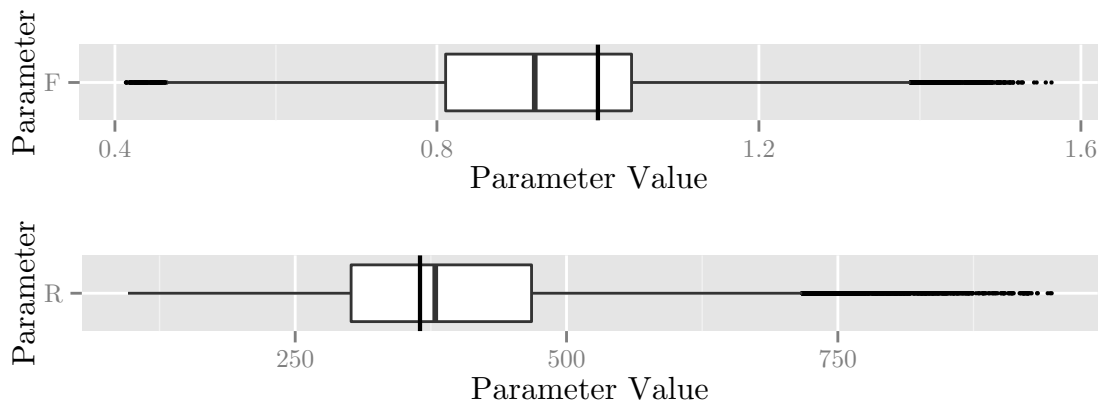
The resultant posterior summaries are shown in Table 4.4 and Figure 4.6. This example required 100,000 MCMC iterations to reduce the standard error due to much higher autocorrelation. The results are very encouraging and in line with the smaller example: the repair rate proves difficult to learn with similar spread to the prior, but the failure rate none the less has satisfactory posterior updating with the 95% credible interval incorporating the ground truth and narrower than the prior interval. This is encouraging given only 25 observations were simulated.

Even in this difficult example, the extended methodology performance was acceptable by MCMC standards on a 5 year old laptop, completing 2.75 iterations per second (for a total runtime of around 10 hours). If one uses the model reformulation, but retains the Metropolis-Hastings latent process sampling instead of the new Exact Conditional Sampling, then even performing only 10 Metropolis-Hastings iterations per latent process results in execution at a rate of around 0.044 iterations per second (leading to a total runtime of around 26 days<sup>2</sup>).

---

<sup>2</sup>Estimated run time based on less than 100,000 iterations.





**Figure 4.6:** Boxplot for posterior simulations from known ground truth using extended methodology for 5 component bridge system. Solid vertical lines show ground truth values.

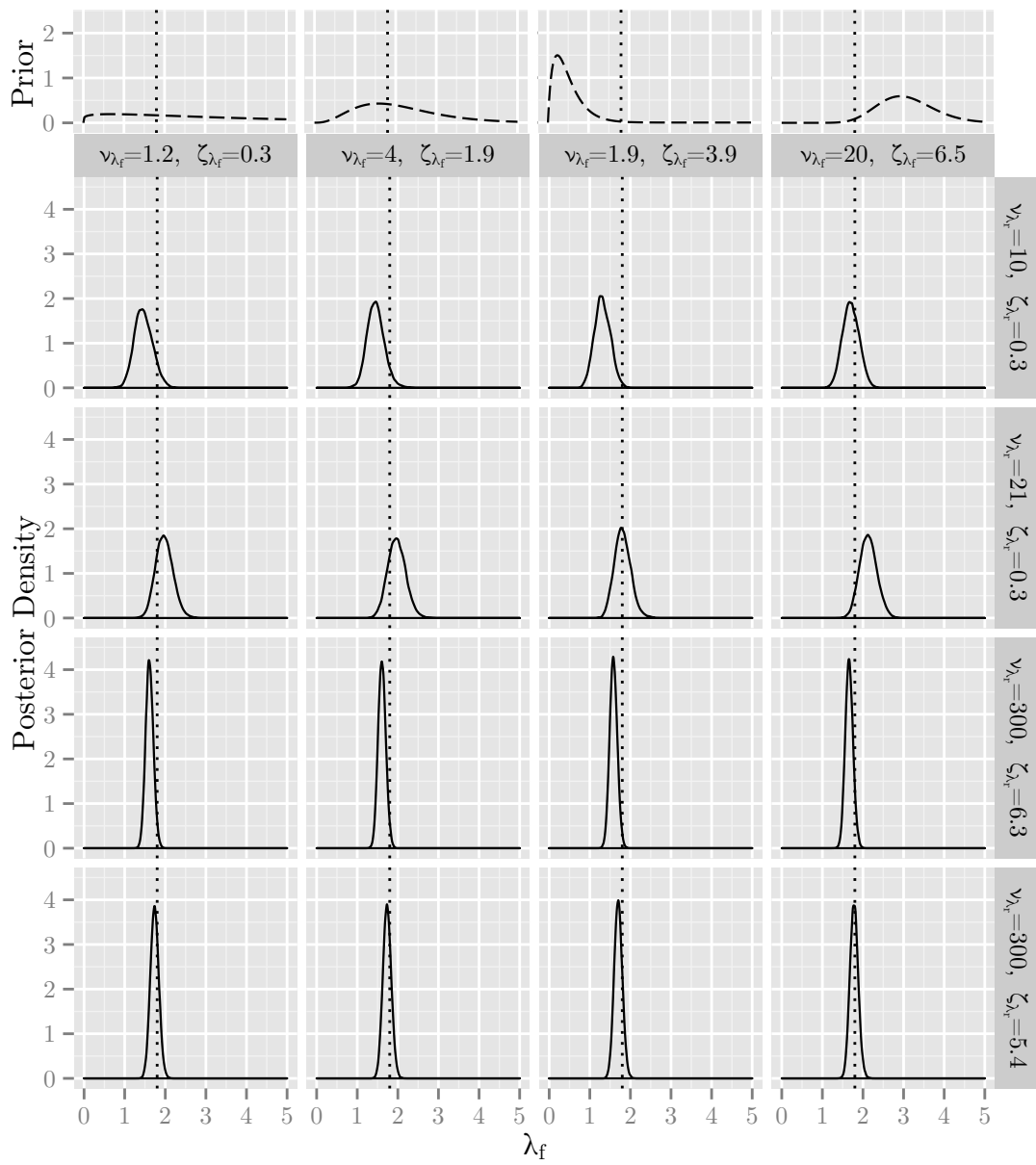
## 4.4 Sensitivity to the prior

The final aspect to consider in evaluating the extended methodology is some gauge of the sensitivity in posterior results to prior specification. There are at least two reasons this is desirable to know: to increase confidence that the method is not too sensitive to prior mis-specification; and in the event of limited resources to learn where prior elicitation efforts would most beneficially be directed.

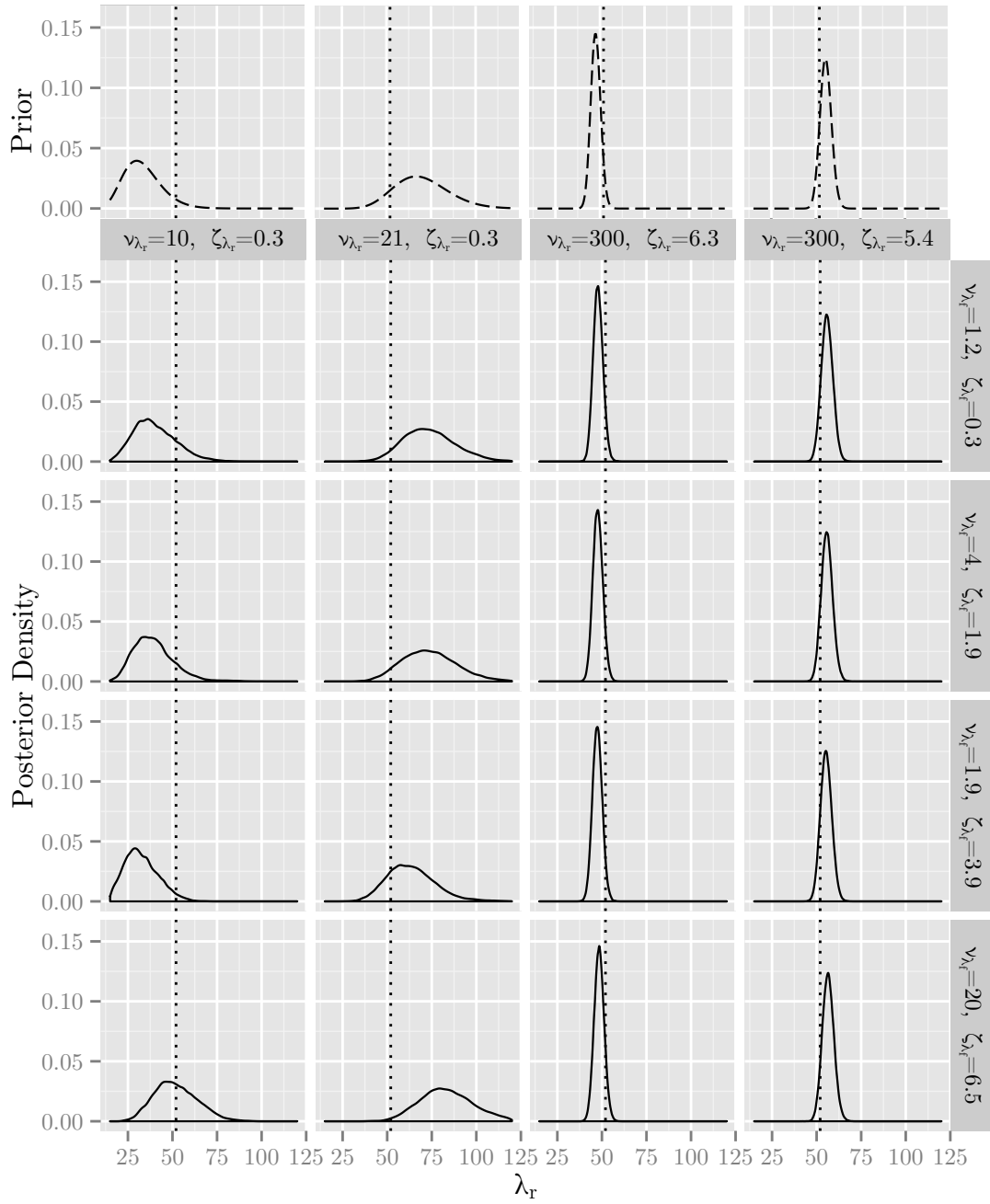
To informally explore the sensitivity, 100 failure times were simulated for a dual repairable redundant system with parameters  $\lambda_f = 1.8 \text{ yr}^{-1}$  and  $\lambda_r = 52 \text{ yr}^{-1}$ . Four priors each were chosen for failure and repair rate (providing 8 priors in total): two relatively vague and two which are stronger but incorrect (one overestimating and the other underestimating the rate). All possible pairs of these priors were combined for  $4 \times 4 = 16$  MCMC runs, the results being depicted in Figures 4.7 and 4.8.

The failure rate posteriors show reasonable robustness to prior mis-specification. The most interesting feature is that the vagueness or otherwise of the repair rate prior greatly influences the variance of the failure rate posterior, whilst the failure rate prior itself has negligible effect. As expected, the repair rate posteriors show little learning taking place and, where it seems to occur (row 3, column 2 and row 4, column 1 repair posteriors), it can easily be explained by the corresponding failure rate priors being informative in the complementary direction.

This is just one illustration for the prevalent two parallel component subsystem de-



**Figure 4.7:** Marginal failure rate posteriors for all combinations of 4 different failure and repair rate priors. Prior parameters shown in the dark grey boxes (failure rate across top, repair rate down the side), with failure rate prior densities depicted by dashed line at the top. Ground truth is dotted vertical line.



**Figure 4.8:** Marginal repair rate posteriors for all combinations of 4 different failure and repair rate priors. Prior parameters shown in the dark grey boxes (repair rate across top, failure rate down the side), with repair rate prior densities depicted by dashed line at the top. Ground truth is dotted vertical line.

sign with realistic parameter values. However, when using this methodology in practice it would be prudent to conduct prior sensitivity simulations for the particular system structure and type of parameters under consideration. In particular, note that for values of the failure and repair rate much closer together than is typical in highly reliable systems this will certainly be a greater concern. Additionally, the techniques of this chapter are not limited to modelling repairable systems and thus more general models with different numbers of parameters and different generator structure each require careful consideration in application.

## 4.5 Exchangeability

This chapter has been developed under the independent and identically distributed (i.i.d.) systems case, since this is the simpler setting and has little direct bearing on the latent process simulation which was the main preoccupation of this chapter.

However, it is possible to weaken this assumption to that of exchangeability, but the development is deferred to Chapter 6 (see §6.3.4, page 137) where it is naturally synthesised with the work of that chapter.

## 4.6 Generalisations

The work of Cano *et al.* (2010) and the work of this chapter represent the two extremes of available information on a failure and repair (or general) process. Between the two extremes there are many partial information scenarios, including for example:

1. Some part of the process may have been observed and the remainder be unobserved.
2. It may be known how many failures and repairs took place, but not the various times to failure or repair.
3. Only failure (or repair) times may be known through logging, but not the repair (or failure) times.
4. The final state of the system at failure may be known (i.e. which component failures had caused system failure and which were still working). This is pertinent

for non-parallel systems such as the bridge system.

Some of these partial information scenarios should, in principle, be possible to accommodate with only minor changes to the methodology. The first example simply requires simulation of part of the process, so that the starting state will not necessarily be full operation and the ending state not necessarily failure. Similarly, the fourth example can be accomplished by conditioning on the restricted set of possible pre-absorption states as alluded to in the final pilot idea on page 79: in other words, those states from which a move to the known failure state is possible.

Scenarios 2 and 3 may require more work and would be interesting avenues of future research.

## 4.7 Summary

A detailed study of the algorithm was presented, including exact conditions for convergence of the latent process simulation for arbitrarily large generators with two non-zero absorbing move rates. As a result of this study, it became evident that model reformulation, facilitation of censored observations and computational speedups were required.

Three major contributions of the thesis were then put forward: model reformulation to enable structured generator matrices; computational improvements to the latent process simulation; and incorporation of censored data, including computational aspects.

Finally, the chapter concluded with a comparison of the original and extended methodologies from the perspective of parameter estimates and computational speed, together with a more complicated example demonstrating the generality of the method in a reliability setting.

# Chapter 5

## Networks, Simply Connected Systems & Their Signature

This short chapter develops some prosaic but essential tools for use in the remainder of the thesis. There is a degree of consensus<sup>1</sup> in the probability and statistics literature that when component failure is the event of interest and the links between components are perfectly reliable the term ‘system’ is commonly used. In contrast, when components are perfectly reliable and link failure is the event of interest the term ‘network’ is commonly used.

Shaked and Suarez-Llorens (2003) provide a catalogue of all cut sets and signatures of coherent systems up to order 4. Navarro and Rubio (2009) provided an algorithm to compute the collection of all cut sets and all signatures of coherent systems of arbitrary order, supplying a catalogue for the 180 coherent systems of order 5 in that paper. In personal correspondence, Professor Jorge Navarro kindly provided electronic copies of the same for the 16,145 coherent systems of order 6.

Samaniego (2007, ch.6) alludes to the equal applicability of the signature to networks as opposed to systems, but there does not appear to have been a systematic derivation of network signatures.

This chapter commences by disambiguating a slight clash in terminology between computer/telecommunications experts and probabilists/statisticians. This motivates a restricted class of ‘simply connected coherent systems’ which are pertinent to analysis

---

<sup>1</sup>A notable exception is the recent monograph by Gertsbakh and Shpungin (2011) which looks at ‘network reliability’ where at varying points the nodes and/or links are unreliable but no rigidly distinguishing terminology is used.

of telecommunications and computer networks. The first contribution is an algorithm which computes the network graph, cut sets and signatures of all ‘simply connected coherent systems’ of arbitrary order. The chapter continues with a second contribution which likewise computes the network graph, cut sets and signatures of all coherent networks. Both algorithms are implemented providing the final contributions of the chapter.

## 5.1 Colliding nomenclatures

Hereinafter, the terms component and node are used interchangeably as appropriate for the context.

In computer science, the term ‘network’ commonly refers to an interconnected collection of computers (or, more generically, nodes) which communicate by using direct or indirect links. There is a natural correspondence to systems and networks in reliability theory when components or links, respectively, are unreliable. However, there are at least two key distinctions from the definitions of system and network as commonly used by probabilists and statisticians:

1. A communications ‘network’ is not necessarily meant to imply that the links are less reliable than the nodes.
2. Design patterns of communications ‘networks’ mean that not every coherent system is representative of a possible network. In particular, the  $k$ -out-of- $n$  systems<sup>2</sup> are not common computer network designs<sup>3</sup>.

The author’s primary reliability interests relate to computer and telecommunications networks. Two simplifying assumptions are made: that there are no repeated edges and that either links or nodes are unreliable but not both. Thus, if link failure is of interest, the term ‘network’ is broadly applicable in both communities. When node (component) failure is of interest, the communications network is to be treated using system reliability, but the class of systems must be restricted, motivating the following definition:

---

<sup>2</sup>A  $k$ -out-of- $n$  system ( $k \leq n$ ) is any system comprising of  $n$  components which functions if and only if there are  $k$  or more of the components functioning.

<sup>3</sup>An exception to this being when end-to-end link capacity is taken into account, which could correspond to a complex consecutive  $k$ -out-of- $n$  type system. This is not considered here.

**Definition 5.1 (Simply connected coherent system)** *A coherent system is said to be simply connected if a graph representation of the system exists in which each component/node appears only once.*

As a concept this is not novel, though the choice of wording may be. Agrawal and Barlow (1984) refer to simply connected coherent systems as systems with a ‘network graph’. However, mixing of terms in this way is undesirable since in the sequel link and node failure are each of interest.

It is acknowledged that this set-up represents a highly idealised notion of a ‘communications network’, but it enables potentially ‘useful’ models (in the sense of George E. P. Box) to be developed.

Some very elementary graph theory terminology and concepts are used below which were not covered in the literature review because it would require too great a diversion. The unfamiliar reader is directed to a respected introductory text such as Bondy and Murty (2008).

## 5.2 Simply connected systems

Consideration of only simply connected coherent systems enables the study of only those systems which represent ‘networks’ most commonly used by computer scientists, but where node failure is of interest. The following algorithm provides a way of cataloguing all simply connected coherent systems of order  $n$ .

First, note that it is often convenient to introduce two artificial components (often labelled  $s$  and  $t$  in graph theory) representing either end of the graphical representation of the system. These components cannot fail, but serve to determine whether a failure has occurred: if  $s$  and  $t$  belong to the same connected component then the system is still operational. Thus, when examining order  $n$  systems, the corresponding graph has  $n + 2$  vertices.

The following algorithm is a contribution of this thesis.

**Algorithm 5.1** *Let  $\mathcal{S}$  be the (initially) empty collection of minimal cut set representations of simply connected coherent systems of order  $n$  and  $\mathcal{G}$  the collection of corresponding graphs.*

*For all integer values  $x$  from 0 to  $2^{((n+2)^2-n-2)/2} - 1$ , perform the following steps:*



1. Let  $\mathbf{b}^{(x)}$  be the  $n + 2$  dimensional binary vector representation of  $x$ . If  $\sum_i b_i^{(x)} < n + 1$ , or if  $b_{n+1}^{(x)} = 1$ , proceed to the next value of  $x$ .
2. Create an  $(n+2) \times (n+2)$  symmetric adjacency matrix,  $A$ , whose upper triangular entries are filled row-wise by the binary representation of  $x$ .  
  
 $A$  defines an undirected graph  $G_x(V, E)$ . Label the first vertex  $s$  and the last vertex  $t$ .
3. If the algebraic connectivity of the graph is zero, proceed to the next value of  $x$ .
4. Compute the collection  $\mathcal{C}_x$  of all minimal  $(s, t)$  vertex cut sets of  $G_x$ . If the union of all sets in  $\mathcal{C}_x$  does not equal  $V \setminus \{s, t\}$ , proceed to the next value of  $x$ .
5. For each possible permutation  $\sigma(2, \dots, n + 1)$ , if  $\sigma(\mathcal{C}_x) \in \mathcal{S}$  then proceed to the next value of  $x$ .
6. If this step is reached, add  $\mathcal{C}_x$  to  $\mathcal{S}$  and correspondingly  $G_x$  to  $\mathcal{G}$ , then proceed to the next value of  $x$ . □

The following Theorem establishes the veracity of the algorithm and is a contribution of this thesis.

**Theorem 5.1** *Algorithm 5.1 produces the cut-sets and network graph representations of all simply connected coherent systems.*

*Proof:* Three things must be proven: (i) that every possible ‘network graph’ representation is checked; (ii) that no non-coherent systems are added to  $\mathcal{S}$ ; and (iii) that all coherent systems are added to  $\mathcal{S}$ .

First, every ‘network graph’ is an undirected graph. Moreover, every undirected graph without self-loops or repeated edges can be represented by a symmetric binary adjacency matrix with zeros on the diagonal and consequently is uniquely determined by the binary upper triangular entries.

An  $(n + 2) \times (n + 2)$  matrix has  $\frac{(n+2)^2 - n - 2}{2}$  upper triangular elements. Therefore, the row-wise vector of upper triangular elements is an  $\frac{(n+2)^2 - n - 2}{2}$  dimensional binary vector. Thus, the binary representation of the integers 0 to  $2^{((n+2)^2 - n - 2)/2} - 1$  indexes all possible adjacency matrices for undirected graphs with  $n + 2$  nodes as required.

Secondly, assume a system which is not coherent reaches step 4. Then by Theorem 2.4, page 16, the union of all sets in the collection  $\mathcal{C}_x$  will not consist of all (actual) component vertices,  $V \setminus \{s, t\}$ . Thus any system which is not coherent is eliminated at step 4, does not reach step 6 and so does not appear in  $\mathcal{S}$ .

Clearly, only step 4 (and to avoid duplication, step 5) is necessary, but the other steps aid computationally by reducing the number of systems for which the costly operation of cut set computation is performed. Nonetheless, it remains to prove the third item: that no coherent systems are eliminated in the process.

So, finally any order  $n$  coherent system must successfully pass all steps of the algorithm:

1. If  $\sum_i b_i^{(x)} < n + 1$  then at least one row,  $i$ , of the adjacency matrix will be zero. But then vertex  $i$  contains no links to the other vertices. If  $i$  is 1 or  $n + 2$ , then  $s$  and  $t$  are not connected directly or indirectly and the system never functions irrespective of component state, so all components are irrelevant (Definition 2.5, page 15) and the system is not coherent. If  $i$  is any other vertex, then the component in question may be operational or failed and will have no impact on any possible paths from  $s$  to  $t$ , so it is irrelevant. Therefore the system is not coherent.

Similarly, if  $b_{n+1}^{(x)} = 1$ , then when the adjacency matrix is created in step 2 this will mean  $A_{1,n+2} = A_{n+2,1} = 1$ . But then  $s$  and  $t$  are directly linked and the system always functions irrespective of component state, so all components are irrelevant and the system is not coherent.

Hence, rejection at step 1  $\implies$  the system is not coherent. Thus, the contrapositive holds: the system is coherent  $\implies$  step 1 will be passed.

2. Step 2 simply sets up the graph based on the adjacency matrix.
3. The algebraic connectivity of the graph is zero if and only if the graph is connected. Thus, step 3 also pertains to component relevance: if the algebraic connectivity is zero, then either  $s$  and  $t$  are in different connected components (and the system never functions) and/or some component is not relevant because it is not in the connected component containing  $s$  and  $t$ .

4. Theorem 2.4 means all coherent systems of order  $n$  will pass step 4.
5. Step 5 simply eliminates isomorphic coherent systems so that there is no duplication when recording them in the final step. □

In terms of implementation, the non-trivial steps are 3 and 4.

Whether to pass step 3 can be determined by inspecting the second smallest eigenvalue of the Laplacian matrix of the graph, or alternatively by simply performing some brute force exhaustive search from any initial node, say  $s$ .

Minimal cut set discovery can be achieved using Berry *et al.* (1999) (implemented in Csárdi and Nepusz, 2006), which establishes all vertex separators of all vertex pairs, then testing each set for  $(s, t)$  separation.

### 5.3 Coherent networks

The algorithm to catalogue all networks with a given number of nodes closely mirrors Algorithm 5.1. However, coherency is now determined by link cut sets and equivalence of networks equates to equivalence of graphs.

For clarity, when talking of a network with  $n$  nodes, this will be referred to as the node order.

The following algorithm is a contribution of this thesis.

**Algorithm 5.2** *Let  $\mathcal{S}$  be the (initially) empty collection of minimal cut set representations of coherent networks with node order  $n$  and  $\mathcal{G}$  the collection of corresponding graphs.*

*For all integer values  $x$  from 0 to  $2^{((n+2)^2-n-2)/2} - 1$ , perform the following steps:*

1. *Let  $\mathbf{b}^{(x)}$  be the  $n + 2$  dimensional binary vector representation of  $x$ . If  $\sum_i b_i^{(x)} < n + 1$ , or if  $b_{n+1}^{(x)} = 1$ , proceed to the next value of  $x$ .*
2. *Create an  $(n+2) \times (n+2)$  symmetric adjacency matrix,  $A$ , whose upper triangular entries are filled row-wise by the binary representation of  $x$ .*

*$A$  defines an undirected graph  $G_x(V, E)$ . Label the first vertex  $s$  and the last vertex  $t$ . In addition, ‘colour’  $s$  and  $t$  black and ‘colour’  $V \setminus \{s, t\}$  red, say.*

3. If the algebraic connectivity of the graph is zero, proceed to the next value of  $x$ .
4. Compute the collection  $\mathcal{C}_x$  of all minimal  $(s, t)$  edge cut sets of  $G_x$ . If the union of all sets in  $\mathcal{C}_x$  does not equal  $E$ , proceed to the next value of  $x$ .
5. If there exists a coloured isomorphism from  $G_x$  to any member of the collection  $\mathcal{G}$ , proceed to the next value of  $x$ .
6. If this step is reached, add  $\mathcal{C}_x$  to  $\mathcal{S}$  and correspondingly  $G_x$  to  $\mathcal{G}$ , then proceed to the next value of  $x$ . □

**Theorem 5.2** *Algorithm 5.2 produces the cut-sets and network graph representations of all coherent networks.*

The full proof of this Theorem appreciably mirrors the proof of Theorem 5.1 due to the essential duality between component/link failure. Step 5 is phrased in terms of coloured isomorphisms because there is extensive existing literature on computational methods to check for these. Indeed, it is implementation of steps 3, 4 and 5 which are non-trivial.

Step 3 is as discussed for Algorithm 5.1.

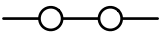
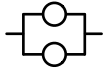
Computation of the edge cut sets in step 4 has not been implemented in publicly available R (R Core Team, 2012) code. The author therefore implemented Shin and Koh (1998) which is available as a distinct utility function in Aslett (2012), and is a contribution of this thesis.

The search for the isomorphisms of step 5 can be performed using Cordella *et al.* (2001) (implemented in Csárdi and Nepusz, 2006).

It is important to note that the resultant signatures will be of varying dimension — unlike for coherent systems — since for a fixed node order there are variable numbers of links.

## 5.4 Representations and their signature

The author implemented both Algorithm 5.1 and Algorithm 5.2. These were used to catalogue the network graphs, signatures and cut sets of varying orders of both simply connected coherent systems and coherent networks.

Number	System Topology	Signature
1		(1, 0)
2		(0, 1)

**Table 5.1:** All simply connected systems of order 2, together with system signature. Contained in data set `sccs02` of `ReliabilityTheory` package (see §7.2.1, page 150).

The ordering of signatures is an interesting area of active research in its own right. A variety of approaches including stochastic ordering, hazard rate ordering and likelihood ratio ordering are considered in Block *et al.* (2006). However, the necessary and sufficient conditions for two systems to be comparable are not always met and Samaniego (2007) explores the stochastic precedence ideas in Arcones *et al.* (2002) as a possible alternative.

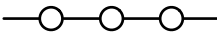
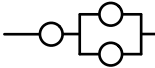
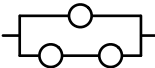
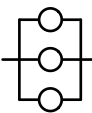
Here, for comparability a similar approach to that taken in the primary other signature cataloguing work (Navarro and Rubio, 2009) is taken, computing the expected lifetimes when assuming independent and identically Exponentially distributed component lifetimes. Since  $\mathbb{E}[Y_{i:n}] = \lambda^{-1} \sum_{j=1}^i 1/(n - j + 1)$  (Samaniego, 2007, p.105), ordering is on  $\mathbb{E}_\tau[T] = \sum_{i=1}^n s_i \mathbb{E}[Y_{i:n}]$ .

### 5.4.1 Simply connected coherent system signatures

A contribution of this thesis is full cataloguing of the network graphs and signatures of all simply connected coherent systems of order 2, 3, 4 and 5. Tables 5.1 to 5.3 show orders 2 to 4, whilst the longer order 5 list is relegated to Appendix A, Table A.2.

It is interesting to note how many coherent systems are excluded by the simple connectivity requirement:

- There is no change to the number of coherent systems of order 2.
- There is one less of order 3, down to 4 from 5.
- There are 9 fewer of order 4, down to 11 from 20.
- There are 145 fewer of order 5, down to 35 from 180.

Number	System Topology	Signature
1		$(1, 0, 0)$
2		$(\frac{1}{3}, \frac{2}{3}, 0)$
3		$(0, \frac{2}{3}, \frac{1}{3})$
4		$(0, 0, 1)$

**Table 5.2:** All simply connected systems of order 3, together with system signature. Contained in data set `sccs03` of `ReliabilityTheory` package (see §7.2.1, page 150).

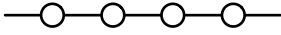
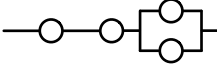
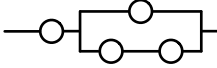
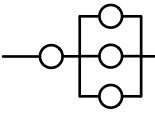
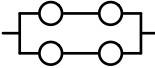
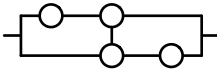
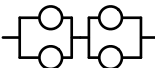
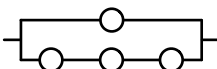
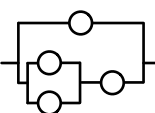
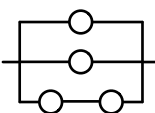
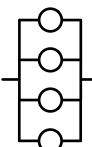
These reductions in the quantity of relevant systems will prove highly advantageous in the next chapter.

### 5.4.2 Coherent network signatures

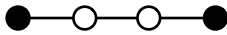
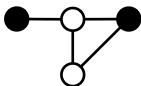
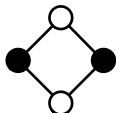
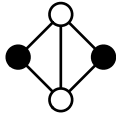
A contribution of this thesis is full cataloguing of the network graphs and signatures of all coherent networks of node order 2 and 3. Tables 5.4 shows node order 2 coherent networks, whilst node order 3 is in Appendix A, Table A.1.

## 5.5 Summary

This short chapter has clarified the connection between telecommunications networks and those systems and networks to which they correspond in reliability theory. Algorithms to enumerate these were provided and implemented, complete with their signature. Although prosaic, these are important tools for the remainder of the thesis.

Number	System Topology	Signature
1		$(1, 0, 0, 0)$
2		$(\frac{1}{2}, \frac{1}{2}, 0, 0)$
3		$(\frac{1}{4}, \frac{7}{12}, \frac{1}{6}, 0)$
4		$(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0)$
5		$(0, \frac{2}{3}, \frac{1}{3}, 0)$
6		$(0, \frac{1}{2}, \frac{1}{2}, 0)$
7		$(0, \frac{1}{3}, \frac{2}{3}, 0)$
8		$(0, \frac{1}{2}, \frac{1}{4}, \frac{1}{4})$
9		$(0, \frac{1}{6}, \frac{7}{12}, \frac{1}{4})$
10		$(0, 0, \frac{1}{2}, \frac{1}{2})$
11		$(0, 0, 0, 1)$

**Table 5.3:** All simply connected systems of order 4, together with system signature. Contained in data set `sccs04` of ReliabilityTheory package (see §7.2.1, page 150).

Number	Network Topology	Signature
1		$(1, 0, 0)$
2		$(\frac{1}{4}, \frac{7}{12}, \frac{1}{6}, 0)$
3		$(0, \frac{2}{3}, \frac{1}{3}, 0)$
4		$(0, \frac{1}{5}, \frac{3}{5}, \frac{1}{5}, 0)$

**Table 5.4:** All coherent networks of node order 2, together with system signature. Black nodes are start/terminal. Contained in data set `cn02` of `ReliabilityTheory` package (see §7.2.1, page 150).



# Chapter 6

## Parametric and Topological Inference with Masked Lifetime Data

The literature on inference for masked system lifetime data is extensive, though interestingly is heavily focused on specific structures (e.g. series/competing risk systems, see Reiser *et al.* (1995) or Kuo and Yang (2000)), specific lifetime distributions (e.g. Exponential, see Gåsemyr and Natvig (2001)) or does not focus on inferring the parameters of the model (e.g. infer hazard, see Ng *et al.* (2012)).

Chapter 4 dealt with inference for Phase-type distributions, providing a very flexible modelling option specifically for repairable systems when the data consists of masked system lifetimes. The original objective for the continuation of the research was to have a system composed of Phase-type distributed subsystems and push the ‘masking’ further: that is, where the ultimate failure time of only a whole system of repairable subsystems is known. Initially, to simplify the problem the Phase-type subsystems are only considered repairable up until their initial failure whereupon they remain failed.

However, the approach which was developed ultimately transpired to have more general applicability. Therefore, in full generality this chapter considers the traditional setting where repair of components in a system is not possible, and a novel signature based data augmentation MCMC scheme is developed for performing Bayesian inference on parameters of the component lifetime distribution in the presence of masked system lifetime data. Crucially, the proposed methodology works for arbitrary coherent

system and network designs and arbitrary (but identical) component lifetime distributions. The signature has been used in inferential work before but not in this way, a recent review of parametric work being Balakrishnan *et al.* (2012), and the first use of signatures with nonparametric predictive inference recently in Coolen and Al-nefaiee (2012).

Furthermore, the proposed methodology enables inference on the topology of the system or network. A collection of candidate systems or networks is specified and posterior probabilities of the structure being each member of the collection can be obtained. This was the motivation for identifying particular structures consistent with telecommunications networks in Chapter 5, along with associated signatures, as will be seen in the sequel.

The chapter begins by formalising the setting of the problem to be addressed under both independent and identically distributed (i.i.d.) and exchangeable systems assumptions. It continues by deriving the signature based data augmentation MCMC scheme for inferring the parameters of the component lifetime distributions in masked lifetime data problems and then extends this to also inferring the topology of the system. Finally, the results of some simple illustrative examples are presented.

## 6.1 The problem setting

The masked system lifetime data inferential problem is examined in two cases: that of i.i.d. systems and that of exchangeable systems.

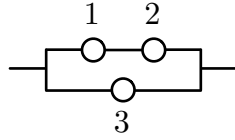
### 6.1.1 Independent systems

The usual route, given a system  $\phi$  with structure function  $\varphi$ , is to establish the system lifetime distribution,  $F_\tau(\cdot)$ , via the reliability function, as in (2.3). Then, in principle inference can follow naturally as described in §3.1:

$$f_{\Psi|\tau}(\psi | \mathbf{t}) \propto L(\psi; \mathbf{t}) f_{\Psi}(\psi)$$

where  $L(\psi; \mathbf{t})$  is the likelihood and  $f_{\Psi}(\psi)$  is the joint prior distribution. The complexity of the likelihood can make this a non-trivial matter. In the i.i.d. case this is:

$$L(\psi; \mathbf{t}) = \prod_{i=1}^m \frac{\partial}{\partial t} F_\tau(t; \psi) \Big|_{t=t_i}$$



**Figure 6.1:** A simple 3 component system.

For example, consider the simple three component system in Figure 6.1. In this instance:

$$\begin{aligned}\varphi(\mathbf{y}) &= 1 - (1 - y_1 y_2)(1 - y_3) \\ \implies \bar{F}_\tau(t) &= 1 - (1 - \bar{F}_{Y_1}(t)\bar{F}_{Y_2}(t))(1 - \bar{F}_{Y_3}(t))\end{aligned}$$

Thus, if the model is that components have Weibull distributed lifetimes,  $Y_i \stackrel{\text{i.i.d.}}{\sim}$  Weibull(scale =  $\alpha$ , shape =  $\beta$ ), so that  $\psi = (\alpha, \beta)$ , a difficult likelihood can result:

$$L(\alpha, \beta; \mathbf{t}) = \prod_{i=1}^m t_i^{-1} \beta (t_i/\alpha)^\beta \exp \{-3(t_i/\alpha)^\beta\} [2 \exp \{(t_i/\alpha)^\beta\} + \exp \{2(t_i/\alpha)^\beta\} - 3]$$

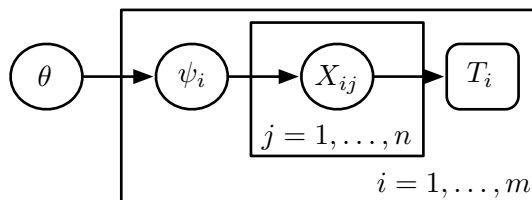
which makes evaluation of the posterior  $f_{\Psi|\tau}(\alpha, \beta | \mathbf{t})$  very cumbersome. Indeed, in modestly more complex situations than this, it may make evaluation of the normalising constant intractable. A Metropolis-Hastings algorithm could be used, but would require careful tuning for every problem.

### 6.1.2 Exchangeable systems

Many models in the literature make the assumption of identically distributed systems just described. Identical distribution of components within a system is a common modelling assumption, since they are often exposed to a common environment. However, assuming identical distribution at the system level is more questionable, since different systems in the field may be subject to wildly different environmental conditions.

A second setting considered in this chapter is where only components within a single system are assumed identically distributed, but systems themselves are considered exchangeable. That is, the model considers systems to be conditionally independent given the component lifetime parameters for components in each system. For example, this corresponds to the hierarchical independence structure described in §3.1.5.

More precisely, suppose that components within system  $i$  follow a common lifetime distribution  $F_Y(y; \psi_i)$ , but where  $\psi_i \neq \psi_j$  for  $i \neq j$ , so that the parameters may vary



**Figure 6.2:** Graphical model representing exchangeable systems.

from system to system. The parameters are realisations from a population distribution  $F_{\Psi}(\psi; \theta)$ . This is depicted in the graphical model in Figure 6.2.

Thus, interest now lies in  $\theta$ , since this determines the distribution of the component parameters random variable  $\Psi$ .

$$f_{\Theta|\tau}(\theta | \mathbf{t}) = f_{\Xi|\tau}(\theta, \boldsymbol{\psi} | \mathbf{t}) / f_{\Psi|\tau}(\boldsymbol{\psi} | \theta, \mathbf{t})$$

This then leads to a posterior predictive distribution for  $\psi$  (recall (3.7), page 39):

$$f_{\Psi^*|\tau}(\psi^* | \mathbf{t}) = \int f_{\Psi}(\psi^*; \theta) f_{\Theta|\tau}(\theta | \mathbf{t}) d\theta \quad (6.1)$$

Clearly this posterior would be substantially more complex than the independent systems model and an alternative approach to direct evaluation is required.

The following section details a novel methodology for achieving this.

## 6.2 Signature based MCMC algorithm

This section contains the majority of the major contributions in this chapter, where a data augmentation based Markov chain Monte Carlo algorithm which makes novel use of the system signature (see Definition 2.8, page 17) is developed.

### 6.2.1 Inferring component lifetime parameters

The first concern, given a system modelled via its component structure plus component lifetime distributions, is to infer the parameters of the lifetime distributions in the presence of masked system lifetime data.

First in the independent systems case, consider instead the simpler situation where all component failure times are known, so that there are  $nm$  i.i.d. observations. Let  $y_{ij}$  be the observed failure time of component  $j$  in system  $i$ . Let  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$

be the collection of all component failure times, where  $\mathbf{y}_i = \{y_{i1}, \dots, y_{in}\}$  are the component failure times for the  $i^{\text{th}}$  system. As already implied, inference would then be straightforward, since:

$$f_{\Psi|Y}(\psi | \mathbf{y}) \propto L(\psi; \mathbf{y}) f_{\Psi}(\psi)$$

where now  $L(\psi; \mathbf{y})$  is just a product of simple component lifetime densities, rather than of complex system lifetime densities. One might then choose  $f_{\Psi}(\psi)$  to be the conjugate prior, leading to a well understood posterior density which can easily be plotted and sampled from.

Given that inference with such ‘complete data’ is simple, the proposal is to introduce the latent component lifetimes as variables and so develop MCMC algorithms which are generally applicable. The schemes now presented fall under the Gibbs sampling via data augmentation methodology, as discussed in §3.2.3. The problem to be solved, and thus the novel aspect of the data augmentation scheme, is specifically how to perform such augmented sampling in this reliability setting.

To be precise, in the independent model MCMC samples are generated from the natural completion of the posterior distribution:

$$f_{\Psi, Y|\tau}(\psi, \mathbf{y}_1, \dots, \mathbf{y}_m | \mathbf{t})$$

by blocked Gibbs sampling using the two full conditional distributions which fully define it (by the Hammersley-Clifford Theorem, page 49):

$$f_{Y|\Psi, \tau}(\mathbf{y}_1, \dots, \mathbf{y}_m | \psi, \mathbf{t}) \tag{6.2}$$

$$f_{\Psi|Y, \tau}(\psi | \mathbf{y}_1, \dots, \mathbf{y}_m, \mathbf{t}) \tag{6.3}$$

Likewise in the case of the exchangeable model with parameters  $\Xi = (\Theta, \Psi)$ , the joint distribution which is the natural completion of the posterior is:

$$f_{\Xi, Y|\tau}(\theta, \psi_1, \dots, \psi_m, \mathbf{y}_1, \dots, \mathbf{y}_m | \mathbf{t})$$

for which blocked Gibbs sampling can be performed, using the two full conditional distributions which fully define it:

$$f_{Y|\Xi, \tau}(\mathbf{y}_1, \dots, \mathbf{y}_m | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \tag{6.4}$$

$$f_{\Xi|Y, \tau}(\theta, \psi_1, \dots, \psi_m | \mathbf{y}_1, \dots, \mathbf{y}_m, \mathbf{t}) \tag{6.5}$$

In this particular application, sampling from (6.3) and (6.5) is a well understood problem for common component lifetime distributions: it is standard Bayesian inference, since  $\psi$  is conditionally independent of  $\mathbf{t}$  given all component failure times. In particular, note that if the ‘components’ are Phase-type repairable subsystems then the contributions of the previous chapter provides the necessary methodology.

However, sampling from (6.2) and (6.4) is the key step which is solved in the sequel with Algorithm 6.1.

Note that in (6.4),

$$f_{Y|\Xi,\tau}(\mathbf{y}_1, \dots, \mathbf{y}_m | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) = f_{Y|\Psi,\tau}(\mathbf{y}_1, \dots, \mathbf{y}_m | \psi_1, \dots, \psi_m, \mathbf{t})$$

since in the presence of  $\psi_i$  the component failure times are conditionally independent of  $\theta$ .

### 6.2.1.1 Sampling Latent Component Failure Times

Sampling the latent component failure times is the crucial step in being able to implement the data augmented Gibbs sampling scheme just proposed in order to perform Bayesian inference for masked system lifetime data. The approach advocated depends on the signature of the system, as described in Definition 2.8, page 17. Recall the signature of the system is the probability vector  $\mathbf{s}$  with elements:

$$s_i = \mathbb{P}(\tau = T_{i:n})$$

There are two scenarios: firstly, when each system is independent, we can consider separately each vector of failure times  $f_{Y|\Psi,\tau}(\mathbf{y}_i | \psi, t_i)$ ; and, secondly when each system is exchangeable, meaning system  $i$  is conditionally independent of the other systems given its lifetime distribution parameter  $\psi_i$ , we can consider separately each vector of failure times:  $f_{Y|\Psi,\tau}(\mathbf{y}_i | \psi_i, t_i)$ . Therefore, when writing  $\psi$  hereinafter, it is meant as appropriate in the context of independence or exchangeability.

The core result which makes the data augmentation algorithm feasible is the statement and proof of the following Lemma which is a contribution of this thesis.

**Lemma 6.1** *Consider a coherent system with signature  $\mathbf{s} = (s_1, \dots, s_n)$ , consisting of components with lifetime distribution  $F_Y(\cdot; \psi)$ . Then, conditioning on the system*

failure time  $t$ , the lifetime density of the components is:

$$\begin{aligned}
& f_{Y|\tau}(y_{i1}, \dots, y_{in}; \psi | t) \\
& \propto \sum_{j=1}^n \left[ f_{Y|Y < t}(y_{i(1)}, \dots, y_{i(j-1)}; \psi) f_{Y|Y > t}(y_{i(j+1)}, \dots, y_{i(n)}; \psi) \mathbb{I}_{\{t\}}(y_{i(j)}) \right. \\
& \quad \left. \times \binom{n-1}{j-1} F_Y(t; \psi)^j \bar{F}_Y(t; \psi)^{n-j+1} s_j \right]
\end{aligned}$$

*Proof:* Since these are assumed to be coherent systems, the failure of a system must coincide with the failure of one of the components, meaning  $y_{ij} = t_i$  for some  $j$ , and the remaining components of that system failed either before or after  $t_i$ .

$$\begin{aligned}
& f_{Y|\tau}(y_{i1}, \dots, y_{in}; \psi | \tau = t) \\
& = \sum_{j=1}^n f_{Y|\tau}(y_{i1}, \dots, y_{in}, \tau = Y_{j:n}; \psi | \tau = t) \\
& = \sum_{j=1}^n f_{Y|\tau}(y_{i1}, \dots, y_{in}; \psi | \tau = t = Y_{j:n}) \mathbb{P}(\tau = Y_{j:n} | \psi, \tau = t) \\
& = \sum_{j=1}^n \left[ f_{Y|Y < t}(y_{i(1)}, \dots, y_{i(j-1)}; \psi) f_{Y|Y > t}(y_{i(j+1)}, \dots, y_{i(n)}; \psi) \mathbb{I}_{\{t\}}(y_{i(j)}) \right. \\
& \quad \left. \times \mathbb{P}(\tau = Y_{j:n} | \psi, \tau = t) \right]
\end{aligned}$$

This means that the component failure times are from a mixture distribution. Moreover, the weights have a tractable form:

$$\begin{aligned}
\mathbb{P}(\tau = Y_{j:n} | \psi, \tau = t) & \propto \mathbb{P}(\tau = t | \tau = Y_{j:n}, \psi) \mathbb{P}(\tau = Y_{j:n} | \psi) \\
& = \mathbb{P}(\tau = t | \tau = Y_{j:n}, \psi) \mathbb{P}(\tau = Y_{j:n}) \\
& = \binom{n-1}{j-1} F_Y(t; \psi)^j \bar{F}_Y(t; \psi)^{n-j+1} s_j \tag{6.6}
\end{aligned}$$

as required. □

Thus, the system signature enables determination of how many of the remaining components failed before and how many after  $t_i$ , via the mixing probabilities (6.6). The probability densities which are mixed are trivial to sample from so long as the left and right censored conditional distributions of the component lifetime distributions are.

This leads to the following algorithm to simulate the latent component failure times.

**Algorithm 6.1 (Signature based data augmentation)** For each of the systems  $i = 1, \dots, m$ :

1. Sample  $j \in \{1, \dots, n\}$  from the discrete probability distribution

$$\mathbb{P}(j) \propto \binom{n-1}{j-1} F_Y(t_i; \psi)^j \bar{F}_Y(t_i; \psi)^{n-j+1} s_j$$

This samples the order statistic indicating that the  $j^{\text{th}}$  component failure caused system failure.

2. Sample:

- $j-1$  values,  $y_{i1}, \dots, y_{i(j-1)}$ , from  $F_{Y|Y < t_i}(\cdot; \psi)$ , the distribution of the component lifetime conditional on failure before  $t_i$
- $n-j$  values,  $y_{i(j+1)}, \dots, y_{in}$ , from  $F_{Y|Y > t_i}(\cdot; \psi)$ , the distribution of the component lifetime conditional on failure after  $t_i$

and set  $y_{ij} = t_i$ .

Each iteration provides  $\mathbf{y}_i$ . and together the  $mn$  values  $y_{..}$  comprise  $\mathbf{y}$ . □

That is, the preceding algorithm can be used to perform (6.2) and (6.4) so long as simulation from the censored conditional distributions is tractable. In principle, the finite support of  $F_{Y|Y < t_i}(\cdot; \psi)$  means it should always be relatively straightforward to sample, since ARMS (Gilks *et al.*, 1995) is very effective for finitely supported univariate distributions.  $F_{Y|Y > t_i}(\cdot; \psi)$  will require more care, but in the worst case simple rejection sampling is in principle an option.

In summary, if the state space is augmented with component failure times, then inference on the rate parameters is trivial and Algorithm 6.1 provides the means of simulating the augmented data. The resulting chain will be vectors of samples  $(\psi, \mathbf{y}_1, \dots, \mathbf{y}_m)$  which can then be used for estimation of quantities of interest related to the posterior distribution as with any MCMC sample.

## 6.2.2 Topological inference

Thus far, the discussion has implicitly assumed that the topology of the system/network in question is known. However, by introducing the signature into the inferential



Type	Order	Signature repetition							Total
		Unique	2	3	4	5	6	7	
Coherent systems	2	2	0	0	0	0	0	0	2
	3	5	0	0	0	0	0	0	5
	4	14	3	0	0	0	0	0	20
	5	43	15	2	6	2	10	1	180
Simply connected coherent systems	2	2	0	0	0	0	0	0	2
	3	4	0	0	0	0	0	0	4
	4	11	0	0	0	0	0	0	11
	5	27	4	0	0	0	0	0	35

**Table 6.1:** Uniqueness of signatures for coherent systems.

procedure it becomes possible to treat the internal structure as unknown and jointly incorporate it into the inference. Thus, in this subsection the assumption is weakened: that is, the connectivity is unknown, save for assuming the connectivity is such that it is a (possibly simply connected) coherent system/network.

There is not necessarily a bijective mapping between signature and topology, but the signature does identify small subsets of topologies. For example, there are 20 unique topologies of systems of order 4, each with a signature. However, only 17 of the 20 signatures are distinct, so successfully identifying the correct signature will in most (but not all) cases identify the topology. When interest is in telecommunications networks (as discussed in Chapter 5), the situation is improved because there are fewer non-unique signatures in the simply connected case, as illustrated in Table 6.1.

The first choice to be made is what collection of topologies,  $\mathcal{M}$ , are of interest. Indeed, Chapter 5 was motivated by the desire to restrict the topologies under consideration (even if, say, a  $k$ -out-of- $n$  system was inferentially credible):

*“Unless there is good reason otherwise, models should obey natural or known constraints even if these lie outside the range of the data.”*

— Cox and Connelly (2011, p.110)

For example, one may know a priori that there are 4 components and that it is a simply connected system since it represents a communications ‘network’, meaning  $\mathcal{M}$

consists of Table 5.3. Alternatively, one may choose to assume simply that there are 5 components or fewer and that it is just a coherent system, in which case  $\mathcal{M}$  consists of all signature results in Shaked and Suarez-Llorens (2003) and Navarro and Rubio (2009). Finally, it is not restricted to systems: one may know a priori there are 3 nodes, but that link failure is of interest so that  $\mathcal{M}$  consists of the coherent networks of Table A.1. This collection will be considered indexed by the element number,  $j$ .

With the collection of topologies of interest,  $\mathcal{M}$ , decided, the posterior becomes:

$$f_{\mathcal{M},\Xi,Y|\tau}(\mathcal{M}_j, \theta, \boldsymbol{\psi}, \mathbf{y} | \mathbf{t})$$

One might initially propose simply extending the simple Gibbs sampler to a systematic or random scan Gibbs sampler (page 47) between the full conditionals:

$$f_{Y|\mathcal{M},\Xi,\tau}(\mathbf{y}_{1\cdot}, \dots, \mathbf{y}_{m\cdot} | \mathcal{M}_j, \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \quad (6.7)$$

$$f_{\Xi|\mathcal{M},Y,\tau}(\theta, \psi_1, \dots, \psi_m | \mathcal{M}_j, \mathbf{y}_{1\cdot}, \dots, \mathbf{y}_{m\cdot}, \mathbf{t}) \quad (6.8)$$

$$f_{\mathcal{M}|\Xi,Y,\tau}(\mathcal{M}_j | \theta, \psi_1, \dots, \psi_m, \mathbf{y}_{1\cdot}, \dots, \mathbf{y}_{m\cdot}, \mathbf{t}) \quad (6.9)$$

where (6.7) and (6.8) are effectively unchanged from the discussion above, except that the signature involved may vary on each iteration. However, an obvious problem is that (6.9) would experience extremely poor mixing. To see this, consider the situation where the chain is started with topology  $\mathcal{M}_j$  of a purely parallel system. Now only those systems with signature such that  $s_n > 0$  can be proposed. Should a move to one of those topologies occur (where necessarily  $s_n < 1$ ), (6.7) will then likely produce simulations of  $\mathbf{y}_i$  st  $\exists i, y_{in} \neq t_i$ , making a subsequent move back to a parallel topology impossible.

Much more seriously, this also draws attention to the fact that the positivity condition (Definition 3.5, page 48) is not satisfied. Thus the Hammersley-Clifford Theorem does not hold and the full conditionals are not necessarily adequate to describe the joint distribution of interest. Indeed, not considering the positivity condition is one of the most common pitfalls when using Gibbs samplers (Robert and Casella, 2011).

The problem can be avoided by using the following full conditionals instead:

$$f_{\mathcal{M},Y|\Xi,\tau}(\mathcal{M}_j, \mathbf{y}_{1\cdot}, \dots, \mathbf{y}_{m\cdot} | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \quad (6.10)$$

$$f_{\Xi|\mathcal{M},Y,\tau}(\theta, \psi_1, \dots, \psi_m | \mathcal{M}_j, \mathbf{y}_{1\cdot}, \dots, \mathbf{y}_{m\cdot}, \mathbf{t}) \quad (6.11)$$

since the block marginals are concordant with positivity.

A natural choice for sampling (6.10) is to sequentially sample:

$$f_{\mathcal{M}|\Xi,\tau}(\mathcal{M}_j | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \quad (6.12)$$

$$f_{Y|\mathcal{M},\Xi,\tau}(\mathbf{y}_1, \dots, \mathbf{y}_m | \mathcal{M}_j, \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \quad (6.13)$$

(6.13) is again unchanged from the previous discussion, except that the signature may vary on each iteration. Note that (6.12) can be written:

$$\begin{aligned} f_{\mathcal{M}|\Xi,\tau}(\mathcal{M}_j | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) &= f_{\mathcal{M}|\Psi,\tau}(\mathcal{M}_j | \psi_1, \dots, \psi_m, \mathbf{t}) \\ &\propto f_{\tau|\Psi,\mathcal{M}}(\mathbf{t} | \psi, \mathcal{M}_j) f_{\Psi|\mathcal{M}}(\psi | \mathcal{M}_j) f_{\mathcal{M}}(\mathcal{M}_j) \\ &\propto \left\{ \prod_{i=1}^m f_{\tau|\Psi,\mathcal{M}}(t_i | \psi_i, \mathcal{M}_j) \right\} f_{\mathcal{M}}(\mathcal{M}_j) \end{aligned}$$

since the component/link lifetime parameters  $\psi$  are unaffected by the particular topology in which they are laid out<sup>1</sup>. The term  $f_{\tau|\Psi,\mathcal{M}}(t_i | \psi_i, \mathcal{M}_j)$  is computed using (2.7) in Corollary 2.6, page 18, for absolutely continuous component/link lifetime distributions. The term  $f_{\mathcal{M}}(\mathcal{M}_j)$  is the prior probability of topology  $j$ .

When the candidate collection of topologies is not too large, this discrete probability mass function can be trivially sampled directly. In the event the collection is very large, it is possible to perform Metropolis-Hastings sampling to avoid having to compute the above for all topologies in order to find the normalising constant — in this instance, if a non-uniform proposal is used then the ordering of elements of  $\mathcal{M}$  alluded to in §5.4 becomes an issue of serious interest.

### 6.2.3 Summary of the MCMC algorithm

For clarity, the above discussion is now crystallised into the following algorithm which summarises how to proceed, and is a contribution of this thesis.

**Algorithm 6.2 (MCMC for masked lifetime data)** *Before commencing the algorithm, the following prerequisites are required:*

- (i) *A collection  $\mathcal{M}$  of candidate system/network signatures and a discrete prior distribution  $F_{\mathcal{M}}(\cdot)$  on the elements of the collection. If topological inference is not required, this collection may contain just one signature vector.*

---

<sup>1</sup>This is a simplifying assumption: there are naturally situations where it is conceivable that topology could impact lifetime parameters.

- (ii) A component lifetime distribution  $F_Y(\cdot; \psi_i)$ , a prior on the exchangeable population of system/network parameter(s) of that distribution  $F_\Psi(\psi_i; \theta)$  and a hyperprior  $F_\Theta(\cdot)$ .
- (iii) A means of performing standard Bayesian inference with the component lifetime distribution.
- (iv) A method of sampling from  $F_{Y|Y<t}(\cdot; \psi)$  and  $F_{Y|Y>t}(\cdot; \psi) \forall t \in [0, \infty)$ .

The following steps should be iterated as many times as required to bring the MCMC standard error of the functional of interest to an acceptable level. Before starting, specify starting values for  $\psi$ , perhaps as random draws from  $F_\Psi(\cdot)$

1. If the collection  $\mathcal{M}$  contains more than one signature, then draw a signature,  $\mathcal{M}_j$ , from the discrete probability mass function:

$$f_{\mathcal{M}|\Xi, \tau}(\mathcal{M}_j | \theta, \psi_1, \dots, \psi_m, \mathbf{t}) \propto \left\{ \prod_{i=1}^m f_{\tau|\Psi, \mathcal{M}}(t_i | \psi_i, \mathcal{M}_j) \right\} f_{\mathcal{M}}(\mathcal{M}_j)$$

Otherwise take  $\mathcal{M}_j$  as the singleton element.

2. Using Algorithm 6.1, sample from  $f_{Y|\Psi, \tau, \mathcal{M}}(\mathbf{y}_i | \psi_i, t_i, \mathcal{M}_j)$  for each observation  $i$ , until the full vector  $\mathbf{y}$  of augmented component lifetimes is simulated.
3. Perform standard Bayesian inference to sample updated parameter values from:

$$f_{\Xi|Y}(\theta, \psi_1, \dots, \psi_m | \mathbf{y}_1, \dots, \mathbf{y}_m)$$

and then loop to step 1. □

It is important to note that if the method used in step 3 allows a jump to any part of the parameter space then by Lemma 3.4 the MCMC algorithm is Harris ergodic.

The algorithm above is for the exchangeable systems case, the independent case being a trivial simplification.

## 6.3 Examples

The generality of this approach means that some clarity will be afforded by a few thoroughly worked examples. The most lucid exposition is provided first with the

canonical reliability example of Exponentially distributed components for both i.i.d. and exchangeable systems. Then the full generality of the approach is established in the final examples where components are Phase-type distributed for both i.i.d. and exchangeable systems, also providing a highly pertinent link between this portion of the thesis and Chapter 4. Indeed, the exchangeable systems example provides a direct extension of the work in Chapter 4.

The broad applicability of the signature based MCMC scheme developed here means that the details of implementation — that is fulfilment of prerequisites (i)–(iv) of Algorithm 6.2 — for each of these examples also form contributions of this thesis.

### 6.3.1 IID systems with Exponential components

#### 6.3.1.1 Model set-up

The component lifetime distributions here are taken to be  $\text{Exponential}(\lambda)$ , so that  $\psi = (\lambda)$  and  $F_Y(y) = 1 - e^{-\lambda y}$ . Availing of the conjugacy result on page 36 of §3.1.2, a  $\text{Gamma}(\text{shape}=\nu, \text{scale}=1/\zeta)$  prior for  $\lambda$  is taken, so that one may immediately write down:

$$\lambda | \mathbf{y} \sim \text{Gamma}(\text{shape} = \nu + nm, \text{scale} = 1/(\zeta + nm\bar{y}))$$

where  $n$  is the order of the system/network and  $m$  is the number of masked system/network lifetimes observed. This satisfies prerequisites (ii) and (iii) of Algorithm 6.2.

Additionally, it is not difficult to derive:

$$F_{Y|Y<t}(y; \lambda) = \begin{cases} 0 & y < 0 \\ \frac{1 - e^{-\lambda y}}{1 - e^{-\lambda t}} & 0 < y < t \\ 1 & y > t \end{cases}$$

$$F_{Y|Y>t}(y; \lambda) = \begin{cases} 0 & y < t \\ 1 - e^{\lambda(y-t)} & y > t \end{cases}$$

so that one may generate random realisations of  $Y | Y < t$  as:

$$y = -\lambda^{-1} \log[U(e^{-\lambda t} - 1) + 1] \quad \text{where} \quad U \sim \text{Uniform}(0, 1)$$

and, random realisations of  $Y | Y > t$  as:

$$y = L + t \quad \text{where} \quad L \sim \text{Exponential}(\lambda)$$

by the memoryless nature of the Exponential. This satisfies prerequisites (iv) of Algorithm 6.2.

The set of signatures to be considered is the only thing remaining to be specified, together with a prior on the collection,  $F_{\mathcal{M}}(\cdot)$ . In all the examples to be presented here,  $F_{\mathcal{M}}(\cdot)$  is simply taken to be a discrete uniform over the collection of signatures in question.

### 6.3.1.2 Parametric inference only

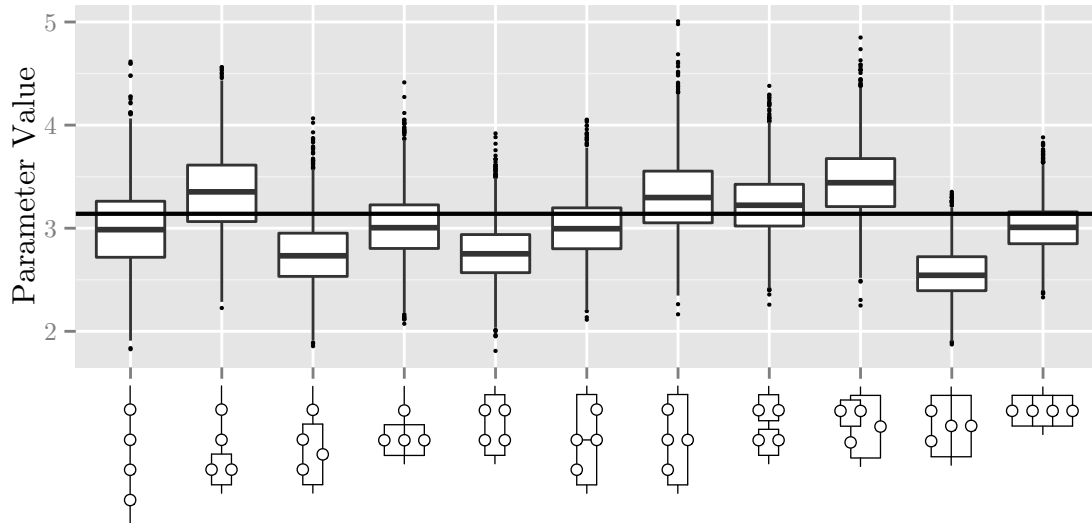
In the first instance, the structure of the system/network is assumed known and only the parameter inferred. A ground truth of  $\lambda = 3.14$  was chosen and  $50 \times 4$  Exponential random draws simulated to represent component failure times. For each of the 11 simply connected coherent systems of order 4, there were 50 order values,  $\{1, \dots, 4\}$ , drawn according to the signature of each system. Then, using the component failure times and order values, 50 system lifetimes were generated for each simply connected coherent system of order 4.

Finally, this was used as the data in 11 separate MCMC runs as outlined in the previous section, with prior parameters  $\nu = 9, \zeta = 2$ . The resultant posterior of  $\lambda$  is shown in Figure 6.3 for each of the systems. Autocorrelation in these simple examples was low so only 2,000 iterations were required to reduce the MCMC standard error enough to provide between one and two decimal places accuracy in the upper/lower 95% quantiles. In most instances, the results are satisfactory.

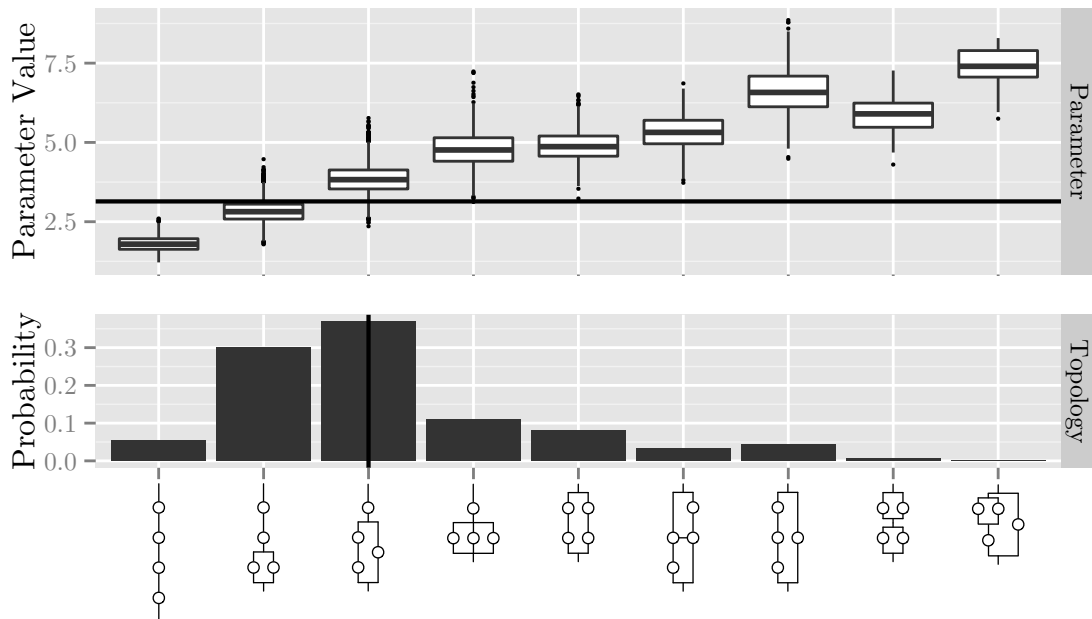
### 6.3.1.3 Parametric and topological inference

The Exponentially distributed i.i.d. component set-up was used again, this time the topology also being considered unknown. 50 masked system lifetimes were simulated from topology number 3 in Table 5.3, page 112, and the whole of that table was taken as the collection of candidate topologies,  $\mathcal{M}$ . A slightly more informative prior with  $\nu = 6, \zeta = 2$  was used and 10,000 MCMC iterations drawn, sufficient to provide over 2,000 iterations in each of the two highest a posteriori topologies.

Figure 6.4 shows the results. The histogram is the marginal posterior density of the topologies, whilst the box plots are the posterior of the rate parameter conditional on the given topology.



**Figure 6.3:** Boxplot for posterior simulations for each of 11 different simply connected coherent system from known ground truth,  $\lambda = 3.14$ . Solid horizontal line shows ground truth.



**Figure 6.4:** Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with  $\lambda = 3.14$ .  $\mathcal{M}$  consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed. Solid lines show ground truth.

This arrangement was explored for other ground truth topologies and the highest a posteriori topology was predominantly correct, though sometimes slipping to second or third highest a posteriori. Anecdotally, in the author’s tests it seemed that the closer to pure series or parallel the more accurate the inference appeared to be, with the systems in the middle of Table 5.3 being more frequently mixed up. This appears to be logical, since one would expect the differences between lifetimes of the systems with quite similar signatures will be smaller. However, for an order of magnitude more observations the posterior appeared to be asymptotically approaching the ground truth topologies.

The state of affairs becomes less clear as the size of the candidate set  $\mathcal{M}$  increases. Including all systems in Tables 5.1, 5.2, 5.3 and A.2 results in a candidate collection of 52 topologies. High relative weight of probability for the true topology was common, but the prevalence of the actual highest a posteriori topology being correct was greatly diminished. An example of this limitation can be seen when taking exactly the same simulated data with the larger collection of 52 topologies, the results being shown in Figure 6.5 — only the topologies with marginal posterior probability exceeding 0.04 are shown. Although the posterior is less satisfactory, it must be borne in mind that this is a large collection of topologies and there are only 50 masked lifetime observations, so the successful restriction to this subset of topologies is an encouraging outcome.

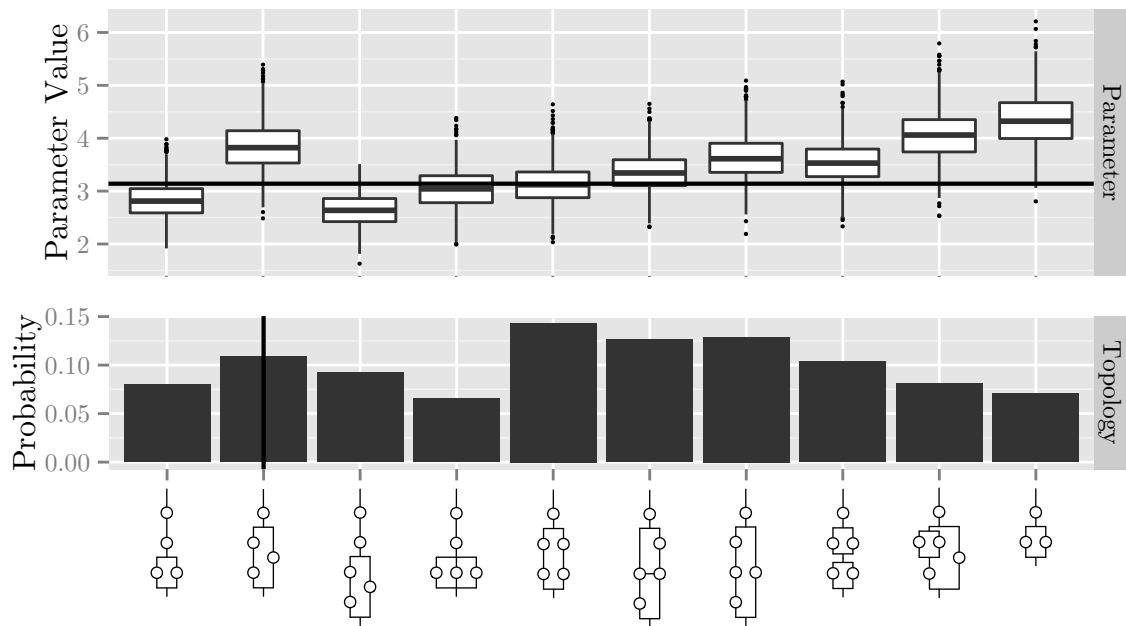
It is interesting to note that the Exponential distribution could arguably be among the harder distributions to consider in this situation, because as noted in the footnote on page 24 the minimum of a collection of Exponential random variables is Exponentially distributed. Consequently, it is likely to be harder to distinguish parts of the system which are in series from a single component.

## **6.3.2 Exchangeable systems with Exponential components**

### **6.3.2.1 Model set-up**

The setting is similar to the previous subsection. Components within systems are independent and identically Exponentially distributed. However, the weaker assumption of exchangeability is made at the system level: the failure rate of components may vary between systems according to a common population distribution. Hierarchically, this





**Figure 6.5:** Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with  $\lambda = 3.14$ .  $\mathcal{M}$  consisted of all 52 order 2, 3, 4 and 5 simply connected coherent systems. Topologies with posterior probability below 0.04 suppressed. Solid lines show ground truth.

is initially tackled as:

$$\begin{aligned}
Y | \lambda &\sim \text{Exponential}(\lambda) \\
\lambda | \nu, \zeta &\sim \text{Gamma}(\text{shape} = \nu, \text{scale} = \zeta) \\
\nu &\sim \text{Log-Normal}(\mu_\nu, \sigma_\nu) \\
\zeta &\sim \text{Log-Normal}(\mu_\zeta, \sigma_\zeta)
\end{aligned}$$

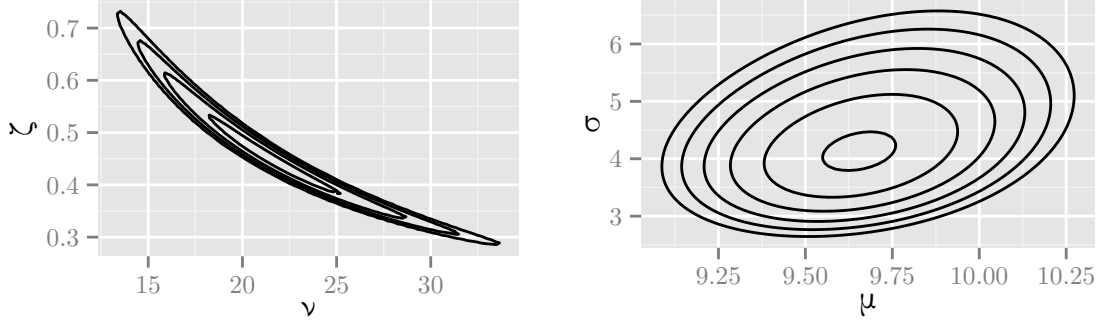
which fulfils prerequisite (ii) of the specification in the algorithm. Prerequisite (iii) is a little more involved. It is necessary to sample from:

$$f_{\Xi|Y}(\nu, \zeta, \lambda_1, \dots, \lambda_m | \mathbf{y}_1, \dots, \mathbf{y}_m)$$

This is achieved by sampling first from  $f_{\Theta|Y}(\nu, \zeta | \mathbf{y})$  and then from  $f_{\Psi|\Theta, Y}(\boldsymbol{\lambda} | \nu, \zeta, \mathbf{y})$ . The latter has a nice conjugate form for each conditionally independent component of  $\boldsymbol{\lambda}$ , but the former requires some work:

$$\begin{aligned}
f_{\Theta|Y}(\nu, \zeta | \mathbf{y}) &= \int \cdots \int f_{\Xi|Y}(\nu, \zeta, \boldsymbol{\lambda} | \mathbf{y}) d\lambda_1 \dots d\lambda_m \\
&\propto \int \cdots \int f_{\Xi, Y}(\nu, \zeta, \boldsymbol{\lambda}, \mathbf{y}) d\lambda_1 \dots d\lambda_m \\
&= \int \cdots \int f_{\Theta}(\nu, \zeta) f_{\Psi|\Theta}(\boldsymbol{\lambda} | \nu, \zeta) f_{Y|\Psi}(\mathbf{y} | \boldsymbol{\lambda}) d\lambda_1 \dots d\lambda_m \\
&= \prod_{i=1}^m \left[ \int f_{\Psi|\Theta, Y}(\lambda_i | \nu, \zeta, \mathbf{y}_{i \cdot}) d\lambda_i \right] f_{\Theta}(\nu, \zeta) \\
&\propto \prod_{i=1}^m \left[ \int (\Gamma(\nu)\zeta^\nu)^{-1} \lambda_i^{n+\nu-1} \exp \left\{ -\lambda_i \left( \zeta^{-1} + \sum_{j=1}^n y_{ij} \right) \right\} d\lambda_i \right] f_{\Theta}(\nu, \zeta) \\
&= \prod_{i=1}^m \left[ \frac{\Gamma(\nu+n)}{\zeta^\nu (\zeta^{-1} + \sum_{j=1}^n y_{ij})^{\nu+n} \Gamma(\nu)} \right] f_{\Theta}(\nu, \zeta) \\
\implies \log f_{\Theta|Y}(\nu, \zeta | \mathbf{y}) &\propto -m\nu \log \zeta + m \log \Gamma(\nu+n) - m \log \Gamma(\nu) - \frac{(\mu_\nu - \log \nu)^2}{2\sigma_\nu^2} \\
&\quad - \log \nu - \frac{(\mu_\zeta - \log \zeta)^2}{2\sigma_\zeta^2} - \log \zeta - (\nu+n) \sum_{i=1}^m \log \left( \zeta^{-1} + \sum_{j=1}^n y_{ij} \right)
\end{aligned}$$

It is now possible, in principle, to sample from  $\log f_{\Theta|Y}(\nu, \zeta | \mathbf{y})$ . However, the form above does not lend itself to easy sampling due to the shape the density may sometimes take. For example, Figure 6.6 (left plot) shows a diagnostic output with a distinctive banana shaped density. The following reparameterisation improves the sampling considerably in conjunction with a change to Log-Normal priors on the mean



**Figure 6.6:** Diagnostic plots of  $\log f_{\Theta|Y}(\nu, \zeta | \mathbf{y})$  before (left) and after (right) reparameterisation.

and variance (denoted by just  $\sigma$  for convenience) of the Gamma, which has the added advantage of being more intuitive to define than that on the shape and scale. Thus:

$$\begin{aligned} \mu &= \nu\zeta & \sigma &= \nu\zeta^2 \\ \implies \nu &= \frac{\mu^2}{\sigma} & \zeta &= \frac{\sigma}{\mu} \\ \implies |J| &= \left| \begin{array}{cc} \frac{\partial}{\partial \mu} \left( \frac{\mu^2}{\sigma} \right) & \frac{\partial}{\partial \sigma} \left( \frac{\mu^2}{\sigma} \right) \\ \frac{\partial}{\partial \mu} \left( \frac{\sigma}{\mu} \right) & \frac{\partial}{\partial \sigma} \left( \frac{\sigma}{\mu} \right) \end{array} \right| = \frac{1}{\sigma} \end{aligned}$$

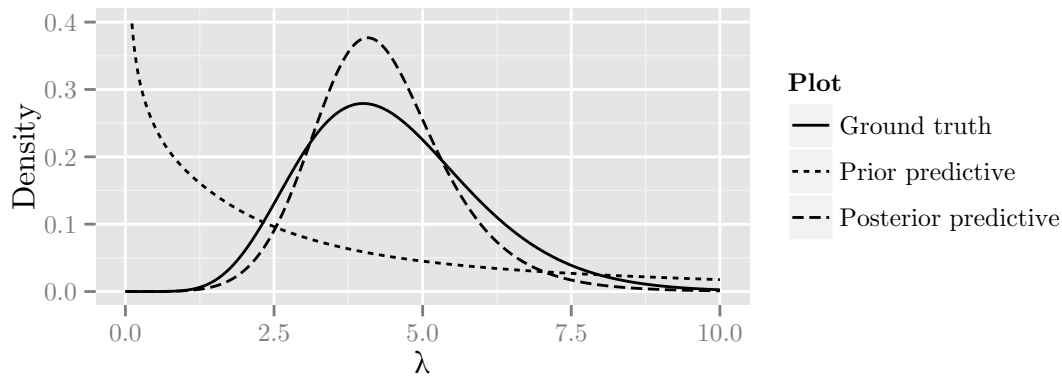
Consequently, taking hyperpriors on the mean and variance of the Gamma prior:

$$\begin{aligned} \mu &\sim \text{Log-Normal}(\mu_\mu, \sigma_\mu) \\ \sigma &\sim \text{Log-Normal}(\mu_\sigma, \sigma_\sigma) \end{aligned}$$

$$\begin{aligned} \implies \log f_{\Theta|Y}(\mu, \sigma | \mathbf{y}) &\propto -m\mu^2\sigma^{-1} \log(\mu^{-1}\sigma) + m \log \Gamma(\mu^2\sigma^{-1} + n) - m \log \Gamma(\mu^2\sigma^{-1}) \\ &\quad - \frac{(\mu_\mu - \log \mu)^2}{2\sigma_\mu^2} - \log \mu - \frac{(\mu_\sigma - \log \sigma)^2}{2\sigma_\sigma^2} - \log \sigma \\ &\quad - (\mu^2\sigma^{-1} + n) \sum_{i=1}^m \log \left( \mu\sigma^{-1} + \sum_{j=1}^n y_{ij} \right) - \log \sigma \end{aligned}$$

An importance sampling regime was tried but exhibited poor tail characteristics. A random-walk Metropolis-Hastings scheme (see page 46) with covariance matrix determined by the Hessian at the mode (numerically estimated) provided good results requiring few iterations. This being so, prerequisite (iii) of the algorithm is provided.

Finally, prerequisite (iv) will be exactly the same as the previous i.i.d. example, but with the rate varying for each system.



**Figure 6.7:** Ground truth, prior predictive and posterior predictive densities for the exchangeable population density of  $\lambda$ .

### 6.3.2.2 Parametric inference only

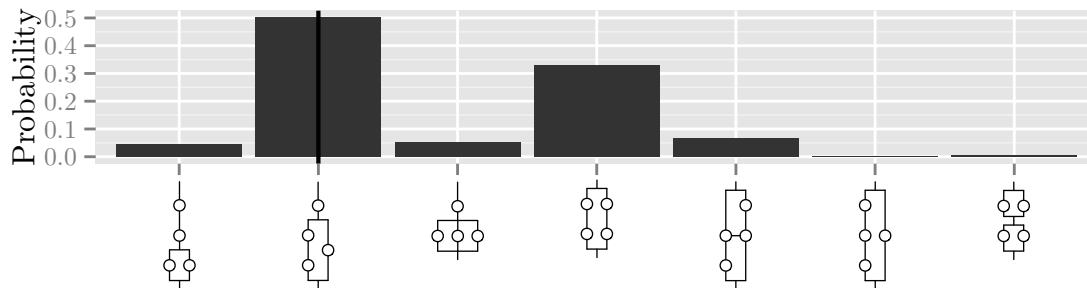
Initially the structure of the system/network is assumed known again. With this Exponential exchangeable example, a natural way of investigating how well the method performs for a known ground truth is to compare the posterior predictive distribution (as per (6.1)) of  $\lambda$  to the true exchangeable population distribution of  $\lambda$ .

A ground truth of  $\lambda \sim \text{Gamma}(\text{shape} = 9, \text{scale} = 0.5)$  was taken and 100 masked system lifetimes generated for order 4 simply connected coherent system number 3 (see Table 5.3). The hyper prior parameters were taken as  $\mu_\mu = 1, \sigma_\mu = 0.5$  (giving Log-Normal hyperprior with mean of  $\approx 3.5$  and variance  $\approx 2.7$ ) and  $\mu_\sigma = 1, \sigma_\sigma = 0.7$  (giving Log-Normal hyperprior with mean of  $\approx 3.9$  and variance  $\approx 7.6$ ).

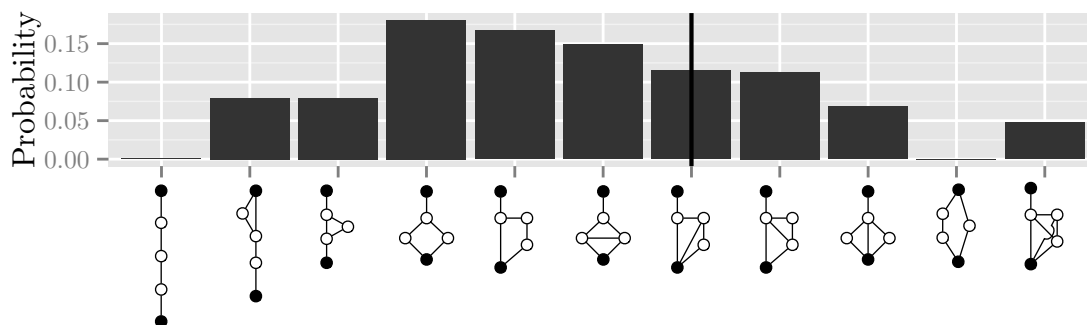
The result of an MCMC run of length 20,000 is shown in Figure 6.7. This shows the grievously incorrect prior predictive has been updated to achieve encouraging levels of agreement in the posterior predictive, although it is slightly under-dispersed in this instance. Given only 50 observations though, failure to capture tail behaviour is somewhat unremarkable.

### 6.3.2.3 Parametric and topological inference

Again the problem was extended to demonstrate inferring both topology and parameters. In both examples which follow, 100 observations were used and slightly more informative hyperpriors ( $\mu_\mu = 1.2, \sigma_\mu = 0.2, \mu_\sigma = 0.8, \sigma_\sigma = 0.2$ ). 20,000 MCMC iterations were performed.



**Figure 6.8:** Topology posterior for ground truth system 3.  $\mathcal{M}$  consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed.



**Figure 6.9:** Topology posterior for ground truth network 7.  $\mathcal{M}$  consisted of all 24 node order 3 coherent networks. Zero probability topologies suppressed.

For comparability with the i.i.d. case, the inference over the candidate collection of all order 4 simply connected coherent systems with ground truth of system 3 was performed. The results of this are in Figure 6.8. The posterior predictive density for  $\lambda$  conditional on system 3 was in even closer agreement than Figure 6.7 and is therefore not plotted separately.

By way of variety and to show the application to coherent networks, a second example with ground truth of coherent network number 7 from Table A.1 was used, with candidate collection of all node order 3 coherent networks. Mirroring the i.i.d. situation, the larger candidate collection clearly affected the accuracy of the inference, the results being visible in Figure 6.9. While identifiability is clearly an issue, there was success in reducing the plausible candidate set of networks considerably from the initial 24.

### 6.3.3 IID systems with Phase-type components

#### 6.3.3.1 Model set-up

The generality of the technique is realised when considering the components or nodes to have a Phase-type lifetime distribution. There are two ways to view such a setting: either as components being repairable systems themselves (up to initial failure), as per discussion in §2.2.2 and the developments of Chapter 4; or simply using Phase-types as a highly flexible lifetime distribution, since they are theoretically dense in the space of all distributions on  $[0, \infty)$  (Asmussen, 2000).

With the material presented thus far, the prerequisites of Algorithm 6.2 are easily addressed. Prerequisite (ii) sees the lifetime distribution being Phase-type with collection of parameters  $\lambda_1, \dots, \lambda_m$ , each having Gamma prior as set out fully in §4.2.1. The base methodology of Bladt *et al.* (2003) together with all of the advances of Chapter 4 provide the means of performing Bayesian inference called for by prerequisite (iii).

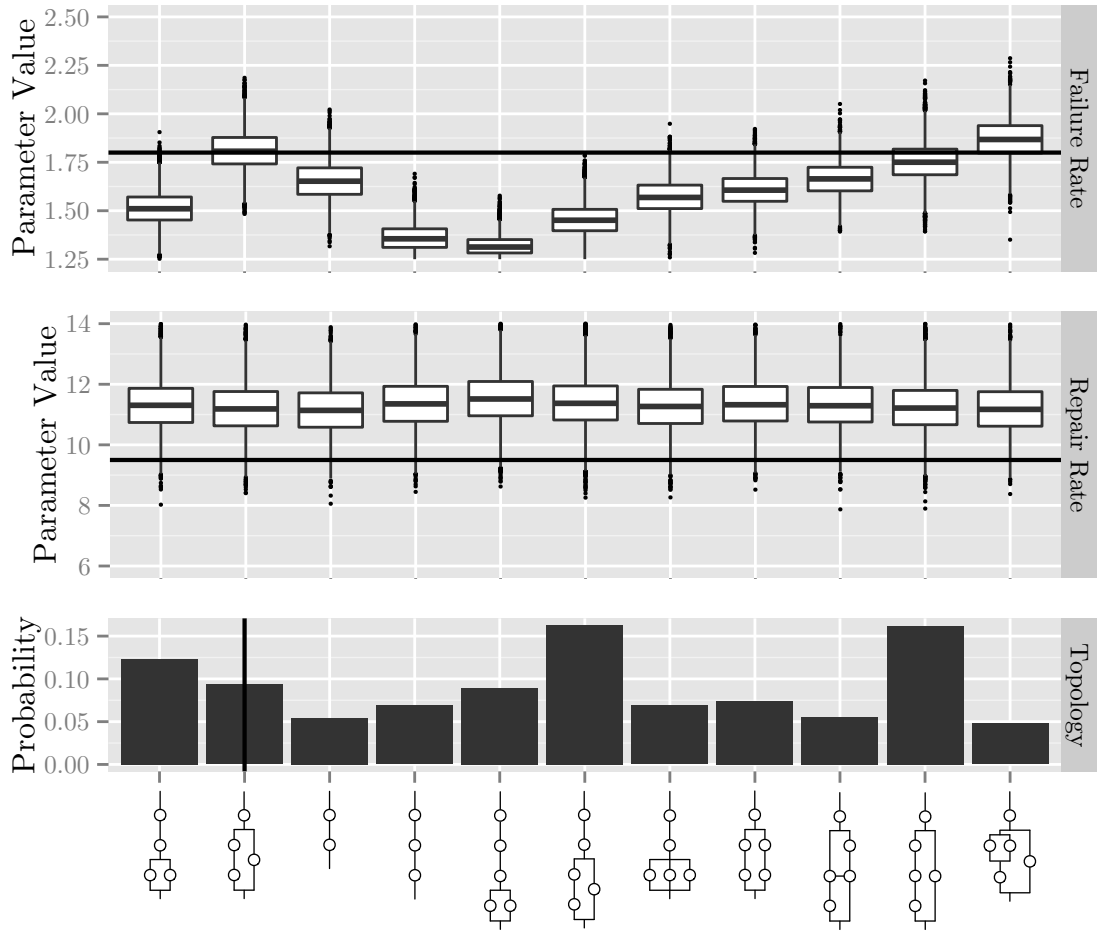
The sampling required by prerequisite (iv) can be handled easily. Algorithm 4.2, page 89, provides a computationally attractive means of sampling  $Y | Y > t$ , whilst due to the finite support a simple technique such as ARMS (Gilks *et al.*, 1995) can be used to sample  $Y | Y < t$ . If found necessary in some application not encountered here, an analogue of Algorithm 4.2 could be developed, but the performance was adequate in all the examples tried during development of this approach.

Having carefully shown the variety of results for the simple Exponential examples, the general case is handled straight away here.

#### 6.3.3.2 Parametric and topological inference for simply connected systems

Exactly the same prior specification was used as in §4.3.1, page 90. 100,000 MCMC iterations were run, with 75 MCMC iterations of the inner Phase-type MCMC (Chapter 4) on each iteration. Figure 6.10 shows the results of this.

It is particularly interesting to note that the difficulty in learning the repair rate results in posteriors which are essentially the same for this parameter across all topologies. The closest to ground truth for the failure rate is in the true system and, as may be expected, the highest a posteriori topology has decreased failure rate posterior offsetting the overestimated repair rate. Thus, although the true topology does well,



**Figure 6.10:** Marginal topology posterior and conditional parameter posteriors, for the ground truth topology being order 4 simply connected coherent system number 3 with  $\lambda_f = 1.8$  and  $\lambda_r = 9.5$ .  $\mathcal{M}$  consisted of all 52 order 2, 3, 4 and 5 simply connected coherent systems. Only topologies with posterior probability  $> 0.04$  shown. Solid lines show ground truth.

ranked 4th from 52 candidate systems, it appears that a sound prior knowledge of the repair rate is more important to accurately inferring the topology than it is to accurately inferring the failure rate, which is intuitively logical given the weak learning of repair rates discussed in Chapter 4.

### 6.3.4 Exchangeable systems with Phase-type components

#### 6.3.4.1 Model set-up

The fullest generality of the method presented is realised in the final example, where the independence of systems assumption is replaced by exchangeability.

There is some work to be done before implementation is possible, although a great deal of the work of Chapter 4 applies. The ideas from the exchangeable system Exponential component lifetime example are useful. Given that the elements of the generator  $\mathbf{G}$  are all 0 or  $\lambda_i$  for some  $i$  (as in §4.2.1) then hierarchically this problem is initially tackled as:

$$\begin{aligned}
 Y \mid \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_l) &\sim \text{PHT}(\boldsymbol{\pi}, \mathbf{G}) \\
 \lambda_i \mid \nu_i, \zeta_i &\sim \text{Gamma}(\text{shape} = \nu_i, \text{scale} = 1/\zeta_i) \\
 \nu_i &\sim \text{Log-Normal}(\mu_{\nu_i}, \sigma_{\nu_i}) \\
 \zeta_i &\sim \text{Log-Normal}(\mu_{\zeta_i}, \sigma_{\zeta_i})
 \end{aligned}$$

That is, if using the Phase-type to model a repairable subsystem then  $\boldsymbol{\lambda} = (\lambda_f, \lambda_r)$  and there are two Gamma priors, with four Log-Normal hyper-priors. However, the development here is completely general in case one merely wishes to exploit the flexibility of Phase-type distributions, rather than impose a repairable subsystem interpretation. This fulfils prerequisite (ii) of the specification in the algorithm.

For prerequisite (iii), one must sample from the posterior:

$$f_{\Xi \mid Y}(\boldsymbol{\nu}, \boldsymbol{\zeta}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m \mid \mathbf{y}_1, \dots, \mathbf{y}_m)$$

where  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_l)$ ,  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_l)$  and  $\boldsymbol{\lambda}_i$  is the vector of  $l$  rate parameters for system  $i$ .

This may be tackled using much of the machinery of Chapter 4, though some simple changes are required. First, note that  $y_{ij}$  is the failure time of ‘component’  $j$  in system



$i$ , but where each ‘component’ is now itself a repairable subsystem (or general Phase-type). Therefore, for clarity, hereinafter these are called nodes. Thus, let  $\phi_{ij}$  be the latent processes of node  $j$  in system  $i$ . This means that the data augmentation scheme of Chapter 4 is adapted for sampling the natural completion:

$$f_{\Xi, \Phi | Y}(\boldsymbol{\nu}, \boldsymbol{\zeta}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m, \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_m | \mathbf{y}_1, \dots, \mathbf{y}_m)$$

This involves using Exact Conditional Sampling (Algorithm 4.1) entirely unchanged to sample:

$$f_{\Phi | \Psi, Y}(\phi_{ij} | \boldsymbol{\lambda}_i, y_{ij})$$

for each  $i, j$ . Then, some additional work is required to sample the parameters in the data augmentation scheme:

$$f_{\Xi | \Phi}(\boldsymbol{\nu}, \boldsymbol{\zeta}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m | \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_m)$$

A similar strategy to that taken for sampling in the exchangeable Exponential case can be taken, where the exchangeable distribution parameters are sampled marginally first, followed by the model parameters conditionally. The detail differs subtly, so must be re-derived carefully as follows.

Recall that the latent process enters the likelihood only via the sufficient statistics of the chain:  $\mathbf{N}$  and  $\mathbf{z}$ . As shown in §4.2.1, these reduce to  $N_k^*$  and  $z_k^*$  for each parameter  $\lambda_k$ . With differing parameters due to system exchangeability, this is notationally extended to  $N_{ik}^*$  and  $z_{ik}^*$  for the sufficient statistics of the latent processes of the nodes in system  $i$  for parameter  $\lambda_{ik}$ . Armed with this notation:

$$\begin{aligned} & f_{\Xi | \Phi}(\boldsymbol{\nu}, \boldsymbol{\zeta}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m | \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_m) \\ & \propto f_{\Phi | \Psi}(\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_m | \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m) f_{\Psi | \Theta}(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m | \boldsymbol{\nu}, \boldsymbol{\zeta}) f_{\Theta}(\boldsymbol{\nu}, \boldsymbol{\zeta}) \\ & \propto \prod_{i=1}^m \left[ \prod_{k=1}^l \lambda_{ik}^{N_{ik}^*} \exp(-\lambda_{ik} z_{ik}^*) \right] \prod_{i=1}^m \left[ \prod_{k=1}^l \frac{\zeta_k^{\nu_k} \lambda_{ik}^{\nu_k - 1}}{\Gamma(\nu_k)} \exp(-\zeta_k \lambda_{ik}) \right] f_{\Theta}(\boldsymbol{\nu}, \boldsymbol{\zeta}) \\ & \propto \prod_{i=1}^m \left[ \prod_{k=1}^l \frac{\zeta_k^{\nu_k} \lambda_{ik}^{N_{ik}^* + \nu_k - 1}}{\Gamma(\nu_k)} \exp(-(z_{ik}^* + \zeta_k) \lambda_{ik}) \right] f_{\Theta}(\boldsymbol{\nu}, \boldsymbol{\zeta}) \end{aligned}$$

Thus, marginally:

$$\begin{aligned}
f_{\Theta|\Phi}(\boldsymbol{\nu}, \boldsymbol{\zeta} | \boldsymbol{\phi}) &= \int \cdots \int f_{\Xi|\Phi}(\boldsymbol{\nu}, \boldsymbol{\zeta}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m | \boldsymbol{\phi}) d\lambda_{11} \cdots d\lambda_{ml} \\
&\propto \prod_{k=1}^l \left[ f_{\Theta}(\nu_k, \zeta_k) \prod_{i=1}^m \int \frac{\zeta_k^{\nu_k} \lambda_{ik}^{N_{ik}^* + \nu_k - 1}}{\Gamma(\nu_k)} \exp(-(z_{ik}^* + \zeta_k) \lambda_{ik}) d\lambda_{ik} \right] \\
&= \prod_{k=1}^l \left[ f_{\Theta}(\nu_k, \zeta_k) \prod_{i=1}^m \frac{\zeta_k^{\nu_k} \Gamma(N_{ik}^* + \nu_k)}{\Gamma(\nu_k) (z_{ik}^* + \zeta_k)^{N_{ik}^* + \nu_k}} \right]
\end{aligned}$$

Implying that each pair  $(\nu_k, \zeta_k)$  has marginal log posterior:

$$\begin{aligned}
\log f_{\Theta|\Phi}(\nu_k, \zeta_k | \boldsymbol{\phi}) &\propto m\nu_k \log \zeta_k - m \log \Gamma(\nu_k) \\
&\quad + \sum_{i=1}^m [\log \Gamma(N_{ik}^* + \nu_k) - (N_{ik}^* + \nu_k) \log(z_{ik}^* + \zeta_k)] \\
&\quad - \frac{(\mu_{\nu_k} - \log \nu_k)^2}{2\sigma_{\nu_k}^2} - \log \nu_k - \frac{(\mu_{\zeta_k} - \log \zeta_k)^2}{2\sigma_{\zeta_k}^2} - \log \zeta_k
\end{aligned}$$

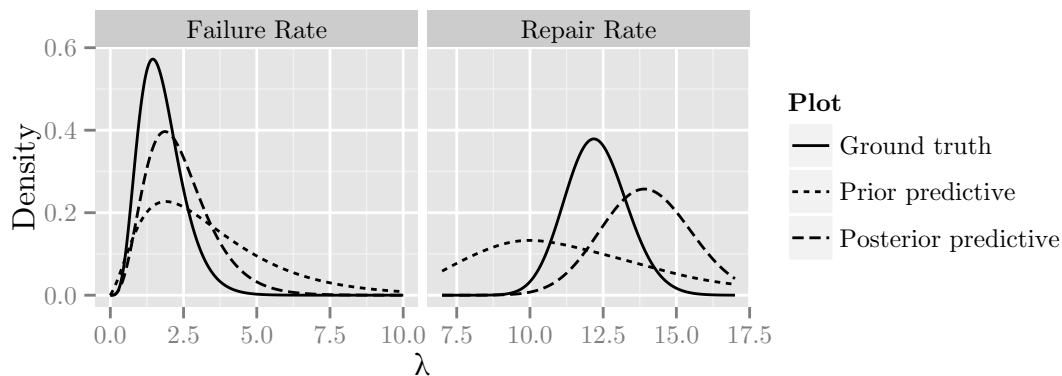
However, in practice it again seemed prudent to transform to the mean and variance, as per the previous exchangeable example:

$$\begin{aligned}
\mu_k &= \frac{\nu_k}{\zeta_k} & \sigma_k &= \frac{\nu_k}{\zeta_k^2} \\
\implies \nu_k &= \frac{\mu_k^2}{\sigma_k} & \zeta_k &= \frac{\mu_k}{\sigma_k} \\
\implies |J| &= \begin{vmatrix} \frac{\partial}{\partial \mu_k} \left( \frac{\mu_k^2}{\sigma_k} \right) & \frac{\partial}{\partial \sigma_k} \left( \frac{\mu_k^2}{\sigma_k} \right) \\ \frac{\partial}{\partial \mu_k} \left( \frac{\mu_k}{\sigma_k} \right) & \frac{\partial}{\partial \sigma_k} \left( \frac{\mu_k}{\sigma_k} \right) \end{vmatrix} = \left| -\frac{\mu_k^2}{\sigma_k^3} \right| = \frac{\mu_k^2}{\sigma_k^3}
\end{aligned}$$

Consequently, taking hyperpriors on the mean and variance (denoted by  $\sigma_k$  for convenience) of the Gamma prior:

$$\begin{aligned}
\mu_k &\sim \text{Log-Normal}(\mu_{\mu_k}, \sigma_{\mu_k}) \\
\sigma_k &\sim \text{Log-Normal}(\mu_{\sigma_k}, \sigma_{\sigma_k})
\end{aligned}$$

$$\begin{aligned}
\implies \log f_{\Theta|\Phi}(\mu_k, \sigma_k | \boldsymbol{\phi}) &\propto m\mu_k^2 \sigma_k^{-1} \log(\mu_k \sigma_k^{-1}) - m \log \Gamma(\mu_k^2 \sigma_k^{-1}) \\
&\quad + \sum_{i=1}^m [\log \Gamma(N_{ik}^* + \mu_k^2 \sigma_k^{-1}) - (N_{ik}^* + \mu_k^2 \sigma_k^{-1}) \log(z_{ik}^* + \mu_k \sigma_k^{-1})] \\
&\quad - \frac{(\mu_{\mu_k} - \log \mu_k)^2}{2\sigma_{\mu_k}^2} - \log \mu_k - \frac{(\mu_{\sigma_k} - \log \sigma_k)^2}{2\sigma_{\sigma_k}^2} - \log \sigma_k \\
&\quad + 2 \log \mu_k - 3 \log \sigma_k
\end{aligned}$$



**Figure 6.11:** Ground truth, prior predictive and posterior predictive densities for the exchangeable population density of  $\lambda_f$  and  $\lambda_r$ .

Sampling is again by a random-walk Metropolis-Hastings scheme (see page 46) with covariance matrix determined by the Hessian at the mode (numerically estimated).

Consequently, the bulk of Chapter 4 still applies, the difference being that instead of just simple draws from a conjugate Gamma distribution on each iteration, there is the additional step of sampling  $\nu$  and  $\zeta$  first. This provides prerequisite (iii) of the algorithm.

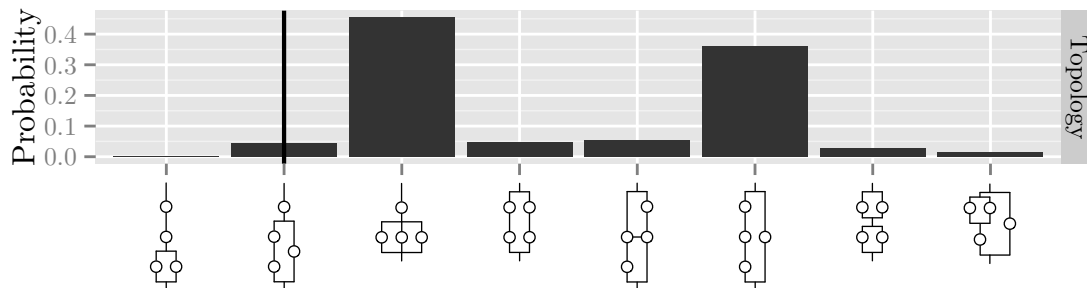
Finally, prerequisite (iv) will be exactly the same as the i.i.d. Phase-type example, but with the generator varying for each system.

### 6.3.4.2 Parametric inference

100 system lifetime observations were simulated from order 4 simply connected coherent system number 3, with a ground truth of  $\lambda_f \sim \text{Gamma}(\text{shape} = 5.5, \text{rate} = 3.1)$  and  $\lambda_r \sim \text{Gamma}(\text{shape} = 135, \text{rate} = 11)$ , with exchangeability at the system level as described above. The hyper prior parameters were taken as follows:

$$\begin{array}{ll}
 \mu_{\mu_f} = 1.1 & \sigma_{\mu_f} = 0.2 \\
 \mu_{\sigma_f} = 0.4 & \sigma_{\sigma_f} = 0.2 \\
 \mu_{\mu_r} = 2.4 & \sigma_{\mu_r} = 0.1 \\
 \mu_{\sigma_r} = 0.1 & \sigma_{\sigma_r} = 0.1
 \end{array}$$

The result of an MCMC run of length 10,000 is shown in Figure 6.11. It is singularly interesting that the repair rate has been learned in some fashion in this exchangeable case, unlike in the i.i.d. setting. The slight right bias of the repair posterior predictive



**Figure 6.12:** Topology posterior for ground truth system 3.  $\mathcal{M}$  consisted of all 11 order 4 simply connected coherent systems. Low or zero probability topologies suppressed.

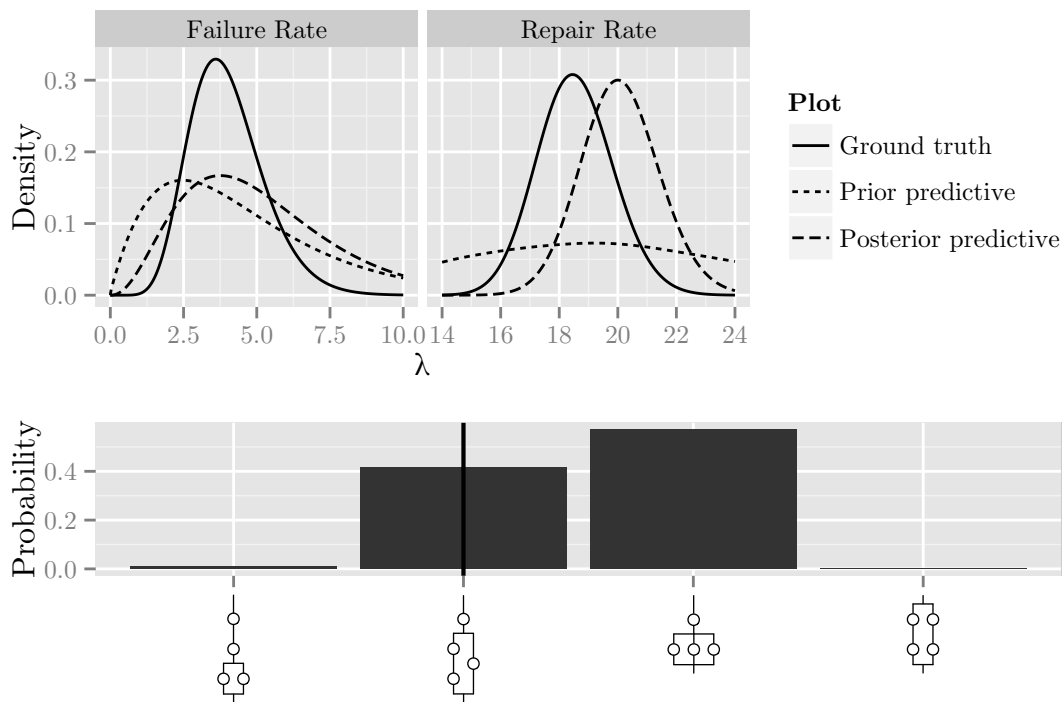
can be explained by the presence of the excess tail weight to the right of the failure rate posterior predictive.

Recall that there are two layers of unobserved data here: the individual Phase-type node lifetimes, and the component failure and repair process within each of the Phase-type nodes. That is, there are 800 components undergoing an unobserved process of failure and repair, leading to 400 dual redundant node failure times which are also unobserved, resulting in 100 system failure times which are observed (but where the node causing failure is unknown). In light of this it seems reasonable to be greatly encouraged at how much learning has been possible. Regrettably it is not all good news as will now be seen.

### 6.3.4.3 Topological inference

When the topological inference is added, taking exactly the same specification as above results in less stellar posterior learning. It would seem at first that there is simply too little information once system design is removed on top of node lifetime and component failure and repair schedule. Figure 6.12 shows the posterior of the topology observed in this instance.

However, in the variety of examples run while testing the method it was intriguing to note that when the prior variance on the exchangeable failure rate distribution was increased substantially, the detection of topology tended to become as good as in the simple Exponential cases, even though the parametric inference was notably weaker in those situations. Figure 6.13 shows this behaviour. Variance parameters in hierarchical



**Figure 6.13:** Example parameter and topology posterior for high variance prior.  $\mathcal{M}$  consisted of all 11 order 4 simply connected coherent systems. Zero probability topologies suppressed.

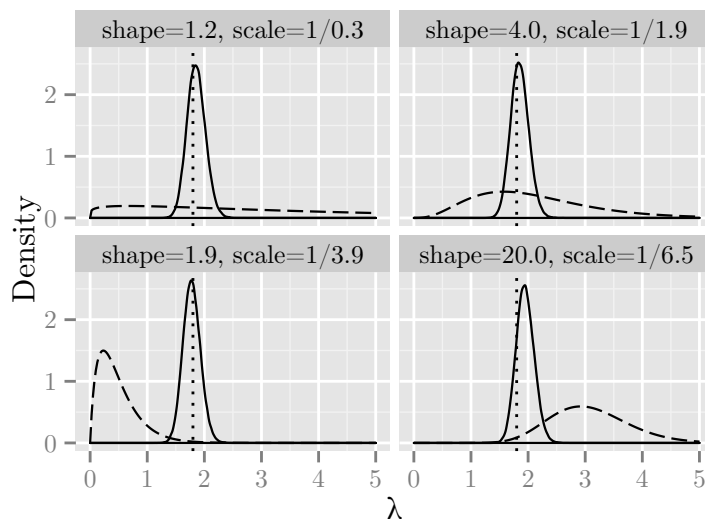
models are known to sometimes cause interesting behaviour (e.g. Gelman, 2006) and a detailed study of the behaviour in this model would be interesting future work.

### 6.3.5 Sensitivity to the prior

It is again important to give some consideration to the sensitivity posterior results may have to prior specification.

The i.i.d. Exponential case is taken as an example, with 100 failure times simulated for topology number 3 in Table 5.3, page 112, with ground truth  $\lambda = 1.8$ . Four priors were chosen for the failure rate (matching the choices in §4.4): two relatively vague and two which are stronger but incorrect (one overestimating and the other underestimating the rate). The prior and kernel density estimates of the posteriors are depicted in Figure 6.14. This shows that the results are very robust to prior specification.

The same analysis was run with the same priors, but where the topology was jointly inferred. The results of the topology posterior for each prior are shown in Figure 6.15. Here, the posterior is certainly affected by prior specification but is not too dramatically



**Figure 6.14:** Marginal failure rate posteriors in solid line for different priors in dashed line. Prior parameters shown in the dark grey boxes, with failure rate prior densities depicted by dashed line. Ground truth is dotted vertical line.

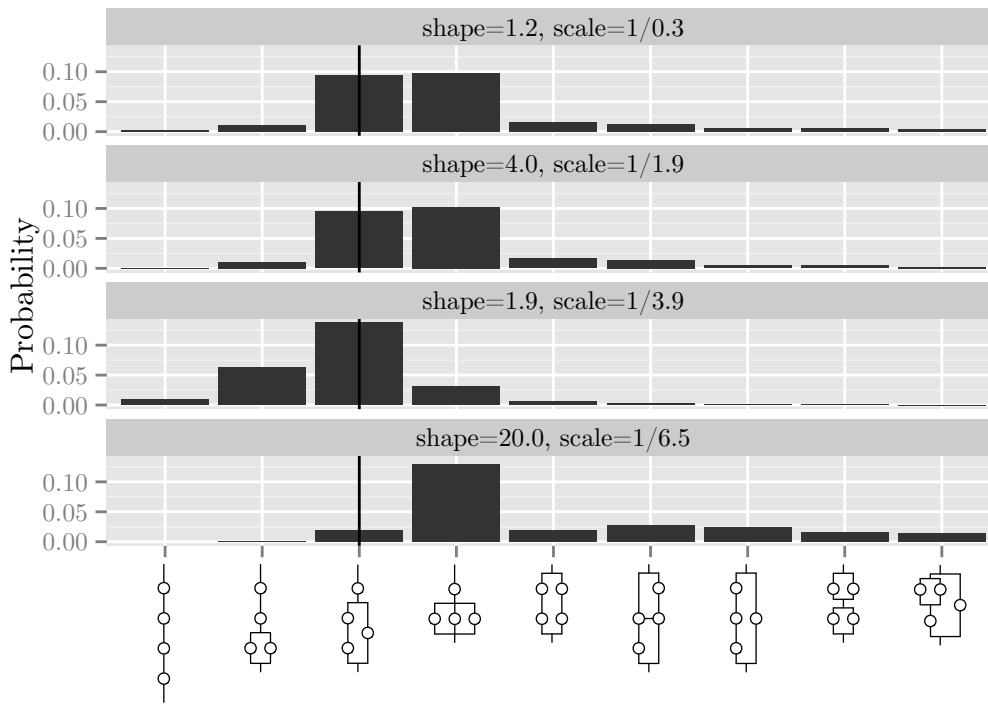
changed, the most noticeable discrepancy occurring for the more informative prior which overestimates the failure rate.

The exchangeable Exponential example can be seen to also be robust to prior misspecification from Figure 6.7. The discussion from §4.4 applies to the i.i.d. Phase-type example, since this is the key full conditional sampling component in the MCMC algorithm. Finally, as mentioned in the section on exchangeable Phase-type lifetime distributions, there was some unusual behaviour pertaining to the prior and this is an avenue which requires further active research.

### 6.3.6 Summary

Examples of parametric inference in isolation, and parametric and topological inference together have been presented for both the simple Exponential lifetime distribution and the flexible Phase-type lifetime distribution in i.i.d. and exchangeable system settings.

The parametric inference has been relatively uniformly effective: it is striking to note that even in those cases where the topology was hard to identify, the marginal parameter posteriors for those topologies with high posterior probability were in all cases concordant with the ground truth parameter values.



**Figure 6.15:** Marginal topology posteriors for different failure rate priors. Prior parameters shown in the dark grey boxes. Ground truth is solid vertical line.

Successful highest a posteriori identification of the topology has been more mixed, showing the difficulty of this as a learning problem. However, in all except the final example the candidate set of topologies has seen dramatic reduction to sets which include the ground truth. Anecdotally, from the variety of examples run during testing this methodology, three things seemed to have the biggest impact on topological inference: large candidate sets, varying numbers of components and extremely weak information such as in the Phase-type case.

In addition, the development of §6.3.4 is important in its own right, extending the ideas of Chapter 4 from an i.i.d. to an exchangeable setting.

# Chapter 7

## Statistical Computing

As alluded to near the start of §3.2, the shift away from entirely closed form solutions toward algorithms has led to an increase in the importance of statistical computing. To an increasing extent the following exhortation is being heeded:

*“It seems remarkable to me how little effort statisticians put into getting their favourite statistical methods used [...] One solution is to expect a reference implementation, some code that is warranted to give the authors’ intended answers in a moderately sized problem.”*

— Ripley (2005)

A step further than this is to provide a general purpose reference implementation which works for arbitrary non-pathological and moderately sized problems. Indeed, since the above quote, publication of general purpose R (R Core Team, 2012) packages alongside original theory has become more common and, particularly for complex algorithms, forms an important contribution to the literature by enabling ready usage of novel methodology.

To this end, two R packages have been produced to provide an easily accessible interface so as to enable others to use the main contributions of this thesis in their own problems. This very short chapter serves as a brief demonstration of these packages since they form major contributions.



## 7.1 ‘PhaseType’ package

The PhaseType package (Aslett, 2011) provides highly optimised implementations of the methodology of Bladt *et al.* (2003) and the extended methodology of Chapter 4. The actual implementation consists of some 3,300 lines of C code, with convenient interfaces for easy use from within R. A lot of emphasis was placed on a small memory footprint, minimal memory creation or copying and on interfacing to high performance LAPACK routines for all linear algebra operations. The memory usage and access patterns are such that there are effectively zero level 2 cache misses when running on an Intel Core 2 Duo 2.4GHz CPU with 4MB L2 cache for modest size generator matrices.

### 7.1.1 Methodology of Bladt *et al.* (2003)

The primary function for performing inference using the original methodology of Bladt *et al.* (2003) is called `phtMCMC`. The function prototype is:

```
phtMCMC(x, states, beta, nu, zeta, n, mhit=1, resume=NULL,  
        silent=FALSE)
```

In brief, the main arguments are:

`x` is the vector of absorption times, `y`.

`states` is an integer specifying the dimension of the non-absorbing part of the Phase-type generator matrix, `G`, to be fitted.

`beta` is the  $n$ -dimensional vector parameter for the Dirichlet prior, that is  $\beta$ .

`nu` is the vector representing the Gamma prior shape parameters on each generator element, filled columnwise,  $\nu_{ij}$ .

`zeta` is the vector representing the Gamma prior shape parameters on each generator row,  $\zeta_i$ .

`n` is the integer specifying the number of MCMC iterations to perform.

`mhit` is the integer specifying the number of inner Metropolis-Hastings MCMC iterations to perform when sampling the latent path (determined, perhaps, using arguments like those in Theorem 4.1).

More detail is available in the package help files.

Thus, one may easily repeat the first example in Bladt *et al.* (2003) where an order 4 Phase-type distribution (4-dimensional non-absorbing sub-generator) is fitted to 250 simulations from a Log-Normal distribution. If the 250 simulations are already stored in the variable `y`, then one simply executes the following (note that no information is given in the original paper about the prior parameters used so those are concocted here):

```
> library(PhaseType)
> # Prior on starting state
> beta <- c(1, 0, 0, 0)
> # Gamma prior shape (one per matrix element, columnwise)
> nu <- rep(3, 4*4)
> # Gamma prior reciprocal scale (one per matrix row)
> zeta <- rep(0.5, 4)
>
> # Execute 10,000 MCMC iterations
> res <- phtMCMC(y, 4, beta, nu, zeta, 1000, mhit=200)
...
> res$samples[1:2,]
      S12      S13      S14      S21      S23      S24      S31
[1,] 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000
[2,] 7.573901 6.57875 3.388483 3.678352 5.327703 5.098657 2.275127
      S32      S34      S41      S42      S43      s1      s2
[1,] 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000
[2,] 3.023175 4.824938 2.93782 2.657661 7.389131 1.789335 2.102437
      s3      s4
[1,] 4.000000 4.000000
[2,] 2.409501 2.20482
```

Note that where a mode of the Gamma prior exists (when  $\nu_{ij} > 1$ ), this is taken as the starting value and otherwise a random draw from the prior is made.

The output `S23`, for example, refers to the sub-generator matrix entry  $S_{23}$  and `s2`, for example, refers to the vector of exit rates entry  $s_2$ , as per (2.11), page 29.

The next subsection is of more interest here, covering the advances in this thesis.

## 7.1.2 Methodology of Chapter 4

The extended methodology of Chapter 4 is provided in `phtMCMC2`. The function prototype is:

```
phtMCMC2(x, T, beta, nu, zeta, n, censored = rep(FALSE, length(x)),
         C=matrix(1.0, nrow=dim(T)[1], ncol=dim(T)[2]), method = "ECS",
         mhit = 1, resume = NULL, silent = FALSE)
```

Full details are in the help files, but interest here is in the following arguments which are described along with their mathematical representation as per Chapter 4:

`x` is the vector of absorption times,  $\mathbf{y}$ .

`T` is an  $(n + 1) \times (n + 1)$  character matrix indicating the structure to impose on the model: that is, specifying the variable position structure for the generator  $\mathbf{G}$ , as for example in (2.10), page 29.

`beta` is the  $n$ -dimensional vector parameter for the Dirichlet prior,  $\beta$ .

`nu` is the pairlist vector representing the Gamma prior shape parameters,  $\nu_i$ .

`zeta` is the pairlist vector representing the Gamma prior reciprocal scale parameters,  $\zeta_i$ .

`n` is the integer specifying the number of MCMC iterations to perform.

Thus, to conduct the analysis of §4.3.1, page 90, for lifetime data already stored in the variable `y`:

```
> library(PhaseType)
> # Define the structure of the Phase-type generator
> T <- matrix(c(0, "R", "R", 0, "F", 0, 0, 0, "F", 0, 0, 0, 0, "F", "F", 0), 4)
```

```

> T
      [,1] [,2] [,3] [,4]
[1,] "0"  "F"  "F"  "0"
[2,] "R"  "0"  "0"  "F"
[3,] "R"  "0"  "0"  "F"
[4,] "0"  "0"  "0"  "0"
> # Prior on starting state
> beta <- c(1, 0, 0)
> # Gamma prior shape (one per model parameter)
> nu <- c("R"=180, "F"=24)
> # Gamma prior reciprocal scale (one per model parameter)
> zeta <- c("R"=16, "F"=16)
>
> # Execute 10,000 MCMC iterations
> res <- phtMCMC2(y, T, beta, nu, zeta, 10000)
...
> res$samples[1:5,]
              F              R
[1,] 1.437500 11.187500
[2,] 1.600102 10.502070
[3,] 1.699556  9.967606
[4,] 1.756052  9.751569
[5,] 1.789740  9.614427

```

Hence, the full machinery of Chapter 4 is available by simply defining the generator structure via a simple character matrix and the prior parameters. The example above executes at a speed of over 1,200 MCMC iterations per second on a 2007 laptop computer, so that even in highly correlated runs it is not unduly time consuming to draw sufficient samples to reduce the standard error.

The package is available with full documentation on the official Comprehensive R Archive Network (CRAN) website:

<http://cran.r-project.org/package=PhaseType>

## 7.2 ‘ReliabilityTheory’ package

The ReliabilityTheory package (Aslett, 2012) provides highly flexible implementations of the methodology of Chapters 5 and 6, with specific routines for different component lifetime distributions being developed.

### 7.2.1 Systems and networks

All simply connected coherent systems of order 2, 3, 4 and 5 are available in the loadable data sets `sccs02`, `sccs03`, `sccs04` and `sccs05` respectively, which are lists ordered by expected lifetimes when assuming i.i.d. Exponentially distributed component lifetimes (see §5.4). Each list element contains: an `igraph` (Csárdi and Nepusz, 2006) object representing the topology (in `$graph`); the signature vector (in `$signature`); and the collection of minimal cut-sets (in `$cutsets`).

Likewise, all coherent networks of order 2 and 3 are available in the loadable data sets `cn02` and `cn03`.

The implementations of Algorithms 5.1 and 5.2 used to generate these lists are included within the two private namespace functions `coherentSystemsOfOrder` and `coherentNetworksOfOrder` respectively. They are not publicly visible because it is not intended that they be routinely run directly (since the results are stored in the named loadable variables above)<sup>1</sup>.

Once the ReliabilityTheory package is loaded, simply running `data(sccs04)`, for example, makes graphs, minimal cut-set collections and signatures of all simply connected coherent systems of order 4 available in the list variable `sccs04`.

### 7.2.2 Making PhaseType easier for reliability problems

The ReliabilityTheory package includes a function which makes the PhaseType package easier to use in reliability problems, the main obstacle to making use of the advances of Chapter 4 via `phtMCMC2()` being specification of the generator matrix. Thus an important contribution of the ReliabilityTheory package is the automatic generation of these by using a simple formula to specify the system structure via an `igraph` object.

---

<sup>1</sup>Additionally, a bug was discovered in the coloured graph isomorphism and  $(s, t)$  vertex separator code of `igraph` v0.6.1, so one must currently manually compile the nightly version of `igraph` to get the correct results until the fix is released to CRAN.

Thus, the generator in (4.12), page 96, can be found by simply running:

```
> library(ReliabilityTheory)
> graph <- graph.formula(s -- 1 -- 2 -- t, s -- 3 -- 4 -- t,
                        1:2 -- 5 -- 3:4)
> G <- systemGraphToGenerator(graph, 1, 365)
> G
```

The double dashes, --, indicate undirected edges and the specification must include perfectly reliable dummy start and terminal nodes, named `s` and `t`. Thus:

- `s -- 1 -- 2 -- t` specifies system number 1, Table 5.1;
- `s -- 3 -- 4 -- t` adds another such system in series, resulting in system number 5, Table 5.3;
- `1:2 -- 5 -- 3:4` finalises the bridge system by inserting the bridging component 5 connected to 1 and 2 on one side and 3 and 4 on the other, resulting in system number 18, Table A.2 — or the 5 component bridge system numbered in the required form of Figure 2.3.

After creating the graph and then running `systemGraphToGenerator()`, `G` contains a list with both a numeric generator matrix (in `$G` with the failure rate,  $\lambda_f = 1$ , and repair rate,  $\lambda_r = 365$ ) and a symbolic matrix (in `$structure$G`), along with a matrix of the constants  $c_{ij}$  (in `$structure$C`, as per §4.2.1, page 76).

If the specific component numbering is not of interest then of course the data described in the previous subsection saves even more work. The bridge system is system number 18 in Table A.2, so an isomorphic result is obtained with:

```
> library(ReliabilityTheory)
> data(sccs05)
> G <- systemGraphToGenerator(sccs05[[18]]$graph, 1, 365)
> G
```

The variables `G$structure$G` and `G$structure$C` are in precisely the form required by `phtMCMC2()` for the arguments `T` and `C` respectively. Thus, after setting prior parameters the inference of §4.3.3 can be performed by:

```
> res <- phtMCMC2(y, G$structure$G, beta, nu, zeta, 100000,
                  C=G$structure$C)
```

### 7.2.3 Parametric and topological inference

The inferential approach developed in Chapter 6 is quite general, being appropriate for a wide range of component lifetime distributions. Consequently, in raw form there is an onus on the user of the technique to provide more of the implementation. To mitigate this, there are two strands to the inferential part of the `ReliabilityTheory` package: at the lowest level, only the general signature based data augmentation scheme is implemented with the user supplying the necessary additional infrastructure; at a higher level, there are convenient and simpler interfaces for specific lifetime distributions which hide the complexity of implementation.

#### 7.2.3.1 Simple interface

Wrappers around the low level code have been provided which implement all of the prerequisites of Algorithm 6.2 for the settings in the previous chapter: Exponential and Phase-type component lifetimes with i.i.d. or exchangeable systems assumptions.

For brevity just the simple example of i.i.d. systems with Exponential component lifetimes is discussed. The function prototype is:

```
maskedInferenceIIDExponential(t, signature, iter, priorShape,
                              priorScale)
```

These arguments may be understood as follows:

`t` a vector of masked system lifetimes.

`signature` a single signature vector (for parametric inference only) or a list of signature vectors (to jointly infer topology).

`iter` the number of MCMC iterations to perform.

`priorShape` the shape of the Gamma prior on the Exponential rate.

`priorScale` the scale of the Gamma prior on the Exponential rate.

Note that for convenience the `signature` argument may also be a single object or list of objects of the same type as the loadable data sets described in the previous subsection.

Thus, conducting the analysis of §6.3.1, starting page 126, for masked lifetime data already stored in the variable `y` involves only the following simple command:

```
> data(sccs04)
> res <- maskedInferenceIIDExponential(y, sccs04[[3]]$signature, 10000,
                                       priorShape=9, priorScale=0.5)
```

when the signature is treated as known to be order 4 simply connected system number 3. If the desire is to infer over all order 4 simply connected coherent systems, this is as simple as replacing `sccs04[[3]]$signature` with `sccs04`. Likewise, to infer over all orders 2–5, simply replace `sccs04[[3]]$signature` with `c(sccs02, sccs03, sccs04, sccs05)`.

Full details for the exchangeable case and for other distributions are in the package help files.

### 7.2.3.2 General framework

If the user wishes to model with component lifetime distributions which are not currently implemented, it is still possible to save substantial work by calling into the general low-level framework which was developed in the package. The i.i.d. case is discussed here, but the exchangeable case mirrors it strongly and is documented in the package help files.

The core signature based MCMC algorithm is implemented in the following function:

```
maskedInferenceIIDCustom(t, signature, cdfComp, pdfComp,
                          rParmGivenData, rCompGivenParm, startParm, iter, ...)
```

This is structured so that when implementing a new component lifetime distribution, one need merely supply the prerequisites of Algorithm 6.2. The detail of the arguments is as follows:

`t`, `signature`, `iter` are as for the previously described function.



`cdfComp` should be a function which computes  $F_Y(\cdot)$ . The first argument should be the value at which to evaluate the CDF, the second argument should be a parameter vector in the same form as `startParm` and `...` will be passed through in the event any additional arguments are required.

`pdfComp` should be a function which computes  $f_Y(\cdot)$ . The argument list is as for the CDF.

`rParmGivenData` should be a function which will generate a random draw from

$$f_{\Psi|Y}(\psi | \mathbf{y}_{1.}, \dots, \mathbf{y}_{m.})$$

That is, a random draw from the standard Bayesian posterior in the full component data case — prerequisite (iii). The first argument should accept a vector of the component lifetimes and `...` will be passed through in the event any additional arguments are required.

`rCompGivenParm` should be a function which will generate a random draw from

$$F_{Y|Y < t_i}(\cdot; \psi) \text{ or } F_{Y|Y > t_i}(\cdot; \psi)$$

That is, a random draw from the conditional component lifetime distribution — prerequisite (iv). The first argument should be a parameter vector in the same form as `startParm`, the second argument should be a scalar for the time  $t_i$ , the third argument should be a censoring indicator:  $-1$  for  $Y < t_i$ ,  $0$  for exact (i.e.  $t_i$  should be returned), and  $+1$  for  $Y > t_i$ . Finally `...` will be passed through in the event any additional arguments are required.

`startParm` is a vector of starting values of named parameters. This also sets the prototype which the routine will use when calling the above user's functions with parameter values.

Anyone intending to use this lower level routine is highly recommended to inspect the source code in the file `src/MaskedLifetimeInference_Exponential.R` for simple examples.

The package is available with full documentation on the official Comprehensive R Archive Network (CRAN) website:

<http://cran.r-project.org/package=ReliabilityTheory>

## 7.3 Summary

This chapter has introduced some of the statistical computing contributions of this thesis, which ensure that the theoretical elements of the thesis are practicably usable with minimal additional work on the part of the end user.

# Chapter 8

## Conclusion

Inference on the component lifetime distribution parameters of a system in the presence of only masked system lifetime data can be a challenging problem. The theory presented here provides new options to reliability practitioners in the case of both repairable and non-repairable components, together with software to enable immediate use.

In the repairable case, the model reformulation to enable structured generator matrices enables better modelling since there is knowledge about the failure and repair process. In addition, the computational advances make the MCMC algorithm possible to use in such problems.

In the non-repairable case, a flexible signature based data augmentation scheme has been presented which enables a broad class of component lifetime distributions to be used. In addition, it opens the door to interesting possibilities of topological inference on system design. All of this flexible framework has been made concrete with the derivation of the necessary prerequisites for the algorithm in both i.i.d. and exchangeable system settings for Exponential and Phase-type distributed component lifetimes.

Prosaic but essential work has been presented to add to the existing catalogue of coherent systems in the literature with catalogues of simply connected coherent systems and coherent networks.

Finally, the research presented is complemented by two R packages: `PhaseType` and `ReliabilityTheory`. This software provides reference implementations which can be used on general non-pathological examples immediately.

More broadly, the contributions of Chapter 4 have uses far beyond those presented in this thesis. Phase-type distributions make natural models for first passage times in a number of scientific modelling settings (e.g. disease progression). Moreover, the ECS algorithm alone is potentially useful in both theory and applications. For example, diffusion bridging involves conditional sampling of latent paths and the ideas in ECS may be of interest for advancing theoretical aspects of inference there. An example application of ECS would be in synthetic biology where there are phases through which a cell may passage before death (or absorption may represent apoptosis, extinction, failure) and latent paths concordant with given absorption times would themselves be of direct interest.

## 8.1 Future work

*“We live on an island of knowledge surrounded by a sea of ignorance. As our island of knowledge grows, so does the shore of our ignorance.”*

— John A. Wheeler

The modicum of land this work contributes to our island of knowledge indeed forms its own new shores of ignorance. These questions beckon further research and include:

1. Can computational speed be improved even further in exchange for an approximate answer, by finding functional approximations to the distributions of the sufficient statistics of the latent processes in Chapter 4? This was alluded to in §4.2.2.1.
2. A study of the sensitivity of the Chapter 4 methodology to the Exponential component lifetime assumption would be of interest. That is, if component lifetimes are actually non-Exponential then how sensitive are functionals such as mean time to failure/repair times which may be inferred?
3. How practical is it to extend component failure and repair distributions beyond the Exponentially distributed assumption which implicitly underlies the continuous-time Markov chain model in Chapter 4?

Specifically:

- one may arbitrarily closely approximate any lifetime distribution by a Phase-type distribution since they are theoretically dense in the space of all distributions on  $[0, \infty)$ .
- Neuts and Meier (1981) consider units with Phase-type lifetime distribution, and Phase-type distributed repairs. They show that a particular single repair facility queuing model may be formed by constructing a large generator matrix through Kronecker products and sums of the Phase-type subgenerators.

Can these two ideas be used to construct instead a Phase-type system representation when the components are not Exponentially distributed?

4. It would be interesting to see if recent parameter expanded data augmentation techniques (Hobert, 2011) can be implemented in the methodology of Chapter 4 to reduce autocorrelation in large generator problems such as (4.12).
5. Are there further pre-cut set checks which would increase the computational performance of Algorithms 5.1 and 5.2, to allow cataloguing of larger order systems and networks?
6. For topological inference in Chapter 6, could one infer the signature vector directly, as opposed to merely assigning posterior weight to a predefined collection of signatures? To do so may appear meaningless, but one may either look for the ‘nearest’ corresponding coherent system signature, or else treat it as a mixture of coherent systems as discussed by Samaniego (2007). Could this also be incorporated into a reversible jump MCMC scheme (Green, 1995) to move between different lengths of signature?
7. In the case of very large collections of candidate topologies, as suggested it is possible to perform a Metropolis-Hastings move instead of exhaustively computing all move probabilities. This was done in the course of the research work, but there are interesting problems related to signature ordering which warrant further work: unless the ordering of signatures is sensible the Metropolis-Hastings proposal of ‘nearby’ signatures can result in the sampler becoming stuck, as is

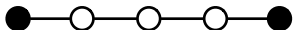
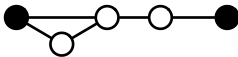
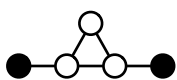
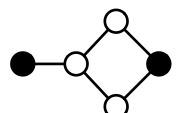
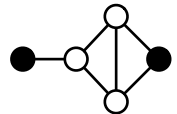
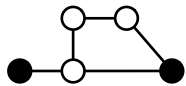
readily seen would happen with some of the marginal posterior topology plots of Chapter 6.

8. It should in principle be possible to accommodate censored data into the methodology of Chapter 6. Deriving the necessary results and examining the effectiveness of the inference would be an interesting avenue of research.
9. At present, the most general case in Chapter 6 is exchangeability at the system level. Is it possible to consider exchangeable rather than i.i.d. components too?
10. The two R packages have much scope for expansion. Indeed potential collaborators have already been in contact with a view to adding more tools. It would also be useful to incorporate the catalogue of coherent system signatures from Shaked and Suarez-Llorens (2003) and Navarro and Rubio (2009). Finally, deriving the prerequisites of Algorithm 6.2 for more common reliability lifetime distributions and adding their implementation to the packages would enhance its usefulness.

# Appendix A

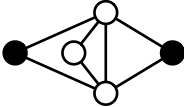
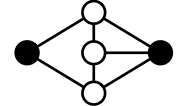
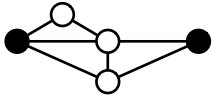
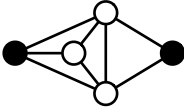
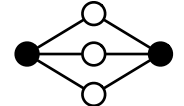
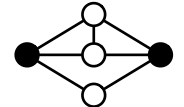
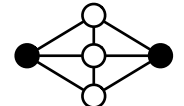
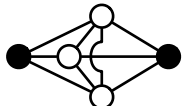
## Signatures

### A.1 Coherent network signatures

Number	Network Topology	Signature
1		$(1, 0, 0, 0)$
2		$(\frac{2}{5}, \frac{1}{2}, \frac{1}{10}, 0, 0)$
3		$(\frac{2}{5}, \frac{1}{2}, \frac{1}{10}, 0, 0)$
4		$(\frac{1}{5}, \frac{3}{5}, \frac{1}{5}, 0, 0)$
5		$(\frac{1}{6}, \frac{3}{10}, \frac{13}{30}, \frac{1}{10}, 0, 0)$
6		$(\frac{1}{5}, \frac{1}{2}, \frac{1}{5}, \frac{1}{10}, 0)$

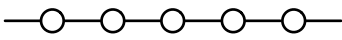
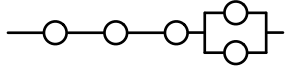
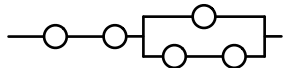
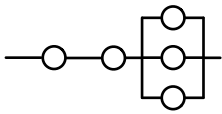
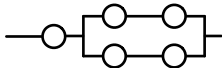
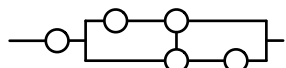
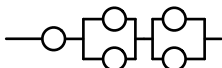
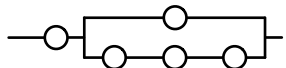
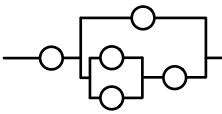
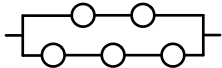
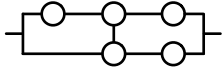
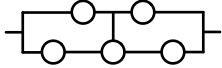
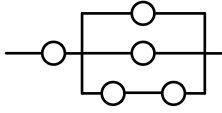
7		$(\frac{1}{6}, \frac{7}{30}, \frac{7}{20}, \frac{11}{60}, \frac{1}{15}, 0)$
8		$(\frac{1}{6}, \frac{7}{30}, \frac{7}{20}, \frac{11}{60}, \frac{1}{15}, 0)$
9		$(\frac{1}{6}, \frac{1}{6}, \frac{11}{30}, \frac{7}{30}, \frac{1}{15}, 0)$
10		$(0, \frac{3}{5}, \frac{3}{10}, \frac{1}{10}, 0)$
11		$(\frac{1}{7}, \frac{1}{7}, \frac{1}{5}, \frac{11}{35}, \frac{16}{105}, \frac{1}{21}, 0)$
12		$(0, \frac{4}{15}, \frac{13}{30}, \frac{7}{30}, \frac{1}{15}, 0)$
13		$(0, \frac{1}{5}, \frac{1}{2}, \frac{7}{30}, \frac{1}{15}, 0)$
14		$(0, \frac{2}{21}, \frac{26}{105}, \frac{3}{7}, \frac{19}{105}, \frac{1}{21}, 0)$
15		$(0, \frac{2}{15}, \frac{7}{15}, \frac{4}{15}, \frac{2}{15}, 0)$
16		$(0, \frac{2}{15}, \frac{5}{12}, \frac{19}{60}, \frac{2}{15}, 0)$

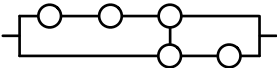
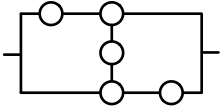
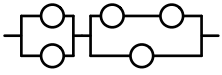
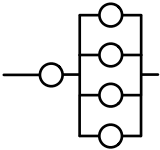
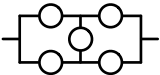
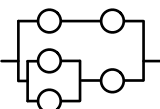
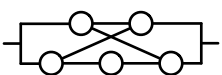
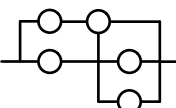
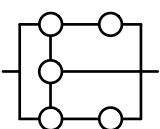
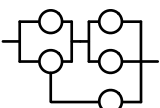
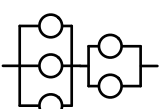
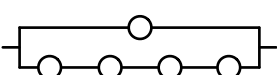


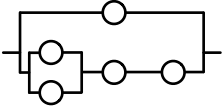
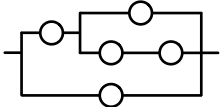
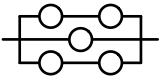
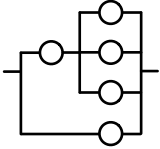
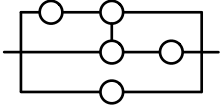
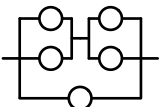
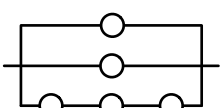
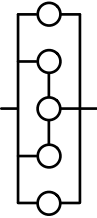
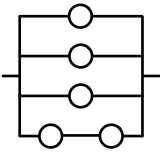
17		$(0, \frac{2}{21}, \frac{4}{21}, \frac{13}{35}, \frac{26}{105}, \frac{2}{21}, 0)$
18		$(0, \frac{1}{21}, \frac{19}{105}, \frac{3}{7}, \frac{26}{105}, \frac{2}{21}, 0)$
19		$(0, \frac{1}{21}, \frac{19}{105}, \frac{2}{5}, \frac{29}{105}, \frac{2}{21}, 0)$
20		$(0, \frac{1}{28}, \frac{5}{56}, \frac{57}{280}, \frac{27}{70}, \frac{3}{14}, \frac{1}{14}, 0)$
21		$(0, 0, \frac{2}{5}, \frac{2}{5}, \frac{1}{5}, 0)$
22		$(0, 0, \frac{4}{35}, \frac{2}{5}, \frac{12}{35}, \frac{1}{7}, 0)$
23		$(0, 0, \frac{1}{28}, \frac{23}{140}, \frac{57}{140}, \frac{2}{7}, \frac{3}{28}, 0)$
24		$(0, 0, \frac{1}{42}, \frac{1}{14}, \frac{4}{21}, \frac{11}{28}, \frac{5}{21}, \frac{1}{12}, 0)$

**Table A.1:** All coherent networks of node order 3, together with system signature. Black nodes are start/terminal. Contained in data set `cn03` of `ReliabilityTheory` package (see §7.2.1, page 150).

## A.2 Simply connected coherent system signatures

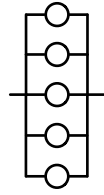
Number	System Topology	Signature
1		$(1, 0, 0, 0, 0)$
2		$(\frac{3}{5}, \frac{2}{5}, 0, 0, 0)$
3		$(\frac{2}{5}, \frac{1}{2}, \frac{1}{10}, 0, 0)$
4		$(\frac{2}{5}, \frac{3}{10}, \frac{3}{10}, 0, 0)$
5		$(\frac{1}{5}, \frac{3}{5}, \frac{1}{5}, 0, 0)$
6		$(\frac{1}{5}, \frac{1}{2}, \frac{3}{10}, 0, 0)$
7		$(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}, 0, 0)$
8		$(\frac{1}{5}, \frac{1}{2}, \frac{1}{5}, \frac{1}{10}, 0)$
9		$(\frac{1}{5}, \frac{3}{10}, \frac{2}{5}, \frac{1}{10}, 0)$
10		$(0, \frac{3}{5}, \frac{3}{10}, \frac{1}{10}, 0)$
11		$(0, \frac{1}{2}, \frac{2}{5}, \frac{1}{10}, 0)$
12		$(0, \frac{2}{5}, \frac{1}{2}, \frac{1}{10}, 0)$
13		$(\frac{1}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{5}, 0)$

14		$(0, \frac{2}{5}, \frac{2}{5}, \frac{1}{5}, 0)$
15		$(0, \frac{3}{10}, \frac{1}{2}, \frac{1}{5}, 0)$
16		$(0, \frac{3}{10}, \frac{1}{2}, \frac{1}{5}, 0)$
17		$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{2}{5}, 0)$
18		$(0, \frac{1}{5}, \frac{3}{5}, \frac{1}{5}, 0)$
19		$(0, \frac{1}{5}, \frac{1}{2}, \frac{3}{10}, 0)$
20		$(0, \frac{1}{5}, \frac{1}{2}, \frac{3}{10}, 0)$
21		$(0, \frac{1}{5}, \frac{2}{5}, \frac{2}{5}, 0)$
22		$(0, \frac{1}{10}, \frac{1}{2}, \frac{2}{5}, 0)$
23		$(0, \frac{1}{10}, \frac{2}{5}, \frac{1}{2}, 0)$
24		$(0, \frac{1}{10}, \frac{3}{10}, \frac{3}{5}, 0)$
25		$(0, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$

26		$(0, \frac{1}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$
27		$(0, \frac{1}{10}, \frac{2}{5}, \frac{3}{10}, \frac{1}{5})$
28		$(0, 0, \frac{2}{5}, \frac{2}{5}, \frac{1}{5})$
29		$(0, \frac{1}{10}, \frac{1}{5}, \frac{1}{2}, \frac{1}{5})$
30		$(0, 0, \frac{3}{10}, \frac{1}{2}, \frac{1}{5})$
31		$(0, 0, \frac{1}{5}, \frac{3}{5}, \frac{1}{5})$
32		$(0, 0, \frac{3}{10}, \frac{3}{10}, \frac{2}{5})$
33		$(0, 0, \frac{1}{10}, \frac{1}{2}, \frac{2}{5})$
34		$(0, 0, 0, \frac{2}{5}, \frac{3}{5})$

---

35



$(0, 0, 0, 0, 1)$

---

**Table A.2:** All simply connected systems of order 5, together with system signature. Contained in data set `sccs05` of `ReliabilityTheory` package (see §7.2.1, page 150).

# References

- Agrawal, A. and Barlow, R. E. (1984), ‘A survey of network reliability and domination theory’, *Operations Research* **32**(3), 478–492.
- Arcones, M. A., Kvam, P. H. and Samaniego, F. J. (2002), ‘Nonparametric estimation of a distribution subject to a stochastic precedence constraint’, *Journal of the American Statistical Association* **97**(457), 170–182.
- Aslett, L. J. M. (2011), *PhaseType: Inference for Phase-type Distributions*. R package.
- Aslett, L. J. M. (2012), *ReliabilityTheory: Tools for structural reliability analysis*. R package.
- Asmussen, S. (2000), ‘Matrix-analytic models and their analysis’, *Scandinavian Journal of Statistics* **27**(2), 193–226.
- Asmussen, S., Nerman, O. and Olsson, M. (1996), ‘Fitting Phase-type distributions via the EM algorithm’, *Scandinavian Journal of Statistics* **23**(4), 419–441.
- Balakrishnan, N., Navarro, J. and Samaniego, F. J. (2012), Signature representation and preservation results for engineered systems and applications to statistical inference, in A. Lisnianski and I. Frenkel, eds, ‘Recent Advances in System Reliability: Signatures, Multi-state Systems and Statistical Inference’, Springer Series in Reliability Engineering, Springer, chapter 1, pp. 1–22.
- Barlow, R. E. and Proschan, F. (1965), *Mathematical Theory of Reliability*, New York: John Wiley.
- Barlow, R. E. and Proschan, F. (1981), *Statistical Theory of Reliability and Life Testing*, To Begin With Press.

- Bayes, T. (1763), ‘An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.’, *Philosophical Transactions* **53**, 370–418.
- Bernardo, J. M. (1979), ‘Reference posterior distributions for Bayesian inference’, *Journal of the Royal Statistical Society, Series B* **41**(2), 113–147.
- Bernardo, J. M. and Smith, A. F. M. (2007), *Bayesian Theory*, 2nd edn, Wiley.
- Berry, A., Bordat, J.-P. and Cogis, O. (1999), Generating all the minimal separators of a graph, in P. Widmayer, G. Neyer and S. Eidenbenz, eds, ‘Graph-Theoretic Concepts in Computer Science’, Vol. 1665 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 167–172.
- Besag, J. (1974), ‘Spatial interaction and the statistical analysis of lattice systems’, *Journal of the Royal Statistical Society, Series B* **36**(2), 192–236.
- Birnbaum, Z. W., Esary, J. D. and Saunders, S. C. (1961), ‘Multi-component systems and structures and their reliability’, *Technometrics* **3**(1), 55–77.
- Bladt, M., Esparza, L. J. R. and Nielsen, B. F. (2011), ‘Fisher information and statistical inference for Phase-type distributions’, *Journal of Applied Probability* **48A**, 277–293.
- Bladt, M., Gonzalez, A. and Lauritzen, S. L. (2003), ‘The estimation of Phase-type related functionals using Markov chain Monte Carlo methods’, *Scandinavian Actuarial Journal* **2003**(4), 280–300.
- Block, H., Dugas, M. R. and Samaniego, F. J. (2006), Characterizations of the relative behavior of two systems via properties of their signature vectors, in N. Balakrishnan, J. M. Sarabia and E. Castillo, eds, ‘Advances in Distribution Theory, Order Statistics, and Inference’, *Statistics for Industry and Technology*, Birkhäuser Boston, chapter 18, pp. 279–289.
- Bondy, J. A. and Murty, U. S. R. (2008), *Graph Theory*, Springer.
- Box, G. E. P. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

- Brémaud, P. (1999), *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer.
- Brooks, S. P., Gelman, A., Jones, G. L. and Meng, X., eds (2011), *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC.
- Butterworth, R. W. (1972), ‘A set theoretic treatment of coherent systems’, *SIAM Journal on Applied Mathematics* **22**(4), 590–598.
- Cano, J., Moguerza, J. M. and Ríos Insua, D. (2010), ‘Bayesian reliability, availability, and maintainability analysis for hardware systems described through continuous time Markov chains’, *Technometrics* **52**(3), 324–334.
- Casella, G. and George, E. I. (1992), ‘Explaining the Gibbs sampler’, *The American Statistician* **46**(3), 167–174.
- Chib, S. and Greenberg, E. (1995), ‘Understanding the Metropolis-Hastings algorithm’, *The American Statistician* **49**(4), 327–335.
- Coolen, F. P. A. and Al-nefaiee, A. H. (2012), ‘Nonparametric predictive inference for failure times of systems with exchangeable components’, *Journal of Risk and Reliability* **226**(3), 262–273.
- Coolen, F. P. A., Coolen-Schrijner, P. and Yan, K. (2002), ‘Nonparametric predictive inference in reliability’, *Reliability Engineering & System Safety* **78**(2), 185–193.
- Cordella, L. P., Foggia, P., Sansone, C. and Vento, M. (2001), An improved algorithm for matching large graphs, in ‘Proc. 3rd IAPR-TC15 Workshop Graph-Based Representations in Pattern Recognition’, pp. 149–159.
- Cox, D. R. (1955), ‘A use of complex probabilities in the theory of stochastic processes’, *Mathematical Proceedings of the Cambridge Philosophical Society* **51**(2), 313–319.
- Cox, D. R. (1972), ‘Regression models and life-tables’, *Journal of the Royal Statistical Society, Series B* **34**(2), 187–220.
- Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press.



- Cox, D. R. and Connelly, C. A. (2011), *Principles of Applied Statistics*, Cambridge University Press.
- Csárdi, G. and Nepusz, T. (2006), ‘The igraph software package for complex network research’, *InterJournal Complex Systems*, 1695.
- Cumani, A. (1982), ‘On the canonical representation of homogeneous Markov processes modelling failure-time distributions’, *Microelectronics Reliability* **22**(3), 583–602.
- Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989), ‘Risk analysis of the space shuttle: Pre-Challenger prediction of failure’, *Journal of the American Statistical Association* **84**(408), 945–957.
- Daneshkhah, A. and Bedford, T. (2008), Sensitivity analysis of a reliability system using gaussian processes, *in* T. Bedford, J. Quigley, L. Walls, B. Alkali, A. Daneshkhah and G. Hardman, eds, ‘Advances in Mathematical Modeling for Reliability’, IOS Press, chapter 2, pp. 46–62.
- de Finetti, B. (1974), *Theory of Probability*, New York: John Wiley.
- Dehon, M. and Latouche, G. (1982), ‘A geometric interpretation of the relations between the Exponential and generalized Erlang distributions’, *Advances in Applied Probability* **14**(4), 885–897.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), ‘Maximum likelihood from incomplete data via the EM algorithm’, *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38.
- Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer. Available at <http://luc.devroye.org/rnbookindex.html>.
- Diaconis, P. (1996), ‘The cutoff phenomenon in finite Markov chains’, *Proceedings of the National Academy of Sciences of the United States of America* **93**(4), 1659–1664.
- Diaconis, P. (2009), ‘The Markov chain Monte Carlo revolution’, *Bulletin of the American Mathematical Society* **46**(2), 179–205.
- Fisher, R. A. (1922), ‘On the mathematical foundations of theoretical statistics’, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

- Fisher, R. A. (1930), ‘Inverse probability’, *Mathematical Proceedings of the Cambridge Philosophical Society* **26**, 528–535.
- Flegal, J. M., Haran, M. and Jones, G. L. (2008), ‘Markov chain Monte Carlo: Can we trust the third significant figure?’, *Statistical Science* **23**(2), 250–260.
- Flegal, J. M. and Hughes, J. (2012), *mcmcse: Monte Carlo Standard Errors for MCMC*, Riverside, CA and Minneapolis, MN. R package version 1.0-1.
- Garthwaite, P. H., Kadane, J. B. and O’Hagan, A. (2005), ‘Statistical methods for eliciting probability distributions’, *Journal of the American Statistical Association* **100**(470), 680–701.
- Gåsemyr, J. and Natvig, B. (2001), ‘Bayesian inference based on partial monitoring of components with applications to preventative system maintenance’, *Naval Research Logistics* **48**(7), 551–577.
- Gelfand, A. E. and Smith, A. F. M. (1990), ‘Sampling-based approaches to calculating marginal densities’, *Journal of the American Statistical Association* **85**(410), 398–409.
- Gelman, A. (2006), ‘Prior distribution for variance parameters in hierarchical models’, *Bayesian Analysis* **1**(3), 515–533.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn, Chapman & Hall/CRC.
- Gelman, A. and Rubin, D. B. (1992), ‘Inference from iterative simulation using multiple sequences’, *Statistical Science* **7**(4), 457–472.
- Geman, S. and Geman, D. (1984), ‘Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.
- Gertsbakh, I. and Shpungin, Y. (2011), *Network Reliability and Resilience*, Springer.
- Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in ‘Bayesian Statistics 4’, pp. 169–193.

- Geyer, C. J. (2010), *mcmc: Markov chain Monte Carlo*. R package version 0.8.
- Geyer, C. J. (2011), Introduction to Markov chain Monte Carlo, *in* ‘Handbook of Markov Chain Monte Carlo’, Chapman & Hall/CRC, chapter 1, pp. 3–48.
- Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995), ‘Adaptive rejection Metropolis sampling within Gibbs sampling’, *Journal of the Royal Statistical Society, Series C* **44**(4), 455–472.
- Green, P. J. (1995), ‘Reversible jump Markov chain Monte Carlo computation and Bayesian model determination’, *Biometrika* **82**(4), 711–732.
- Grimmett, G. R. and Stirzaker, D. R. (2001), *Probability and Random Processes*, 3 edn, Oxford University Press.
- Häggström, O., Asmussen, S. and Nerman, O. (1992), EMPHT – a program for fitting Phase type distributions, Technical report, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden.
- Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Methuen & Co Ltd.
- Hastings, W. K. (1970), ‘Monte Carlo sampling methods using Markov chains and their applications’, *Biometrika* **57**(1), 97–109.
- Heidelberger, P. and Welch, P. D. (1983), ‘Simulation run length control in the presence of an initial transient’, *Operations Research* **31**(6), 1109–1144.
- Higham, N. J. (2008), *Functions of Matrices: Theory and Computation*, Society for Industrial & Applied Mathematics.
- Hobert, J. P. (2011), The data augmentation algorithm: theory and methodology, *in* ‘Handbook of Markov Chain Monte Carlo’, Chapman & Hall/CRC, chapter 10, pp. 253–293.
- Hobolth, A. and Jensen, J. L. (2011), ‘Summary statistics for endpoint-conditioned continuous-time Markov chains’, *Journal of Applied Probability* **48**(4), 911–924.

- Hobolth, A. and Stone, E. A. (2009), ‘Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution’, *Annals of Applied Statistics* **3**(3), 1204–1231.
- Jeffreys, H. (1939), *Theory of Probability*, 1st edn, The Clarendon Press.
- Jeffreys, H. (1946), ‘An invariant form for the prior probability in estimation problems’, *Proceedings of the Royal Society of London, Series A* **186**(1007), 453–461.
- Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006), ‘Fixed-width output analysis for Markov chain Monte Carlo’, *Journal of the American Statistical Association* **101**(476), 1537–1547.
- Jones, G. L. and Hobert, J. P. (2001), ‘Honest exploration of intractable probability distributions via Markov chain Monte Carlo’, *Statistical Science* **16**(4), 312–334.
- Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley.
- Kaplan, E. L. and Meier, P. (1958), ‘Nonparametric estimation from incomplete observations’, *Journal of the American Statistical Association* **53**(282), 457–481.
- Kochar, S., Mukerjee, H. and Samaniego, F. J. (1999), ‘The “signature” of a coherent system and its application to comparisons among systems’, *Naval Research Logistics* **46**(5), 507–523.
- Kuo, L. and Yang, T. Y. (2000), ‘Bayesian reliability modeling for masked system lifetime data’, *Statistics & Probability Letters* **47**(3), 229–241.
- Laplace, P. S. (1774), ‘Mémoire sur la probabilité des causes par les évènements’, *Mémoires de Mathématique et de Physique* **6**, 621–656.
- Lawless, J. F. and Fredette, M. (2005), ‘Frequentist prediction intervals and predictive distributions’, *Biometrika* **92**(3), 529–542.
- Lindley, D. V. (1958), ‘Fiducial distributions and Bayes’ theorem’, *Journal of the Royal Statistical Society, Series B* **20**(1), 102–107.

- Mangel, M. and Samaniego, F. J. (1984), ‘Abraham Wald’s work on aircraft survivability’, *Journal of the American Statistical Association* **79**(386), 259–267.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953), ‘Equation of state calculations by fast computing machines’, *Journal of Chemical Physics* **21**(6), 1087–1092.
- Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, 1st edn, Springer. Available at <http://probability.ca/MT/>.
- Meyn, S. P. and Tweedie, R. L. (2009), *Markov Chains and Stochastic Stability*, 2nd edn, Cambridge University Press.
- Moler, C. and Van Loan, C. (2003), ‘Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later’, *SIAM Review* **45**(1), 3–49.
- Navarro, J. and Rubio, R. (2009), ‘Computations of signatures of coherent systems with five components’, *Communications in Statistics – Simulation and Computation* **39**(1), 68–84.
- Neuts, M. F. (1975), ‘Probability distributions of Phase type’, *Liber Amicorum Prof. Emeritus H. Florin, Dept. Math, Univ. Louvain, Belgium* pp. 173–206.
- Neuts, M. F. (1994), *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Dover Publications.
- Neuts, M. F. and Meier, K. S. (1981), ‘On the use of Phase type distributions in reliability modelling of systems with two components’, *OR Spectrum* **2**(4), 227–234.
- Neuts, M. F. and Pagano, M. E. (1981), ‘Generating random variates from a distribution of Phase type’, *WSC ’81 Proceedings of the 13th Conference on Winter Simulation* **2**, 381–387.
- Neuts, M. F., Pérez-Ocón, R. and Torres-Castro, I. (2000), ‘Repairable models with operating and repair times governed by Phase type distributions’, *Advances in Applied Probability* **32**(2), 468–479.

- Ng, H. K. T., Navarro, J. and Balakrishnan, N. (2012), ‘Parametric inference from system lifetime data under a proportional hazard rate model’, *Metrika* **75**(3), 367–388.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006), *Uncertain Judgements: Eliciting Experts’ Probabilities*, Wiley.
- R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at <http://www.R-project.org/>.
- Reiser, B., Guttman, I., Lin, D. K. J., Guess, F. M. and Usher, J. S. (1995), ‘Bayesian inference for masked system lifetime data’, *Journal of the Royal Statistical Society, Series C* **44**(1), 79–90.
- Ripley, B. D. (2005), How computing has changed statistics, in A. C. Davison, Y. Dodge and N. Wermuth, eds, ‘Celebrating Statistics: Papers in honour of Sir David Cox on his 80th Birthday’, Oxford University Press, chapter 10, pp. 197–211.
- Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, 2nd edn, Springer.
- Robert, C. P. and Casella, G. (2011), ‘A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data’, *Statistical Science* **26**(1), 102–115.
- Roberts, G. O., Gelman, A. and Gilks, W. R. (1997), ‘Weak convergence and optimal scaling of random walk Metropolis algorithms’, *The Annals of Applied Probability* **7**(1), 110–120.
- Roberts, G. O. and Rosenthal, J. S. (2004), ‘General state space Markov chains and MCMC algorithms’, *Probability Surveys* **1**, 20–71.
- Roman, S. (2008), *Advanced Linear Algebra*, Graduate Texts in Mathematics, 3rd edn, Springer.
- Samaniego, F. J. (1985), ‘On closure of the IFR class under formation of coherent systems’, *IEEE Transactions on Reliability* **R-34**(1), 69–72.

- Samaniego, F. J. (2007), *System Signatures and Their Applications in Engineering Reliability*, Springer.
- Savage, L. J. (1954), *The Foundations of Statistical Inference*, New York: John Wiley.
- Shaked, M. and Suarez-Llorens, A. (2003), ‘On the comparison of reliability experiments based on the convolution order’, *Journal of the American Statistical Association* **98**(463), 693–702.
- Shin, Y. Y. and Koh, J. S. (1998), ‘An algorithm for generating minimal cutsets of undirected graphs’, *Journal of Applied Mathematics and Computing* **5**(3), 681–693.
- Singpurwalla, N. D. and Wilson, S. P. (1995), ‘The exponentiation formula of reliability and survival: Does it always hold?’, *Lifetime Data Analysis* **1**(2), 187–194.
- Singpurwalla, N. D. and Wilson, S. P. (1999), *Statistical Methods in Software Engineering*, Springer.
- Tanner, M. A. and Wong, W. H. (1987), ‘The calculation of posterior distributions by data augmentation’, *Journal of the American Statistical Association* **82**(398), 528–540.