Bayesian modelling of recurrent pipe failures in urban water systems using non-homogeneous Poisson processes with latent structure

Submitted by

Theodoros Economou

to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics, September 2010.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred upon me.

Theodoros Economou
Abstract

Recurrent events are very common in a wide range of scientific disciplines. The majority of statistical models developed to characterise recurrent events are derived from either reliability theory or survival analysis. This thesis concentrates on applications that arise from reliability, which in general involve the study about components or devices where the recurring event is failure.

Specifically, interest lies in repairable components that experience a number of failures during their lifetime. The goal is to develop statistical models in order to gain a good understanding about the driving force behind the failures. A particular counting process is adopted, the non-homogenous Poisson process (NHPP), where the rate of occurrence (failure rate) depends on time. The primary application considered in the thesis is the prediction of underground water pipe bursts although the methods described have more general scope.

First, a Bayesian mixed effects NHPP model is developed and applied to a network of water pipes using MCMC. The model is then extended to a mixture of NHPPs. Further, a special mixture case, the zero-inflated NHPP model is developed to cope with data involving a large number of pipes that have never failed. The zero-inflated model is applied to the same pipe network.

Quite often, data involving recurrent failures over time, are aggregated where for instance the times of failures are unknown and only the total number of failures are available. Aggregated versions of the NHPP model and its zero-inflated version are developed to accommodate aggregated data and these are applied to the aggregated
version of the earlier data set.

Complex devices in random environments often exhibit what may be termed as state changes in their behaviour. These state changes may be caused by unobserved and possibly non-stationary processes such as severe weather changes. A hidden semi-Markov NHPP model is formulated, which is a NHPP process modulated by an unobserved semi-Markov process. An algorithm is developed to evaluate the likelihood of this model and a Metropolis-Hastings sampler is constructed for parameter estimation. Simulation studies are performed to test implementation and finally an illustrative application of the model is presented.

The thesis concludes with a general discussion and a list of possible generalisations and extensions as well as possible applications other than the ones considered.
Acknowledgements

It was a privilege and honour to have studied at a department which consisted of people whose support was only matched by their expertise. In particular I am indebted to both my supervisors Trevor Bailey and Zoran Kapelan for their trust, guidance and most importantly for their sense of humour.

To Yehuda Kleiner for kindly providing the data for the thesis.

Also to Renato whose friendship and advice kept me sane most of the time and to Rachel for her support and understanding. I would also like to thank Yiwei for making me smile. Also Pete, Graham and Matt for their computing support.

Special thanks to Les Grossman without whom the last few years would be significantly more miserable.

To Dimitris who was always there for me and to the many people and friends who made this experience enjoyable and unforgettable.

Finally, I wish to dedicate this to my parents and sister for their unconditional support and love, and to Victoria for giving me strength and for completing me.
Contents

Acknowledgements 4

Contents 5

List of Figures 11

List of Tables 14

1 Introduction 16

1.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
1.2 Aims . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
1.3 Structure of Thesis . . . . . . . . . . . . . . . . . . . . . . . . . . 20
1.4 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

2 Background 22

2.1 Recurrent Event Modelling . . . . . . . . . . . . . . . . . . . . . . . 22

2.1.1 Counting Processes . . . . . . . . . . . . . . . . . . . . . . . . . 23
## Contents

2.1.2 Reliability Theory and Survival Analysis ........................................ 26

2.2 Counting Processes in Reliability ...................................................... 29
  2.2.1 Renewal Processes ................................................................. 29
  2.2.2 Non-Homogeneous Processes ................................................... 30
  2.2.3 Processes with Dependency Structures ....................................... 31

2.3 Recurrent Failures in Water Pipe Networks ....................................... 32

2.4 Summary ......................................................................................... 37

3 Data ................................................................................................. 38
  3.1 Basic Summary Analysis of the Data ............................................... 38
  3.2 Tests for temporal trends ............................................................... 42
  3.3 Summary ......................................................................................... 43

4 Mixed Effect NHPP Models .................................................................. 45
  4.1 Non-Homogeneous Poisson Process Models ..................................... 45
    4.1.1 Failure Rate ............................................................................ 46
    4.1.2 Likelihood Function ............................................................... 50
    4.1.3 Left Truncation ..................................................................... 52
    4.1.4 Random Effects ..................................................................... 53
    4.1.5 Bayesian Framework and MCMC .......................................... 55
  4.2 Mixed Effects NHPP model ............................................................ 56
## Contents

6.2.1 Model Formulation ............................................. 101
6.2.2 Simulation Experiments ............................................. 102
6.2.3 Model Application ............................................. 105

6.3 Summary ........................................................ 112

7 Hidden Semi-Markov NHPP ............................................. 114

7.1 Motivation .................................................. 114
7.2 Hidden Markov Models ............................................. 117
7.3 Markov Modulated Poisson Processes ............................................. 119
7.4 Hidden Markov NHPP Model ............................................. 121
    7.4.1 Model Formulation ............................................. 121
    7.4.2 Likelihood .................................................. 123
    7.4.3 Recursive Algorithms: Forward ............................................. 124
    7.4.4 Recursive Algorithms: Backward ............................................. 127
    7.4.5 E-M Algorithm ............................................. 128
    7.4.6 State Holding Times ............................................. 129
7.5 Review of Hidden Semi-Markov Models ............................................. 130
7.6 Hidden Semi-Markov NHPP Model ............................................. 133
    7.6.1 Semi-Markov Chains ............................................. 133
    7.6.2 Likelihood Formulation ............................................. 136
List of Figures

2.1 Daily flow levels for river Severn in 2005 . . . . . . . . . . . . . . . 24

3.1 Pipe age vs pipes . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

3.2 Histogram of total number of failures per pipe . . . . . . . . . . . . 40

3.3 Failure counts vs pipe length . . . . . . . . . . . . . . . . . . . . . 40

3.4 Monthly number of failures vs month (1962-2003) . . . . . . . . . . 41

3.5 Monthly number of failures vs pipes (1-1349) . . . . . . . . . . . . . 42

3.6 Laplace test for 865 pipes . . . . . . . . . . . . . . . . . . . . . . . 43

4.1 Bathtub-shaped failure rate function . . . . . . . . . . . . . . . . . 47

4.2 Examples of log-linear failure rate . . . . . . . . . . . . . . . . . . 48

4.3 Examples of power-law failure rate . . . . . . . . . . . . . . . . . . 49

4.4 Actual vs predicted no. of failures . . . . . . . . . . . . . . . . . . 60

4.5 Samples of the log-posterior . . . . . . . . . . . . . . . . . . . . . . 63

4.6 Deviance samples - actual (red) and simulated data (black) . . . . . 64

4.7 Posterior means and Cr.I. for $\theta_i$ . . . . . . . . . . . . . . . . . 65
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.8</td>
<td>Estimated ranks vs actual ranks</td>
<td>66</td>
</tr>
<tr>
<td>4.9</td>
<td>Estimated number of failures vs individual pipes</td>
<td>68</td>
</tr>
<tr>
<td>5.1</td>
<td>Rank of $p_i$ vs rank of $\hat{p}_i$</td>
<td>79</td>
</tr>
<tr>
<td>5.2</td>
<td>Actual vs predicted no. of failures</td>
<td>80</td>
</tr>
<tr>
<td>5.3</td>
<td>$\hat{R}$ for 1355 model parameters</td>
<td>82</td>
</tr>
<tr>
<td>5.4</td>
<td>Samples of the log-posterior</td>
<td>83</td>
</tr>
<tr>
<td>5.5</td>
<td>Deviance samples - actual (red) and simulated data (black)</td>
<td>83</td>
</tr>
<tr>
<td>5.6</td>
<td>Posterior means and Cr.I. for $\theta_i$ and $p_i$</td>
<td>85</td>
</tr>
<tr>
<td>5.7</td>
<td>Estimated ranks vs actual ranks</td>
<td>86</td>
</tr>
<tr>
<td>5.8</td>
<td>Estimated number of failures vs individual pipes</td>
<td>86</td>
</tr>
<tr>
<td>6.1</td>
<td>Actual vs predicted no. of failures</td>
<td>95</td>
</tr>
<tr>
<td>6.2</td>
<td>Samples of the log-posterior</td>
<td>96</td>
</tr>
<tr>
<td>6.3</td>
<td>Deviance samples - actual (red) and simulated data (black)</td>
<td>97</td>
</tr>
<tr>
<td>6.4</td>
<td>Posterior means and Cr.I. for $\theta_i$</td>
<td>98</td>
</tr>
<tr>
<td>6.5</td>
<td>Estimated ranks vs actual ranks</td>
<td>99</td>
</tr>
<tr>
<td>6.6</td>
<td>Estimated number of failures vs individual pipes</td>
<td>99</td>
</tr>
<tr>
<td>6.7</td>
<td>Rank of $p_i$ vs rank of $\hat{p}_i$</td>
<td>105</td>
</tr>
<tr>
<td>6.8</td>
<td>Actual vs predicted number of failures</td>
<td>105</td>
</tr>
<tr>
<td>6.9</td>
<td>$\hat{R}$ for 1355 model parameters</td>
<td>106</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>6.10</td>
<td>Samples of the log-posterior</td>
<td>107</td>
</tr>
<tr>
<td>6.11</td>
<td>Deviance samples - actual (red) and simulated data (black)</td>
<td>107</td>
</tr>
<tr>
<td>6.12</td>
<td>Posterior means and Cr.I. for $\theta_i$ and $p_i$</td>
<td>109</td>
</tr>
<tr>
<td>6.13</td>
<td>Estimated ranks vs actual ranks</td>
<td>110</td>
</tr>
<tr>
<td>6.14</td>
<td>Estimated number of failures vs individual pipes</td>
<td>110</td>
</tr>
<tr>
<td>7.1</td>
<td>Exponentially increasing intensity function</td>
<td>115</td>
</tr>
<tr>
<td>7.2</td>
<td>Clustered/Repellent failures</td>
<td>116</td>
</tr>
<tr>
<td>7.3</td>
<td>Hidden Markov NHPP model</td>
<td>122</td>
</tr>
<tr>
<td>7.4</td>
<td>Hidden Markov NHPP model</td>
<td>130</td>
</tr>
<tr>
<td>7.5</td>
<td>A semi-Markov chain</td>
<td>134</td>
</tr>
<tr>
<td>7.6</td>
<td>Semi-Markov modulated NHPP</td>
<td>136</td>
</tr>
<tr>
<td>7.7</td>
<td>Monthly failure counts vs month</td>
<td>156</td>
</tr>
<tr>
<td>7.8</td>
<td>Samples of the log-posterior</td>
<td>157</td>
</tr>
<tr>
<td>7.9</td>
<td>Deviance samples - actual (red) and simulated data (black)</td>
<td>158</td>
</tr>
<tr>
<td>7.10</td>
<td>Observed and estimated cumulative number of failures</td>
<td>158</td>
</tr>
</tbody>
</table>
# List of Tables

3.1 Network description .............................................. 39

3.2 Summary statistics for monthly failures ....................... 41

4.1 Simulated data ..................................................... 58

4.2 Simulated NHPP results .......................................... 59

4.3 NHPP parameter estimates ....................................... 64

4.4 NHPP confusion matrices ....................................... 67

5.1 Simulated ZINHPP results ..................................... 78

5.2 Simulated ZINHPP results ..................................... 78

5.3 ZINHPP parameter estimates .................................. 84

5.4 ZINHPP confusion matrices - observation period ............ 87

5.5 ZINHPP confusion matrices - prediction period ............... 88

6.1 Simulated aggNHPP results .................................... 94

6.2 $\hat{R}$ values ..................................................... 96
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>aggNHPP parameter estimates</td>
<td>97</td>
</tr>
<tr>
<td>6.4</td>
<td>aggNHPP confusion matrices - observation period</td>
<td>100</td>
</tr>
<tr>
<td>6.5</td>
<td>aggNHPP confusion matrices - prediction period</td>
<td>101</td>
</tr>
<tr>
<td>6.6</td>
<td>Simulated aggZINHPP results</td>
<td>104</td>
</tr>
<tr>
<td>6.7</td>
<td>Simulated aggZINHPP results</td>
<td>104</td>
</tr>
<tr>
<td>6.8</td>
<td>aggZINHPP parameter estimates</td>
<td>108</td>
</tr>
<tr>
<td>6.9</td>
<td>aggZINHPP confusion matrices - observation period</td>
<td>111</td>
</tr>
<tr>
<td>6.10</td>
<td>aggZINHPP confusion matrices - prediction period</td>
<td>112</td>
</tr>
<tr>
<td>7.1</td>
<td>Priors, input values and estimates</td>
<td>154</td>
</tr>
<tr>
<td>7.2</td>
<td>Pipe information</td>
<td>155</td>
</tr>
<tr>
<td>7.3</td>
<td>Priors and estimates</td>
<td>157</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This chapter introduces motivations behind this thesis and presents a discussion of the primary aims and applications which are the development of flexible and general statistical models for recurrent failures in underground water pipes. The structure of subsequent chapters in the thesis is then clearly laid out.

1.1 Motivation

Recurrent event data refers to observed occurrences of the same event over time in an individual or unit of interest and can be found in a wide range of applications. In medicine for instance, recurrent events often constitute some kind of disease or medical condition, a particular example being recurrent mammary gland tumours (Freedman et al., 1993). In car manufacturing, warranty claims on automobiles may constitute recurrent events whose analysis will ultimately reflect the quality and reliability of cars (Cook and Lawless, 2007). In the debugging process of new software, the underlying process behind recurrences of faults (bugs) may be of interest in order to optimise the stopping time of fault testing (Dalal and Mallows, 2008).
1.1. Motivation

The primary motivation in this thesis, arises from recurrent failures in underground water pipes. As will be discussed later in this chapter, data involving networks of water pipes tend to have certain features which raise problems that are specific to the application. At the same time, pipes are repairable components and thus may be categorised as a typical application arising from reliability theory, a perspective that will be adopted in this thesis.

Improved understanding of the failure mechanism in pipes and modelling the stochastic behaviour of the recurrences is highly beneficial to end-users such as water companies. Useful models for pipe failures may be used both for diagnostic purposes as well as optimisation tools for developing replacement strategies. In addition, they can be utilised in conjunction with economic assessment models to act as decision making tools for water managers (O’Day, 1982).

Both physically based and statistical pipe failure models are reported in the literature. A fairly recent comprehensive review paper on the structural deterioration of water mains (Kleiner and Rajani, 2001a) concludes that pipe failures occur due to environmental and operational stresses acting upon pipes but also due to the fact that the structural integrity of the pipes has been altered by degradation, corrosion, unsuccessful installation and manufacturing defects. The article reviews several complex physical models that have been utilised in order to capture the effects that these stresses have on the structural properties of the pipes. The main findings were that these models typically address only a single aspect of those interactions and that in order to describe the entire failure process one would need to combine some of these models. More importantly, the authors state that even the best physical model is only as good as the available data and in the water industry good quality data sets are expensive to obtain and often unavailable.

A companion paper, (Kleiner and Rajani, 2001b) provides a comprehensive review of statistical models applied to water networks. They conclude that physical mechanisms that give rise to failures often require data that are not readily available which is why statistical models provide an efficient and ‘cheap’ surrogate.
1.1. Motivation

Underground water pipes offer a special scenario in terms of understanding the mechanism behind recurrent failures. On one hand, there exists the natural deterioration or ageing and on the other, there are external acting forces due to the fact that the pipes are buried. Boxall et al. (2007) state that although ageing is the major driver behind pipe bursts, there are many other factors affecting the failure process such as external corrosion, the amount of loading (pressure, ground movement), pipe length and diameter, the type of material pipes are made of (e.g. cast iron, asbestos cement), the quality of installation and workmanship and even the burst history itself.

In addition, the deterioration mechanism in the pipes is quite complicated. Kleiner and Rajani (2001b) state that there are two groups into which the deterioration of pipes may be classified. First is the structural deterioration which reduces the resistance of pipes to the various stresses acting upon them. Second, is the alteration in the inner surfaces which result in diminished hydraulic capacity but at the same time also decreased resiliency. In other words, the ageing process is quite complicated so that the relationship between the rate of failures and time can be highly non-linear.

Boxall et al. (2007) also make an important point, that models attempting to predict the behaviour of water pipes are not only limited by both the quality and quantity of available data, but also by the inadequate number of available and applied statistical techniques.

Data quality is a serious issue as many data sets contain uncertainty, e.g. due to unreliable recording of failure times or inaccurate measurements of the confounding factors or even the lack of the actual failure times (aggregation).

In addition, data quantity is also a problem since many data sets include information about failures for a time interval which is only a small fraction of the actual lifetime (left truncation). For this reason, many modelling approaches consider the grouping of data sets into classes of similar or ‘homogeneous’ pipes that, for
instance, are made of the same material or have the same diameter.

Despite the fact that grouping pipes together in a sensible way often aids the analysis, it would be more sensible to work with individual pipes as this would be more relevant to water companies. Also, a pipe specific analysis is more intuitive since ultimately each pipe in a network is unique. However, a pipe specific model may have extra problems to deal with. Lack of data quantity for instance, can result in many pipes in the network which have zero recorded failures making it particularly awkward to capture the failure process.

Another issue, which is specifically relevant to water pipes, is the inability to observe all possible factors and processes which either directly or indirectly affect the failure mechanism. Boxall et al. (2007) state that the evolution of pipe bursts is a complex function of numerous variables, many of which are unknown or not directly quantifiable. Examples vary from random anthropogenic acts such as driving heavy vehicles loaded over the appropriate limit, to complex continuous processes such as extreme climate conditions. Both these scenarios may result in elevated failure occurrence, relative to the underlying failure mechanism.

In conclusion, the specific features of water pipe failure data, will inevitably test the limitations of current methodology. Therefore, the use of non-standard statistical methodologies is necessary. A second but equally important motivation in this thesis, is the development of appropriate statistical methodology which is relevant to failures in pipes but which may also be applied to other areas involving recurrent events.

### 1.2 Aims

The main objective of this thesis is the development of predictive statistical models for recurrent failures in water pipes which may also have a more general application. The specific aims are listed below:
1.3 Structure of Thesis

- To develop models that explicitly capture the underlying mechanisms driving the failure process and at the same time to utilise the stochastic nature of models to account for both natural and induced uncertainty in the data.
- To consider pipe specific models that allow for the practical reality that pipes in a network are heterogeneous.
- To allow, in the modelling framework, for particular issues with water pipe data such as data quality and quantity.
- To develop general innovative and flexible models that specifically deal with unobserved processes that may be influencing the failure mechanism.
- To investigate feasible and efficient mechanisms for model fitting.

1.3 Structure of Thesis

In Chapter 2, a brief review of recurrent event literature is discussed. In particular, the different but equivalent ways of studying recurrent events in reliability theory and survival analysis are distinguished. Furthermore, a literature review of statistical modelling for water pipes is presented and analysed.

A specific data set from a network consisting of 1349 pipes is presented in Chapter 3. This data originates from a North American municipality and will be used as the principal application throughout the thesis.

In Chapter 4, a Bayesian non-homogeneous Poisson process (NHPP) mixed effects model is formulated. The model is implemented using MCMC and tested on simulated data. It is then applied to the North American data set. In Chapter 5, the NHPP model is extended to account for zero-inflation and implementation is tested using synthetic data. The zero-inflated model is also applied to the North American data.
Chapter 6 extends models from Chapters 4 and 5 so that they can cope with aggregation in data, where number of failures within time periods are given instead of individual failure times. The implementation of each aggregated model is tested on simulated data used in previous chapters, and then applied to the North American network. Finally, the relative fit of the various models is discussed.

The possibility of unobserved processes influencing the behaviour of the deterioration mechanism in water pipes is investigated in Chapter 7. A hidden semi-Markov NHPP model is formulated to cope with this particular issue. Furthermore, an algorithm for efficient evaluation of the likelihood is developed and coupled to a Metropolis-Hastings sampler to construct a feasible fitting mechanism. The model is tested on simulated data and then applied to selected pipes from the North American network.

Finally, Chapter 8 provides conclusions and further considerations for modelling pipe failures and more general recurrent events. Potential areas for future work are also highlighted.

1.4 Summary

This chapter has introduced the underlying motivations behind this thesis: The study of flexible statistical models to capture the behaviour of recurrent water pipe failures, which are general enough to apply in other applications. In addition, the aims and structure of the thesis were clearly outlined.

In the subsequent Chapter 2, statistical models for recurrent events are reviewed and specifically a literature review of modelling recurrent water pipe failures is provided.
Chapter 2

Background

In Chapter 1 it was made clear that the motivation in this thesis is two-fold. One refers to the particular issues with failure data from water pipe networks while the other relates to the development of innovative methodology. The aim in this chapter is to review associated key literature and to establish general concepts relevant to subsequent chapters.

2.1 Recurrent Event Modelling

Although the primary application here refers to recurrent water pipe failures, it is sensible to start with general statistical concepts about recurrent events which underpin the modelling of water pipe failures.

Recurrent event data are very common and potentially arise in a wide range of applications such as medicine and engineering. A scientific purpose involved in collecting such data, is to study the mechanisms that give rise to the recurrences through modelling. In pipe failures for instance, one is usually interested in modelling the rate of failures per unit time. Statistical modelling of recurrent events is beneficial both for better understanding of the mechanisms that drive the recur-
2.1. Recurrent Event Modelling

The following sections introduce basic models, concepts and quantities which will be used throughout the thesis.

2.1.1 Counting Processes

The reasoning behind the idea of applying statistical models to recurrent events will depend on the specific application. In general, events occur in time according to some underlying and possibly non-stationary process which has an element of randomness. In other words, the mechanism that drives these recurrences has a well defined structure (e.g. based on some physical law) and the actual events evolve in time according to this structure but with some uncertainty or randomness. This uncertainty may be occurring for reasons such as natural variability or external factors affecting the process. Statistical models are appropriate because of their ability to allow for such randomness.

The literature on statistical analysis of recurrent event data has grown significantly over the past twenty years (Cook and Lawless, 2007) in many different but often parallel streams. The basic premise for most of these applications is one described above, where events in time occur according to some stochastic process. Specifically, a natural and extensively used modelling framework for recurrent event data which behave stochastically in time is the one involving counting processes (Cox and Isham, 1980).

Counting processes belong to the general family of stochastic processes which describe the evolution of time dependent random variables according to some specific stochastic structure. An example of a continuous stochastic process is the Wiener process $Y(t)$ where the stochastic structure of the process comes from the property that $Y(t_b) - Y(t_a) \sim N(0, t_b - t_a)$ for $0 \leq t_a < t_b$. A counting process is effectively a discrete stochastic process where a random variable $N(t_a, t_b)$ is utilised, referring
2.1. Recurrent Event Modelling

to the count of events in the time interval \((t_a, t_b]\) where \(0 \leq t_a < t_b\). A commonly used counting process is the Poisson process where \(N(t_a, t_b)\) is distributed as Poisson with a mean that depends on the interval \((t_a, t_b]\) (Chapter 4 discusses Poisson processes in more detail). Typically, the inherent randomness in a stochastic process is modelled by well defined distributions, like the Poisson, which posses ‘nice’ properties (Lindsey, 2004). However, other distributions may be used.

For the purposes of this thesis, \(N(t)\) will denote a counting process for the number of events from \(t = 0\) up until time \(t\), otherwise the interval will be explicitly specified i.e. \(N(t_a, t_b)\). Note also that \(N(t_a, t_b) = N(t_b) - N(t_a)\).

The plot in Figure 2.1 shows the levels of river flow (in cubic metres per second) in each day for year 2005, in river Severn in the UK. This may be modelled by a continuous stochastic process such as the Wiener process. However, if one was interested in high river flows (i.e. floods), then the points above a threshold, say 160, could be considered and these can now be modelled by a counting process such as the Poisson process (points are shown as solid circles).

![Figure 2.1: Daily flow levels for river Severn in 2005](image-url)
2.1. Recurrent Event Modelling

For consistency, events will referred to as failures throughout the thesis since this is more relevant to water pipes.

A counting process is characterised through the intensity function \( \lambda(t) \) defined as the limit of the probability of one or more failures in a small time period divided by the length of that interval, and it is given by:

\[
\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(N(t, t + \Delta t) \geq 1)}{\Delta t} \quad (2.1)
\]

Under certain assumptions mentioned later, \( \lambda(t) \) is equivalent to the failure rate. The mean function of a counting process is defined as:

\[
\Lambda(t_a, t_b) = E[N(t_a, t_b)] \quad 0 \leq t_a < t_b
\]

and it is defined as the expected number of failures in the time interval \((t_a, t_b]\). As will be shown later, under certain conditions:

\[
\Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda(u)du
\]

For consistency, \( \Lambda(t) \) will denote the average number of occurrences in the interval \((0, t]\). When the lower bound is not 0, the interval will be explicitly given i.e. \( \Lambda(t_a, t_b) \). In addition, it is clear that \( \Lambda(t_a, t_b) = \Lambda(t_b) - \Lambda(t_a) \).

In the case of water pipes which are considered repairable systems (see Chapter 4 for details), \( \lambda(t) \) can be considered as a measure of reliability. A related quantity also reflecting reliability, is the rate of occurrence of failures (ROCOF) (Rigdon and Basu, 2000) which is defined as:

\[
\text{ROCOF} = \frac{d}{dt} \Lambda(t)
\]

Clearly, \( \lambda(t) \) and ROCOF seem to be describing the same thing, and in fact they are equivalent when no simultaneous failures are assumed possible (Thompson, 1981), since if the opposite were true then \( \Lambda(t) \) would not be continuous. For the purposes of this thesis we assume that simultaneous failures are not possible so that:

\[
\lim_{\Delta t \to 0} \frac{\Pr(N(t, t + \Delta t) \geq 2)}{\Delta t} = 0 \quad (2.2)
\]
and from here on $\lambda(t)$ will be referred to as the failure rate.

In the context of counting processes, it is also interesting to consider the distribution of the time intervals $T$ between failures, i.e. the lifetimes. One can then study quantities such as the survivor function:

$$S(t) = \Pr(T > t)$$

and the hazard function:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

which specifies the limit of the probability of failure in a small time interval given survival to the beginning of the interval. It is important here to distinguish between the hazard function, which can be considered as the instantaneous rate of failure, and the intensity function $\lambda(t)$ which is the ‘global’ failure rate.

### 2.1.2 Reliability Theory and Survival Analysis

There is an established fundamental duality between two different ways of viewing a point process (Thompson, 1988). One is to consider the random variable $N(t_a, t_b)$ introduced earlier which counts the number of failures in the arbitrary interval $(t_a, t_b]$. The other way is to consider the random amount of time during which a fixed number of failures will occur, for instance by looking at the times between each individual failure. The two ways are perfectly equivalent and is clearly a matter of which one is more convenient to use, a choice which will depend on the model specification itself as well as the application. Assuming failures occur at times $t_k$ where $k = 1, 2, \ldots$ then the aforementioned equivalence can be expressed as:

$$\Pr(N(t) \geq k) = \Pr(t_k \leq t)$$

In general, $N(t)$ is often used by reliability theory type applications whereas looking at the distribution of times between failures is favoured by survival analysis.
2.1. Recurrent Event Modelling

Therefore, the areas of research where recurrent event data arise frequently can be categorised as: survival analysis and reliability theory. Both these areas focus on modelling time to a certain failure. The differences between them are the applications and research questions that arise. Survival analysis relates to biological/medical issues where the event of interest is death or onset of a disease/clinical condition whereas reliability data come from applications such as engineering where time to failure of a component or piece of machinery is the subject of study.

Survival analysis concerns the study of the effect of certain treatments on survival of individuals. Usually, little attention is paid to the actual times between recurrences of a disease or illness and rarely is the actual structure of the system of much interest (Newby, 1993). A typical example of recurrent event data within survival analysis would be the study on the effect of a certain drug on the recurrence rate of tumours in humans/animals. Often the effect of the treatment is of interest rather than the underlying process causing the tumours. A simple but effective model for the instantaneous rate of occurrence of tumours or the hazard function is the proportional hazards model (Cox, 1972a):

\[ h(t, x) = h_0(t) \exp\{\beta x'\} \]  

(2.3)

where \( x = (x_1, \ldots, x_p) \) is a vector of covariates including treatment and \( \beta = (\beta_1, \ldots, \beta_p) \) is vector of associated parameters. In addition, \( h_0(t) \) is an arbitrary and unspecified baseline hazard function. This is a typical survival model since it is defined in terms of the hazard function which represents the instantaneous rate of failure (e.g. a tumour). The hazard function uniquely determines the distribution of inter-failure times and can be used to formulate the likelihood. Specifically, for the hazard in Equation (2.3), the corresponding survivor function is given by (Kalbfleisch and Prentice, 2002):

\[ S(t, x) = \exp \left\{ - \int_0^t h_0(u) \exp\{\beta x'\} du \right\} \]

so that the density function of time is given by:

\[ f(t, x) = h(t, x)S(t, x) \]
2.1. Recurrent Event Modelling

The clever thing about this well-established model is that $h_0(t)$ need not be specified and is considered a nuisance function so that it is integrated out and by definition only the partial likelihood model parameters are utilised. Intuitive modifications to this model do exist, such as making $h_0(t)$ different for each interval between recurrences. Here the main idea is that this model will give an indication as to whether the effect of the drug is significant or not.

Consider on the other hand, mechanical equipment which experience failures, e.g. stops functioning. Assuming that the equipment may be repaired to working condition, these failures can naturally be considered recurrences of the same event. Suppose also that certain covariates $\mathbf{x}$ that are known to affect the way this component fails are available and that the objective is to quantify the deterioration of this equipment over time. Similar to the example above, it is sensible to consider the failure rate which is a natural reliability measure and a possible model for that would be:

$$\lambda(t, \mathbf{x}) = \exp\{\theta t\} \exp\{\beta \mathbf{x}'\}$$

Assuming a Poisson counting process, the model is properly defined as:

$$N(t_a, t_b) \sim \text{Pois}(\Lambda(t_a, t_b))$$

$$\Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda(u, \mathbf{x}) du = \theta^{-1} e^{\beta \mathbf{x}'} [e^{\theta t_b} - e^{\theta t_a}]$$

Parameter $\theta$ is of interest which relates to the deterioration of the component given the effect of the covariates. If $\theta$ is close to zero, then the covariates are mostly responsible for the behaviour of $\lambda(t, \mathbf{x})$ whereas if $\theta$ is large and positive one may conclude that this piece of equipment tends to deteriorate rapidly with usage.

These simple examples are typical of the way that recurrent event data are modelled in the disciplines of survival analysis and reliability theory. The two fields of study have many overlapping features in terms of models and methodology. In the context of this thesis, the reliability theory paradigm is favourable since the particular application of underground water pipes is a typical reliability application.
Recent texts on survival analysis include the ones by Kalbfleisch and Prentice (2002) and Kleinbaum and Klein (2005) both of which contain chapters on analysing recurrent event data. Furthermore, the books by Meeker and Escobar (1998) and Wolstenholme (1999) provide a good overview on reliability modelling and give an idea of how recurrent event data usually arise and how they are commonly treated.

2.2 Counting Processes in Reliability

Statistical methods applied to reliability have progressively advanced and are now well established. Some useful papers include the one by Lawless (1983) who provides an account of the advances in statistical reliability up to that time and dedicates a small section to recurrent event data emphasising the lack of publications in that respect. It was around that time that more work evolved on stochastic processes applied to recurrent event data in reliability. Thompson (1981) discusses counting process models in reliability and criticises the assumptions of each particular model considered. Since then, many models have been developed in reliability, based on counting processes (Ansell and Phillips, 1989; Newby, 1993; Singpurwalla, 1995; Pena, 2006; Krivtsov, 2007).

For the remainder of this section, some of the most established counting process models within reliability are introduced with associated references, as well as discussions concerning the underlying assumptions for each one.

2.2.1 Renewal Processes

An important concept when dealing with occurrences of failures in reliability is that of a repairable versus a non-repairable system. A non-repairable system is one where after a failure has occurred the system stops functioning and either cannot be repaired (e.g. light bulb) or the cost for repairing is too great (this is analogous to death in survival analysis). In a non-repairable system, interest generally lies
2.2. Counting Processes in Reliability

in the time until failure. In contrast, repairable systems can be brought back to working condition after a failure has occurred so that other failures may possibly follow. Recurrent event data may be then naturally associated with repairable systems.

An issue that arises in studies of repairable systems is the degree to which the system may be repaired. A perfect repair for example means that the system is brought back to the same working condition as before (this is sometimes referred to as the “good-as-new” assumption). In that case, the times between subsequent failures must necessarily be identically distributed. If in addition these times are also independent, then this defines a renewal process which is effectively a time homogeneous process where the reliability of the system stays the same through time given the repairs.

Renewal processes are simple in their assumption of identically and independently distributed (i.i.d.) inter-failure times which makes them easy to work with. A particularly useful renewal process is the homogeneous Poisson process (HPP) where the times between failures are exponentially distributed, the failure rate \( \lambda(t) = \lambda \) is constant in time and the number of failures in any time interval \((t_a, t_b]\) follows a Poisson distribution with mean \( \Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda du = \lambda t_b - \lambda t_a. \)

2.2.2 Non-Homogeneous Processes

Ascher (1968) introduced a model whose reliability is allowed to decrease with time, calling this the “bad-as-old” concept. The idea is that the inter-failure times are dependent in time (but not on the failures themselves).

Perhaps the most well-established example of a model which involves the bad-as-old property is the non-homogeneous Poisson process (NHPP) which in contrast to the HPP has intensity function \( \lambda(t) \) which depends on time but still retains the property that the number of failures in an arbitrary time interval follow a Poisson
2.2. Counting Processes in Reliability

distribution.

A natural extension to renewal processes would be to drop the assumption of identical distributions for the times between failures, to obtain the model of a system whose reliability changes after each failure. Sen and Bhattacharyya (1993), for example, utilise a piecewise exponential model where the inter-failure times follow independent exponential distributions with a mean that depends on the time interval itself. The difference between this model and the NHPP is that it assumes constant reliability along the time intervals between the failures whereas the NHPP assumes that the reliability changes continuously. In a similar way of thinking, Zequeira and Valdes (2001) consider a model where the failure rate is constant in each interval but constrained to have a value greater than its value in previous intervals.

A generalisation of both renewal processes and the NHPP is the modulated renewal process proposed by Cox (1972b) where the intensity function is given (Rigdon and Basu, 2000) by:

\[
\lambda(t) = f(B_t) \exp\{\beta_1 g_1(t) + \cdots + \beta_p g_p(t)\}
\]

where \( f() \) is a function of the backward recurrence time \( B_t \) (time since the most recent occurrence). The \( g_i(t) \) are functions of time \( t \) and \( \beta_i \) are parameters to be estimated. When all \( \beta_i \) are zero then \( \lambda(t) \) depends only on the time of the last failure this is a renewal process. If \( f() \) is a constant and \( g_i(t) \) are known functions then the process is a NHPP.

2.2.3 Processes with Dependency Structures

Non-homogeneous processes can account for any relationship that the failure mechanism may have with time but do not allow for any kind of dependency between the failures themselves. A particular example of such behaviour might be that some (but not necessarily all) of the repairs are inadequate causing a number of subsequent failures, so that the overall process would appear to have occurrences
2.3. Recurrent Failures in Water Pipe Networks

clustered together. A model that can capture such behaviour is the so called
branching Poisson process (BPP), introduced by Bartlett (1963), which assumes
that ‘primary’ failures occur as a Poisson process with constant failure rate. After
each primary failure, a renewal process is evoked with some probability \((1 - p)\)
which causes a number of subsequent failures (that number is modelled by a dis-
crete random variable). Note that this assumes that the probability of a ‘proper’
repair \(p\) is always the same. Interestingly, Hansen and Ascher (2002) utilise a
modified BPP to model recurrent unsuccessful repairs of the same failure in a new
automobile.

Another very interesting model, is the doubly stochastic Poisson process (Cox and
Isham, 1980) which in essence is a Poisson process whose intensity function is
governed by another stochastic process. A typical example of such model is the
Markov modulated Poisson process (MMPP) where the intensity function depends
on a discrete Markov chain. Scott and Smyth (2003) use an MMPP to model click
rate data for individual computer users browsing through the world wide web. The
same idea of modulating a process with a Markov chain to formulate a model for
failure data is considered in Özekici and Soyer (2003) where the ‘mother’ process
is a Bernoulli. Allowing a Poisson process to depend on another stochastic model,
results in a complex process with an implicit dependency structure.

Many models involving stochastic processes have been developed within reliability
that can effectively capture the behaviour of repairable systems and so far only
a selection of these models have been presented. This section has outlined some
concepts and assumptions about the nature of the failure processes that need to
be taken into account when modelling recurrent failures in repairable systems.

2.3 Recurrent Failures in Water Pipe Networks

After introducing basic concepts about recurrent failures in reliability, the discus-
sion can now be directed towards the specific application of water pipe failures.
Having established the role of counting processes in recurrent event modelling, the objective in this section is to investigate existing work on water pipe failures involving such models.

Underground water pipes inevitably experience a number of failures over their useful lifetime meaning that these need to be repaired after each failure (either burst or blockage) until it is deemed necessary to replace them. In that sense, water pipes are considered a typical example of a repairable system. Clearly, the cost of repair is less than the cost of replacement unless the frequency of failures is too high. Therefore, predictive models for the failure mechanism of pipes is vital to water companies.

Statistical methodology has been frequently used to model recurrent failures in water pipes. A considerable proportion of such models relates to counting processes or equivalently to modelling times between failures. Some key review articles include the ones by Kleiner and Rajani (2001b) who provide an up to date detailed overview of statistical models for pipe failures and Loganathan et al. (2002) who give a detailed account of statistical pipe replacement analyses. In addition, Engelhardt et al. (2000) review rehabilitation strategies for distribution networks from a UK perspective and also provide a reasonable summary of statistical models that attempt to capture the failure mechanism.

A key point identified in the literature is that the ideal modelling technique would be one that simultaneously describes the physical mechanisms in the pipes and quantifies the accumulated uncertainties in the failure process. The literature may be loosely divided in two mainstreams of models: regression methods and counting processes.

An early example of modelling the rate of failure against time was considered by Shamir and Howard (1979) in a Normal regression model where the number of pipe breaks per unit length per year, was regressed against an exponential function of time and pipe age. The model was later on extended by Walski and Pellicia
2.3. Recurrent Failures in Water Pipe Networks

(1982) to include factors that allow for differences in pipes due to diameter and also take into account previous failures. Work has also been done with linear models assuming linear relationship with time, see for instance Kettler and Goulter (1985) and references within Kleiner and Rajani (2001b).

Giustolisi and Savic (2006) and Giustolisi et al. (2006) developed a Genetic Algorithm technique called evolutionary polynomial regression (EPR) which seeks the best subset of polynomial formulae that describe the mean of a Normal linear model. This technique is effective for finding the optimal way which covariates may be affecting the mean of continuous responses such as failure frequency.

Another interesting piece of work which considers regression methods is the one by Clark et al. (1982) where the authors model two response variables, namely the time until the first failure and the total number of failures after the first one has occurred.

A modelling framework that has often been adopted by many researchers for the statistical modelling of water pipes is one that involves counting processes which can naturally describe the occurrence of failures in time by modelling the failure rate $\lambda(t)$. This explicitly allows for temporal structures and possible effects from external factors in the form of covariates.

Andreou et al. (1987) propose a very interesting model in which the failure process is divided in two separate stages, by combining a typical survival analysis model and a counting process. The idea is that in the early stages in the lifetime of pipes, times between failures are modelled using a proportional hazards model which is adjusted to accommodate the observed fact that the time to next failure becomes shorter after each failure. The cut-off point of the two stages is right after the third break, where they assume that failures occur according to a Poisson process with a constant rate of failure.

Mailhot et al. (2000) consider a model where the time until the first failure follows a Weibull distribution while the rest of the inter-failure times are identically
2.3. Recurrent Failures in Water Pipe Networks

exponentially distributed. The model is applied to a network consisting of 2096 pipe segments which are assumed homogeneous (pipe failures occur with common rate). In addition, the authors consider situations where there is an unknown time period between pipe installation date and the start of the observation period, so they develop a methodology which uses the observed data to interpolate the past, given the assumed model. Dridi et al. (2005) use a Bayesian formulation of this model for two hypothetical data sets.

In an effort to try and describe the behaviour of underground water pipe failures in the city of Trondheim in Norway, Lei and Saegrov (1998) considered an accelerated lifetime model which models the logarithm of time to failure $T$ as:

$$\log(T) = \beta x' + Z$$

where $\beta x'$ a linear combination of covariates $x$ and $Z$ follows a log-Weibull distribution. The idea is to capture the effect of $Z$ which will either accelerate or decelerate the time to failure.

A time dependent renewal process was utilised by Gat and Eisenbeis (2000) to describe the occurrence of failures in a network of 1212 pipes. Heterogeneity between pipes was allowed through covariates such as pipe length, soil type and, perhaps more importantly, the actual number of previous failures (NOPF). Inter-failure times follow an extreme value distribution which is different for each interval based on the NOPF.

In a paper by Goulter et al. (1993), the NHPP is used to model the recurrence of failures in the sense that the failure rate is a function of both time and distance from the previous failure. The data considered, comes from a Canadian city and it was observed that the empirical probability of pipe-breaks was highest within a short time and distance from a previous break. The authors used an appropriate parametric form for the failure rate so that when integrated over space and time, it provided a well defined form for the mean number of failures in a given time interval and distance. Since the process is NHPP, this mean number of failures
2.3. *Recurrent Failures in Water Pipe Networks*

in time and space is the mean of a Poisson distribution and so the corresponding likelihood is used for model fitting.

Saldanha et al. (2001) consider a NHPP for failures in service water pumps in a nuclear plant where they model the failure rate with both a power law and a log-linear formulation (discussed later in Section 4.1.1). The NHPP is also utilised by Kleiner and Rajani (2008) where they consider both pipe dependent and time dependent covariates in the characterisation of the failure rate. In addition, a Bayesian random effects NHPP with a power law characterisation of \( \lambda(t) \) is applied to a network of underground water pipes in New Zealand by Watson (2005) and the fitted model is used in conjunction with cost functions to present a feasible pipe replacement policy. More recently, Park et al. (2008) have also considered a global (i.e. same failure rate assumed for all pipes) power law and log-linear based NHPP model for a network of 9649 pipes observed over a 70 year period.

It is also worth mentioning that some work involving stochastic process models for pipe failures, includes Markov processes. The idea is that the states of a semi-Markov chain represent deterioration states of water pipes, see for instance (Li and Haims, 1992). Other semi-Markovian models and relevant references are mentioned in Kleiner and Rajani (2001b).

From the above review it is clear that counting processes (modelling the failure rate) provide a natural modelling framework for recurrent failures in underground water pipes (and also in many similar repairable systems). The NHPP is identified as a particularly useful model for many reasons. First, it has desirable properties, e.g. the number of failures in any time interval follows a Poisson distribution. Second, the NHPP will by definition naturally allow for ageing (deterioration) which is an inherent problem of water pipes. Also, as will be discussed in Chapter 4, there is much flexibility in modelling the failure rate \( \lambda(t) \) of a NHPP, which in turn allows a lot of flexibility in choosing a suitable model to reflect known physical behaviour.
2.4 Summary

Applications concerning recurrent event data are very common within survival analysis and reliability theory. The two disciplines use different modelling techniques which are nevertheless based on exactly the same concepts. The particular application in this thesis concerns water pipe failures, which is typical of reliability theory and a discussion on recurrent events within reliability identified counting process models as a natural framework.

Furthermore, it was shown that much of the literature on statistical modelling of water pipe failures involves counting process models and in particular the NHPP seems to be favoured by many authors. The plan for the remainder of this thesis is to adopt and explore the idea of using NHPP models for recurrent pipe failures. Specifically Chapters 4-7 investigate flexible Bayesian models based on the NHPP. In the next chapter we introduce an illustrative data set which will be used in subsequent analyses.
Chapter 3

Data

In this chapter a data set which will form the basis for investigations in subsequent chapters is presented and explored. The data set refers to a network of 1349 underground water pipes in a large North American town. For confidentiality reasons, little can be said about the geographical location or origin of the data, but a full description can be given as to the content of the data.

3.1 Basic Summary Analysis of the Data

All 1349 pipes in the network are made of the same material, namely cast iron, which means that the pipes are expected to have relatively similar behaviour in terms of deterioration. The times of failures, which for this data refer to pipe bursts, are given in individual months for every year. All pipes were observed until the end of 2003, however the start of the observation periods differ. Specifically, each pipe in the network was observed since installation but not all pipes were installed at the same year. Therefore at the end of 2003, the oldest pipe in the network has been observed for 58 years while the youngest has been observed for 43. A plot of pipe age is shown in Figure 3.1.
3.1. Basic Summary Analysis of the Data

A summary of pipe failures is given in Table 3.1.

Table 3.1: Network description

<table>
<thead>
<tr>
<th>Number of pipes</th>
<th>1349</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of failures</td>
<td>5425</td>
</tr>
<tr>
<td>Earliest failure on record</td>
<td>1962</td>
</tr>
<tr>
<td>Latest failure on record</td>
<td>2003</td>
</tr>
<tr>
<td>Earliest installed pipe</td>
<td>1945</td>
</tr>
<tr>
<td>Latest installed pipe</td>
<td>1960</td>
</tr>
</tbody>
</table>

In addition, a histogram of the total number of failures for each pipe is shown in Figure 3.2 where a (kernel) density estimate is also plotted. The distribution is heavily skewed with considerable mass around zero but also with some quite high extreme values. Out of the 1349 pipes, 297 (∼22%) have never experienced failure during their lifetime.

As an additional covariate, the length of each pipe in metres is available. Pipe length has been identified as a very influential variable to the failure mechanism by many authors (e.g. Boxall et al. (2007)). Kleiner and Rajani (2007) considered modelling a homogeneous set of pipes (same material, diameter, vintage, cathodic...
3.1. Basic Summary Analysis of the Data

Figure 3.2: Histogram of total number of failures per pipe

(protection) and have identified pipe length to have a strong relationship with the number of failures so that it may be used as a surrogate for exposure. A plot of failure counts per pipe against length is shown in Figure 3.3. As expected, there seems to be a positive relationship between the two since a longer pipe is more exposed to external factors affecting the failure process.

Figure 3.3: Failure counts vs pipe length
3.1. Basic Summary Analysis of the Data

Furthermore, Figure 3.4 shows the total number of failures in each month for the whole network in the time period 1945-2003 (note that the first failure in the network occurred in 1962). In terms of failures, the total number rises up to about mid 1980s where the number seems to drop. A scatter plot smoother (kernel regression) is also plotted, in an effort to describe the overall shape which confirms that the total number of failures in each month drops slightly towards the end of the observation period. This drop at the end is somewhat surprising with regards to the nature of water pipes since at best, the overall failure rate would be expected to stay constant. There may be some unknown feature hidden in the data such as some kind of maintenance performed to the network. On the other hand this may just be due to the natural variation in the failure process. Summary statistics of monthly failures are given in Table 3.2.

![Figure 3.4: Monthly number of failures vs month (1962-2003)](image)

Table 3.2: Summary statistics for monthly failures

<table>
<thead>
<tr>
<th>Statistic</th>
<th>mean</th>
<th>median</th>
<th>s.d.</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly Failures</td>
<td>7.66</td>
<td>3.0</td>
<td>10.38</td>
<td>0</td>
<td>57</td>
</tr>
</tbody>
</table>

It would also be sensible to look at an indicative measure for the failure rate and in this case, a reasonable quantity would be the number of failures per month. Figure 3.5 plots failure counts per month against each pipe. The variation in the plot is an indication of how different the failure mechanism for each pipe is.
3.2 Tests for temporal trends

When dealing with the issue of modelling data involving recurrent failures, it is always useful to decide beforehand whether there is any evidence that the failure rate (referred to as $\lambda(t)$ in Chapter 2) is varying with time or if it is constant. For instance, the use of a NHPP model instead of the conventional HPP is more reasonable if indeed the failure rate depends on time.

Many tests for temporal trends are available and most can be found in Bain et al. (1985). Furthermore, Vaurio (1999) presents tests on whether any apparent trend in $\lambda(t)$ is monotonic or not. For the purposes of this thesis, the Laplace test for temporal trend (Wolstenholme, 1999) is considered.

Briefly, if we assume that $n$ failures have been observed at times $t_1, t_2, \ldots, t_n$ then conditional on $t_n$, the times $t_1, \ldots, t_{n-1}$ are order statistics from a Uniform distribution $U(0, t_n)$ if the failure rate is constant. If indeed no trend is present, the test statistic:

$$T = \frac{\sum_{j=1}^{n-1} \frac{t_j}{n-1} - \frac{t_n}{2}}{t_n \sqrt{\frac{1}{12(n-1)}}}$$

approximately follows a $N(0, 1)$ distribution under the null hypothesis of no trend ($H_0$) and can therefore be used to test $H_0$ against the alternative hypothesis of
there being a temporal trend.

Performing the Laplace test in the data set presented here, identified 82 pipes for which the null hypothesis can be rejected at the 95% level (only 865 pipes out of 1349 could be tested in the sense that they have failed two or more times in total). The tests for each pipe are plotted in Figure 3.6 where for reference two lines are plotted at $T = \pm 1.96$ reflecting the 95% critical region of the standard Normal distribution.

![Figure 3.6: Laplace test for 865 pipes](image)

3.3 Summary

From these exploratory analyses, it is clear that the failure mechanism varies considerably from pipe to pipe and the modelling framework should take this into account. The explanatory variable length should prove useful in a statistical model since it appears positively correlated with the failure count for each pipe.

As was mentioned earlier, data availability is an issue when modelling water pipe failures. Although failure times exist since the installation date for each pipe, information on potentially useful variables such as pipe diameter, ground pressure
and soil type, is not available. Furthermore, there appears to be a considerable number of pipes that have never failed. This may cause issues with models that attempt to capture the behaviour of individual pipes since no actual information exist relating to the failing mechanism.

In the following chapter, the NHPP is investigated as modelling framework and specifically a mixed effects NHPP model is formulated and applied to the data set discussed here.
Chapter 4

Mixed Effect NHPP Models

As discussed in Chapter 1, the general theme in this thesis is the statistical modelling of recurrent events with a specific focus on water pipe failures. In Chapter 2 it was argued that the process driving such failures has an element of randomness and counting processes are a natural choice for modelling this type of situation. Specifically the NHPP was introduced as a flexible model that offers the opportunity of explicitly incorporating and characterising the non-linear relationship between failure rate and time as well as allowing for the inclusion of covariates.

In this chapter, we investigate a NHPP model for a set of pipes and specifically incorporate random effects to allow for heterogeneity between them. The model is applied to the North American data presented in Chapter 3.

4.1 Non-Homogeneous Poisson Process Models

As discussed in Chapter 2, a counting process may be described as a random variable $N(t_a, t_b)$ for $0 \leq t_a < t_b$ which represents the count of failures in time period $(t_a, t_b]$. Recall also that $\Lambda(t_a, t_b)$ is the mean number of failures in the time interval $(t_a, t_b]$ so $\Lambda(t_a, t_b) = E[N(t_a, t_b)]$. The HPP and the NHPP are specific
counting processes where $N(t_a, t_b)$ is Poisson distributed.

Formally, a counting process $\{N(t), t \geq 0\}$ with intensity function $\lambda(t)$ is a NHPP (Cook and Lawless, 2007) if:

- $N(0) = 0$

- For any intervals $(t_a, t_b]$ and $(t_c, t_d]$ such that $t_a < t_b \leq t_c < t_d$, the random variables $N(t_a, t_b)$ and $N(t_c, t_d)$ are independent. This is often termed the independent increment property

- The number of failures in a time interval $(t_a, t_b]$ follows a Poisson distribution with mean $\Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda(u) du$ so that:

$$
\Pr(N(t_a, t_b) = m) = \frac{\exp\{-\Lambda(t_a, t_b)\}\Lambda(t_a, t_b)^m}{m!} \quad (4.1)
$$

The main difference between the NHPP and the HPP is that the HPP has stationary increments. This means that for the HPP, the $\Pr(N(t_a, t_b) = m)$ is independent of time and simply depends on the length of the time interval.

As a result of the above defining properties of the NHPP, the distribution of times between failures depends on the time that the immediate previous failure has occurred (Pulcini, 2001). So the conditional probability distribution of the $k^{th}$ failure time given that the previous failure occurred at $t_{k-1}$ is:

$$
f(t_k | t_{k-1}) = \lambda(t_k) \exp\left\{- \int_{t_{k-1}}^{t_k} \lambda(u) du \right\} = \lambda(t_k) \exp\{-\Lambda(t_{k-1}, t_k)\} \quad (4.2)
$$

### 4.1.1 Failure Rate

A typical repairable system that experiences ageing will have the characteristic that during the early part of its life it will fail more than expected due to defects or installation errors. This period is sometimes called the burn-in or the debugging period. After that, the ‘useful’ life of the pipe begins where the failure rate is
approximately constant and smaller than that of the early failures period. Towards the end of its lifetime, ageing has a significant effect on deterioration and the failure rate increases sharply. As a result, the overall shape of the failure rate curve resembles a bathtub. Figure 4.1 depicts the typical shape of a bathtub-shaped $\lambda(t)$.

![Bathtub-shaped failure rate function](image)

Figure 4.1: Bathtub-shaped failure rate function

In theory, the failure rate of an underground water pipe would be expected to have a bathtub shape given its nature as a typical repairable system. Pragmatically, one would expect the failure rate to be less simple than that, since water pipes are buried underground for many years rendering it difficult to observe properly in terms of collecting data. A simple scenario may be one where a pipe has only been observed during the later years of its useful lifetime so that only the deterioration phase of the curve needs to be modelled. A more complicated picture may involve external unobserved processes which have an effect on $\lambda(t)$ giving it a completely unexpected shape (this point is discussed in more detail in Chapter 7).

As far as the literature is concerned, there are two parametric models for the failure rate of the NHPP which have dominated: the log-linear model and the power law model. Lee (1980) gives a good description of both and proposes a test for their adequacy based on a model which generalises both, whereas Coetzee (1997) offers a very good review and classical inference for both models. It is worth noting
that both of these models, in their simplest form, are monotone with respect to time. This means that each of the two models can effectively characterise each (but only one) of the three phases in the bathtub curve. In addition, they are capable of capturing the rare situation where a system may be improving with time, i.e. decreasing $\lambda(t)$.

The log-linear model, which is sometimes referred to as the exponential model, was initially discussed by Cox and Lewis (1966) and as its name suggests, it relates $\lambda(t)$ and time $t$ with an exponential function:

$$\lambda(t) = \alpha e^{\psi t} \quad t \geq 0; \alpha > 0$$

Parameter $\psi$ controls the rate of increase or decrease, depending on its sign and $\alpha$ represents the failure rate at $t = 0$. When $\psi = 0$ the rate is constant so that the process can become HPP if necessary. Figure 4.2 shows some instances of the log-linear $\lambda(t)$ with arbitrary choices for $\alpha$ and $\psi$. This model is quite effective,

![Figure 4.2: Examples of log-linear failure rate](image)

when used in conjunction with the NHPP, in representing an ageing system which deteriorates quite rapidly. For some applications of log-linear NHPP models see Lindsey (1995) and Shin et al. (1996). Possible covariates $x = (x_1, x_2, \ldots, x_p)$
maybe be incorporated as a scaling effect on $\lambda(t)$:

$$\lambda(t; x) = \alpha e^{\psi t} e^{\beta x'}$$

(4.3)

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$.

The other popular choice for $\lambda(t)$ is the so called power law model which is sometimes also known as the Weibull model or the Duane model. A NHPP with power law intensity is also often referred to as the Power Law Process (PLP). As the name suggests, this model describes $\lambda(t)$ by raising time to a power:

$$\lambda(t) = \gamma \theta t^{\theta - 1} \quad t \geq 0; \gamma > 0, \theta > 0$$

(4.4)

In this model, $\gamma$ is a scale parameter and $\theta$ is a shape parameter that controls the shape of the curve. $\theta = 1$ implies that the rate is constant (HPP); $\theta > 1$ means that the function is increasing and $\theta < 1$ gives a decreasing $\lambda(t)$. By definition, $\lambda(t) = 0$ at $t = 0$. Figure 4.3 depicts three examples for values of $\theta$ around 1. Covariates may be included in the same manner as for the log-linear model.

![Figure 4.3: Examples of power-law failure rate](image)

It is worth mentioning that when using a power law $\lambda(t)$ for the NHPP, the distribution of time to the first failure is Weibull and the distributions of inter-failure times (Equation (4.2)) are truncated Weibull distributions where the truncation...
occurs at the time of the immediate previous failure. For that reason, the NHPP with a power law intensity is also referred to as the Weibull process. Some examples of Bayesian NHPP models with a power law intensity can be found in Sen (2002), Yu et al. (2007) and Kim et al. (2008). Classical procedures for model fitting and prediction are considered by Lee and Lee (1978), Engelhardt and Bain (1978) and Rigdon and Basu (2000) while Crow (1982) considers computation of confidence intervals for the power-law NHPP.

In general, the form of $\lambda(t)$ is not restricted to these two parametric models and the choice for it quite flexible. A paper by Lawless (1987) considers the NHPP with an intensity function $\lambda(t) = \lambda_0(t) \exp\{\beta x'\}$ where $\lambda_0(t)$ is either parametric or semi-parametric. The author also discusses inclusion of random effects in terms of a random intercept in $\beta x'$. Landers et al. (2001) use the semi-parametric proportional intensity model introduced by Prentice et al. (1981) and compare it to a log-linear NHPP. Other examples include Pulcini (2001) who uses a superposition of two power law functions to capture the complete bathtub shape; Ryan (2003) who explores a NHPP with a time dependent mixture intensity function and Krivtsov (2007) who presents a family of parametric models to represent $\lambda(t)$.

### 4.1.2 Likelihood Function

One straightforward way to formulate the NHPP likelihood is by considering the distributions $f(t_k|t_{k-1})$ of the time intervals between each recurrent failure, given in Equation (4.2). Assume that a single water pipe has been observed in the time period $(T_0, T_{\text{end}}]$ having experienced $n$ failures at times $T_0 < t_1, t_2, \ldots, t_n \leq T_{\text{end}}$. $T_0$ is not necessarily the time of installation of the pipe but the starting time of observation. One can then divide the observation period into intervals of single failures, i.e. $(T_0, t_1], (t_1, t_2], \ldots, (t_n, T_{\text{end}}]$. If the sampling process is failure truncated then $t_n = T_{\text{end}}$ and the last interval will be $(t_{n-1}, t_n]$ whereas if it is time truncated then $t_n < T_{\text{end}}$ and the last interval $(t_n, T_{\text{end}}]$ will contain no observed failures. In the failure truncated case, the likelihood of a NHPP with failure rate
\[ \lambda(t; \theta) \text{ is derived by simply multiplying the conditional inter-failure distributions:} \]

\[
L_{PP}(T_0, t_1, \ldots, t_n; \theta) = \prod_{j=1}^{n} \lambda(t_j; \theta) \exp\{-\Lambda(T_0, t_n; \theta)\} = \prod_{j=1}^{n} \lambda(t_j; \theta) \exp\{-\Lambda(T_0, t_n; \theta)\} \]  

(4.5)

where \( \theta \) is a vector of parameters in \( \lambda(t; \theta) \). In the time-truncated case, we need to multiply the failure truncated likelihood (4.5) by the probability of zero failures in the interval \((t_n, T_{end})\). Using the Poisson property of the NHPP in Equation (4.1) this probability is simply given by \( \exp\{-\Lambda(t_n, T_{end}; \theta)\} \) so the time-truncated likelihood is:

\[
L_{PP}(T_0, t_1, \ldots, t_n, T_{end}; \theta) = \prod_{j=1}^{n} \lambda(t_j; \theta) \exp\{-\Lambda(T_0, t_n; \theta)\} \exp\{-\Lambda(t_n, T_{end}; \theta)\} = \prod_{j=1}^{n} \lambda(t_j; \theta) \exp\{-\Lambda(T_0, T_{end}; \theta)\} \]  

(4.6)

Note that in the specification of the NHPP, a failure is not allowed to occur at \( T_0 \) and for the purposes of the thesis, any interval that we consider is left-open so that a failure cannot have occurred at the beginning of it.

Furthermore, if one is modelling more than one pipe, the situation may arise where no failures were recorded for a specific pipe in the observation period. In that case, the likelihood of the NHPP simply becomes the probability of zero failures in \((T_0, T_{end})\) which is given by:

\[
\Pr(N(T_0, T_{end}) = 0) = \exp\{-\Lambda(T_0, T_{end}; \theta)\} \]

For completion purposes, note that it is possible to derive the likelihood by considering the random variable \( N(t_a, t_b) \) which is Poisson distributed with mean \( \Lambda(t_a, t_b) \). Consider the number of recurrences \( y_1, y_2, \ldots, y_n \) in the non-overlapping intervals \((t_0, t_1], (t_1, t_2], \ldots, (t_{n-1}, t_n]\) where for clarity \( t_0 = T_0 \) and \( t_n = T_{end} \). From the independent increment property of the NHPP, the likelihood is then given by mul-
4.1. Non-Homogeneous Poisson Process Models

Multiplying the Poisson probability of each $N(t_a, t_b)$:

$$L(y_1, \ldots, y_n; \theta) = \prod_{j=1}^{n} \frac{(\Lambda(t_{j-1}, t_j; \theta))^{y_j} \exp\{-\Lambda(t_a, t_b; \theta)\}}{y_j!} \tag{4.7}$$

Now as the number of intervals increases and therefore the size of each interval tends to zero, one can use a limiting argument (Meeker and Escobar, 1998) to show that the likelihood in Equation (4.7) is approximately given by Equation (4.6).

4.1.3 Left Truncation

In general, left truncation is quite common in data sets involving recurrent failures arising from repairable systems. In the water pipes context, data are often limited to only a few years in relation to the age of a pipe (Gat and Eisenbeis, 2000), especially in the UK, where only recently interest has grown in collecting data on pipe failures. However, many pipe networks have been in existence for much longer.

Specifically, left truncation here refers to a situation where a pipe has been observed in the interval $(T_0, T_{end}]$ and has failed at times $(t_1, \ldots, t_n)$ within that interval but has actually been installed at a time $T_{ins}$ so that $T_{ins} < T_0$. So there exists a period $(T_{ins}, T_0]$ in which there is no knowledge about how many failures have occurred and when. In that case, one can make use of the only information available, that at least one failure has occurred in $(T_{ins}, t_1]$ (where $t_1 =$ time of first failure) and utilise the Poisson property of the NHPP to derive $\Pr(N(t_{ins}, t_1) \geq 1)$, i.e. the probability of at least 1 failure in that time interval:

$$\Pr(N(t_{ins}, t_1) \geq 1) = 1 - \Pr(N(t_{ins}, t_1) = 0) = 1 - \exp\{-\Lambda(T_{ins}, t_1)\}$$

The likelihood in Equation (4.6) can then be modified to include left truncation:

$$L_{PPT}(T_{ins}, t_1, \ldots, t_n, T_{end}; \theta) = \left(1 - e^{-\Lambda(T_{ins}, t_1; \theta)}\right) \left[\prod_{j=2}^{n} \lambda(t_j; \theta)\right] e^{-\Lambda(t_1, T_{end}; \theta)} \tag{4.8}$$

By using Equation (4.8) the model is able to extrapolate the behaviour of the failure
4.1. Non-Homogeneous Poisson Process Models

process during the unobserved period taking into account the observed data set. This is then particularly useful when the unknown history of a pipe is of interest.

4.1.4 Random Effects

So far, the NHPP model has been formulated with a general failure rate \( \lambda(t) \) for a single pipe. However, water distribution networks usually consist of a collection of pipes which means it is more sensible to consider a model which accommodates that scenario. Suppose that a network consists of \( N \) pipes and that each pipe \( i = 1, 2, \ldots, N \) has been observed to fail at times \((t_{i,1}, t_{i,2}, t_{i,n_i})\) in the time period \((T_{i,0}, T_{i,end})\). The likelihood for a single pipe is still given by Equation (4.6) introducing an \( i \) where appropriate whereas the likelihood for the whole network, assuming no correlation structure between the pipes, is given by multiplying each of the \( N \) likelihoods:

\[
\prod_{i=1}^{N} \left\{ \prod_{j=1}^{n_i} \lambda(t_{i,j}; \theta, x_i) \right\} \exp \left\{ -\Lambda(T_{i,0}, T_{i,end}; \theta, x_i) \right\}
\]

(4.9)

where \( x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p}) \) is a vector of possible covariates for each pipe. The parameters \( \theta \) in \( \lambda(t_{i,j}; \theta, x_i) \) are common for all pipes and the only difference between the failure rate for each pipe comes from the covariates. Effectively, this model formulation is assuming a homogenous set of pipes whose failure mechanism is driven by the same NHPP and any variability that may exist between pipes is left to be explained by the covariates and age of each pipe.

This kind of assumption is in many cases undesirable since even pipes of the same material are pragmatically expected to be heterogenous because of external uncontrolled or unobserved processes acting on them. A lorry, for instance, that is passing on a road designed for lighter vehicles may indeed affect the failure process of one or more pipes in a network. Therefore it is sensible to allow for possible heterogeneity by using a model that distinguishes between each pipe (pipe specific model).
This kind of within subject variability is very efficiently handled with the use of random effects. The basic idea behind random effects, is that selected parameters of the model become pipe specific and therefore they effectively become random variables (Pawitan, 2001). Furthermore, in a random effects model, any parameter can be considered as a random variable or as a fixed value and in the cases where there are mixtures of both, models are termed as mixed effect models.

In the context of water pipes, it is clear that the NHPP model in Equation (4.9) would benefit from the inclusion of random effects since many pipes would in reality be expected to be heterogenous but also because analysis of the random effects may flag oddly behaving pipes. For example, better installation sometimes means that pipes that ought to behave similarly exhibit diverse behaviour in terms of failures. Boxall et al. (2007) clearly state that age alone is not enough to reflect the deterioration in the pipes and perform a study of important factors that may help in characterising the failure process more adequately. It turns out that many potential covariates are sometimes unavailable but their combined effect may be accounted for by the inclusion of random effects.

Consider for instance a NHPP process where \( \lambda(t_{i,j}; \theta, x_i) \) obeys the power law in Equation (4.4). A sensible idea would be to make the scale parameter \( \gamma \) pipe specific so that all pipes have the same ageing behaviour (common \( \theta \)) but with different magnitudes:

\[
\lambda_i(t_{i,j}; \theta, x_i|\gamma_i) = \gamma_i \theta t_{i,j}^{\theta-1} \quad \gamma_i > 0, \theta > 0
\]

In that case, a possible model for the random effects might be \( \gamma_i \sim \text{log-Normal}(0, \sigma^2) \) where \( \sigma^2 \) is an unknown parameter.

On the other hand, if we believed that some pipes may be deteriorating in a different way than others (e.g. pipes of the same material, diameter, age etc. may be affected differently by weather) then we could also force \( \theta \) to be pipe specific which is what Watson (2005) successfully implemented in his PhD thesis. For example, assuming that \( \theta_i \sim \text{Gam}(\alpha, \kappa) \) is a plausible model since we need \( \theta_i > 0 \).
Conditional on the random effects, the likelihood in Equation (4.9) remains structurally unaffected, with the exception that $\lambda(t_{i,j}; x_i)$ and $\Lambda(T_{i,0}, T_{i,end}; x_i)$ need to be replaced by $\lambda(t_{i,j}; x_i|\theta_i)$ and $\Lambda_i(T_{i,0}, T_{i,end}; x_i|\theta_i)$ where $\theta_i$ are the random effects.

Finally, the inclusion of random effects in a pipe specific model, allows for the incorporation of dependency structure between the pipes. Take for instance the example where the scale parameter was considered to be random by assuming that $\gamma_i \sim N(0, \sigma^2)$. One could assume a multivariate Normal instead, so that $\gamma_i \sim MVN(0, \Sigma)$ where the variance-covariance matrix $\Sigma$ models the dependency structure of the pipes.

### 4.1.5 Bayesian Framework and MCMC

The implementation of a mixed effects NHPP model is not trivial since it does not fall under a general category such as Generalised Linear Mixed Models (GLMM) which are nowadays included in many statistical software packages such as R (R Development Core Team, 2009). Techniques such as maximum likelihood and penalised likelihood (McCulloch and Searle, 2001) can be used to fit mixed effect models such as the one proposed here.

However, in this thesis, Bayesian Markov chain Monte Carlo (MCMC) methods (Gamerman, 1997) will be considered, in which random effects can be naturally incorporated in the sense that all parameters are random quantities. In addition, as will be discussed in Chapters 5-7, it is considerably easier in a MCMC framework, to consider further modifications to models such as some kind of mixture structure.

To reinforce this point, Pawitan (2001) appropriately states that theoretical analyses of random effects models can become complicated and cumbersome if at all possible so that one needs to rely on computer software to numerically fit such models and this is exactly what MCMC techniques are used for. Some well established
MCMC methods which serve the purposes of this thesis are given in Appendix A along with details about inference, model diagnostics, model comparison and prediction.

4.2 Mixed Effects NHPP model

Although many authors recognise the usefulness of pipe specific models, relatively little work exists in using mixed effects NHPP based models within the water pipe literature. In this section, the NHPP model discussed so far, is extended to a mixed effects version and studied within Bayesian MCMC framework. As this model cannot be directly fitted by most standard software, simulation studies are performed to assess model effectiveness. Furthermore, the model is applied to the water pipes data set introduced in Chapter 3.

4.2.1 Simulation Experiments

As discussed in Appendix A, the practical and user-friendly software package WinBUGS (Spiegelhalter et al., 2003) can be used to fit Bayesian mixed effect models using Gibbs sampling. The likelihood of the mixed effects NHPP model in Equation (4.9) is not built into WinBUGS, however the model can be implemented in a non-standard way using code given in Appendix B.

To verify that this non-standard model formulation works in WinBUGS, simulation experiments were conducted to fit a mixed effects NHPP model to a synthetic data set of water pipe failures. First random samples were generated from a NHPP in terms of failure times $t_1, t_2, \ldots, t_n$. Meeker and Escobar (1998) illustrate a convenient way of constructing samples from a NHPP with failure rate $\lambda(t)$ and mean function:

$$\Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda(u)du = \Lambda(t_b) - \Lambda(t_a)$$

For a monotone increasing $\Lambda(t)$ (which will be true for any $\lambda(t) > 0$), it can be
shown that the random variables $\Lambda(T_0, t_1)$ and $\Lambda(t_{j-1}, t_j)$ for $i = 2, \ldots, n$ are i.i.d. and follow an exponential distribution with mean 1, assuming that the start of the observation period is $T_0 = 0$. One can then use inverse transform sampling (Morgan, 1984), which is straightforward for the exponential distribution, to get the following equations:

$$
\Lambda(t_1) = -\log(u_1)
$$

$$
\Lambda(t_2) - \Lambda(t_1) = -\log(u_2)
$$

$$
\vdots
$$

$$
\Lambda(t_n) - \Lambda(t_{n-1}) = -\log(u_n)
$$

where $u_j \sim U(0, 1)$. Solving for $t_1, \ldots, t_n$ would then give the failure times of a NHPP observed up to $t_n$, the time of the last failure. Meeker and Escobar (1998) also provide a recurrence formula for the times $t_1, \ldots, t_n$:

$$
t_1 = \Lambda^{-1}(-\log(u_1))
$$

$$
t_j = \Lambda^{-1}(\Lambda(t_{j-1}) - \log(u_j)) \quad j = 2, \ldots, n
$$

which in the case of the power law and log-linear $\lambda(t)$ is easy to compute since $\Lambda^{-1}(t)$ is well defined.

50 NHPP processes (pipes) were simulated such that each is influenced by a single covariate $x_i$. For each one, the failure rate is based on the power law so that:

$$
\lambda(t; x_i | \theta_i) = \gamma \theta_i t^{\theta_i-1} e^{\beta_1 x_i} = \theta_i t^{\theta_i-1} e^{\beta_0 + \beta_1 x_i}, \quad i = 1, \ldots, 50
$$

(4.10)

Here, the random effects are modelled through the pipe specific $\theta_i$ but note that $\beta_0$ may also be random if necessary. To complete the model specification, the model for $\theta_i$ is given by:

$$
\theta_i \sim \text{Gam}(\alpha, \kappa)
$$

i.e. Gamma distributed with shape parameter $\alpha$ and rate parameter $\kappa$.

The recurrence formula for generating failure times $t_{i,j}$ for each process $i$ is then:

$$
t_{i,j} = \left[ t_{i,j-1}^{\theta_i} - \log(u_{i,j}) e^{\beta_0 + \beta_1 x_i} \right]^{1/\theta_i}, \quad j = 1, 2, \ldots
$$
where as before, $u_{i,j} \sim U(0, 1)$ and $t_{i,0} = 0$. The age of each pipe at present time is predetermined and set to be the end of the observation period $T_{i,\text{end}}$ so that only $t_{i,j} \leq T_{i,\text{end}}$ are collected. For convenience the time units of the simulations will be referred to as years. Furthermore, in order to test the ability of the model to perform in situations where some pipe history is unknown, left truncation is forced to the data by randomly predefining the start of the observation period for each pipe, i.e. $T_{i,0} \geq 0$, so we disregard any $t_{i,j} \leq T_{i,0}$. A particular instance of 10 such synthetic pipes is shown in Table 4.1. To test the predictive capability of the model, 10 years worth of data were generated for each pipe.

Table 4.1: Simulated data

<table>
<thead>
<tr>
<th>Pipe</th>
<th>$x_i$</th>
<th>$T_0$</th>
<th>Age ($T_{\text{end}}$)</th>
<th>Age at end of prediction</th>
<th>No. of failures, $n_i$</th>
<th>No. of failures in prediction period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76.11</td>
<td>41</td>
<td>78</td>
<td>40</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>72.67</td>
<td>30</td>
<td>71</td>
<td>33</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>53.03</td>
<td>43</td>
<td>87</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>48.28</td>
<td>0</td>
<td>95</td>
<td>28</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>40.81</td>
<td>59</td>
<td>87</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>21.74</td>
<td>33</td>
<td>49</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>46.57</td>
<td>25</td>
<td>99</td>
<td>11</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>53.44</td>
<td>21</td>
<td>56</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>70.03</td>
<td>26</td>
<td>88</td>
<td>41</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>50.72</td>
<td>0</td>
<td>87</td>
<td>11</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

The total number of failures in the observation period is denoted by $n_i$ so that the log-likelihood for pipe $i$ given $\theta_i$ is:

$$
\ell_i(T_{i,0}, t_{i,j}, T_{i,\text{end}}; \beta_0, \beta_1|\theta_i) = n_i \log(\theta_i) + n_i(\beta_0 + \beta_1 x_i) + (\theta_i - 1) \sum_{j=1}^{n_i} \log(t_{i,j}) + \left[T_{\theta_i} - T_{i,\text{end}}^{\theta_i}\right] \exp\{\beta_0 + \beta_1 x_i\} 
$$

(4.11)

and hence the log-likelihood for all 50 pipes is: $\sum_{i=1}^{50} \ell_i(T_{i,0}, t_{i,j}, T_{i,\text{end}}; \beta_0, \beta_1|\theta_i)$. The NHPP model with failure rate given in Equation (4.10) was implemented in
4.2. Mixed Effects NHPP model

WinBUGS where 2 Markov chains were run for 20000 iterations each. After a burn-in of 15000 and a thinning of 5, 2000 posterior samples were collected. The sample size was sufficient to ensure that the chains had converged (both tests described in Section A.1.4 were used) and that there were enough samples after the burn-in to ensure reasonable estimates. To test the reliability of the simulation studies, 50 different data sets were generated using the same parameter inputs and for each of those data sets the model was applied. Results were based on averaging over the fitted models.

\( \beta_0 \) and \( \beta_1 \) were given flat Normal priors as shown in Table 4.2, along with the uninformative Gamma hyperpriors given to \( \alpha \) and \( \kappa \). The actual input values, posterior means, standard errors and 95\% credible intervals (Cr.I.) are also given in the table. The posterior means are comparable to the input values and the Cr.I. are reasonably tight so that no problems are apparent in the MCMC implementation of the model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Actual Value</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>N(0, 1000)</td>
<td>-8</td>
<td>-7.71 (0.11)</td>
<td>[-7.893,-7.539]</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>N(0, 1000)</td>
<td>0.06</td>
<td>0.056 (0.001)</td>
<td>[0.053,0.057]</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Gam(0.5, 0.005)</td>
<td>60</td>
<td>64.51 (3.08)</td>
<td>[58.59,70.82]</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>Gam(0.5, 0.005)</td>
<td>40</td>
<td>44.24 (1.95)</td>
<td>[39.51,47.15]</td>
</tr>
</tbody>
</table>

Considering now the 10 year prediction period for each pipe, Figure 4.4 plots the total number of failures for each \( i = 1, \ldots, 50 \) in that interval, against the mean posterior predicted equivalent of that quantity. A 45\° line and 95\% prediction intervals are also plotted. The predicted values are taken as means from the posterior predictive distributions of failure counts and the intervals are calculated as the percentiles of samples from those distributions. The model predicts rather well as the points lie reasonably on the line with the exception of some counts with high
values where the model appears to be slightly overpredicting.

![Graph](image)

Figure 4.4: Actual vs predicted no. of failures

Summarising, the MCMC implementation of the mixed effects model appears to fit the simulated data adequately in terms of comparing parameter inputs and estimates. In addition, the predictive capability of the models was also adequate with the exception of minor overprediction for pipes with high failure counts. The most probable reason for that, is left truncation which the model does not explicitly account for. The model effectively estimates the failure rate unconditional on any unknown history before the start of the observation period. Nevertheless, the overall performance of the model is adequate and there is nothing in these simulation studies to cause any concern about this particular implementation.
4.2.2 Model Application

The data set introduced in Chapter 3 involves a network consisting of 1349 underground water pipes. The available information consist of failure times from the installation date (which is different for each pipe) until the end of year 2003 and the length of each pipe $i$ in metres ($x_i$). As we have seen in Chapter 3, although the network appears to be deteriorating as a whole (in terms of raw yearly number of failures), individual pipes behave differently from each other. For example some pipes are clearly ageing while others experience no failures at all. As discussed earlier in this chapter, a mixed effects NHPP model is clearly a good choice for modelling such behaviour since it can specifically allow for such heterogeneity between pipes.

The main assumption in such a model is that the failure process of each pipe is independent of other pipes given the random effects and the covariates but note that the model for the random effects is common for all pipes. That means that the model has the ability to pool information which may help in better estimating pipes with no observed failures. A power-law NHPP will be considered for each pipe $i$, much like the one presented in the simulations above (Section 4.2.1). This is because a power-law formulation of $\lambda(t)$ with a different ‘shape’ random effect $\theta_i$ for each pipe results in a versatile model able to capture a situation where each pipe in the system behaves uniquely.

As illustrated in Chapter 3, the data set is quite informative specifically in the sense of knowing individual break times as well as having failure information since the installation date of each pipe. In order to test the performance of the model for situations where data are left-truncated, only data after 1969 will be used (all installation dates vary from 1945 to 1960). Furthermore, only information up to and including year 1998 will be considered which will allow for a prediction period of 5 years, since all pipes were observed up until 2003. The model will then be fitted for the 30 year period 1969-1998 of accessible failure data and then used to

The failure rate function for each pipe is the same as in Equation (4.10), namely:

\[
\lambda(t; x_i | \theta_i) = \theta_i t^{\theta_i - 1} e^{\beta_0 + \beta_1 x_i}
\]

\[
\theta_i \sim \text{Gam}(\alpha, \kappa)
\]

and hence the log-likelihood for each pipe \(i\) given the random effects is the same as in Equation (4.11). Covariate \(x_i\) relates to pipe length in metres although note that in some applications, pipe length is used as an offset so that one is effectively modelling the number of failures per unit time and length. For the purposes of this and subsequent chapters, length will be treated as a covariate whose effect on the failure rate needs to be estimated.

The model was implemented in WinBUGS using the code developed and tested in the previous section. Two MCMC chains with variable initial values were run for 30000 iterations with a burn-in of 15000 and a thinning of 5. This resulted in 3000 posterior samples from each chain. The Geweke diagnostic (discussed in Section A.1.4) was performed on all parameters and each test statistic was in the range of \([-1.96, 1.96]\] meaning that the alternative hypothesis of non-convergence for each MCMC sample was rejected at the 95% level.

An alternative convergence test is the Gelman and Rubin diagnostic (Section A.1.4) which also takes into account that more than one chain was run using dispersed initial values. The diagnostic calculates a value \(\hat{R}\) for each parameter and this must be close to 1 if convergence has been reached. In Table 4.3, \(\hat{R}\) values for each model parameter are given. All values are close to 1 indicating convergence.

In addition, Figure 4.5 shows a plot of an MCMC sample from the ‘log-posterior’, meaning samples from the logarithm of the joint posterior distribution of model parameters \((\beta_0, \beta_1, \alpha, \kappa)\), evaluated at each MCMC iteration. As discussed in Appendix A, this plot is another indication of convergence since the joint posterior distribution is a global summary of all model parameters and its stability indicates convergence.
4.2. Mixed Effects NHPP model

The notion of posterior predictive diagnostics for model fit is also discussed in Appendix A. In particular, the idea of a test quantity $T(y, \theta)$ (where $y$ is the data and $\theta$ the model parameters) is discussed which aims to assess the discrepancy between fitted model and data. In the Bayesian sense, one may collect samples of the test quantity for data $\tilde{y}$ simulated from the posterior predictive distribution of the response and compare these with samples of $T(y, \theta)$ evaluated using the actual data at each sample of the parameter posteriors. To assess the significance of the discrepancy, one may calculate a p-value defined as:

$$p\text{-value} = \Pr(T(\tilde{y}, \theta) \geq T(y, \theta) | y)$$

In MCMC this may be estimated by the proportion of samples that $T(\tilde{y}, \theta)$ exceeds $T(y, \theta)$. A very small p-value would indicate poor model fit indicating that the observed data extreme in relation to data simulated from the model.

A sensible measure for the test quantity $T(y, \theta)$ which is a function of both the data and the parameters, is the deviance of the model (see Appendix A) which is defined as minus twice the log-likelihood. For the NHPP model here, samples of the deviance were already collected but to perform this diagnostic required to predict failure times at each posterior sample to calculate the deviance for each of those generated data sets. The word predict is used since data need to be simulated from the posterior predictive distribution of the response. The plot in Figure 4.6
4.2. Mixed Effects NHPP model

shows the time series plot of the deviance samples for simulated data (in black) and for the observed data (in red). The $p$-value was estimated to be 0.43 meaning that model fit is adequate.

The prior distributions of $\beta_0$, $\beta_1$, $\alpha$ and $\kappa$ along with estimates (posterior means), standard errors and 95% Cr.I. are shown in Table 4.3. All parameters appear significant but particular interest lies on $\beta_1$ relating to the effect of pipe length. The estimate of $\beta_1$ is positive meaning that the effect of pipe length on the failure rate of each pipe, is scaling by $\exp(0.0047) = 1.005$ for every metre of length. In terms of a plot of the failure rate $\lambda_i(t; x_i|\theta_i)$ against $t$, the effect of length is an upwards vertical shift of the curve.

Figure 4.7 shows posterior means for random effects $\theta_i$ along with 95% Cr.I. in red noting that these are quite tight.
4.2. Mixed Effects NHPP model

One possible way of seeing how well the deterioration was captured which is also a useful way of checking model fit, would be to compare the failure ranks of each pipe, both actual and estimated. Ranking the pipes based on the observed data may be done by considering the total number of failures for each, in the observation period 1969-1998. The idea is then to compare observed ranks with estimated ones. The estimated ranks were calculated by ranking failures from posterior predictive distributions of failure counts for each pipe, thus obtaining posterior distributions of the failure ranks.

Given the structure of the power law failure rate, the random effects $\theta_i$ reflect the deterioration in each pipe. So if the ranks according to the model compare well with the actual ones, it may signify the appropriateness of $\theta_i$ in the model. In Figure 4.8 the means of failure rank posteriors are plotted against the actual ranks. 95\% Cr.I. are also shown (in red) along with a 45° line. The information on the plot shows that the model is adequately capturing the ranks except at the lower end which corresponds to the 346 pipes that experienced zero failures during the 30 year period. Zero failures were observed for 25.6\% of the pipes which may be causing problems in terms of accurately capturing the deterioration mechanism. In Chapter 5 this point is addressed in more detail.

A quantity which may be of interest is whether or not a pipe will fail in a certain
4.2. Mixed Effects NHPP model

Figure 4.8: Estimated ranks vs actual ranks

time period, such as the 5 year prediction period chosen for this data. Given posterior samples of the fitted failure times and hence the estimated failure count in a time period, one may estimate the probability of more than one failure by calculating the proportion of MCMC samples for which the estimated count of pipe breaks is greater or equal to 1.

Samples from the posterior distribution of the probability of one or more failures were collected for both the observation and the prediction period. These probabilities can be used in a Bernoulli trial to ‘decide’ whether a pipe will fail or not, i.e. if zero is the outcome of the trial then pipe does not fail and vice versa. A 2x2 ‘confusion’ matrix or contingency table may then be constructed, whose diagonal entries reflect the number of pipes that were correctly predicted to fail or not, whereas off-diagonal entries correspond to the number of wrongly classified pipes.

For the NHPP model applied here, 500 such matrices were constructed and then averaged, both for the 30 year observation period and the 5 year prediction period. These matrices are presented in Table 4.4. The matrix relating to the observation
4.2. Mixed Effects NHPP model

period states that (on average) 141.2 pipes were correctly estimated not to fail and 866 were correctly estimated to fail. In addition, 137 pipes that failed were predicted not to fail whereas 204.8 pipes that had no bursts were estimated to do so. In total 1007.2 out of 1349 pipes were correctly estimated by the model. Similarly, for the prediction period, 864.7 pipes (summing the diagonals) were properly predicted by the model.

Table 4.4: NHPP confusion matrices

<table>
<thead>
<tr>
<th>NHPP</th>
<th>Estimated pipe failures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td>Not Failed</td>
<td>141.2</td>
<td>204.8</td>
<td>346</td>
</tr>
<tr>
<td>Failures</td>
<td>Failed</td>
<td>137.0</td>
<td>866.0</td>
<td>1003</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>278.2</td>
<td>1070.8</td>
<td>1349</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NHPP</th>
<th>Predicted pipe failures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td>Not Failed</td>
<td>716.3</td>
<td>315.7</td>
<td>1032</td>
</tr>
<tr>
<td>Failures</td>
<td>Failed</td>
<td>168.6</td>
<td>148.4</td>
<td>317</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>884.9</td>
<td>464.1</td>
<td>1349</td>
</tr>
</tbody>
</table>

To assess overall model fit, the observed failure count for each pipe can be compared with the estimated analogue. Figure 4.9 shows the actual failure counts for both the observation and prediction period (35 years) and the equivalent fitted values (in blue) taken as posterior means from predictive distributions of counts. For clarity, prediction intervals were not overlaid. The plot itself suggests a rather reasonable fit since the line relating to model estimates follows the points quite closely.

Summarising, the analysis suggests that overall the mixed effects NHPP model provided an adequate fit to the data with a reasonable predictive capability. More importantly, the model was able to distinguish between the different failure processes of each pipe through the random effects $\theta_i$ while allowing for the effects of
4.3 Summary

In this chapter, a Bayesian mixed effects NHPP model with a power-law failure rate was developed and its MCMC implementation was tested on simulated data. The simulations proved that the model provides a good fit to the simulated data meaning that if failure process in the real data does behave similar to the assumed NHPP, the model is able to capture it.

The model was then applied to the North American data discussed in Chapter 3. The fit to the data was good in the sense that the model was able to capture the de-
4.3. Summary

terioration in each pipe and this was assessed by comparing observed and estimated failure ranks. The fitted model was further used to calculate, for each pipe, the probability of failure within the observation (1969-1998) and the prediction period (1998-2003). Using those probabilities, the estimated events of at least one failure were compared with the actual ones and confusion matrices were constructed for each period. These further showed that the model had adequate fit and predictive capability.

Nevertheless, there was indication of model inability to deal with some pipes that never failed during the observed period. This could indicate that in relation to others, some pipes have an extra element of resilience, i.e. resistance to failure. In subsequent Chapter 5, this particular issue is addressed and a possible extension the NHPP is proposed to effectively deal with such situations.

The methodology described in this chapter is presented in both Economou et al. (2007) and Economou et al. (2007) where a mixed effects NHPP model with a power law failure rate is applied to a network of 532 underground water pipes. In both articles the model is used as a baseline and compared with both a zero-inflated and an aggregated version of itself respectively, which relate to subsequent Chapters 5 and 6.
Chapter 5

Mixtures of NHPP models

The mixed effects NHPP model illustrated in Chapter 4 has proven to be a flexible model able to capture the non-linear relationship between failure rate and time. Using random effects, each pipe is effectively modelled by a different NHPP which assumes that ageing is the only major factor controlling the failure mechanism.

In this chapter, it is argued that these assumptions are questionable in practice and that the conventional NHPP model may not be adequate for certain scenarios. For that reason, a mixture model of NHPPs is formulated and more specifically, a zero-inflated NHPP model is applied to the North American data. Although zero-inflated Poisson models have been extensively studied in the past, the concept of zero-inflation in NHPPs has not been investigated in the literature so far.

5.1 Motivation

For repairable components such as water pipes, the assumption of deterioration mainly due to ageing is theoretically reasonable, but may not always be true in practice. A common reason is the fact that possibly unobserved internal or external processes are acting on the pipes distorting the underlying failure process in a way
as to render conventional NHPP models inadequate. In the application of a NHPP model on the North American data in Chapter 4, we have already seen a possible example of this where some pipes exhibit an extra resistance to failure resulting in more zeros, in terms of failure counts, than can be explained by the model.

Consider a situation where repairable components deteriorate with time, but the resulting ageing process may be different in certain time periods of varying size. This may be quantified by asserting that the deterioration mechanism is perhaps changing states in time where these states are dependent upon several factors acting on the component other than age. With reference to water pipes, an example may be that from time to time heavy vehicles are inappropriately driven over roads which have a weight limit resulting in instantaneous increases in pipe breaks over irregular time intervals. Another example with longer term effects may be that a series of exceptionally cold winters results in pipes which are more prone to failure meaning that the failure rate will be higher than usual for those years. This kind of behaviour, is not easily captured by conventional NHPP models with a failure rate which is monotonically related to time, e.g. models with a power law or a log-linear $\lambda(t)$.

5.2 Mixture Models

Mixture models are flexible and general enough to accommodate varying behaviour of the failure process which cannot be explained by a conventional NHPP. For mixture models, the distribution of the response variable is a weighted combination of densities so that the actual probability distribution is multimodal. Mathematically, a mixture distribution for observations $y = (y_1, \ldots, y_n)$ may be defined as:

$$f(y) = p_1 f(y; \theta_1) + p_2 f(y; \theta_2) + \cdots + p_m f(y; \theta_m)$$

where $p_k \geq 0$ such that $\sum_{k=1}^{m} p_k = 1$ and $f(y; \theta_k)$ belong to some parametric family, each with corresponding parameters $\theta_k = (\theta_{k,1}, \ldots, \theta_{k,p})$. Note that the choice about each $f(y; \theta_k)$ is arbitrary, as long as these are proper probability
5.2. Mixture Models

distributions. In that case \( f(y) \) will also be proper, its integral will be 1 (given that \( p_k \) are independent of \( y \)).

A typical example of an application where a mixture model may be applied is one which involves a data set with a continuous response variable concerning a particular mammalian species, which lacks information about gender. If there is a significant difference in the response between the two genders, the response will inevitably have a bimodal density where, for instance, a mixture of two Gaussians may be applied. In general, there is no constraint in the \( f(y|\theta_k) \) being the same or even on how large \( m \) should be but it is sensible to restrict \( m \) to be as small as is necessary otherwise identifiability problems may arise.

5.2.1 Likelihood Function

In the NHPP context, a convenient way which one may use the mixtures idea is by assigning a possibly pipe distinct mixing probability \( p_{i,k} \) to each set of failure times for each pipe \( i \). In that case, failures are generated for each pipe by a different NHPP according to \( p_{i,k} \). Consider a set of pipes \( i = 1, 2, \ldots, N \) where each has failed \( n_i \) times at \( T_{i,0} < t_{i,1}, \ldots, t_{i,n_i} \leq T_{i,end} \). The mixture model proposed here is then:

\[
 f_{\text{mix}}(T_{i,0}, t_{i,1}, \ldots, t_{i,n_i}, T_{i,end}; \Theta_i) = \sum_{k=1}^{m} p_{i,k} L_{PP}(T_{i,0}, t_{i,1}, \ldots, t_{i,n_i}, T_{i,end}; \theta_{i,k})
\]

where \( \theta_{i,k} \) corresponds to possibly pipe specific parameters of each NHPP with likelihood \( L_{PP}(\cdot) \) given in Equation (4.6). Also, \( \Theta_i = (\theta_{i,1}, \ldots, \theta_{i,m}, p_{i,1}, \ldots, p_{i,m}) \).

The formulation of the overall likelihood of the model for \( i = 1, 2, \ldots, N \) is then:

\[
 \prod_{i=1}^{N} f_{\text{mix}}(T_{i,0}, t_{i,1}, \ldots, t_{i,n_i}, T_{i,end}; \Theta_i)
\]

This mixture model is effectively assuming that for each pipe \( i \), failures are generated by either one of \( k \) NHPP processes according to probabilities \( p_{i,1}, \ldots, p_{i,k} \).

Although the \( p_{i,k} \) are pipe specific, these are assumed to be fixed effects, unlike the
5.2. Mixture Models

random effects of models implemented in Chapter 4. Effectively then, the mixing probabilities $p_{i,k}$ are model parameters that need to be estimated.

5.2.2 Label Switching

When using Bayesian methods for fitting mixture models and in particular MCMC, the problem of so called label switching may arise. In essence, the issue arises because of the invariance of the likelihood under relabelling of the mixture components in the model (Celeux et al., 2000). The effects of that invariance directly affect the (joint) posterior distribution of the parameters since the posterior is proportional to the likelihood times the joint prior. Stephens (2000) states that label switching may lead to a highly symmetric and multimodal posterior making it difficult to summarise, especially by using marginal distributions as these will most likely be inappropriate.

Specifically, consider a general mixture model as before:

$$f(y_i) = p_1 f(y_i; \theta_1) + p_2 f(y_i; \theta_2) + \cdots + p_k f(y_i; \theta_m)$$

The likelihood for this model with data $y_1, \ldots, y_n$ is then:

$$L(y; \theta_1, \ldots, \theta_m, p_1, \ldots, p_m) = \prod_{i=1}^{n} [p_1 f(y_i; \theta_1) + p_2 f(y_i; \theta_2) + \cdots + p_k f(y_i; \theta_m)]$$

The label switching problem arises because randomly permuting the identities of $\theta_1, \ldots, \theta_m$ and the corresponding mixing probabilities $p_1, \ldots, p_m$, will not change the value of this likelihood.

Label switching can be diagnosed from time series plots of MCMC samples for each parameter as well as density estimation plots. Signs of jumps in the former coupled with associated multimodality in the latter will indicate label switching issues. This is because at any point of the MCMC run, the sampler may ‘confuse’ the identity of one or more mixture components so it will simulate samples of one from the posterior of the other. We return to the problem of label switching in Chapter 7 with possible ways for preventing it.
5.2.3 Zero-Inflated Poisson Models

A particular issue with data sets involving repairable equipment with relatively long lifetimes, such as underground water pipes, is the fact that data are often inadequate both in terms of monitoring possible covariates and of truncation due to lack of past information. Specifically, data involving water pipes are quite often limited to only a few years in relation to the age of the pipes in the network (Gat and Eisenbeis, 2000). This added to the fact that in general, failures are rare events over the lifespan of a pipe, results in many data sets having a considerable number of zeros in terms of failure counts.

Recall, that a NHPP model for the occurrence of failures in time, is equivalent to assuming a Poisson distribution with a time varying mean (Section 4.1). Using this property, the total number of failures over the whole observation period is also Poisson distributed and this is considered in Chapter 6 which is looking at the application of NHPP to aggregated data.

In addition, data may be viewed at the aggregate level i.e. the number of failures over the whole observation period for each pipe. In situations where a considerable number of pipes have experienced no breaks at all, given the Poisson nature of the counts, the data is zero-inflated. More accurately, the formal definition of zero-inflation is when the proportion of zeros in the data exceeds the proportion that is generated by the fitted model and it is a phenomenon which appears frequently in certain kinds of count data.

To cope with zero-inflation in defect counts on items in manufacturing, Lambert (1992) introduced a Zero-Inflated Poisson (ZIP) regression model. The author stated that a possible explanation for zero-inflation is that the failure process moves randomly between a perfect state in which failures are rare and an imperfect state in which failures are more likely, and this is due to slight, unobserved changes in the environment. This is very similar to the concept motivating this chapter i.e. that reliability of repairable devices in random environments is bound to be affected by
5.2. Mixture Models

unobserved factors.

The ZIP model and its concept have been widely used in a broad range of applications: Gupta et al. (1996), Zorn (1998), Bohning et al. (1999) provide a thorough overview of Poisson models coping with excess zeros in the counts. Angers and Biswas (2003) and Ghosh et al. (2006) utilise the ZIP within the Bayesian framework whereas Agarwal et al. (2002) considers the ZIP in the context of spatial count data.

The main idea behind the ZIP model is that extra probability is allowed for zero counts which is equivalent to attaching a ‘spike’ at zero in terms of the density shape. The ZIP is clearly a special case of a mixture model with two components: one which generates zeros with probability \((1 - p)\) and the second which produces counts (including zero) from a Pois\((\mu)\) distribution with probability \(p\). So considering a random sample of counts \(y_1, \ldots, y_n\), the ZIP model may be formulated in the following way:

\[
f_{ZIP}(y_i) = \begin{cases} 
(1 - p) + pe^{-\mu}, & \text{if } y_i = 0, \\
pe^{-\mu} \frac{\mu^{y_i}}{y_i!}, & \text{if } y_i = 1, 2, \ldots
\end{cases}
\]

whose log-likelihood function is:

\[
l(y; \mu, p) = \sum_{i=0}^{n} I_{(y_i=0)} \log \left( 1 - p + pe^{-\mu} \right) + \sum_{i=0}^{n} I_{(y_i>0)} \left[ \log(p) - \mu + y_i \log(\mu) - \log(y_i!) \right]
\]

where:

\[
I_{(\text{event})} = \begin{cases} 
1, & \text{if event is True}, \\
0, & \text{if event is False}
\end{cases}
\]

The model can be extended to having observation specific \(\mu\) and \(p\), further allowing for the inclusion of covariates and random effects.
5.3 Extending the NHPP for Zero-Inflation

Using the same idea as in the ZIP model, one could extend the conventional NHPP model to a zero-inflated version, where each process is either a NHPP or a zero-generating process depending on a mixing probability. Consider a set of pipes, where each has failed times in the period \([T_{i,0}, T_{i,end}]\) at times \(T_{i,0} < t_{i,1}, \ldots, t_{i,n_i} \leq T_{i,end}\). A Zero-Inflated NHPP (ZINHPP) model may then be formulated as:

\[
\begin{aligned}
    f_{\text{ZIPP}}(\tau_i; p_i, \theta_i) &= \\
    &= (1 - p_i) + p_i e^{-\Lambda(T_{i,0}, T_{i,end})}, \quad \text{if } n_i = 0, \\
    &= p_i L_{\text{PP}}(\tau_i; \theta_i), \quad \text{if } n_i = 1, 2, \ldots
\end{aligned}
\]

where \(e^{-\Lambda(T_{i,0}, T_{i,end})}\) is the probability of zero failures in the time interval \([T_{i,0}, T_{i,end}]\) and for clarity, \(\tau_i = T_{i,0}, t_{i,1}, \ldots, t_{i,n_i}, T_{i,end}\). As usual, the vector \(\theta_i\) denotes the parameters and \(L_{\text{PP}}(\cdot)\) denotes the likelihood of the NHPP. Similar to the ZIP model, the log-likelihood of the ZINHPP model for pipes \(i = 1, \ldots, N\) is given by:

\[
l(\tau; \Theta, p) = \sum_{i=1}^{N} I_{(n_i=0)} \log \left( (1 - p_i) + p_i \exp \left\{ -\Lambda(T_{i,0}, T_{i,end}) \right\} \right)
\]

\[
+ \sum_{i=1}^{N} I_{(n_i>0)} \left[ \log(p_i) + \log(L_{\text{PP}}(\tau_i|\theta_i)) \right]
\]

where \(\Theta = (\theta_1, \ldots, \theta_N), p = (p_1, \ldots, p_N)\) and \(\tau = (\tau_1, \ldots, \tau_N)\).

5.3.1 Simulation Experiments

The aim in this section, is to implement a ZINHPP model in WinBUGS and to test that implementation using simulations. To generate response values from a ZINHPP model one may again utilise the recurrence formula in simulating NHPP values introduced in Section 4.2.1. The difference here, is that one will need to flip a ‘zero-inflation coin’ before deciding whether the process in actually a NHPP or whether it is a zero process. So assuming that \(p_i\) reflects the mixing probability of pipe \(i\) as in Equation (5.1), the random variable \(u_i \sim \text{Bern}(p_i)\) may be used to so
5.3. Extending the NHPP for Zero-Inflation

that when \( u = 1 \) the process is NHPP and recurrence times may be generated as before. When \( u_i = 0 \), the number of failures \( n_i \) for pipe \( i \) in the observation period \((T_{i,0}, T_{i,end}]\) is zero.

As before, the design here is to generate failure times in years, for \( N = 50 \) hypothetical pipes which behave according to a ZINHPP model. Equation (5.1) will reflect the stochastic nature of each pipe \( i = 1, 2, \ldots, 50 \) and again a power law formulation will be assumed for the failure rate of the NHPP with random effects \( \theta_i \) so that:

\[
\lambda(t; x_i|\theta_i) = \theta_i t^{\theta_i-1} e^{\beta_0 + \beta_1 x_i} \\
\theta_i \sim \text{Gam}(\alpha, \gamma)
\]

The parameters of the model are then the intercept \( \beta_0 \), the associated coefficient \( \beta_1 \) of covariate \( x_i \), the shape and rate parameters \( \alpha \) and \( \kappa \), and the mixing probability for each pipe \( p_i \). The start of the observation period \( T_{i,0} \) for \( i \) will be equal to zero relative to its age \( T_{i,end} \) effectively assuming that each pipe was observed since the day it first started being used. This means no left truncation is assumed as the purpose here is to specifically study the zero inflation aspect.

The log-likelihood for pipe \( i \) given the random effects \( \theta_i \) is:

\[
\ell_i(T_{i,0}, t_{i,j}, T_{i,end}; p_i, \beta_0, \beta_1|\theta_i) = \begin{cases} 
\log(p_i) + n_i \log(\theta_i) + n_i(\beta_0 + \beta_1 x_i) \\
+(\theta_i - 1) \sum_{j=1}^n \log(t_{i,j}) + \left[T_{i,0}^{\theta_i} - T_{i,end}^{\theta_i}\right] \exp\{\beta_0 + \beta_1 x_i\} \\
+I(n_i=0)\left\{\log(1 - p_i + p_i \exp\left\{\left[T_{i,0}^{\theta_i} - T_{i,end}^{\theta_i}\right] \exp\{\beta_0 + \beta_1 x_i\}\right\}\right\}
\end{cases}
\] (5.2)

where \( t_{i,j} \) denote the time (year) of failure number \( j \) for pipe \( i \). The log-likelihood of all the pipes is then given by \( \sum_{i=1}^{50} \ell_i(T_{i,0}, t_{i,j}, T_{i,end}; p_i, \beta_0, \beta_1|\theta_i) \).

Failure times were generated for a period of 10 years after each \( T_{i,end} \) so that the predictive strength of the applied model may also be assessed. 100 data sets were generated using the same predetermined parameter values each time and the ZINHPP model was fitted to each of these using WinBUGS (relevant code presented
5.3. Extending the NHPP for Zero-Inflation

in Appendix B). Although the ZINHPP is not a conventional model, it may still be formulated in WinBUGS in the same way as the NHPP but with the additional use of a Bernoulli variable \( u_i \sim \text{Bern}(p_i) \) in order to ‘pick-out’ the right elements in the log-likelihood. Fortunately, label switching is less of a problem when dealing with zero-inflated models since effectively, the model for the zero-generating process, which accounts for the excess of zeros, is a probability of 1 involving no parameters at all.

Model parameters were given uninformative priors which are shown in Table 5.1 and Table 5.2. The priors for the mixing probabilities \( p_i \) were Beta\((a, b)\) distributions for which both shape parameters \( a \) and \( b \) were equal to 1 which is equivalent to a Uniform distribution between 0 and 1.

Table 5.1: Simulated ZINHPP results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Actual Value</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>N(0,1000)</td>
<td>-8</td>
<td>-7.57 (0.111)</td>
<td>[-7.795,-7.359]</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>N(0,1000)</td>
<td>0.06</td>
<td>0.052 (0.002)</td>
<td>[0.049,0.055]</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Gam(0.5, 0.005)</td>
<td>60</td>
<td>104.17 (5.209)</td>
<td>[94.184,114.299]</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>Gam(0.5, 0.005)</td>
<td>40</td>
<td>69.18 (3.269)</td>
<td>[62.931,75.577]</td>
</tr>
</tbody>
</table>

Table 5.2: Simulated ZINHPP results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Min</th>
<th>Max</th>
<th>Posterior Mean</th>
<th>25%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual ( p_i )</td>
<td>-</td>
<td>0.11</td>
<td>1.00</td>
<td>0.73</td>
<td>0.28</td>
<td>0.97</td>
</tr>
<tr>
<td>Estimated ( p_i )</td>
<td>Beta(1,1)</td>
<td>0.37</td>
<td>0.67</td>
<td>0.56</td>
<td>0.44</td>
<td>0.64</td>
</tr>
</tbody>
</table>

For each fitted model, two MCMC chains were run for 50000 iterations with a burn-in of 25000 and a thinning of 5. This resulted in 10000 posterior samples for each parameter which was adequate to ensure convergence.

Samples of the joint posterior distribution were collected for each fitted model and
then averaged across models. Table 5.1 shows the posterior mean and standard error for $\beta_0$, $\beta_1$, $\alpha$ and $\kappa$ as well as the original input values which compare rather well with the estimates. In addition, the 95% Cr.I. for the four parameters appear to be adequately narrow. Although $\beta_0$ and $\beta_1$ were accurately captured by the model, the estimates for two parameters relating to the random effects $\theta_i$ were not very reasonable. This is noticeable when comparing posterior means and the actual values. However, the random effects are Gamma distributed implying that $E(\theta_i) = \alpha/\kappa$. The estimated mean for the random effects ($104.17/19.18 = 1.50$) is then comparable the input mean ($60/40 = 1.5$). Therefore, the variance of $\theta_i$ is underestimated ($0.0218$ compared to $0.0375$).

Table 5.2 shows some summary statistics for both the inputs as well as the estimates of $p_i$. The posterior mean, the minimum and maximum value and the 25% and 75% sample percentile show that there is some discrepancy between input and estimated values. In addition, a plot of the ranks of $p_i$ against the ranks of $\hat{p}_i$ (posterior means) is shown in Figure 5.1. The points on the plot corresponding to the $p_i$ ranks lie adequately on the 45° line with the exception of some odd points.

![Figure 5.1: Rank of $p_i$ vs rank of $\hat{p}_i$](image)

Since this is a mixture model, the discrepancies between actual values and estimates
5.3. Extending the NHPP for Zero-Inflation

for both $p_i$ and the parameters of the random effects $\theta_i$, may be attributed to lack of identifiability. For instance, a pipe $i$ with zero recorded failures may have a low failure rate due small value for $\theta_i$ but also a small value of $p_i$. This means that the pipe has a naturally low failure rate but also an extra attribute which allows extra probability of not failing and this may lead to lack of identifiability.

For each generated data set, the failure times for a 10 year prediction period were also generated and the total number of failures in that period was calculated for each pipe. Samples of the posterior predictive distribution of the failure count (in the prediction period) for each pipe were collected and averaged between models. The actual average of the total number of failures in the 10 year period is plotted against the means of the posterior predictive distributions in Figure 5.2. Included in the plot are the 95% prediction intervals calculated as the 2.5% and 97.5% quantiles of the sample. From this, one may argue that the predictive capability of the model is reasonably adequate since the points are close enough to the 45° line and well within the prediction intervals.

![Figure 5.2: Actual vs predicted no. of failures](image)

Overall, the model fits adequately with regards to estimating the parameters. However, the fit is not very accurate for the mixing probabilities $p_i$ and the variance
of the random effects, although the ranks of the estimated $p_i$ agree well enough with the inputs. A possible reason for that may be model identifiability. Nevertheless, the simulations have shown that at the very least, the model has the ability to capture data generated from the ZINHPP itself, suggesting that there are no fundamental problems with the MCMC implementation of the model.

5.3.2 Model Application

In this section, we return to the data set described in Chapter 3 which was already used as an application for the mixed effects NHPP model in Chapter 4. That fit was quite reasonable, with the exception that given the model, excess zeros in failures counts was flagged as a possible issue which the model was not capable of dealing with. In this section, the ZINHPP model is applied to the same data to see whether model fit can be improved in any way.

The model will be identical to the one used in the ZINHPP simulations earlier in Section 5.3.1, where the failure rate of the NHPP is modelled as:

$$
\lambda(t; x_i | \theta_i) = \theta_i t^{\theta_i - 1} e^{\beta_0 + \beta_1 x_i} \\
\theta_i \sim \text{Gam}(\alpha, \kappa)
$$

where $x_i$ refers to pipe length. The log-likelihood given $\theta_i$ for each of the 1349 pipes is thus given by Equation (5.2).

Other than the pipe specific random effects $\theta_i$, each pipe has also an associated parameter $p_i$ reflecting the mixing parameter in the model. The quantity $p_i$ reflects the probability that an observation comes from a NHPP instead of a zero generating process. This means that $(1 - p_i)$ may be considered as a quantity reflecting resistance to failure. In other words, a small value of $p_i$ indicates that a pipe will have extra resistance to failure on top of the underlying breakage resistance which is modelled by the NHPP.

Therefore, it may be sensible to characterise $p_i$ in terms of pipe length but also
5.3. Extending the NHPP for Zero-Inflation

in terms of pipe age (at the end of the observation period). This is because some ‘young’ pipes may have more durability than expected. In addition, some pipes with short length may exhibit more resistance than others since their short length exposes them to less risk. A sensible way to model $p_i$ is:

$$\logit(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \gamma_{i,0} + \gamma_1 x_i + \gamma_2 T_{i,\text{end}}$$

where $x_i$ is the length of each pipe and $T_{i,\text{end}}$ the age of each pipe. Note that $\gamma_{i,0}$ are pipe specific fixed effects meaning that this model is structurally equivalent to the one used in the simulations.

In WinBUGS, the Gibbs sampler was used to run two chains, which were given dispersed initial values, for a total of 60000 iterations. Burning-in half of those and thinning by 10 resulted in 6000 posterior samples. $\hat{R}$ values from the Gelman and Rubin diagnostic were calculated for each of the 1355 unknowns in the model $(\beta_0, \beta_1, \alpha, \kappa, \gamma_{i,0}, \gamma_1, \gamma_2)$ and shown in Figure 5.3. Again all values are very close to 1 meaning that the samples converged adequately.

![Figure 5.3: $\hat{R}$ for 1355 model parameters](image)

MCMC samples of the log-posterior are shown in Figure 5.4 where the apparent stability also indicates convergence.

Recall that a posterior predictive diagnostic for model fit was used in Chapter 4 which was based on a test quantity $T(y, \theta)$ chosen to be the model deviance. The
idea was to compare the deviance of the data with the deviance of simulated data, at each posterior sample. The plot in Figure 5.5 depicts deviance samples for the actual data (in red) and the ones for the simulated data (in black). The $p$-value was estimated as 0.45 meaning that observed data are not extreme in relation to the fitted model indicating an acceptable model fit.

In Table 5.3, the posterior mean, standard error and credible intervals for $\beta_0$ and $\beta_1$ are shown. Both parameters appear to be significant since zero is not included in the credible intervals. As before, pipe length has a positive effect on the mean break count. Significance is also evident for parameters $\gamma_1$ and $\gamma_2$ of the $p_i$ which reflects the probability that failures come from a NHPP or equivalently, the probability of
5.3. Extending the NHPP for Zero-Inflation

not being extra resistant to failure. Parameter $\gamma_2$ is negative, which means that the effect of age is lowering the odds $\frac{p_i}{1-p_i}$. In other words, an older pipe has less (extra) resistance to failure. Surprisingly, parameter $\gamma_1$ which relates to pipe length, has an positive effect on the odds meaning that longer pipes tend to have an element of resistance to failure which is greater than shorter pipes would have.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior Mean</th>
<th>s.e.</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>N(0,1000)</td>
<td>-5.463</td>
<td>0.236</td>
<td>$[-5.947, -5.034]$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>N(0,1000)</td>
<td>0.0037</td>
<td>0.00023</td>
<td>$[0.0032, 0.0041]$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Gam(0.5,0.005)</td>
<td>98.87</td>
<td>9.001</td>
<td>$[83.479, 117.700]$</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Gam(0.5,0.005)</td>
<td>200.822</td>
<td>7.820</td>
<td>$[87.320, 118.001]$</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>N(0,1000)</td>
<td>0.212</td>
<td>0.033</td>
<td>$[0.158, 0.286]$</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>N(0,1000)</td>
<td>-0.0158</td>
<td>0.0041</td>
<td>$[-0.0241, -0.0129]$</td>
</tr>
</tbody>
</table>

Posterior means and 95% credible intervals for random effects $\theta_i$ and posterior means for pipe specific parameters $p_i$ are plotted in Figure 5.6. The intervals for $\theta_i$ are reasonably tight. The $p_i$ plot does not include intervals as these render the plot very unclear but it does suggest that the model has picked up a number of pipes with low values of $p_i$ (i.e. extra resilient pipes).

Having already looked at how the fitted failure ranks for each pipe compare with the actual ranks for the model in Chapter 4, we can now do the same for the ZINHPP model. The purpose was to try and get a better understanding of how well the model captures the deterioration in each pipe. The estimated ranks, which are taken as the posterior means of the ranks, are plotted against actual ranks in Figure 5.7 along with 95% Cr.I. (in red) and a 45° line for reference. The plot looks similar to the one in Figure 4.8 which shows the corresponding ranks from model in Chapter 4, with the exception that this model captures the lower end of the ranks better, which unsurprisingly refers to the pipes with zero recorded failures. Thus, in the sense of capturing the ranks of each pipe, the ZINHPP is performing
5.3. Extending the NHPP for Zero-Inflation

Figure 5.6: Posterior means and Cr.I. for $\theta_i$ and $p_i$

relatively better than the NHPP model.

Furthermore, the observed number of failures for each pipe in the overall 35 year period is compared with the equivalent estimation from the model in Figure 5.8. The blue line represents the estimates which are calculated as the means of corresponding posterior predictive distributions. The plot does not differ much from the analogous one for the NHPP model.

Another useful measure considered in the application of this data set in Chapter 4, was the probability of whether a pipe will fail or not. Confusion matrices were constructed for both observation and prediction periods looking at how well the model was performing in capturing this specific aspect. These are shown in Tables 5.4 and
5.3. Extending the NHPP for Zero-Inflation

Figure 5.7: Estimated ranks vs actual ranks

Figure 5.8: Estimated number of failures vs individual pipes

5.5. For comparison purposes, the corresponding confusion matrices for the NHPP are also given. The observation period table, shows that the zero-inflated model performs better as both off-diagonal entries are smaller in value and in particular the top right entry. This comes as no surprise since that entry refers to the pipes
which are predicted by the model to have zero failures. However, the prediction matrix for the ZINHPP model is only marginally better which is somewhat disappointing. Recall though, the number of failures in the network during the last few years of the original observation period, experiences an unexpected drop (Figure 3.4). Most of this drop is within the chosen prediction period which may be why both models fail to perform as well as they do in the observation period.

Table 5.4: ZINHPP confusion matrices - observation period

<table>
<thead>
<tr>
<th>ZINHPP</th>
<th>Estimated pipe failures</th>
<th>Not Failed</th>
<th>Failed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Failures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>188.8</td>
<td>157.2</td>
<td></td>
<td>346</td>
</tr>
<tr>
<td>Failed</td>
<td>119.0</td>
<td>884.0</td>
<td></td>
<td>1003</td>
</tr>
<tr>
<td>Total</td>
<td>307.8</td>
<td>1041.2</td>
<td></td>
<td>1349</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NHPP</th>
<th>Estimated pipe failures</th>
<th>Not Failed</th>
<th>Failed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Failures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>141.2</td>
<td>204.8</td>
<td></td>
<td>346</td>
</tr>
<tr>
<td>Failed</td>
<td>137.0</td>
<td>866.0</td>
<td></td>
<td>1003</td>
</tr>
<tr>
<td>Total</td>
<td>278.2</td>
<td>1070.8</td>
<td></td>
<td>1349</td>
</tr>
</tbody>
</table>

5.4 Summary

The mixed effects model in Chapter 4 performed reasonably well, however the observed vs estimated failure ranks plot and the confusion matrices, indicated that there may be underachievement in capturing the behaviour of individual pipes that never failed. For that reason, a zero-inflated NHPP was developed in this chapter and applied to the same data set for comparison. The failure ranks plot was improved in the sense that the modified model was able to better classify pipes with zero failures. Furthermore, the confusion matrices showed increased model performance suggesting that for this particular data set, adjusting for zero-inflation
5.4. Summary

Table 5.5: ZINHPP confusion matrices - prediction period

| Model | Predicted pipe failures | | | |
|---|---|---|---|
| | Not Failed | Failed | Total |
| ZINHPP Actual Pipe Failures | Not Failed | 722.8 | 309.2 | 1032 |
| | Failed | 168.7 | 148.3 | 317 |
| | Total | 891.5 | 457.5 | 1349 |
| NHPP Actual Pipe Failures | Not Failed | 716.3 | 315.7 | 1032 |
| | Failed | 168.6 | 148.4 | 317 |
| | Total | 884.9 | 464.1 | 1349 |

affects model performance positively.

Initial work related to the discussions in this chapter, is presented in Economou et al. (2007) where a mixed effects ZINHPP model is successfully applied to an underground water network of 532 asbestos cement pipes in Auckland, New Zealand. A mixed effects NHPP model is also applied to the same data and a comparison of the results showed that the deterioration in the pipes is more precisely estimated when zero-inflation is allowed for. Furthermore, in Economou et al. (2008) a ZINHPP model is applied to the North American data although failures were considered at yearly level instead of monthly level showing that there was relative improvement in model performance when zero-inflation is accounted for. Using the idea of zero-inflation in NHPP models for recurrent event data has not been thus far considered in the literature.

The level of data quality here, in terms of failure information is excellent both because the timescale is small (monthly level) but also because information exists since the installation date for each pipe. However, this level of information is not often available and water pipe data, at least in the UK, are often aggregated. In
5.4. Summary

In many cases, the actual failure times are unknown so that data consist of failure counts over large periods of time. In the next chapter, the Poisson property of the NHPP is utilised so that the model can be applied to aggregated data.
Chapter 6

NHPP for Aggregated Data

A particular issue with data sets involving water pipe failures is quality and availability. One particular problem that arises is aggregation where individual failure times are unknown for large intervals of time, often for the entire observation period.

Capturing the relationship between time and the failure mechanism of a pipe is then particularly difficult because of the lack of information regarding failure times. Aggregation is not an issue that has been investigated in the literature so far, specifically when employing counting process models such as the NHPP. In this chapter, the applicability of the NHPP model on aggregated data is investigated and models from Chapters 4 and 5 are revisited and applied to such data.

6.1 NHPP model for Aggregated Data

In this thesis, the NHPP model has been introduced, explained and extended to include random effects. This was also formulated as a mixture model, for which the ZINHPP is a special case. The data required to derive the NHPP likelihood for a single pipe are the observed failure times. Nevertheless, when dealing with more
than one pipe, the possibility that some of them have not experienced any failures was allowed. This meant that instead of using the actual likelihood, the equation $\exp \{-\Lambda(T_0, T_{end})\}$ was used which is the Poisson probability of zero failures in the interval $(T_0, T_{end}]$ (also used in the formulation of the ZINHPP). This is due to the defining property of the NHPP given in Equation (4.1) stating that the number of failures in any time interval is Poisson distributed. Meeker and Escobar (1998) use this property to derive the NHPP likelihood in an alternative way to the one presented in this thesis.

Therefore it is possible to replace the NHPP likelihood with the Poisson probability of the total number of failures (Crowder et al., 1991). This means disregarding the failure times which will result in loss of information, but the model will still be a NHPP with failure rate $\lambda(t)$. When dealing with only one pipe, this will likely result in poor parameter estimates since we would be using the minimum amount of information from the data. However, as the number of pipes increases the information in the data is pooled (especially if reasonable covariates are available) in order to get as good an understanding of the deterioration process as possible.

This can then be used in situations where the only information available are the total number of failures in each pipe. It can also can be used when data are available only as aggregated failure counts at particular times instead of individual failure times, thus violating the assumption in Equation (2.2) which disallows simultaneous failures. An example would be a situation where the data consists of failure counts per year whereas ideally the actual days (or some other appropriate time unit) that each failure has occurred are desired.

### 6.1.1 Model Formulation

It is theoretically possible to employ the NHPP to data at any level of aggregation and use the minimal information in the data to get some understanding of the failure mechanism. In general, suppose that pipe $i$ has been observed in the time
6.1. NHPP model for Aggregated Data

period \((T_{i,0}, T_{i,\text{end}})\) which is further divided into \(k = 1, 2, \ldots, m_i\) time periods, i.e. \((T_{i,0}, T_{i,1}], (T_{i,1}, T_{i,2}], \ldots, (T_{i,m_i-1}, T_{i,m_i}]\) with \(T_{i,m_i} = T_{i,\text{end}}\), such that in each time period a number of failures \(n_{i,k} \geq 0\) has been recorded. Clearly, if all \(n_{i,k} = 1\) the NHPP can be applied in its conventional form given in Equation (4.6). Assuming that not all \(n_{i,k} = 1\), we employ the Poisson distribution in each interval so that the likelihood for pipe \(i\) is given by:

\[
L_{\text{agg}}(D_i; \theta_i) = \text{Pois}(\Lambda(T_{i,0}, T_{i,1}; \theta_i)) \times \cdots \times \text{Pois}(\Lambda(T_{i,m_i-1}, T_{i,m_i}; \theta_i)) = \prod_{k=1}^{m_i} \frac{e^{-\Lambda(T_{i,k-1}, T_{i,k}; \theta_i)} \Lambda(T_{i,k-1}, T_{i,k}; \theta_i)^{n_{i,k}}}{n_{i,k}!} = \exp \left\{ -\Lambda(T_{i,0}, T_{i,m_i}; \theta_i) \right\} \prod_{k=1}^{m_i} \frac{\Lambda(T_{i,k-1}, T_{i,k}; \theta_i)^{n_{i,k}}}{\prod_{k=1}^{m_i} n_{i,k}!} \right) (6.1)
\]

where \(D_i = (T_{i,0}, \ldots, T_{i,m_i}, n_{i,1}, \ldots, n_{i,m_i})\) represents the data and \(\theta_i\) are possibly pipe specific parameters of the NHPP.

Recall that:

\[
\Lambda(t_a, t_b) = \int_{t_a}^{t_b} \lambda(u)du
\]

so the function \(\Lambda(t_a, t_b)\) will contain all parameters \(\theta_i\) of the failure rate \(\lambda_i(t; \theta_i)\) including any random effects and possible covariates \(x_i\) which for neatness were omitted in the notation. The corresponding likelihood for a network composed of \(N\) such pipes is then given by \(\prod_{i=1}^{N} L_{\text{agg}}(D_i; \theta_i)\).

6.1.2 Simulation Experiments

A sensible way of testing how the aggregated model compares with the non-aggregated one is to apply it to the same data set that was generated by simulations in Chapter 4 and to compare parameter estimates. In those simulations, failure times were generated for 50 NHPP processes, influenced by a single covariate \(x_i\) with a failure rate based upon the power law:

\[
\lambda(t; x_i|\theta_i) = \theta_i t^{\theta_i-1}e^{\beta_0 + \beta_1 x_i}, \quad i = 1, \ldots, 50
\]
implying that the mean number of failures in arbitrary time interval \((t_a, t_b]\) is given by:

\[
\Lambda(t_a, t_b; x_i|\theta_i) = \left[t_b^\theta_i - t_a^\theta_i\right] e^{\theta_0 + \theta_1 x_i}
\]  \hspace{1cm} (6.2)

so that \(\theta_i\) are random effects. For each of the 50 simulated pipes, the same observation period is being used as before with left truncation being induced on purpose. In addition, data were generated for a further 10 years to be used for predictive purposes. As far as the aggregation level is concerned, the worse case scenario will be assumed, where the only information available is the total number of failures for each pipe within the observation period.

The data for each pipe is then the number of failures \(n_i\) and the observation period \((T_{i,0}, T_{i,end}]\). The model simplifies to:

\[
n_i|\theta_i \sim \text{Pois}(\Lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)) = e^{-\Lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)} \Lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)^{n_i} / n_i!
\]

with \(\Lambda(\cdot, \cdot)\) given by Equation (6.2). This aggregated NHPP (aggNHPP) model can be implemented in WinBUGS. This is straightforward since only the Poisson likelihood needs to be utilised. Relevant WinBUGS code, is given in Appendix B.

Flat priors were used for each parameter and two chains were run for 110000 samples with a burn-in of 100000, to ensure convergence. Applying a thinning of 5 resulted in 4000 posterior samples for each model parameter. As before, the model was applied to 50 simulated data sets generated using the same input values. The results were then based on averaging posterior samples over each fitted model.

Table 6.1 shows the actual values along with posterior means, standard errors and credible intervals for the four model parameters. For comparison purposes, the same results are shown for the non-aggregated model in Chapter 4. The posterior means are comparable between the two models meaning that the fit of the aggNHPP is similar to the NHPP. Note however that the standard errors are greater in the aggNHPP which may be reflecting the loss of information in the data due to aggregation.
### 6.1. NHPP model for Aggregated Data

**Table 6.1: Simulated aggNHPP results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Prior</th>
<th>Actual Value</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aggNHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>-8</td>
<td>-7.05 (0.216)</td>
<td>[-7.480,-6.618]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.06</td>
<td>0.056 (0.001)</td>
<td>[0.053,0.059]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>60</td>
<td>68.67 (5.529)</td>
<td>[58.345,79.863]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>40</td>
<td>47.87 (2.667)</td>
<td>[42.764,53.316]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>-8</td>
<td>-7.71 (0.11)</td>
<td>[-8.618,-8.178]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.06</td>
<td>0.056 (0.001)</td>
<td>[0.065,0.071]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>60</td>
<td>64.51 (3.08)</td>
<td>[58.59,70.82]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>40</td>
<td>44.24 (1.95)</td>
<td>[39.51,47.15]</td>
</tr>
</tbody>
</table>

Samples from the posterior predictive distribution of the number of failures within the prediction period were also collected and the means of those are used in Figure 6.1 which compares them with the actual failure counts. 95% prediction intervals are also plotted. The plot indicates adequate predictive capability since the majority of points lie on the 45° line and the intervals are reasonably tight. More importantly, the plot looks very similar to the equivalent plot for the non-aggregated version of the model (Figure 4.4).

In summary, the aggNHPP described here is equivalent to the NHPP model in Chapter 4, but applied to aggregated data. The simulation studies have shown that although information on individual failure times were discarded, the aggregated model can still perform adequately in terms of parameter estimates and prediction. The only difference was that parameter posteriors for the aggNHPP had slightly larger variance, which may be attributed to elevated uncertainty due to loss of information. The aggregated model was able to absorb enough information about the failure process and its relationship with time since overall the results match the ones from the original model quite closely.
6.1. NHPP model for Aggregated Data

In order to see how the aggregated NHPP model performs on real data, we return to the water pipe data set described in Chapter 3 and considered in Chapters 4 and 5 as an application for both the NHPP and the ZINHPP. The mixed effects NHPP model had a failure rate based on the power law:

\[ \lambda(t; x_i | \theta_i) = \theta_i t^{\theta_i - 1} e^{\beta_0 + \beta_1 x_i}, \quad i = 1, \ldots, 1349 \]

\[ \theta_i \sim \text{Gam}(\alpha, \kappa) \]

where \( \theta_i \) is a random effect and covariate \( x_i \) refers to length. The mean number of failures \( \Lambda(t_a, t_b; x_i | \theta_i) \) in an arbitrary time interval \( (t_a, t_b] \) is then given by Equation (6.2) above.

The data will be totally aggregated so that no failure times are assumed known. Therefore, the model here is identical to the one used in the simulation studies earlier.

Figure 6.1: Actual vs predicted no. of failures

6.1.3 Model Application
6.1. NHPP model for Aggregated Data

In previous analyses of this data set, information about failures was assumed unknown before the year 1969 as a way of explicitly inducing left truncation. The same is assumed here for comparison purposes. Also, the last 5 years worth of data for each pipe was ignored so that it may be considered as a prediction period. Hence, the observation period for each pipe is \((T_{i,0}, T_{i,end}] = (1969, 1998]\) whereas the prediction period is \((1998, 2003]\).

The aggNHPP was fitted in WinBUGS and in the usual manner, two MCMC chains were considered. 3000 samples were collected from each one after a burn-in of another 3000 and thinning by 10. To monitor convergence, the Gelman and Rubin diagnostic \((\hat{R})\) was calculated for each parameter (see Table 6.2). None of the \(\hat{R}\) values is substantially above 1 indicating convergence. In addition, a trace plot of the samples from the log-posterior is shown in Figure 6.2 which also confirms convergence in the MCMC posterior samples.

As before, MCMC samples of the deviance for simulated and observed data are compared to assess model fit. Figure 6.3 shows samples of the deviance calculated

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Parameter} & \beta_0 & \beta_1 & \alpha & \kappa \\
\hline
\hat{R} & 1.61 & 1.01 & 1.33 & 1.78 \\
\hline
\end{array}
\]

Figure 6.2: Samples of the log-posterior
6.1. NHPP model for Aggregated Data

using the actual data (red) and simulated data (black). The $p$-value was estimated as 0.524 meaning that there is no significant discrepancy between model and data.

![Figure 6.3: Deviance samples - actual (red) and simulated data (black)](image)

Posterior means, standard errors and 95% Cr.I. for $\beta_0$, $\beta_1$, $\alpha$ and $\kappa$ are given in Table 6.3. The estimates and standard errors for $\beta_1$ between the NHPP and the aggNHPP model are similar. However, the estimates for the other three parameters are different between the two models. This may be due to omission of temporal structure information of the deterioration process which in turn means less infor-

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior</th>
<th>s.e.</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>5.05</td>
<td>0.533</td>
<td>[3.968, 5.943]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.0049</td>
<td>0.00025</td>
<td>[0.0044, 0.0053]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>1.56</td>
<td>0.129</td>
<td>[1.340, 1.635]</td>
</tr>
<tr>
<td>aggNHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>284.15</td>
<td>125.128</td>
<td>[106.498, 368.4250]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>-5.82</td>
<td>0.342</td>
<td>[-6.454, -5.197]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.0047</td>
<td>0.00022</td>
<td>[0.0042, 0.0051]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>78.74</td>
<td>8.789</td>
<td>[63.400, 97.020]</td>
</tr>
<tr>
<td>NHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>79.99</td>
<td>6.550</td>
<td>[68.119, 93.580]</td>
</tr>
</tbody>
</table>
6.1. NHPP model for Aggregated Data

Information in modelling $\theta_i$. In this case, the model will give less ‘emphasis’ on $\theta_i$ and more on $\beta_0$ mainly because of the definition for the mean failure count:

$$\Lambda(t_a, t_b; x_i | \theta_i) = \left[ t_b^\theta_i - t_a^\theta_i \right] e^{\beta_0 + \beta_1 x_i} \quad (6.3)$$

Lack of information for estimating $\theta_i$ makes it harder to distinguish between $\theta_i$ and $\beta_0$. This is illustrated in Figure 6.4 which shows the posterior mean for each $\theta_i$ along with 95% Cr.I. in red. The range of $\theta_i$ for the aggNHPP is much smaller (values close to zero) than for $\theta_i$ from the NHPP (see Figure 4.7). This means that each $\theta_i$ has less impact on the mean number of failures (Equation (6.3)). Combined with the fact the all $\theta_i$ are considerably less that 1, the network is effectively assumed to ‘improve’. This is counterintuitive because otherwise this behaviour would be apparent in the exploratory plots in Chapter 3 (see Figure 3.4). Therefore it may be that the $\theta_i$ have so small values because the aggNHPP model does not have enough information to properly estimate them. Hence effectively, the aggNHPP is emulating a Poisson model with a mean that varies with pipe length.

A plot comparing the estimated failure ranks for each pipe with the actual ones is shown in Figure 6.5. The plot is quite similar to Figure 4.8 which refers to the ranks of the original model. The aggNHPP model is as capable of capturing this aspect of the data as well as the NHPP model.
6.1. NHPP model for Aggregated Data

In addition, the blue line in Figure 6.6 represents the estimated failure counts plotted against individual pipes. Black points correspond to the observed counts. The similarity of this plot to the equivalent one for the NHPP model, reinforces the conclusion that the aggNHPP model can capture the behaviour of the pipes in terms of failure counts just as well. However, the aggNHPP fitted values do not capture all observations exactly and this is not the case in the original NHPP model.

Figure 6.5: Estimated ranks vs actual ranks

Figure 6.6: Estimated number of failures vs individual pipes
Confusion matrices reflecting the ability of the model to capture whether or not a pipe will fail in either the observation or the prediction period are shown in Tables 6.4 and 6.5. For reference, the corresponding tables for the original NHPP model are also shown. The aggNHPP model performs relatively better than the NHPP model in this binary event for both prediction and observation period. Although this is unexpected, it is sensible since these matrices show an ‘aggregate’ measure over a period of time and this is how the aggNHPP is defined. On the other hand, the conventional NHPP model is trying to also capture the failure process for each pipe explicitly, therefore providing better understanding about the actual relationship between failure rate and time.

Table 6.4: aggNHPP confusion matrices - observation period

<table>
<thead>
<tr>
<th>aggNHPP Model</th>
<th>Estimated pipe failures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>176.6</td>
<td>169.4</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>Failed</td>
<td>136.5</td>
<td>866.5</td>
<td>1003</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>313.1</td>
<td>1035.9.8</td>
<td>1349</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NHPP Model</th>
<th>Estimated pipe failures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>141.2</td>
<td>204.8</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>Failed</td>
<td>137.0</td>
<td>866.0</td>
<td>1003</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>278.2</td>
<td>1070.8</td>
<td>1349</td>
<td></td>
</tr>
</tbody>
</table>

6.2 Aggregated Zero-Inflated NHPP

In this section, we return to the issue introduced in Chapter 5 where many data sets involving water pipes are zero-inflated in the Poisson sense. A ZINHPP was formulated where failures are either generated by a NHPP or by a zero-generating process, according to a mixing probability.
Table 6.5: aggNHPP confusion matrices - prediction period

<table>
<thead>
<tr>
<th>aggNHPP</th>
<th>Predicted pipe failures</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>836.1</td>
<td>195.9</td>
<td>1032</td>
</tr>
<tr>
<td>Failed</td>
<td>215.0</td>
<td>102.0</td>
<td>317</td>
</tr>
<tr>
<td>Total</td>
<td>1051.1</td>
<td>297.9</td>
<td>1349</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NHPP</th>
<th>Predicted pipe failures</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
<td>Failed</td>
<td>Total</td>
</tr>
<tr>
<td>Actual Pipe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Failed</td>
<td>716.3</td>
<td>315.7</td>
<td>1032</td>
</tr>
<tr>
<td>Failed</td>
<td>168.6</td>
<td>148.4</td>
<td>317</td>
</tr>
<tr>
<td>Total</td>
<td>884.9</td>
<td>464.1</td>
<td>1349</td>
</tr>
</tbody>
</table>

6.2.1 Model Formulation

To formulate an aggregated, zero-inflated NHPP (aggZINHPP) model, consider dividing the observation period for pipe \( i \) into one or more possibly non-equal time intervals \( (T_{i,0}, T_{i,1}], (T_{i,1}, T_{i,2}], \ldots, (T_{i,m_i-1}, T_{i,m_i}] \) with \( T_{i,m_i} = T_{i,end} \) such that in each interval, more than one failures \( n_{i,k} \) for \( k = 1, \ldots, m_i \) have occurred. Using previous notation, the aggZINHPP model is given by:

\[
f_{ZINHPP}(n_i; p_i, \theta_i) = \begin{cases} 
(1 - p_i) + p_i e^{-\Lambda(T_{i,0}, T_{i,end}; \theta)}, & \text{if } n_i = 0, \\
p_i L_{agg}(D_i; \theta_i), & \text{if } n_i = 1, 2, \ldots
\end{cases}
\]

where \( L_{agg}(D_i; \theta_i) \) is given by Equation (6.1) representing the likelihood of the aggregated NHPP. \( D_i = (T_{i,0}, \ldots, T_{i,m_i}, n_{i,1}, \ldots, n_{i,m_i}) \) represents the data with \( \sum_{k=1}^{m_i} n_{i,k} = n_i \). The mixing parameter \( p_i \) reflects the probability that the counts are generated by the NHPP and not a zero-generating process.

The simulations in Section 6.1.2 of this chapter, indicated that even in the worst case of aggregation, the deterioration mechanism in the pipes is adequately captured by the aggregated model. Therefore it can be assumed that the zero-inflated version of the aggregated model will be able to characterise both the failure process.
and zero-inflation just as effectively. This is because zero inflation is assumed at
the pipe level (not the failure level) so aggregation should not significantly affect
estimation of the mixing probabilities.

6.2.2 Simulation Experiments

To test the implementation of the aggZINHPP model, the same simulated data are
used, as in Chapter 5, for the ZINHPP model. The mixing probabilities \( p_i \) were
used as parameters in the distribution of \( u_i \sim \text{Bern}(p_i) \) so that when \( u_i = 1 \), failure
times are generated from a NHPP process, with a power law intensity function:

\[
\lambda(t; x_i|\theta_i) = \theta_i t^{\theta_i - 1} e^{\beta_0 + \beta_1 x_i}
\]

\( \theta_i \sim \text{Gam}(\alpha, \kappa) \)

where \( x_i \) is a single covariate. This means that:

\[
\Lambda(t_a, t_b; x_i|\theta_i) = \int_{t_a}^{t_b} \lambda(u; x_i|\theta_i)du = \left[ t_b^{\theta_i} - t_a^{\theta_i} \right] \exp\{\beta_0 + \beta_1 x_i\}
\]

When \( u_i = 0 \), the total number of failures \( n_i \) for pipe \( i \) in the observation period
are zero.

The data involve 50 hypothetical pipes, each with a different ‘shape’ random effect
\( \theta_i \). As before, the worst case of aggregation is assumed where no failure times are
actually recorded so that the only available data for each pipe consist of the total
number of failures \( n_i \) in the observation period \((T_{i,0}, T_{i,end})\). No left-truncation is
assumed so that \( T_{i,0} = 0 \) for each pipe. The relevant WinBUGS code is given in
Appendix B.

The model for each hypothetical pipe \( i \) is given by:

\[
\begin{align*}
\text{f}_{\text{ZINAPP}}(n_i; x_i|\theta_i) = & \begin{cases} 
(1 - p_i) + p_i e^{-\Lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)}, & \text{if } n_i = 0, \\
p_i \frac{\lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)}{\theta_i^\lambda} \Lambda(T_{i,0}, T_{i,end}; x_i|\theta_i)^{n_i}, & \text{if } n_i = 1, 2, \ldots 
\end{cases} 
\end{align*}
\]

(6.4)
6.2. Aggregated Zero-Inflated NHPP

This gives the log-likelihood for each pipe given the random effects as:

\[
\ell_i(n_i, T_{i,0}, T_{i,end}; p_i, \beta_0, \beta_1|\theta_i) = I_{(n_i>0)} \{ \log p_i - \log n_i! \\
- \left[ T_{i,0}^\theta - T_{i,end}^\theta \right] \exp\{\beta_0 + \beta_1 x_i\} + n_i \log \left( \left[ T_{i,0}^\theta - T_{i,end}^\theta \right] \exp\{\beta_0 + \beta_1 x_i\} \right) \}
+ I_{(n_i=0)} \{ \log (1 - p_i + p_i \exp \left( \left[ T_{i,0}^\theta - T_{i,end}^\theta \right] \exp\{\beta_0 + \beta_1 x_i\} \right)) \}
\]

The log-likelihood for all 50 pipes is then given by:

\[
\sum_{i=1}^{50} \ell_i(n_i, T_{i,0}, T_{i,end}; p_i, \beta_0, \beta_1|\theta_i)
\]

The aggZINHPP was applied to the same 100 data sets generated for the original ZINHPP in Chapter 5. As before, parameter posteriors samples were collected for each of the 100 fitted models and then averaged across all models. Recall also that failures were simulated for 10 years after each \(T_{i,end}\) so that the predictive capability of the model may be tested. Two MCMC chains were run for each fitted model, each with dispersed initial values. In total, each chain was run for 50000 iterations with a burn-in of 25000 and a thinning of 5 to ensure convergence and adequate mixing.

The parameters of the model were given the same uninformative priors as before and these are shown in Table 6.6 and Table 6.7. For comparison purposes, the appropriate results from the original ZINHPP model are also given. Posterior means for all parameters and especially the \(p_i\) are comparable for the two models. The standard errors of the parameters in Table 6.6 are larger for the aggregated model. Again, this probably reflects increased model uncertainty due to aggregation.

In Figure 6.7, the ranks of \(p_i\) are plotted against the ranks of \(\hat{p}_i\), which are calculated as the posterior means of ranks. The plot is almost identical to the equivalent from the non-aggregated model, meaning that the aggZINHPP captures zero-inflation just as well as the ZINHPP.

Figure 6.8 shows the actual number of failures in the 10 year prediction period against the predicted number of failures (taken as the mean of the posterior pre-
6.2. Aggregated Zero-Inflated NHPP

Table 6.6: Simulated aggZINHPP results

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Prior</th>
<th>Actual Value</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aggZINHPP</td>
<td>$\beta_0$</td>
<td>$\text{N}(0, 1000)$</td>
<td>-8</td>
<td>-7.45 (0.305)</td>
<td>[-8.038,-6.847]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\beta_1$</td>
<td>$\text{N}(0, 1000)$</td>
<td>0.06</td>
<td>0.049 (0.001)</td>
<td>[0.046,0.052]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\alpha$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>60</td>
<td>97.29 (10.262)</td>
<td>[78.319,118.343]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\kappa$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>40</td>
<td>65.434 (5.035)</td>
<td>[56.263,76.025]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\beta_0$</td>
<td>$\text{N}(0, 1000)$</td>
<td>-8</td>
<td>-7.57 (0.111)</td>
<td>[-7.795,-7.359]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\beta_1$</td>
<td>$\text{N}(0, 1000)$</td>
<td>0.06</td>
<td>0.052 (0.002)</td>
<td>[0.049,0.055]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\alpha$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>60</td>
<td>104.17 (5.209)</td>
<td>[94.184,114.299]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\kappa$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>40</td>
<td>69.18 (3.269)</td>
<td>[62.931,75.577]</td>
</tr>
</tbody>
</table>

Table 6.7: Simulated aggZINHPP results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Min</th>
<th>Max</th>
<th>Posterior Mean</th>
<th>25%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>$p_i$</td>
<td>0.11</td>
<td>1.00</td>
<td>0.73</td>
<td>0.28</td>
<td>0.97</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$p_i$</td>
<td>Beta(1,1)</td>
<td>0.37</td>
<td>0.66</td>
<td>0.56</td>
<td>0.43</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$p_i$</td>
<td>Beta(1,1)</td>
<td>0.37</td>
<td>0.67</td>
<td>0.56</td>
<td>0.44</td>
</tr>
</tbody>
</table>

dictive distribution). The $45^\circ$ line is also plotted for reference as well as 95% prediction intervals.

The predictions compare rather well with the non-aggregated model (see Figure 5.2). This, together with the earlier comments on parameter estimates, shows that the aggZINHPP model can capture enough of the deterioration mechanism embedded in the data to perform just as well as the ZINHPP both in terms of model fit and predictive capability.
6.2. Aggregated Zero-Inflated NHPP

In this section, we return to the North American data set which relates to an underground network of 1349 water pipes in a North American municipality. The objective is to implement the aggZINHPP model to this data and compare with previous results.

Figure 6.7: Rank of $p_i$ vs rank of $\hat{p}_i$

Figure 6.8: Actual vs predicted number of failures

6.2.3 Model Application

In this section, we return to the North American data set which relates to an underground network of 1349 water pipes in a North American municipality. The objective is to implement the aggZINHPP model to this data and compare with previous results.
6.2. Aggregated Zero-Inflated NHPP

In Chapter 5, a ZINHPP model was applied to this data set, where we recall that the failure rate was defined as:

$$\lambda(t; x_i|\theta_i) = \theta_i t^{\theta_i - 1} e^{\beta_0 + \beta_1 x_i}$$

$$\theta_i \sim \text{Gam}(\alpha, \kappa)$$

with $x_i$ corresponding to pipe length. Here the same model is applied to a fully aggregated version of the data used earlier in Section 6.1.3 to apply the aggNHPP model.

The count of failures $n_i$ for each pipe $i = 1, \ldots, 1349$ in the time period $(T_i, 0, T_{i,\text{end}}] = (1969, 1998]$, is distributed according to Equation (6.4). Recall also that the mixing probability $p_i$ was modelled as:

$$\logit(p_i) = \gamma_{i,0} + \gamma_1 x_i + \gamma_2 T_{i,\text{end}}$$

where $\gamma_{i,0}$ are fixed effects. The model is then effectively the same as in the simulation studies earlier except for the particular parameterisation of $p_i$. Like before, the 5 year period (1998, 2003] is considered for prediction.

Two MCMC chains were run in WinBUGS for 90000 iterations with a burn-in of 60000 and thinning of 10. This resulted in 3000 samples from each chain. The $\hat{R}$ diagnostic was used to monitor convergence. All 1355 values are close to 1 (see Figure 6.9) implying convergence.

![Figure 6.9: $\hat{R}$ for 1355 model parameters](image-url)
6.2. Aggregated Zero-Inflated NHPP

Convergence is also indicated by a trace plot of log-posterior samples shown in Figure 6.10.

![Figure 6.10: Samples of the log-posterior](image)

Figure 6.10: Samples of the log-posterior

Figure 6.11 depicts samples of the deviance for simulated data (in black). Also plotted, are samples of the deviance for the actual data (in red). The $p$-value is estimated as 0.292 indicating adequate model fit.

![Figure 6.11: Deviance samples - actual (red) and simulated data (black)](image)

Parameter estimates of the six global parameters $\beta_0, \beta_1, \alpha, \kappa, \gamma_1$ and $\gamma_2$ are shown in Table 6.8 along with their standard errors and 95% Cr.I. For reference, the equivalent summaries are also given for the ZINHPP model from Chapter 5. All parameter estimates between the two models are comparable, unlike the case with the NHPP and aggNHPP (see Table 6.3). This indicates that the aggZINHPP
model is better than the aggNHPP is the sense that it has a ‘closer’ fit to its non-aggregated counterpart, at least in terms of posterior means. However, there is a slight discrepancy between the posterior means of $\beta_0$ and the parameters $\alpha$ and $\kappa$ for the random effects $\theta_i$, between the two models. In addition, the standard errors for those three parameters are higher for the aggZINHPP. Once more, this can be attributed to the difficulty in distinguishing between $\beta_0$ and $\theta_i$ when no data are available for actual burst times.

The conclusion that the aggZINHPP fits ‘closer’ to its non-aggregated counterpart than the aggNHPP is reinforced by the first plot in Figure 6.12 which relates to posterior means for $\theta_i$. The values appear much closer to the ones from the ZINHPP model (see Figure 5.3). This is certainly not the case when comparing values of $\theta_i$ between the NHPP and aggNHPP models. The posterior mean estimates of $p_i$ also appear to be very similar to the equivalent estimates from the ZINHPP model.

What these results seem to suggest, is that taking account of zero inflation in the

### Table 6.8: aggZINHPP parameter estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior Mean</th>
<th>s.e.</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aggZINHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>-6.460</td>
<td>0.928</td>
<td>[-7.609, -5.399]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.0036</td>
<td>0.00023</td>
<td>[0.0032, 0.0041]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>127.33</td>
<td>21.088</td>
<td>[94.179, 170.301]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>112.089</td>
<td>10.560</td>
<td>[93.229, 133.503]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\gamma_1$</td>
<td>N(0, 1000)</td>
<td>0.203</td>
<td>0.034</td>
<td>[0.143, 0.276]</td>
</tr>
<tr>
<td>aggZINHPP</td>
<td>$\gamma_2$</td>
<td>N(0, 1000)</td>
<td>-0.0147</td>
<td>0.0043</td>
<td>[-0.0232, -0.0070]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\beta_0$</td>
<td>N(0, 1000)</td>
<td>-5.463</td>
<td>0.236</td>
<td>[-5.947, -5.034]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\beta_1$</td>
<td>N(0, 1000)</td>
<td>0.0037</td>
<td>0.00023</td>
<td>[0.0032, 0.0041]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\alpha$</td>
<td>Gam(0.5, 0.005)</td>
<td>98.87</td>
<td>9.001</td>
<td>[83.479, 117.700]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\kappa$</td>
<td>Gam(0.5, 0.005)</td>
<td>200.822</td>
<td>7.820</td>
<td>[87.320, 118.001]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\gamma_1$</td>
<td>N(0, 1000)</td>
<td>0.212</td>
<td>0.033</td>
<td>[0.158, 0.286]</td>
</tr>
<tr>
<td>ZINHPP</td>
<td>$\gamma_2$</td>
<td>N(0, 1000)</td>
<td>-0.0158</td>
<td>0.0041</td>
<td>[-0.0241, -0.0129]</td>
</tr>
</tbody>
</table>
6.2. Aggregated Zero-Inflated NHPP

Figure 6.12: Posterior means and Cr.I. for $\theta_i$ and $p_i$

aggregated model has helped greatly in capturing as much information about the temporal structure as possible, through the random effects $\theta_i$. A possible reason for that may be due to the fact that pipes with ‘unexpected’ zero failures have been accounted for by zero-inflation. Therefore it is easier for the model to adjust parameter estimation given the effect of pipe length.

Samples from the posterior predictive distributions of failure ranks for each pipe were collected and in Figure 6.13 the means of those distributions are plotted against the observed ranks. 95% prediction intervals are also plotted in red. The plot is almost identical to the corresponding plot for the ZINHPP model (Figure 5.7) showing that the aggregated model can capture the ranks just as well.

Furthermore, in Figure 6.14 the actual failure counts for each pipe during the
6.2. Aggregated Zero-Inflated NHPP

Figure 6.13: Estimated ranks vs actual ranks

The whole 35 year period (both observation and prediction period) are shown along with estimated analogue (in blue). The plot is quite similar to the one for the ZINHPP (Figure 5.7) although some points are noticeably not captured as well.

Figure 6.14: Estimated number of failures vs individual pipes
Confusion matrices are now considered which compare the (observed) event of more than one failure per pipe with the estimated analogue for each of the two time periods (observation and prediction). The matrices for the aggZINHPP model are given in Tables 6.9 and 6.10. For reference, the corresponding tables for the ZINHPP model are also provided. Comparing these for the observation period reveals that the performance of the two models is almost equivalent with the aggregated version slightly underachieving. The same can be said for the prediction period which means that the aggregated model has a predictive capability which is roughly equal to the ZINHPP model.

### Table 6.9: aggZINHPP confusion matrices - observation period

<table>
<thead>
<tr>
<th>aggZINHPP Model</th>
<th>Estimated pipe failures</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Pipe Failures</td>
<td>Not Failed</td>
<td>Failed</td>
</tr>
<tr>
<td>Not Failed</td>
<td>197.6</td>
<td>148.4</td>
</tr>
<tr>
<td>Failed</td>
<td>123.8</td>
<td>879.2</td>
</tr>
<tr>
<td>Total</td>
<td>321.4</td>
<td>1027.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ZINHPP Model</th>
<th>Estimated pipe failures</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Pipe Failures</td>
<td>Not Failed</td>
<td>Failed</td>
</tr>
<tr>
<td>Not Failed</td>
<td>188.8</td>
<td>157.2</td>
</tr>
<tr>
<td>Failed</td>
<td>119.0</td>
<td>884.0</td>
</tr>
<tr>
<td>Total</td>
<td>307.8</td>
<td>1041.2</td>
</tr>
</tbody>
</table>

In summary, the aggZINHPP performed very similarly to the ZINHPP. The analysis has also revealed another interesting fact which is that zero-inflation in the aggregated model, resulted in parameter estimates which were comparable to the non-aggregated model. This was not true for the aggNHPP which had quite different parameter estimates to the NHPP in Chapter 4. Namely, the lack in distinguishing between parameters $\beta_0$ and random effects $\theta_i$ is much less of an issue when zero-inflation is accounted for.
Table 6.10: aggZINHPP confusion matrices - prediction period

<table>
<thead>
<tr>
<th>aggZINHPP</th>
<th>Predicted pipe failures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
</tr>
<tr>
<td>Actual Pipe Failures</td>
<td>Not Failed</td>
</tr>
<tr>
<td></td>
<td>Failed</td>
</tr>
<tr>
<td></td>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ZINHPP</th>
<th>Predicted pipe failures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Failed</td>
</tr>
<tr>
<td>Actual Pipe Failures</td>
<td>Not Failed</td>
</tr>
<tr>
<td></td>
<td>Failed</td>
</tr>
<tr>
<td></td>
<td>Total</td>
</tr>
</tbody>
</table>

6.3 Summary

In this chapter, the concept of aggregation in data sets involving recurrent failure times was introduced. This presents an issue when the goal is to capture the deterioration process. This particular problem has not been investigated in the literature and the specific goal of this chapter was to investigate this using the NHPP. The defining NHPP property stating that failure counts in disjoint intervals is Poisson distributed was utilised to formulate a NHPP model for aggregate data.

In the first part of the chapter, the mixed effects NHPP model from Chapter 4 is revisited and adjusted to account for aggregation. Simulation experiments suggested that the aggNHPP model is able to capture the failure mechanism almost as well as the original model, with the exception that there may be some identifiability issues between the scale parameter $\beta_0$ and the random effects $\theta_i$ of the power law based failure rate. This problem becomes more evident in the application of the model to the North American data where the aggNHPP behaves quite differently from the NHPP in terms of parameter estimates.
6.3. Summary

In the second part of the chapter, the aggregated model is further extended to account for zero-inflation, as in Chapter 5. Simulation experiments showed that the aggZINHPP model performs reasonably closely to the ZINHPP. The application of the model on the real data, resulted in good model fit but more importantly in similar parameter estimates between ZINHPP and aggZINHPP. The identifiability between $\beta_0$ and $\theta_i$ was still apparent but to a significantly lesser extend.

Work related to this chapter, was presented in Economou et al. (2007) where an aggregated mixed effects NHPP model is applied to a network of 532 water pipes in New Zealand. A non-aggregated model was also implemented for comparison and the ability of the aggregated model to perform as well as the conventional NHPP model was demonstrated.

NHPP models are key in this thesis and results from Chapters 4-6 are encouraging. However, the conventional NHPP is unable to account for a specific characteristic which is sometimes present in recurrent event data. That phenomenon is a dependency structure not in time, but in the failures themselves. A possible solution would be to add an autoregressive term in the failure rate (e.g. number of previous failures). A better solution would be to induce a more structured dependency configuration using an appropriate probability model and this is done in the following chapter where the possibility of allowing the failure rate to depend on Markov models is explored.
Chapter 7

Hidden Semi-Markov NHPP

So far in the thesis, the NHPP has been central in modelling deterioration in water pipes. However, even with the zero-inflation modification, the NHPP model is unable to account for some features commonly found in recurrent event data.

In this chapter, the possibility of extending the NHPP model to include latent or hidden semi-Markov structure is explored and the implementation of such a model is discussed. Model application includes simulation studies as well as selected pipes from the North American data set.

Formulating a Bayesian hidden semi-Markov NHPP has not been discussed in the literature and here it is presented as a novel way of handling state changes in the process giving rise to recurrent failures.

7.1 Motivation

The NHPP model has been used in this thesis, mainly because it is ideal for modelling repairable components such as underground water pipes which are expected to deteriorate with time. The definition and properties of the NHPP render it easy to work with, as was shown in Chapters 5 and 6 where the model was extended to
account for zero-inflation and aggregation both separately and simultaneously.

However, Thompson (1981) states that in a NHPP, the entire process is determined by the distribution of time to first failure. Although desirable in some cases, this property is in general a drawback since it limits the model from capturing situations where the failure process behaves unexpectedly. For example consider the pipe network in the North American data where total monthly failure counts increased but then unexpectedly decreased in the last 5 years or so (Figure 3.4).

Even though the flexibility of the NHPP is increased considerably by formulating it as a mixture as in Chapter 5, the mixture occurs at the pipe level and not at the failure level. A possible remedy would be to actually force the mixing probability to depend on time so that this will introduce an extra level of temporal structure other that the one from the NHPP. This results in a process which is even more flexible that the ZINHPP.

In addition, recurrent failures may exhibit clustering due to a dependency structure of the failures. The conventional NHPP model does not allow for such behaviour which potentially limits model flexibility. Furthermore, in the context of repairable systems in complex environments, there may be unobserved processes influencing the underlying failure mechanism. For example periods of severe weather, are a possible factor affecting the ageing in water pipes.

Consider the failures in Figure 7.1 which were caused by a process whose rate is increasing exponentially with time. The power law based NHPP model considered so far is capable of capturing this relatively easily.

![Figure 7.1: Exponentially increasing intensity function](image)
7.1. Motivation

Quite often though, one may find that the rate is not only dependent on time but on other quantities or even processes which in many cases are unobserved. For instance, the failures may appear to be repelling each other in some time intervals and to be clustered in some others as in Figure 7.2. In many applications, this behaviour can be attributed to state changes over time in the component of interest, meaning that the failure rate is different during each state.

Whether the recurrent failures arise from complex devices or from patients undergoing treatment, it is quite common that the individuals of interest exhibit such state changes due to possibly unobserved factors which may also be time dependent. Newby (1993) states that engineers have always been more interested in models that deal with structural dependence, both in a deterministic sense but also through Markov models. He goes on to say that such models can capture the essence of system behaviour and at the same time allow for the exploration of different scenarios.

A framework which introduces temporal state changes as well as implicitly inducing a dependency structure in the failures, are Markov modulated models (e.g. Özekici and Soyer (2003); Scott and Smyth (2003)) where the rate of occurrence is governed by a Markov process. In addition, if the Markov process is latent then this refers to hidden Markov models (HMM), e.g. in Betro et al. (2008) the authors model daily rainfall data using a HMM where there is an underlying correspondence between the hidden Markov states and the concept of discrete weather states.

Markov modulation of the NHPP will introduce a dependency structure in the failures. Furthermore, the problem of the NHPP being driven by the distribution
of time to first failure is eliminated.

In the remainder of this chapter, the possibility of allowing the failure rate of the NHPP to depend on a hidden Markov chain in order to account for irregular behaviour in the failure process is investigated. Furthermore, a hidden semi-Markov NHPP is formulated and implemented using MCMC.

### 7.2 Hidden Markov Models

In this section a short review of recent work involving application of HMMs is presented.

HMMs were introduced in the late 1960s and since then they have been extensively studied and used. A paper by Rabiner (1989) provides a good description of the history of HMMs. That article can be considered as fundamental in its role as a detailed tutorial on HMMs and their particular utilisation in speech recognition, which has always been the primary use of these models. The monograph by MacDonald and Zucchini (1997) is one of the few available texts which offers a detailed description of HMMs and examples of their applications.

The basic premise of HMMs is simple. A model is allowed to depend on an unobserved discrete time Markov chain, usually through its parameters, and this results in a ‘joint’ model which is more complex but also more flexible. An example would be one where the ‘mother’ model is a Poisson with mean $\mu(S)$ where $S$ relates to one of two possible states of an unobserved Markov chain. $\mu(1)$ can be different than $\mu(2)$ so that the resulting model may capture complicated situations where for instance counts exhibit temporal clustering.

Kozumi (2000) considered a model for discrete survival data, where the hazard function is made up of a sequence of baseline hazard functions and allowed these to vary according to a discrete time HMM. As a result, each baseline hazard function is different in separate time intervals depending on how long each hidden state is
7.2. Hidden Markov Models

held for. The author shows how to implement a Bayesian model using Metropolis-Hastings and presents an example where the data consist of information on head-and-neck cancer patients in 1-month intervals.

In the article by Betro et al. (2008), the authors consider a discrete time HMM and relate the hidden states to the concept of discrete weather changes. Their aim is to model extreme rainfall events in a small area in Italy and the latent chain is the underlying process whereas the ‘mother’ model is a Weibull distribution. Model parameters are estimated using the EM algorithm.

Another application where the hidden states of a latent discrete time Markov chain have a physical interpretation is in the monitoring of CD4 cell count which helps in the monitoring of health status and disease progression of HIV-infected patients (Jouyaux et al., 2000). The authors assume a hidden chain to model the true underlying stages of the disease and use the ‘mother’ model to account for the noise in the resulting observed process. The model is Bayesian and Gibbs sampling is used for estimation.

The flexibility of HMMs has also allowed their use in econometrics and the paper by Rydén et al. (1998) considers an HMM in the context of daily returns from several industry portfolios. The model is estimated using the EM algorithm which is the most conventional way of fitting HMMs (discussed later in Section 7.4.5).

Estimation and inference for HMMs using maximum likelihood can be found in MacDonald and Zucchini (1997) and references therein. Also, Rydén and Titterington (1998) present and theoretically discuss a Gibbs sampler for analysing HHMs as well as hidden Markov fields where a pseudolikelihood is used instead of the likelihood itself. In addition, a thorough paper by Scott (2002) discusses Bayesian MCMC methods for hidden Markov models. The author discusses methods for evaluating model likelihood using matrix versions of existing recursive algorithms. The article addresses issues such as label switching, explores estimation techniques such as maximum a posteriori estimation, Gibbs sampling and Metropolis-Hastings,
Markov Modulated Poisson Processes

Doubly stochastic counting processes whose rate depends on a Markov process have been studied in the past and in general these are called Markov modulated processes (MMP). In general, the Markov process in an MMP is assumed to be unobserved but whether it is discrete or continuous usually depends on the application. Özekici (1997) for instance considers the probabilistic formulation as well as ergodic and transient analysis of a Markov modulated Bernoulli process (MMBP) where the probability of success depends on a discrete Markov chain. That MMBP is further considered in Özekici and Soyer (2003) where the authors investigate model inference within a Bayesian framework. They derive posterior distributions analytically for cases where the Markov process in observable, but also describe a Gibbs sampler applicable when the modulating process is hidden.

Furthermore, Özekici and Soyer (2006) investigate a semi-Markov modulated Poisson process where the intensity function depends on a continuous time semi-Markov process. The authors consider transient and ergodic properties as well as Bayesian
model inference where they use suitable priors to analytically derive parameter posteriors and predictive distributions. For clarity, the acronym MMPP will be used to denote a general Markov modulated Poisson process and whether the Poisson process conditional on the latent chain it is homogeneous or not, will be specified explicitly.

In a MMPP, the difference between assuming an observed or a hidden Markov process is considerable since in the latter case, evaluation of the likelihood needs to take into account all the possible trajectories of the latent process. For discrete state Markov chains, this usually implies computationally costly combinatoric summations and equivalent sets of integrals for continuous Markov processes. For the discrete case (HMMs), the so called ‘forward-backward’ algorithm has been introduced by Baum and Egon (1967) and Baum and Sell (1968) to solve the issue of efficient evaluation of the likelihood. The algorithm makes use of the Markovian structure of the model, and at each discrete time step, accumulates the joint probability of the model and the latent chain given the observed data up to that point. This is investigated further in Section 7.4.3.

A homogeneous MMPP model with two hidden states was applied by Scott (1999) where the latent process was allowing for illegal access to phone accounts and the recurrent events referred to phone calls. Although the author assumed a continuous latent process, the forward-backward algorithm was still utilised since under certain assumptions, the continuous Markov process may be discretised. The model is implemented using Gibbs sampling.

Scott and Smyth (2003) consider a non-homogeneous MMPP model and show that this model may be viewed as a superposition of two latent Poisson processes that are governed by an unobserved continuous Markov process, which can be discretised under certain conditions. The model is then ultimately a HMM so that a forward-backward algorithm may be utilised in conjunction with Gibbs sampling. More recently, Fearnhead and Sherlock (2006) developed a generalisation of the forward-backward algorithm for a nonhomogeneous MMPP with a continuous time hidden
process and defined a Gibbs sampler to implement the model on data involving the
evolution of genomes.

Rydén (1996) discusses methods for fitting MMPPs and specifically presents an
EM algorithm along with several ways of implementation.

### 7.4 Hidden Markov NHPP Model

The idea of allowing the failure rate of a NHPP to depend on another probabilistic
model was introduced in Cox and Isham (1980) as a general class of models termed
doubly stochastic Poisson processes. Here, we develop on that idea and consider
a situation where the failure $\lambda(t)$ of the NHPP, is allowed to vary according to
a discrete Markov chain. Note that HMMs are considered as a natural extension
of mixture models especially in cases where there is a temporal dimension in the
data. In that sense, the work in this chapter is an extension of the discussions
about mixtures of NHPPs in Chapter 5. In addition, HMMs implicitly introduce
correlation in the recurrent failures which is a desired effect in many applications
including underground water pipes.

#### 7.4.1 Model Formulation

Consider a NHPP process with an intensity function $\lambda(t, S)$ whose parameters
are varying according to a latent Markov chain with discrete state space $S \in
\{1, 2, \ldots, M\}$. We label this model using the acronym NHPP-HMM. The hidden
chain, is defined by an initial distribution $\pi = (\pi(1), \ldots, \pi(M))$ and a transition
matrix $P = \{p_{i,j}\}$ where $\sum_j p_{i,j} = 1$, such that:

$$
\{p_{i,j}\} = 
\begin{pmatrix}
  p_{1,1} & p_{1,2} & \cdots & p_{1,M} \\
  p_{2,1} & p_{2,2} & \cdots & \vdots \\
  \vdots & \vdots & \ddots & \vdots \\
  p_{M,1} & \cdots & \cdots & p_{M,M}
\end{pmatrix} = \Pr(S = j|S = i \text{ at previous time step})
$$
7.4. Hidden Markov NHPP Model

where \( S \in \{1, 2, \ldots, M\} \) with \( M \) being the number of possible discrete states.

Suppose further, that the observation period is \((T_0, T_{end}]\) and that \( n \) failures have been observed at times \( T_0 < t_1, t_2, \ldots, t_n \leq T_{end} \). Assume also that the observation period can be divided into \( T_{end} - T_0 \) discrete time steps \((T_1, T_2, \ldots, T_{end})\). In the background, the chain which operates in discrete time, starts at \( T_0 \) at an arbitrary state \( S_1 \) according to the initial distribution \( \pi \). So for the time period \((T_0, T_1]\) the chain is at state \( S_1 \). The chain will then jump to the next state (which could be the same as the old) according to the row of \( P \) corresponding to the current state, and this carries on until \( T_{end} \).

Supposing that a total of \( Q \) state changes have occurred, Figure 7.3 shows a particular instance of the NHPP-HMM process with failure rate \( \lambda(t, S) \). The picture shows that the chain starts at state \( S_1 \) and holds that state for the first 5 time steps, then jumps to state \( S_2 \) which is held for 3 time steps and so on. The mother NHPP remains the same during the time period that a particular state is held for.

![Figure 7.3: Hidden Markov NHPP model](image-url)
7.4. Hidden Markov NHPP Model

7.4.2 Likelihood

The likelihood of a NHPP-HMM is effectively the conditional likelihood of the NHPP given the latent chain. To formulate this, assume that the chain has in fact been observed in \([T_0, T_{\text{end}}]\) so that available data consist of failure times \((t_1, \ldots, t_n)\) and chain states \(S_1, S_2, \ldots, S_m\) at each discrete time step \(T_1, T_2, \ldots, T_m = T_{\text{end}}\). Also, recall that the independent increments property of the NHPP allows the likelihood to be expressed as independent Poisson counts at disjoint time intervals.

For consistency with previous notation (Equation (4.6)), let the NHPP likelihood in the arbitrary time interval \((t_a, t_b]\) be expressed as \(L_{\text{PP}}(N(t_a, t_b); \theta|S)\) where the Poisson random variable \(N(t_a, t_b)\) denotes the count of failures that have occurred in \((t_a, t_b]\) and \(S\) corresponds to the state of the chain during that interval. \(\theta\) refers to parameters of the NHPP. The likelihood of the recurrent failures and the chain is then:

\[
L_{\text{MPP}}(D; \Theta) = \pi(S_1)L_{\text{PP}}(N(T_0, T_1); \theta|S_1) \prod_{k=2}^{m} p_{S_{k-1}, S_k} L_{\text{PP}}(N(T_{k-1}, T_k); \theta|S_k)
\]

where \(D = (T_0, t_1, \ldots, t_n, T_{\text{end}}; S_1, \ldots, S_Q)\) and \(\Theta = (\theta, P, \pi)\)

When the discrete time steps of the chain match the time scale at which failures \(t_1, \ldots, t_n\) were recorded it is more convenient that \(L_{\text{PP}}(N(t_a, t_b); \theta|S)\) denotes the Poisson likelihood at each time step (the data need to be re-expressed in terms of failure counts in each discrete time step). However, if the failures were recorded at a smaller timescale than the one corresponding to the chain, then \(L_{\text{PP}}(N(t_a, t_b); \theta|S)\) should represent the appropriate likelihood contribution of the NHPP relating to the failures that occurred in each discrete time step. Either way, the independent increments property of the NHPP allows for both ways of expressing the NHPP likelihood.

Although \(L_{\text{MPP}}\) is well defined, it was derived under the assumption that the states \(S_1, \ldots, S_m\) have actually been observed. To define the likelihood of the NHPP-HMM one needs to assume that no data are available about the latent process.
so $L_{MPP}$ needs to be summed over all possible states at each time step. This
twill reflect the fact that the chain has not been observed so the likelihood of the
NHPP-HMM is given by:

$$L_{HMPP}(T_0, t_1, \ldots, t_n, T_{end}; \Theta) = \sum_{S_1} \cdots \sum_{S_m} L_{MPP}(D; \Theta)$$ (7.1)

The combinatoric summation in Equation (7.1) is known to cause computational
issues especially for data concerning large observation periods. For that reason,
recursive algorithms have been developed (Rabiner, 1989) which effectively utilise
the Markovian structure of general HMMs to accumulate the joint probability of
the hidden states at each time step.

### 7.4.3 Recursive Algorithms: Forward

The concept behind forward-backward recursive algorithms is that a forward ‘pass’
over the data, collects the probability of being in a particular state at each time
step. Formally, a forward variable may be defined as

$$v_T(j) \equiv \Pr(\text{data up to } T \text{ and chain is at state } S_T = j | \theta)$$

where $\theta$ refers to parameters of the mother process (NHPP). The variable is cal-
culated recursively as each one accumulates information up to and including the
specific time step $T$.

For the NHPP-HMM model discussed above, a forward variable may be defined as:

$$v_T(j) = \sum_{i=1}^{M} \Pr(S_{T-1} = i, S_T = j, \text{data up to } T | \theta)$$

$$= L_{PP}(N(T-1, T); \theta) | S_T = j \sum_{i=1}^{M} p_{i,j} v_{T-1}(i)$$

where $v_1(j) = \pi(j) L_{PP}(N(T_0, T_1); \theta) | S_1 = j$. Clearly, forward variables can be
used to calculate the likelihood as they accumulate the necessary information. In
particular, summing the last variable \( v_{T_{\text{end}}}(j) \) over all possible states will effectively give the likelihood exactly as in Equation (7.1).

Working with forward variables presents an elegant way of evaluating the joint distribution of the data and the hidden chain. However, it also means working with multiplications of probabilities which will rapidly get smaller leading to potential computer underflow. Within the HMM literature (see Devijver (1985) and references therein), an efficient way of dealing with underflow is by scaling the forward probabilities at each time step.

The argument behind scaling, is that one works with the conditional distribution of the states instead of the joint distribution of the states and the data. Therefore, at each time step, one works with the probability distribution of the states, which will properly sum to one. However, care is needed when the intention is to evaluate the likelihood and one must keep track of the scaling factors for proper likelihood calculation at the last time step.

Scott (2002) gives a matrix representation of the forward (and backward) algorithm. More intuitive and convenient calculations are used which are especially useful for coding the algorithms in a computer language. The idea is to work with matrices \( A_T = \{a_{T,i,j}\} \) with \( T = T_1, \ldots, T_{\text{end}} \), where each element:

\[
a_{T,i,j} = \Pr(S_{T-1} = i, S_T = j| \text{data up to } T, \theta)
\]

reflects the conditional probability of going to state \( j \) at \( T \) from state \( i \) at time \( T - 1 \) given the data up to time \( T \). These matrices are more ‘intuitive’ since they resemble transition matrices. Summing the rows of \( A_T \) for instance, gives vector:

\[
\alpha_T(j) = \Pr(S_T = j| \text{data up to } T, \theta)
\]

with \( j = 1, \ldots, M \) elements, each one corresponding to the probability of being in a particular state at time \( T \) given the data up to \( T \). Note that \( v_T(j) \) and \( \alpha_T(j) \) are related - the latter is just a scaled version of the former, so that its elements add up to 1.
7.4. Hidden Markov NHPP Model

Using these matrices, the likelihood may still be effectively calculated as a byproduct. The advantage of using $A_T$ is that each of the matrices is scaled to sum to 1, to avoid underflow. The matrices can be computed recursively and specifically for the NHPP-HMM these are given by:

$$a_{T,i,j} \propto \Pr(S_{T-1} = i, S_T = j, \text{ data at } T| \text{ data up to } T-1, \theta)$$

$$= \alpha_{T-1}(i)p_{i,j}L_{PP}(N(T-1,T); \theta)|S_T = j)$$

(7.2)

where $a_{T_1,i,i} \propto \pi(i)L_{PP}(N(T_0,T_1); \theta|S_{T_1} = i)$ and $a_{T_1,i,j} = 0$

where proportionality disappears by dividing with a scaling factor so that $A_T$ sums to 1. For the purposes of likelihood calculation, one needs to keep track of those scaling factors at each step since these are proportional to the likelihood itself.

The likelihood at time step $T$ is evaluated by multiplying the scaling factor at $T$ with the likelihood at $T - 1$. This means that before dividing Equation (7.2) with the scaling factor at $T$, one needs to adjust (or ‘re-scale’) by multiplying with the likelihood at $T - 1$. The benefit of doing this is that the scaling factor involves summing Equation (7.2) over $i$ and $j$ and this can be done avoiding underflow. This is because only scaled values are involved: the conditional distribution of states at the previous time step, probabilities from $P$ and probabilities (or density values) from the mother process.

Note that ‘re-scaling’ is not necessary at each time step and this is shown in Devijver (1985). In that case, one simply needs to multiply all the scaling factors at the last time step to get the likelihood. Here the re-scaling method is preferred since this can be directly generalised to hidden semi-Markov models (HSMMs), which is one of the main purposes of this chapter.

For increased stability, logs may used wherever possible, for instance when summing Equation (7.2) over $i$ (summing the rows of $A$), the quantity $L_{PP}(N(T - 1,T); \theta)|S_T = j)$ can be factorised out so that logs may be employed (this point is revisited later in Section 7.6).
7.4. Hidden Markov NHPP Model

7.4.4 Recursive Algorithms: Backward

Backward recursion, which starts at the end of the observation period, uses the already calculated forward variables to update the joint distribution at each time step $T$ with the rest of the information in the data (i.e. data corresponding to the time period $[T + 1, T_{\text{end}}]$). A backward variable may be defined as

$$z_T(j) \equiv \Pr(\text{all data and chain is at } S_T = j | \theta)$$

Although backward recursion in not required to calculate the likelihood, it is necessary when interest lies in the distribution of the hidden chain which may be used to investigate quantities such as the most likely state trajectory or if one wishes to generate response values from the fitted model.

Scott (2002) describes how forward matrices $A_T$ may be updated to derive backward matrices $A'_T = \{a'_{T,i,j}\}$ such that:

$$a'_{T,i,j} = \Pr(S_{T-1} = i, S_t = j | \text{data up to } T_{\text{end}}, \theta)$$

reflecting the conditional probability of going from state $i$ at $T - 1$ to state $j$ at $T$ given the whole data. The idea is to start at the end where intuitively, $A'_{T_{\text{end}}} = A_{T_{\text{end}}}$. Summing the columns of $A'_{T_{\text{end}}}$ gives the vector $\alpha'_{T_{\text{end}}}(j)$ which is defined as:

$$\alpha'_{T_{\text{end}}}(j) = \Pr(S_{T_{\text{end}}-1} = j | \text{data up to } T_{\text{end}}, \theta) \quad (7.3)$$

To derive the next matrix in the backward recursion $A'_{T_{\text{end}}-1}$, the idea is to consider each column $j$ of $A_{T_{\text{end}}-1}$ which when divided by its margin $\alpha_{T_{\text{end}}-1}(j)$ gives a vector with elements $j = 1, \ldots, M$ summing to 1 corresponding to:

$$\Pr(S_{T_{\text{end}}-2} = i, S_{T_{\text{end}}-1} = j | \text{data up to } T_{\text{end}} - 1, \theta) \quad (7.4)$$

Multiplying each element $j$ of this vector with $\alpha'_{T_{\text{end}}}(j)$ means that one is multiplying Equation (7.4) with:

$$\Pr(S_{T_{\text{end}}-1} = j | \text{data up to } T_{\text{end}}, \theta)$$
and since these are independent due to the conditioning of the latter on the whole data, this multiplication results in:

$$\Pr(S_{T_{\text{end}}-2} = i, S_{T_{\text{end}}-1} = j | \text{data up to } T_{\text{end}}, \theta)$$

Note that summing the rows of $A'_{T_{\text{end}}-1}$ will give the same vector resulting from summing the columns of $A'_{T_{\text{end}}}$ which will automatically sum to 1 due to the definition of the forward matrices. Summing the rows of $A'_{T_{\text{end}}-1}$ gives a vector with $j = 1, \ldots, M$ elements reflecting the probability of being in state $j$ at $T_{\text{end}} - 1$ given all the data. The next backward matrix is calculated analogously and recursively.

Formally, the algorithm may be described as follows. Each element of $a'_{T,i,j}$ of the backward matrix $A'_{T}$ is calculated as:

$$a'_{T,i,j} = \frac{a_{T,i,j}}{\alpha_T(j)} \times \alpha'_{T+1}(j)$$

$$= \Pr(S_{T-1} = i | S_T = j, \text{data up to } T, \theta) \times \Pr(S_T = j | \text{data up to } T_{\text{end}}, \theta)$$

where $\alpha'_{T+1}(j)$ is defined in Equation (7.3) and is calculated as the sum of columns of $A'_{T+1}$.

A stochastic version of backward recursion is discussed in Scott (2002). The idea is that one starts by generating a state $S_{T_{\text{end}}}$ at $T_{\text{end}}$ from $\alpha_{T_{\text{end}}}(j)$ and then generates state $S_{T_{\text{end}}-1}$ by sampling from the distribution proportional to the $(S_{T_{\text{end}}-1})$th column of $A_{T_{\text{end}}}$ and so on. One may use this at each MCMC iteration, to sample a chain trajectory which will result in samples from the posterior distribution of the state sequence.

### 7.4.5 E-M Algorithm

One of the most commonly used implementation frameworks for HMMs is the EM algorithm. This is a general method for fitting statistical models that in this context allows for unobserved latent variables so that HMMs can be estimated with the use of ‘missing’ indicator variables (MacDonald and Zucchini, 1997). One
7.4. Hidden Markov NHPP Model

can use forward-backward algorithms to obtain the conditional expectation of the missing variables given the data or in other words, to obtain the expected value of the hidden chain.

The idea of using the EM algorithm in this way for HMMs, was first proposed in Baum et al. (1970) and is often termed as the re-estimation technique. This is because one may derive formulae for each of the HMM parameters which maximise the conditional likelihood (given the chain) and use these in an iterative procedure. The authors show that the likelihood of an HMM (under some mild assumptions on the density function of the mother process) has some special properties which allow it to be locally maximised using the iterative procedure.

For the purposes of this thesis, a Bayesian MCMC framework is adopted which has recently become well established as an alternative approach for fitting HMMs. Note that both the EM algorithm and MCMC, require evaluation of the likelihood so that forward-backward algorithms need to be utilised.

7.4.6 State Holding Times

An alternative but equivalent way of describing the stochastic behaviour of a Markov chain, is by considering the distribution \( h_i(\tau) \) of the holding times \( \tau \) for each state \( i \). In other words we are asking the question: while in state \( i \), what is the distribution of time that will elapse before there is a state change? For a discrete Markov chain (Howard, 1971a) the length of time \( \tau \) that a state \( i \) occupies is implicitly distributed as:

\[
h_i(\tau) = (p_{ii})^{\tau-1}(1 - p_{ii})
\]

So when the chain enters state \( i \), it will hold that state for a random (discrete) amount of time \( \tau \) which is geometrically distributed.

For the NHPP-HMM, one needs to consider the time intervals \( \tau_1, \tau_2, \ldots, \tau_Q \) where as before, \( Q \) denotes the total number of state changes. The relevant diagram is
shown in Figure 7.4 which relates to the same imaginary instance of a NHPP-HMM as in Figure 7.3.

The fact that the holding times for the hidden chain are geometrically distributed, can be a potential limitation to the model in terms of flexibility, especially if one is interested in modelling the temporal behaviour of the hidden process. A possible remedy, is to consider a semi-Markov chain which is essentially a Markov chain with temporal structures.

This implies a more flexible model, but it also allows explicit modelling of the duration time between states (Dong and He, 2007). In a hidden semi-Markov model (HSMM), the mother process is dependent on a latent semi-Markov chain and in the remainder of this chapter, we consider the formulation and implementation of a hidden semi-Markov NHPP model.

### 7.5 Review of Hidden Semi-Markov Models

Considering the discussions so far in this chapter, it should be clear that hidden Markov models are natural extensions of mixture models since they allow stochastic temporal behaviour in the mixing probabilities. Furthermore, HMMs are particularly useful (and should theoretically outperform mixture models) when time is
involved in the data. This is why the NHPP-HMM model considered so far is appealing - it combines the temporal structure of NHPP models with the temporal structure of a Markov chain.

However, the fact the state holding times in an HMM are geometrically distributed, constitutes a potential drawback. Guedon (2003) states that such an implicit state occupancy distribution is inappropriate for several applications such as speech segment duration, the length of branching zones in plants and the length of segments of a given C+G content along DNA sequences. HSMMs are a natural extension of HMMs where holding time distributions are defined explicitly while still retaining the Markovian dependency structure. In this section, a brief literature review of HSMMs is presented, and a hidden semi-Markov NHPP model is investigated in subsequent sections.

A point made earlier in this chapter was that an HMM is in fact an HSMM where holding time distributions are geometric. However, despite the fact that the transition between HMMs and HSMMs is a smooth one, the complexity of the model increases considerably, especially in terms of calculating the likelihood. This is because an HSMM has an extra ‘latent level’ which relates to the possible holding times of each different state. Conceptually, one needs to consider all possible state sequences but also all possible holding times for each of those states. This renders HSMMs computationally intensive models and as a result, the literature on HSMM applications is considerably smaller than that relating to HMMs.

HSMMs were first introduced by Ferguson (1980) with an application in speech recognition. A few years after the idea was introduced, HSMMs became more used in research. Levinson (1986) considers an HSMM with continuous holding times following a Gamma distribution. The author considers the use of recursive algorithms for HSMMs (see Rabiner (1989)) which extend the forward-backward algorithms of HMMs and uses the re-estimation techniques discussed earlier in Section 7.4.5. Underflow is clearly flagged as a potential issue and the likelihood is accumulated at each step of the forward algorithm using scaling, as described in
Section 7.4.3. Furthermore, Guedon (1992) provides a very good review of HMMs and HSMMs in speech processing applications.

Sansom and Thomson (2001) utilise a HSMM and apply it to breakpoint rainfall data where breakpoints correspond to the times at which ambient rain rate changes. The mother model is Gaussian where the latent part of the model aims to identify physical causes for each hidden state. The authors use recursive procedures along with the EM algorithm for estimation. Holding times are given both non-parametric (empirical) distributions and mixtures of geometric distributions and the latter is shown to perform considerably better.

Guedon (2003) considers HSMMs where the mother process consists of discrete sequences. The EM algorithm is utilised and re-estimation formulae are explicitly calculated along with a forward-backward algorithm which uses scaling at each time step for stability. A usual assumption, when dealing with HSMMs, is that the last time step in the whole process coincides with a state change. This assumption simplifies the estimation procedure in terms of deriving the re-estimation formulae. For a more general HSMM, this assumption is too simplifying and Guedon (2003) assumes right censoring in terms of the hidden chain so that the holding time distribution for the last state interval is replaced by the survivor function. The author uses the idea of partial likelihood to incorporate that into the estimation procedure. Possible models for the holding times are discussed and specifically the use of Poisson, Binomial and Negative Binomial distributions are mentioned. The model is applied on data sets involving branching and flowering patterns in plants.

The paper by Bulla and Bulla (2006) generalises the work of Rydén et al. (1998) which looks at the application of HMMs on daily returns from industry portfolios. Specifically the authors apply two HSMMs with negative binomial holding times, (each with a different mother model) on the same data and compare the results with the HMM from Rydén et al. (1998). They show that the HSMMs fit better and also manage to capture most of the distributional and temporal properties (stylised facts) of the returns. In addition the HSMMs manage to capture a
specific stylised fact (long term effect) which the HMM fails to do because of the exponentially decaying nature of the geometrically distributed holding times. The model is estimated using the algorithms and methods of Guedon (2003) discussed above and the same authors have recently written an R package (Bulla et al., 2010) which fits HSMMs using those methods. In the package, one may fit HSMMs where the mother model is one of Normal, Poisson, t and Bernoulli distributions where the holding time distributions are either non-parametric (empirical), Geometric, Negative Binomial, Poisson or Logarithmic.

Another recent paper by Dong and He (2007) involves the utilisation of a HSMM on a data set involving degradation of (latent) health states in hydraulic pumps. Re-estimation formulae are derived explicitly so that the EM framework is utilised. Holding time distributions are estimated empirically from the data.

It is worth mentioning that not many authors have considered the use of MCMC methods for HSMMs. In addition, the use of a NHPP as the mother process in an HSMM has also not been given full consideration. In the remainder of this chapter, the formulation and estimation of a Bayesian hidden semi-Markov NHPP (NHPP-HSMM) is investigated.

7.6 Hidden Semi-Markov NHPP Model

7.6.1 Semi-Markov Chains

To simplify the description of the NHPP-HSMM, it makes sense to first consider the definition of a semi-Markov chain. One way of describing a discrete time semi-Markov chain is through an initial distribution for each state $\pi = (\pi(1), \ldots, \pi(M))$,
a transition matrix $P = \{p_{i,j}\}$ where $p_{i,i} = 0$ and $\sum_j p_{i,j} = 1$:

$$
\{p_{i,j}\} = \begin{pmatrix}
0 & p_{1,2} & \cdots & p_{1,M} \\
p_{2,1} & 0 & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
p_{M,1} & \cdots & \cdots & 0
\end{pmatrix}
$$

and a series of holding times distributions $[h_1(\tau; \phi_1), h_2(\tau; \phi_2), \ldots, h_M(\tau; \phi_M)]$ with associated parameters $\phi = (\phi_1, \ldots, \phi_M)$. In words, the chain starts at some state $S_1 \in \{1, 2, \ldots, M\}$ decided by $\pi$ at $T_0$. It will hold that state for a time interval $\tau_1$ say, according to the distribution $h_{S_1}(\tau_1; \phi_1)$. It will then jump to a different state, $S_2$ according to probabilities $p_{S_1}$, the $S_1$th row of $P$ (note that the diagonal of the matrix is zero so that self-transitions are not allowed). The new state will be held for a time $\tau_2$ according to $h_{S_2}(\tau_2)$ and the process goes on accordingly. The reason for not allowing self transitions ($p_{i,i}=0$) is because this would work counter-intuitively with the idea of holding times between state changes.

One could define the chain such that $P$ has an arbitrary diagonal but that would complicate things in terms of holding times since by definition, one state could be held for an amount of time and then go back to itself and hold for another amount of time, making it hard to explicitly define a model for the overall holding time.

Suppose now that this chain has been observed in the interval $(T_0, T_{end}]$ and that $Q$ state changes have occurred with holding time intervals $(\tau_1, \tau_2, \ldots, \tau_Q)$ just like in Figure 7.5. The likelihood of this semi-Markov chain assuming that a state change

![Figure 7.5: A semi-Markov chain](image-url)
7.6. Hidden Semi-Markov NHPP Model

has actually occurred at $T_{end}$ is:

$$L_{MC}(S_1, \ldots, S_Q, \tau_1, \ldots, \tau_Q; \pi, P, \phi) = \pi(S_1)h_{S_1}(\tau_1; \phi_1)\prod_{k=2}^{Q} p_{S_{k-1}, S_k} h_{S_k}(\tau_k; \phi_k)$$

(7.5)

If the data is right censored, i.e. a state change does not actually occur at $T_{end}$, then the survival function $\Pr(\tau > \tau_Q; \phi_Q)$ is used, instead of $h_{S_Q}(\tau_Q; \phi_Q)$. Also note that in the definition of a semi-Markov chain, it is assumed that once a state has been entered, it will hold there for at least one time step implying that any distribution chosen to characterise $h_S()$ must be zero-truncated.

Although the Markov property that the next state depends on the state at a previous lag is distorted in a semi-Markov chain, some of the Markovian structure is retained. In fact, the transitions of a semi-Markov chain behave like a first-order Markov chain where another integer-valued random variable controls the stay at each state (Howard, 1971b). This is called the imbedded Markov chain.

There is an even more general way of characterising a semi-Markov chain. One could allow for $M(M-1)$ holding time distributions $h_{i,j}(\tau; \phi_{i,j})$ where $i \neq j$ so that the amount of time a state is held for is affected by which state the chain was in before the transition occurred. This generalisation is straightforward and the formulations presented here can be easily adjusted to satisfy this.

A semi-Markov chain may be thought of as a Markov chain with a transition matrix that is time varying. An essential statistic of the semi-Markov process is the interval transition probability $\psi_{i,j}(T)$ (Howard, 1971b) which is the probability that the chain will be in state $j$ given that it started at state $i$ at $T = 1$. For $T = 1, 2, \ldots$ this is calculated recursively by:

$$\psi_{i,j}(T) = I_{i,j} \left[ \sum_{k=T+1}^{\infty} \sum_{m=1}^{M} p_{i,m} h_m(k; \phi_k) \right] + \sum_{m=1}^{M} p_{i,m} \sum_{k=1}^{T} h_m(k; \phi_k) \psi_{m,j}(T-k)$$

where $I_{i,j} = 1$ if $i = j$ and $I_{i,j} = 0$ otherwise, $\psi_{i,i}(1) = 1$ and $\psi_{i\neq j}(1) = 0$. In matrix form, $\Psi(T) = \{\psi_{i,j}(T)\}$ can be thought of as a ‘time dependent’ transition matrix whose rows always add to 1. This effectively implies that the way the semi-Markov
7.6. Hidden Semi-Markov NHPP Model

process is described earlier, with holding time distributions $h_i(\tau; \phi_i)$, is equivalent to a Markov chain with time dependent transition probabilities.

### 7.6.2 Likelihood Formulation

Now suppose that a semi-Markov chain $S$ is the underlying process driving a NHPP with some intensity function $\lambda(t, S)$, a process whose particular realisation is depicted in Figure 7.6. Given that the chain was observed, the likelihood of this semi-Markov modulated process is then formulated by combining the likelihood $L_{PP}(\cdot)$ of the mother process and the likelihood of the semi-Markov chain in Equation (7.5):

$$L_{SMPP}(D; \Theta) = \pi(S_1)h_{S_1}(\tau_1; \phi_1)L_{PP}(N(\tau_1); \theta|S_1)$$

$$\times \prod_{k=2}^{Q} p_{S_{k-1}, S_k} h_{S_k}(\tau_k; \phi_k)L_{PP}(N(\tau_k); \theta|S_k)$$

(7.6)

where $D = (T_0, t_1, \ldots, t_n, T_{end}; S_1, \ldots, S_Q, \tau_1, \ldots, \tau_Q)$ and $\Theta = (\theta, \pi, P, \phi)$ with $N(\tau_k)$ being the number of failures in the interval $\tau_k$. Note that for clarity and consistency, data relating to recurrence times are given in the usual format $(t_1, \ldots, t_n)$ so that $L_{PP}(N(\tau_k); \theta|S_k)$ denotes the NHPP likelihood (given the chain) of any $t_j \in \tau_k$.

Figure 7.6: Semi-Markov modulated NHPP
The likelihood in Equation (7.6) is conditional on having observed the latent chain. To formulate the NHPP-HSMM likelihood, one therefore needs to sum Equation (7.6) over all possible states $S \in (1, 2, \ldots, M)$ and all possible time intervals $\tau \in (1, 2, \ldots, T_{\text{end}})$. As before, the observed data for the NHPP-HSMM are just the times of the failure occurrences $t_j$ and the bounds of the observation period $(T_0, T_{\text{end}})$ so the likelihood for the NHPP-HSMM is given by:

$$L(T_0, t_1, \ldots, t_n, T_{\text{end}}; \Theta) = \sum_{\tau_1 + \cdots + \tau_Q = T_{\text{end}}} \sum_{s_1=1}^{M} \cdots \sum_{s_Q=1}^{M} L_{\text{SMPP}}(D; \Theta)$$  \hspace{1cm} (7.7)$$

The NHPP-HSMM formulated here, is effectively a model which can jump to and from a number of parallel and independent ongoing NHPPs with the jumps being controlled by the hidden chain. Note also that for the sake of generalisation, we do not impose an upper limit on the holding time distribution for each state so that it is theoretically possible for a particular state to occupy the whole of the observation period.

Although the likelihood in Equation (7.7) is well-defined, its evaluation can be computationally prohibitive (even more so than for the NHPP-HMM) for any reasonable length of observation period $(T_0, T_{\text{end}})$ and number of states. As with HMMs, recursive forward-backward algorithms need to be utilised to overcome that problem.

### 7.6.3 Forward Algorithm

The nature of recursive algorithms in HMMs relies heavily on the discrete nature of the chain so that in each time step, the joint distribution of chain and data may be calculated. To utilise these algorithms in a HSMM model, one would need to also ‘discretise’ the model in equal time steps of one unit. The holding time distributions in a HSMM are discrete so that the whole process may be thought of in discrete time steps instead of time intervals.

For instance, being at state $j$ at $T_2$ could either mean that the state was held for
two time steps since the start or that any other state $i \neq j$ was held for one time step and the process jumped to state $j$ at $T_1$. This is how the forward-backward variables will function but it is important to note that the mother process needs to also have the ability to be discretised. In the case of the NHPP, it was shown that discretisation is possible.

The discretised version of the NHPP-HSMM can be broadly described as a process which starts at a particular state $i$ at $T_0$, according to the initial distribution $\pi$. It holds that state for at least one time step so that the data at $T_1$ are the failures that occurred in $(T_0, T_1]$. State $i$ is occupied at $T_1$ and the chain will either hold that state until $T_2$ or enter another state $j$ according to $P$, and hold that for at least one time step. The rest follows accordingly.

The idea behind forward recursion in HSMMs bears the same concept as in the case of HMMs but differs slightly in the definition. Formally, a forward variable $v_T(j)$ is considered sequentially at each discrete time step $T = T_1, T_2, \ldots, T_{end}$ (Rabiner, 1989) where:

$$v_T(j) = Pr \text{ (data upto } T \text{ and chain exits } S_T = j | \theta)$$

i.e. the joint probability of the data up to $T$ and that the chain exits state $j$ at time step $T$. As with the forward variables for HMMs, $v_T(j)$ can be computed
7.6. Hidden Semi-Markov NHPP Model

recursively:

\[ v_{T_1}(j) = \pi(j)h_j(1; \phi_j)L_{PP}(N(T_0, T_1); \theta | S = j) \]

\[ v_{T_2}(j) = \pi(j)h_j(2; \phi_j)L_{PP}(N(T_0, T_2); \theta | S = j) + \sum_{i=1}^{M} v_{T_1}(i)p_{i,j}h_j(1; \phi_j)L_{PP}(N(T_1, T_2); \theta | S = j) \]

\[ v_{T_3}(j) = \pi(j)h_j(3; \phi_j)L_{PP}(N(T_0, T_3); \theta | S = j) + \sum_{i=1}^{M} v_{T_2}(i)p_{i,j}h_j(2; \phi_j)L_{PP}(N(T_1, T_3); \theta | S = j) + \sum_{i=1}^{M} v_{T_1}(i)p_{i,j}h_j(1; \phi_j)L_{PP}(N(T_2, T_3); \theta | S = j) \]

\[ v_{T_4}(j) = \pi(j)h_j(4; \phi_j)L_{PP}(N(T_0, T_4); \theta | S = j) + \cdots \]

and so on.

The computation of variables \( v_T(j) \) is not as simple as in HMMs, since at each new time step, an extra term is added which does not depend on \( v_{T-1}(j) \). This term corresponds to the probability that a state has been held since the start of the observation period. For instance, this term for \( v_{T_3}(j) \) is given by Equation (7.8). It is this extra term which prohibits scaling at each step as in HMMs. This is because these extra terms do not contain any scaled components so that re-scaling at the last time step to obtain the likelihood is invalid.

It is therefore necessary to “re-scale before scaling” at each step in order to calculate the HSMM likelihood properly. The idea is to formulate HSMM forward recursion in a similar way as in Section 7.4.3 for the HMM using forward (and hence backward) matrices. A forward matrix was constructed at each step \( T \) whose \( (i,j) \)th elements reflect the conditional probability of occupying state \( j \) at \( T \) given that state \( i \) was occupied at \( T - 1 \), conditional on the data up to time \( T \). Summing the rows of these matrices provided the necessary terms for recursion. Each matrix was scaled so that elements add to 1 but not before re-scaling by multiplying with the scale factor at the previous time step. The re-scaling occurs after summations to avoid underflow.

For HSMMs, things are not as straightforward since not all terms need to be
re-scaled but the idea of performing as many summations as possible before re-
scaling, still remains. As before, consider forward matrices $A_{T} = \{a_{T,i,j}\}$ where each element:

$$a_{T,i,j} = \Pr(S_{T-1} = i \text{ and chain exits } S_{T} = j | \text{data up to } T, \theta)$$

reflects the probability of the exiting state $j$ at $T$ given being in state $i$ at $T - 1$ conditional on the data up to $T$. Summing the rows of this matrix gives vectors:

$$\alpha_{T}(j) = \Pr(\text{chain exits } S_{T} = j | \text{data up to } T, \theta)$$

reflecting the conditional probability of exiting state $j$ at $T$ given the data up to that time step. Clearly, the off-diagonal terms of this $A_{T}$ are ‘easier’ to work with since they indicate a state change and their calculation will be the similar as in HMMs with the exception of having to include the holding time probability of 1 time step (from $T - 1$ to $T$). The diagonal entries on the other hand are more complicated since they reflect exiting state $i$ at $T$ given state $i$ at $T - 1$.

It is sensible to divide the calculation of each forward matrix in two parts. The first part relates to the probability that the process exits state $j$ at time $T$ given that it already was in $j$ at $T - 1$. The other corresponds to the probability that it exits state $j$ at $T$ given that it was in a different state at $T - 1$. For the latter, we introduce quantity $\delta_{i,j,T}$ where $\delta_{i,i,T} = 0$ and:

$$\delta_{i,j,T} = \Pr(S_{T-1} = i \text{ and chain exits } S_{T} = j | \text{data up to } T, \theta)$$

which reflects the off diagonal $(i,j)^{th}$ entries of the forward matrix. Clearly, $\delta_{i,j,T}$ may be calculated recursively using $\alpha_{T}(j)$, much like in the HMM case.

Further, we introduce quantities $\beta_{j}$ and $\gamma_{j}$ which relate to the diagonal entries of the forward matrices:

$$\beta_{T}(j) = \Pr(S_{T-1} = j \text{ and chain exits } S_{T} = j \text{ and at least one state change has occurred before } T - 1 | \text{data up to } T, \theta)$$
7.6. Hidden Semi-Markov NHPP Model

and:

$$\gamma_T(j) = \Pr(S_{T-1} = j \text{ and chain exits } S_T = j \text{ and no state change occurred before } T - 1|\text{data upto } T, \theta)$$

The variable $\gamma_T(j)$ corresponds to the new quantity that needs to be considered at each time step which relates to the probability of holding state $j$ since the start of the observation period. The quantity $\beta_T(j)$ relates to exiting state $j$ at $T$ given that at least one state change has taken place since the $T_0$, meaning that unlike $\gamma_T(j)$, this may be calculated sequentially.

Given the independent increments property of the NHPP, it makes sense to pre-calculate vectors (with length $T_{\text{end}}$):

$$F_j = (L_{PP}(N(T_0, T_1); \theta|S = j), \ldots, L_{PP}(N(T_{\text{end}} - 1, T_{\text{end}}); \theta|S = j))$$

for $j = 1, 2, \ldots, M$, which relate to the NHPP likelihood contributions at each time step. The idea is to utilise the fact that the likelihood of the NHPP in the interval $(T_j, T_{j+2}]$ can be calculated as the product of the likelihood in $(T_j, T_{j+1}]$ and in $(T_{j+1}, T_{j+2}]$.

Precalculating vectors whose elements correspond to ratios of probabilities from holding time distributions:

$$H_j = \left(\frac{h_j(2; \phi_j)}{h_j(1; \phi_j)}, \frac{h_j(3; \phi_j)}{h_j(2; \phi_j)}, \ldots, \frac{h_j(T_{\text{end}} - T_0; \phi_j)}{h_j(T_{\text{end}} - T_0 - 1; \phi_j)}\right)$$

for $j = 1, 2, \ldots, M$

will also aid in gaining computational efficiency when calculating the diagonals of $A_T$. Note that each $H_j$ has length $(T_{\text{end}} - T_0 - 1)$. At each new time step $T$, each $\beta_{T-T_a}(j)$ with $T_a = 1, 2, \ldots, T - 1$ can be multiplied with the appropriate entries of $F_j$ and $H_j$, to accumulate information on the probability of holding state $j$ up to $T$ given that at least one state change has occurred.

Let the likelihood of the NHPP-HSMM at time $T$ be denoted by $\ell_T$. Then, the
7.6. Hidden Semi-Markov NHPP Model

recursive forward algorithm for \( T = T_1, T_2, \ldots, T_{\text{end}} \) is as follows:

\begin{align*}
T_1: \quad \gamma_{T_1}(j) &= v_{T_1}(j) = \pi(j)h_j(1; \phi_j)F_j(1) \\
\beta_{T_1}(j) &= \delta_{i,j,T_1} = 0 \\
\ell_{T_1} &= \sum_{j=1}^{M} \gamma_{T_1}(j) \\
\alpha_{T_1}(j) &= \gamma_{T_1}(j)/\ell_{T_1} \\
A_{T_1} &= \text{diag}(\alpha_{T_1}(j))
\end{align*}

\begin{align*}
T_2: \quad \gamma_{T_2}(j) &= \gamma_{T_1}(j)H_j(1)F_j(2) \\
\delta_{i,j,T_2} &= \alpha_{T_1}(i)p_{i,j}h_j(1; \phi_j)F_j(2) \\
\beta_{T_2}(j) &= \left[ \sum_{i \neq j} \delta_{i,j,T_2} \right] \times \ell_{T_1} \\
v_{T_2}(j) &= \gamma_{T_2}(j) + \beta_{T_2}(j) \\
\ell_{T_2} &= \sum_{j=1}^{M} v_{T_2}(j) \\
\alpha_{T_2}(j) &= v_{T_2}(j)/\ell_{T_2} \\
A_{T_2} &= \frac{1}{\ell_{T_2}} \begin{pmatrix}
\gamma_{T_2}(1) & \ell_{T_1}\delta_{i,j,T_2} \\
& \ddots \\
\ell_{T_1}\delta_{i,j,T_2} & \gamma_{T_2}(M)
\end{pmatrix}
\end{align*}
7.6. Hidden Semi-Markov NHPP Model

\[ T_N \geq T_3 : \]

\[ \gamma_{T_N}(j) = \gamma_{T_{N-1}}(j) H_j(T_{N-1}) F_j(T_N) \]

\[ \delta_{i,j,T_N} = \alpha_{T_{N-1}}(i) p_{i,j} h_j(1; \phi_j) F_j(T_N) \]

\[ \beta_{T_2}(j) = \beta_{T_2}(j) H_j(T_{N-2}) F_j(T_N) \]

\[ \vdots \]

\[ \beta_{T_{N-1}}(j) = \beta_{T_{N-1}}(j) H_j(T_1) F_j(T_N) \]

\[ \beta_{T_N}(j) = \left[ \sum_{i \neq j} \delta_{i,j,T_N} \right] \times \ell_{T_N-1} \]

\[ v_{T_N}(j) = \gamma_{T_N}(j) + \sum_{u=2}^{T_{N-1}} \beta_{u}(j) + \beta_{T_N}(j) \]

\[ \ell_{T_N} = \sum_{j=1}^{M} v_{T_N}(j) \]

\[ \alpha_{T_N}(j) = v_{T_N}(j) / \ell_{T_N} \]

\[ A_{T_N} = \frac{1}{\ell_{T_N}} \begin{pmatrix} \gamma_{T_N}(1) + \sum_{u=2}^{T_{N-1}} \beta_{u}(1) & \cdots & \ell_{T_N-1} \delta_{i,j,T_N} \\ \ell_{T_N-1} \delta_{i,j,T_N} & \cdots & \gamma_{T_N}(M) + \sum_{u=2}^{T_{N-1}} \beta_{u}(M) \end{pmatrix} \]

At the last time step \( T = T_{\text{end}} \), the survivor function of the holding times should be used, instead of \( h_i(\tau; \phi_j) \) itself, unless the assumption of forcing a state change at that last time step is appropriate. If survivor functions are to be used, one should also precalculate vectors that correspond to each \( H_j \), replacing \( h_j(\tau; \phi_j) \) with the survivor. The likelihood of the NHPP-HSMM in Equation (7.7) is given by \( \ell_{T_{\text{end}}} \).

The techniques described above apply more generally than just the NHPP as the mother process. As long as the assumed process enjoys the independent increments property, meaning that it may be ‘discretised’, then the discussion presented so far is still valid. Lévy processes (Applebaum, 2004), examples of which include the Poisson, the Gamma and the Wiener process, are suitable candidates fulfilling the necessary requirements.

Underflow can still present problems when applying the forward algorithm described here and the natural solution is to work on the log scale. Multiplication becomes addition, division becomes subtraction and the only problem remaining...
is how to evaluate \( \log (X + Y) \) from \( \log (X) \) and \( \log (Y) \). One possibility would be to let \( M = \max (\log (X), \log (Y)) \) so that:

\[
\log (X + Y) = \log \left( e^{\log (X) - M} + e^{\log (Y) - M} \right) + M
\]

which is a method immune to underflow since in the worst case it approximates \( \log (X + Y) \) by \( M \).

### 7.6.4 Backward Algorithm

Once the forward matrices \( A_T \) have been calculated, backward recursion may be applied in the same way as described in Section 7.4.4 for the HMMs. The idea is to construct backward matrices \( A'_T = \{a'_{T,i,j}\} \) such that:

\[
a'_{T,i,j} = \Pr(S_{T-1} = i \text{ and chain exits } S_t = j | \text{data up to } T_{\text{end}}, \theta)
\]

and these are calculated as:

\[
a'_{T,i,j} = \frac{a_{T,i,j}}{\alpha_T(j)} \times a'_{T+1}(j) = \Pr(S_{T-1} = i | \text{chain exits } S_T = j, \text{data up to } T, \theta) \times \Pr(\text{chain exits } S_T = j | \text{data up to } T_{\text{end}}, \theta)
\]

where as before, the variables \( a'_{T+1}(j) \) are calculated by summing the columns of \( A'_{T+1} \) and are formally defined as:

\[
a'(j) = \Pr(\text{chain exits } S_{T-1} = j | \text{data up to } T_{\text{end}}, \theta)
\]

The effectiveness of using backward matrices \( A'_T \), both for HMMs and HSMMs, is that these are by definition automatically scaled so that their elements sum to 1. Summing the rows of each \( A'_T \), gives a vector corresponding to the distribution of states given all the data and these may be utilised to find the most likely state trajectory for example.
7.7 MCMC Model Implementation

Upon evaluation of the likelihood from forward recursion, the Metropolis-Hastings (MH) algorithm can then be used directly. This provides a convenient fitting mechanism and in fact only the forward recursion needs to be employed. A Gibbs sampler would not be as straightforward since full conditional distributions need to be calculated. Scott (2002) presents a feasible implementation of Gibbs sampling in the context of HMMs.

7.7.1 Metropolis-Hastings

The difficulty in using MH most often lies in the choice and the tuning of the proposal distribution. For the NHPP-HSMM, the dimension of the parameter space can be large with each parameter taking a different set of values. For instance, parameters of the intensity function $\lambda(t, S)$ may have an infinite or semi-infinite support whereas parameters of $P$ and $\pi$ take values on $[0, 1]$ and in general will not be independent since for example $\pi(1) = 1 - \sum_{j=2}^{M} \pi(j)$. Clearly, a standard multivariate distribution will not suffice as the proposal, in which case a mixture of (multivariate) distributions is more appropriate. Assuming a reasonable choice for the proposal, the variance must be carefully tuned to ensure an acceptance rate which leads to quick convergence and good mixing.

Full details of MH are discussed in Appendix A, but the basic idea behind MH is that at each step $i$ of the sampling, a candidate $\theta^*$ is drawn from a pre-selected proposal distribution $q(\theta^*, \theta^{(i)})$ where $\theta^{(i)}$ denotes the current value of the parameters. Then according to the acceptance probability:

$$
\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{f(\theta^*|y)q(\theta|\theta^*)}{f(\theta|y)q(\theta^*|\theta)} \right\}
$$

the candidate is either accepted to become $\theta^{(i+1)}$ or rejected so that $\theta^{(i+1)} = \theta^{(i)}$.

One of the implementation issues with MH is the choice of proposal distributions.
A widely used technique is the so called random walk MH sampler (Chib and Greenberg, 1995) where the proposal $q_{rw}(\theta^*|\theta^{(i)})$ for $\theta^*$ is centred at $\theta^{(i)}$. Usual choices for the proposal $q(\theta^*|\theta^{(i)})$ are the multivariate Normal and multivariate-$t$ distributions, for instance $\theta^* \sim \mathcal{N}(\theta, \Sigma)$. The term random walk is used for this sampler since effectively, the candidate $\theta^*$ is sampled from the following process:

$$\theta^* = \theta + D$$

where $D$ (the increment variable) has the same multivariate distribution as $q_{rw}(\cdot|\cdot)$ but centred at zero. If, $q_{rw}(\cdot|\cdot)$ is chosen to be symmetric then the acceptance probability reduces, involving only the ratio of posteriors:

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{f(\theta^*|y)}{f(\theta|y)} \right\}$$

Another commonly used MH sampler, is the simple but effective independence sampler. The proposal distribution is defined as $q(\theta^*|\theta^{(i)}) = q_i(\theta^*)$ so that the draws for each candidate value are independent on the current value of the chain. Naturally, $q_i(\cdot)$ is multivariate but unlike the random walk sampler, it does require specifying the location as well as the spread (Chib and Greenberg, 1995). The acceptance probability for the independence sampler is given by:

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{f(\theta^*|y)q_i(\theta)}{f(\theta|y)q_i(\theta^*)} \right\}$$

The difficulty in choosing a proposal distribution lies in the fact that one is looking for a reasonable acceptance rate. In practice, neither a high rate nor a low rate are ideal since the former will lead to slower convergence whereas the latter may lead to inadequate mixing. Gelman et al. (1996) concluded that for high dimensional problems, optimal acceptance rates lie around 24% whereas other authors point towards acceptance rates in the range of 20% to 50% (Gamerman, 1997). These are only meant to be used as rough guidelines and in practice, optimal rates will depend on the application.

The rate of acceptance depends highly on the variance of the proposal which needs to be adjusted accordingly. In the case of a Normal random walk proposal for
instance, one may control the rate of acceptance through the variance parameter of the Gaussian which will be centred around the current location of the chain.

### 7.7.2 Proposal Distribution

A question that arises when trying to choose a suitable proposal distribution for the NHPP-HSMM, is what to do with the parameters $\pi$ and $P$. Recall that the components of $\pi$ need to sum to 1 and so does each row of $P$. Perhaps the easiest choice is to use an independence sampler and a proposal made up of Uniform distributions. Suppose we want a proposal for $\pi = (\pi_1, \pi_2, \ldots, \pi_M)$ such that $\sum_{i=1}^{M} \pi_i = 1$. Then consider:

$$
q(\pi_1) = U(0, 1) \\
q(\pi_2 | \pi_1) = U(0, 1 - \pi_1) \\
q(\pi_3 | \pi_1, \pi_2) = U(0, 1 - \pi_1 - \pi_2) \\
\vdots \\
q(\pi_{M-1} | \pi_1, \ldots, \pi_{M-2}) = U(0, 1 - \pi_1 - \cdots - \pi_{M-2}) \\
\pi_M = 1 - \pi_1 - \cdots - \pi_{M-1}
$$

Using Bayes’ theorem, the joint proposal for $(\pi_1, \ldots, \pi_{M-1})$ is then given by:

$$
q(\pi_1, \ldots, \pi_{M-1}) = q(\pi_1)q(\pi_2 | \pi_1) \times \cdots \times q(\pi_{M-1} | \pi_1, \ldots, \pi_{M-2}) \quad (7.9)
$$

This proposal may be used in the same manner for each row of $P$ and assuming the parameters of the NHPP may be sampled using a random walk sampler, the resulting proposal for all parameters of the latent model is a combination of the random walk sampler and the independence sampler.

An alternative approach is to use an independence sampler for $\pi = (\pi_1, \ldots, \pi_M)$ where the proposal is a Dirichlet distribution:

$$(\pi_1, \pi_2, \ldots, \pi_n) \sim \text{Dir}(\alpha_1, \ldots, \alpha_n) \quad \text{where} \quad \alpha_i = 1 \quad (7.10)$$
7.7. MCMC Model Implementation

giving a flat density with $\sum_{i=1}^{M} \pi_i = 1$.

Although the independence sampler is considerably easy to use, it is harder to ‘control’ and may take longer to converge since each new move is independent of the current one, meaning that the useful information buried in the existing location of the chain is lost. Marin and Robert (1997) propose a random walk sampler which extends the independence Dirichlet sampler in Equation (7.10). The idea is that the proposal for $q(\pi^*|\pi)$ for parameters $\pi = (p_1, \ldots, p_M)$ is given by:

$$q(\pi^*|\pi) = \text{Dir}(\alpha \pi_1, \ldots, \alpha \pi_M) \quad (7.11)$$

where $E[\pi_j^*] = \frac{\alpha \pi_j}{\sum_{j=1}^{M} \alpha \pi_j} = \pi_j$

which implies a proposal centred at the previous value of the chain. Large values of $\alpha$ (which can be considered a way to control the variance) will produce ‘moves’ that are more local, leading to higher acceptance rate. Marin and Robert (1997) suggest to either pick $\alpha$ at random from a predetermined set of values for each MCMC iteration or perform prior small runs to determine a reasonable value.

In addition to the parameters of the hidden chain, the parameters of both the holding time distributions as well as the NHPP may be sampled using a Normal random walk sampler described earlier where the proposal distribution is:

$$q(\theta^*|\theta) = N(\theta, \Sigma_\theta)$$

and $\Sigma_\theta$ is a diagonal matrix, unless of course some components of $\theta$ are correlated. Suitable transformations may be necessary to accommodate certain parameters. For instance, if the support of parameter $\theta$ is $(0, \infty)$ then $\theta^* = \exp(X^*)$ is a suitable candidate where $X^*|X \sim N(X = \log(\theta), \sigma^2)$. In the case of the Normal distribution, the proposal $q(\theta^*|\theta)$ is easy enough to derive using general transformation theory giving:

$$q(\theta^*|\theta) \sim N(\log(\theta^*), \sigma^2) \times \frac{1}{\theta^*}$$

meaning that the ratio of proposals in the acceptance probability calculation simplifies to:

$$\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} = \frac{\theta^*}{\theta}$$
Proposal distributions coming from different samplers may be combined, such as the random walk and the independence, to form a multivariate proposal \( q(\theta^* | \theta) \) where \( \theta \) denotes all model parameters. The drawback of such a convoluted proposal is loss of control in the sense of tweaking the variance components to achieve a desired acceptance rate. A possible remedy would be to use componentwise Metropolis-Hastings where at each iteration, a block of parameters is updated using random walk and another block using the independence sampler (for details see Gamerman (1997)).

### 7.7.3 Label Switching

Given that HMMs and thus HSMMs extend mixture models, the issue of label switching discussed in Section 5.2.2 still presents potential difficulties. The problem is related to identifiability and arises because the likelihood is invariant under random permutations of the hidden states. Richardson and Green (1997) discuss the issue in the context of finite mixture models. Given an observed state sequence, the likelihood in Equation (7.6) of the NHPP-HSMM will have the same value regardless of whether state 1 is labelled as state 2 and vice versa.

Given this and the fact that no specific prior information about the structure of the state labels exists, the posterior distribution is multimodal and symmetric. Each mode of the posterior corresponds to each possible way of permuting the state labels. Sampling from such a posterior is certainly not trivial and it is even harder to find a sensible way of summarising it.

The issue then, lies in some kind of ordering in the states which is not something that may be estimated from the data, due to the latent nature of the model. A common way of dealing with the problem is to constrain parameters in a way so that if label switching occurs, these constraints will be violated. A typical example would be to constrain parameters of the mother process which relate to the mean. In the NHPP-HSMM constraints could be imposed on the parameters
of the intensity function $\lambda(t|S)$ so that:

$$\lambda(t|S = 1) < \lambda(t|S = 2) < \cdots < \lambda(t|S = M)$$

Although constraints may be imposed by using appropriate proposals, it is also possible to use appropriate priors which have a density of zero in the areas where violations occur (Scott, 2002). This will disrupt the symmetry in the posterior by breaking the symmetry in the prior, thus providing a solution to label switching (Stephens, 2000). In the context of MH, using such priors will result in rejection of any proposed candidates outside the range of the constraints since the posterior will be zero. However, this effectively implies using priors that are informative. Other authors have also considered imposing constraints by reparameterising the model, see for example Robert and Titterington (1998) where this is done for models where the mother model is Normal or Poisson.

Although using constraints such as parameter ordering will help the identifiability issue, it is not a ‘perfect’ solution since some label switching may still occur. Celeux et al. (2000) make the point that it is hard to assess the influence of parameter ordering on inference and that the effects are less benign than thought since the design and performance of the MCMC sampler are directly affected. The authors also stress the fact that the true posterior distribution will have $M!$ modes, where $M$ is the number of unknown components (states), which reflect the number of ways that the state labels can be permuted. A constrained model will typically concentrate in a single mode, which is acceptable although strictly speaking it is using informative priors and will not necessarily result in the same inference if the constraints are changed. Celeux et al. (2000) propose a way to utilise MCMC to sample from the true posterior using simulated tempering and also develop ways of deriving estimates from an MCMC sample of such a posterior, using an algorithm to reorder each sample as if they all came from a single mode of the posterior.

Stephens (2000) states that use of constraints through prior distributions needs to be very carefully chosen as different constraints may results in different results. The author also states that label switching can still persist even after imposing
7.7. MCMC Model Implementation

these constraints and suggests that a more reasonable solution is to sample from the unconstrained posterior and then relabel the MCMC output in a way which will make the marginal posteriors as unimodal as possible (also see references therein). The relabelling of the output is based upon minimising the posterior expected loss by utilising suitable loss functions.

In addition to these methods of coping with label switching, any natural ordering that may be implied by knowledge on the hidden chain would certainly be helpful since the restrictions on the posterior would then be natural as opposed to artificial. Prior information on the latent part of the model may be also useful in model design where, for instance, absorbing states may be included in the transition matrix.

7.7.4 Prior distributions

The probability distributions chosen to express prior beliefs about parameters in latent type models such as the NHPP-HSMM should ideally be uninformative. Otherwise, there will be a lack of flexibility in the tuning of the parameters. The priors, however, should reflect the right support for each parameter, especially the ones in $\pi$ and $P$ which refer to probabilities that need to sum to 1. In that case, one may use priors which are similar to the proposal distribution given in (7.9) or use a flat $\text{Dir}(\alpha_i = 1)$ distribution:

$$\pi = (\pi_1, \ldots, \pi_M) \sim \text{Dir}(\alpha_1 = 1, \ldots, \alpha_M = 1)$$

The Dirichlet distribution is perhaps a better choice since the prior is more tractable, i.e. the marginal priors for each $\pi_i$ will have Beta distributions with parameters depending on $\alpha_i$. 151
7.8 Simulation Experiments

For the purposes of both testing the implementation and gaining an understanding of its computational burden, the model was applied to simulated data. Specifically, the simulations relate to synthetic pipes \( i = 1, 2, \ldots, N \) that deteriorate with time according to a NHPP with a non-stationary failure rate that is governed by a hidden semi-Markov chain. The Markov modulation will indirectly introduce correlation between successive interarrival times, a property that the NHPP alone does not have.

Specifically, suppose that each pipe \( i \) has failed \( n_i \) times at \( t_{i,1}, \ldots, t_{i,n_i} \) and that the failure rate obeys the power law such that:

\[
\lambda(t, S; x_i) = \theta_S t^{\theta_S - 1} \exp \{ \beta_0 + \beta_1 x_i \}
\]

where the shape parameter \( \theta_S \) is different for each state \( S \) of the hidden chain. The single covariate \( x_i \) has an associated parameter \( \beta_1 \), and \( \beta_0 \) is the intercept which also relates to the scale parameter of the power law. Recall that depending on values of \( \theta_S \), the rate is either increasing/decreasing non-linearly with time, or it is constant if \( \theta_S = 1 \). The effect of the covariates is to shift the rate curve up or down. Effectively, this formulation assumes that the rate \( \lambda(t, S; x_i) \) is different for each state of the hidden chain through the parameter \( \theta_S \) and that the hidden Markov process is global (has the same effect on all pipes). This scenario could represent a network of underground water pipes where the latent process is common to the whole network.

Also, suppose that the hidden chain has 3 states with initial distribution:

\[
\pi = (\pi_1, \pi_2, \pi_3),
\]

transition matrix:

\[
P = \begin{pmatrix}
0 & p_{1,2} & p_{1,3} \\
p_{2,1} & 0 & p_{2,3} \\
p_{3,1} & p_{3,2} & 0
\end{pmatrix}
\]
and that the holding time distribution for each state is Poisson with mean $\phi_S$:

$$h_S(\tau; \phi_S) \sim \text{Pois}(\phi_S) \quad \text{where} \quad S = 1, 2, 3$$

$N = 50$ pipes were simulated from the 3 state NHPP-HSMM. Table 7.1 shows the priors for each parameter as well as the input values used to simulate the data. Two chains were run for 30000 iterations with a burn-in of 10000 and by thinning of 5 this resulted in 8000 samples from the posterior of each parameter. Both tests described in Section A.1.4 were used to ensure convergence.

To cope with potential label switching, the shape parameters $\theta_S$ were restricted so that $\theta_1 < \theta_2 < \theta_3$. The model was implemented using a combination of 6 random walk samplers for each of $(\theta_S, \beta_0, \beta_1, \phi_S, P, \pi)$. For parameters $\theta_S, \beta_0, \beta_1$ and $\phi_S$ a multivariate Normal proposal distribution was used with necessary transformations for the positive parameters $\theta_S$ and $\phi_S$. For parameters of the $P$ and $\pi$, the Dirichlet proposal in Equation (7.11) was utilised with $\alpha = 150$ for $P$ and $\alpha = 75$ for $\pi$. Appropriate code was written in R (R Development Core Team, 2009) for both the forward-backward algorithm and the MCMC sampler (see Appendix B). The acceptance rate for each sampler was adjusted to be in the range (0.2, 0.3).

Table 7.1 also shows the posterior mean for each fitted parameter along with standard errors and 95% Cr.I. Comparing input values and posterior means it is clear that the model fits the data quite well. In addition, the simulations have indicated that this implementation of the model in R is somewhat computationally intensive. Specifically, running these simulations on a reasonably fast computer took approximately 48 hours for each MCMC chain.

7.9 Model Application

The NHPP-HSMM developed throughout this chapter is clearly a substantially more complex model than the ones discussed in Chapters 4-6. The attention in this chapter has been primarily on methodological development and implementation of
7.9. Model Application

Table 7.1: Priors, input values and estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Input Values</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>0.5</td>
<td>0.503 (0.0018)</td>
<td>[0.499, 0.506]</td>
</tr>
<tr>
<td>( \theta_2 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>1</td>
<td>0.993 (0.0027)</td>
<td>[0.988, 0.506]</td>
</tr>
<tr>
<td>( \theta_3 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>2</td>
<td>1.999 (0.0022)</td>
<td>[1.991, 1.999]</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>\text{N}(0, 1000)</td>
<td>-1</td>
<td>-0.988 (0.0126)</td>
<td>[-1.011, -0.964]</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>\text{N}(0, 1000)</td>
<td>0.03</td>
<td>0.030 (0.0001)</td>
<td>[0.030, 0.030]</td>
</tr>
<tr>
<td>( \pi_1 )</td>
<td>\text{Dir}(1, 1, 1)</td>
<td>0.1</td>
<td>0.055 (0.0409)</td>
<td>[0.002, 0.152]</td>
</tr>
<tr>
<td>( \pi_2 )</td>
<td>\text{Dir}(1, 1, 1)</td>
<td>0.7</td>
<td>0.719 (0.0683)</td>
<td>[0.577, 0.842]</td>
</tr>
<tr>
<td>( \pi_3 )</td>
<td>\text{Dir}(1, 1, 1)</td>
<td>0.3</td>
<td>0.227 (0.0709)</td>
<td>[0.103, 0.373]</td>
</tr>
<tr>
<td>( p_{1, 2} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.7</td>
<td>0.715 (0.0347)</td>
<td>[0.643, 0.781]</td>
</tr>
<tr>
<td>( p_{1, 3} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.3</td>
<td>0.285 (0.0347)</td>
<td>[0.219, 0.357]</td>
</tr>
<tr>
<td>( p_{2, 1} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.2</td>
<td>0.194 (0.0264)</td>
<td>[0.146, 0.248]</td>
</tr>
<tr>
<td>( p_{2, 3} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.8</td>
<td>0.806 (0.0264)</td>
<td>[0.752, 0.854]</td>
</tr>
<tr>
<td>( p_{3, 1} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.65</td>
<td>0.674 (0.0303)</td>
<td>[0.614, 0.732]</td>
</tr>
<tr>
<td>( p_{3, 1} )</td>
<td>\text{Dir}(1, 1)</td>
<td>0.35</td>
<td>0.326 (0.0303)</td>
<td>[0.268, 0.386]</td>
</tr>
<tr>
<td>( \phi_1 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>15</td>
<td>14.378 (0.2692)</td>
<td>[13.854, 14.904]</td>
</tr>
<tr>
<td>( \phi_2 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>10</td>
<td>9.812 (0.2021)</td>
<td>[9.423, 10.209]</td>
</tr>
<tr>
<td>( \phi_3 )</td>
<td>\text{Gam}(0.5, 0.005)</td>
<td>2</td>
<td>1.912 (0.0955)</td>
<td>[1.732, 2.106]</td>
</tr>
</tbody>
</table>

the NHPP-HSMM within the Bayesian MCMC framework, rather than refining the efficiency of the algorithms proposed. The computational burden of the model is further discussed in Chapter 8, but as was seen in the simulations earlier, the proposed implementation of the NHPP-HSMM is very computational even for just 50 simulated pipes.

For that reason it is not a practical feasibility to proceed as in previous chapters and apply the model to the whole North American network of 1349 pipes. Instead, we focus on just a single pipe selected from the network but we return to a discussion of application to a whole network in Chapter 8.

Failures times in the North American data set are recorded at individual months
7.9. Model Application

per year. Here, to illustrate the NHPP-HSMM it is reasonable to select a pipe that has failed a considerable number of times. For consistency with previous chapters, only 30 years of data (1969-1998) will be considered for model fitting whereas the last 5 years (1999-2003) in the observation period will be used for prediction. Table 7.2 shows burst information for the chosen pipe.

<table>
<thead>
<tr>
<th>Table 7.2: Pipe information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installation Date</td>
</tr>
<tr>
<td>Observed Until</td>
</tr>
<tr>
<td>Failure Count (1955-2003)</td>
</tr>
<tr>
<td>Failure Count (1969-1998)</td>
</tr>
<tr>
<td>Failure Count (1999-2003)</td>
</tr>
<tr>
<td>Pipe Length</td>
</tr>
</tbody>
</table>

Furthermore, the number of failures per month are plotted in Figure 7.7 for the period 1969-2003 and the vertical line denotes the start of the prediction period. From the plot it is clear that the prevailing monthly count is zero so it would be sensible to formulate a model similar to the zero-inflated NHPP from Chapter 5. Specifically, we consider a HSMM-NHPP model with two hidden states. One state corresponds to a NHPP with a power law failure rate whereas the other state relates to a zero generating process. This is effectively a zero-inflated model where the mixing component is governed by a hidden semi-Markov chain.

To put such a model in the context of the discussion so far, it may be formulated in the following way: given the hidden chain $S$, the model is NHPP with failure rate:

$$\lambda(t, S) = \begin{cases} 
0, & \text{if } S = 1, \\
\theta t^{\theta-1}, & \text{if } S = 2
\end{cases}$$

Furthermore, we assume Poisson distributions for the holding times of each state:

$$h_S(\tau; \phi_S) \sim \text{Pois}(\phi_S) \quad \text{where} \quad S = 1, 2$$
7.9. Model Application

Figure 7.7: Monthly failure counts vs month

and note that because of the way the NHPP-HSMM was defined earlier, the transition matrix is given by:

$$ P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} $$

so that the only other unknowns in the model are the parameters of the initial distribution:

$$ \pi = (\pi_1, \pi_2), $$

For the implementation of the model, a random walk MH sampler similar to the one used for the simulations earlier was utilised. Two MCMC chains were run and after adequate burn-in and thinning by 5, 10000 posterior samples were collected from each chain. Convergence was assessed using the Gelman and Rubin diagnostic and values of $\hat{R}$ are given in Table 7.3 indicating convergence. A plot of log-posterior samples is given in Figure 7.8 which also indicated convergence in the MCMC chains.

Estimates for model parameters are given in Table 7.3 along with standard errors and 95% credible intervals. All intervals are adequately tight implying significance. Estimates for $\phi_1$ and $\phi_2$ imply that the holding time distribution for the state relating to the zero-process has greater mean the for the one relating to the NHPP.
7.9. Model Application

Figure 7.8: Samples of the log-posterior

This was expected since periods of zero failures are generally longer than periods with failures (see Figure 7.7).

Table 7.3: Priors and estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior Mean (s.e.)</th>
<th>95% Cr.I.</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>0.79 (0.044)</td>
<td>[0.707, 0.879]</td>
<td>1.05</td>
</tr>
<tr>
<td>$\pi_1$</td>
<td>$\text{Dir}(1, 1)$</td>
<td>0.56 (0.282)</td>
<td>[0.04, 0.982]</td>
<td>1.03</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>$\text{Dir}(1, 1)$</td>
<td>0.44 (0.282)</td>
<td>[0.02, 0.963]</td>
<td>1.18</td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>6.63 (0.1475)</td>
<td>[2.939, 9.052]</td>
<td>1.00</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>$\text{Gam}(0.5, 0.005)$</td>
<td>4.68 (1.577)</td>
<td>[2.520, 8.666]</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The plot in Figure 7.9 shows a plot of deviance samples for simulated data (in black) and deviance samples for the actual data (in red). The estimate for the associated $p$-value is 0.51 meaning that model fit is adequate.

A plot of the observed cumulative number of failures per month is given in Figure 7.10 in order to graphically illustrate the fit of the model. The black line represents the observed cumulative counts per month whereas the blue line represents the estimated equivalent taken as the mean of the appropriate posterior predictive distribution from which 95% prediction intervals are also plotted in red. Note that the prediction period 1999-2003 is also included. Looking at the plot, it is clear the model has performed very well in fitting the data as well as in predicting.
7.9. Model Application

Figure 7.9: Deviance samples - actual (red) and simulated data (black)

Figure 7.10: Observed and estimated cumulative number of failures
7.10 Summary

This chapter considered the formulation and application of NHPP models with latent structure, for recurrent failure data. Specifically, a hidden semi-Markov NHPP model was discussed and applied using MCMC to both simulated data and to water pipe failure data. The results have shown that the model preforms quite well and is potentially more useful than the conventional NHPP model.

A particular issue with the NHPP-HSMM was computational burden. This was not unreasonable for single pipe data, however using this particular implementation to fit the model to the whole North American network was infeasible.

An important point raised in this chapter was that the formulation of the latent model may be generalised to other counting processes besides the NHPP. In addition, the model could potentially be used in other applications involving recurrent events other than pipe failures.

Work based on this chapter was presented in Economou et al. (2008) where the particular MCMC implementation of the NHPP-HSMM model is discussed within the context of water pipes. In addition, an application of the NHPP-HSMM on recurrent river floods is given in Economou et al. (2009), where dependence in time was modelled implicitly using time dependent covariates. The latter article is an example of how the latent model discussed in this chapter may be generalised to accommodate other applications.
Chapter 8

Conclusions

In this final chapter, the main findings of the thesis are summarised and then a number of remaining issues including possible directions for future work are discussed.

8.1 Thesis Summary

As mentioned in Chapter 1, the main objective of this thesis was the development of predictive statistical models for recurrent failures in water pipes with a possibly wider scope to other applications.

Chapter 2 presented a general review on recurrent event modelling within reliability theory where counting process models are identified as one of the most common frameworks. A literature review on statistical modelling of water pipe failures using counting processes was also included and the NHPP is identified as a particularly useful model able to naturally characterise ageing. In Chapter 3, the North American data set used throughout the thesis is described.

Chapter 4 then went on to present an MCMC formulation of a mixed effects NHPP model whose implementation was tested on simulated data before applying it to
the North American network. The model was a good fit to the data although there was an indication that excess zeros in failure counts per pipe may be an problem. In relation to the aims of the thesis outlined in Chapter 1, the mixed effects NHPP allows for pipe specific modelling of the failure rate and includes both natural variation (NHPP) and induced uncertainty (random effects). In addition, the random effects in the model offered a way of explicitly modelling a pipe dependency structure, although this was not implemented.

Chapter 5 investigated mixtures of mixed effect NHPP models and in particular, a zero-inflated mixed effects NHPP model was formulated and applied to the North American data, after testing its MCMC implementation using simulations. This ZINHPP model showed increased model performance in relation to the NHPP model in Chapter 4. Using zero-inflation in conjunction with NHPP models is something that has not been investigated so far in the literature. Furthermore, accounting for zero-inflation in the data addressed a specific issue relative to water pipe data sets, which was a particular objective in the thesis.

Chapter 6 went on to discuss the issue of aggregated failure data. Specifically, both the mixed effects NHPP and ZINHPP from previous chapters were modified so that they can be applied to aggregated data. Simulation studies ensured that MCMC implementation of each aggregated model was adequately tested and then applied to an aggregated version of the North American data set. Although the aggNHPP was able to model the failure counts in pipes, it was not successful in capturing the explicit relationship of failure rate and time in the same way as the NHPP model. This was due to loss of information because of aggregation. However, the aggZINHPP provided a fit almost identical to the ZINHPP model, implying that when zero-inflation was accounted for, the loss of information had significantly less impact to model fit. Using NHPP models for aggregated pipe data has not been considered in the literature so far. Dealing with aggregation addressed one of the objectives of the thesis to deal with specific problems related to water pipe failure data.
Chapter 7 considered a more general issue with recurrent event data, where unobserved processes are acting on the underlying recurrence mechanisms. Specifically a scenario was postulated where latent processes can be modelled using discrete Markov chains so that the effect on the assumed mother process is a change in state. A hidden Markov NHPP model was initially explored and then a more flexible hidden semi-Markov NHPP model was formulated. A recursive algorithm to evaluate the likelihood of the NHPP-HSMM was developed and MCMC implementation of the model was tested using simulations. The model was applied to a single pipe from the North American data where model flexibility was illustrated. Using a hidden semi-Markov NHPP model for recurrent pipe failures as well as the particular formulation of the recursive algorithm represent the original contributions from this chapter. This chapter refered to the objective of developing innovative, flexible and generic models that deal with unobserved processes acting on water pipes.

Finally, models discussed in Chapters 4-7 were effectively implemented using MCMC methods. Both R and WinBUGS were utilised and relevant code is given in Appendix B. Developing feasible and efficient model fitting mechanisms was also an objective of the thesis.

8.2 Practical Issues

Throughout the thesis, not much reference was given on the practical engineering implications of applying the models presented. Given that both R and WinBUGS are free software, the implementation of all models presented here is feasible. In addition, all of the inference is Bayesian and as was seen, posterior distributions of useful quantities expressing risk (such as failure ranks) can then be used in conjunction with economic loss functions in decision making or risk management studies.

A considerable amount of work exists in the literature that considers how risks
8.3. Future Considerations

derived from statistical models such as the ones in this thesis, can be used to
develop pipe replacement policies. Shamir and Howard (1979) for instance, used
their model (discussed in Chapter 2) in conjunction with cost data to derive optimal
pipe replacement times. A comprehensive review of pipe replacement analyses is
presented in Loganathan et al. (2002) where in addition, the authors present a
particular formulation of the failure rate which depends on replacement and repair
costs.

Models presented in Chapters 4-6 are proven efficient enough to be applicable to
a wide range of data involving pipe failures. The North American data consisted
of 1349 pipes rendering it of considerable size and although failures were modelled
at the monthly level, the models were implemented within very reasonable time
limits. Although the latent structure model in Chapter 7 was quite efficient for a
single water pipe, applying it to the whole North American network would require
a rather considerable amount of time meaning that this model is inefficient for a
whole pipe network.

8.3 Future Considerations

A particular problem that arose in Chapter 7 was the computational issue with the
NHPP-HSMM model. In the context of water pipes, the model should preferably
be applied to a whole network of pipes, something which is currently too computa-
tionally intensive. One reason for the computational needs of the model is the
nature of MCMC that requires a considerable number of posterior samples. This
multiplies the computation requirement of the forward-backward algorithm by the
number of required MCMC samples.

Another reason behind the computational burden of the NHPP-HSMM is the par-
ticular implementation in R. Utilising a computer language such as C++ may im-
prove the efficiency of such a model considerably so that application to the whole
North American network becomes feasible.
Furthermore, it was mentioned early on in Chapter 7, that the number of hidden states was assumed predetermined. However, a completely flexible hidden Markov or semi-Markov model is one which is allowed to choose the optimal number of states. This can be done by selecting a reasonable range of hidden states and using model selection techniques to choose the model with the best fit. Richardson and Green (1997), for example, use reversible jump MCMC methods to choose the optimal number of components for mixture models. An analogous method could be used for the NHPP-HSMM.

An issue relating to the NHPP models applied to the North American network in Chapters 4-6, was the lack of covariates in the data. Although the specific data set is adequately informative in terms of failures, the only additional information included is pipe length. As was mentioned in Chapter 2 information about covariates such as ground pressure and soil type are very useful in modelling the failure rate in a pipe specific model. Furthermore, use of time dependent covariates, such as climate variables, may be even more useful in trying to capture the underlying failure mechanism. The inclusion of such covariates (time dependent or otherwise) is a straightforward addition to all of the models discussed in this thesis and presents no additional methodological issues.

An additional issue is the lack of spatial information of the specific network studied in this thesis. Latitude/longitude data for a water pipe network would be considerably valuable in pipe specific models such as the ones explored here. This would allow an implicit dependency structure to be incorporated in the model. For example, one could include a spatial random effect to the mixed effects model in Chapter 4 using a suitable proximity matrix. This would enable the model to allow for fluctuations in the failure mechanism for each pipe due to influence from neighbouring pipes.

A related point to the inclusion of spatial random effects, is allowing for a dependency structure in the pipes. As discussed earlier in the thesis, one could do that through the inclusion of correlated random effects. For instance, a random inter-
cept could be included in the specification of the failure rate, which is modelled using a multivariate Normal distribution with zero mean and a variance-covariance matrix which needs to be estimated. Including such a random effect is relatively straightforward for the current implementation of the model in WinBUGS.

Future research related to this study includes:

- Allowing the mixing probability of the zero-inflated NHPP model in Chapter 5 to depend on time
- Considering a formulation of the hidden semi-Markov NHPP model in Chapter 7 where the latent Markov process is continuous
- Include an autoregressive term in the form of a covariate, which related to the cumulative number of failures
- Consider the ‘effective age’ of a pipe. Parameterise the failure rate is a way so that it “jumps” after each failure so that the pipe is ageing according to failures and not time.
Appendix A

Bayesian Framework and MCMC

Here, a discussion of Bayesian MCMC methodology is provided along with techniques and methods that were used throughout the thesis.

A.1 Outline

Very generally, consider a data set \( y = \{y_1, \ldots, y_n\} \) and assume a model for it with unknown parameter vector \( \theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \) such that the likelihood is \( f(y|\theta) \). Also assume that prior knowledge about parameters \( \theta \) can be expressed in terms of a probability distribution \( f(\theta) \), the prior distribution. The posterior distribution of the parameters given the data \( f(\theta|y) \) is then given by using Baye’s theorem (O’Hagan and Foster, 2004):

\[
    f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int_{\theta} f(y|\theta)f(\theta)d\theta} = \frac{f(y|\theta)f(\theta)}{f(y)} \quad (A.1)
\]

So basically, the prior beliefs \( f(\theta) \) about the parameters, which can be as vague or noninformative as we like, are updated by the data (the likelihood) to arrive at the posterior \( f(\theta|y) \) which expresses the uncertainty about the parameters after the information in the data has been taken into account. In this example, \( \theta \) is assumed continuous but it can also be discrete if necessary where the integral in
the denominator will become a sum. An overview of the progress in Bayesian
statistics and a good list of references is offered in Berger (2000) whereas O’Hagan
and Foster (2004) provide an interesting general survey of Bayesian inference.

Although Equation (A.1) is well defined and is in general easy to write down for
almost any type of probability, the (multi-dimensional) integral in the denomina-
tor is often hard to solve analytically if at all possible. Ideas such as conjugate
priors exist where particular models for $f(y|\theta)$ are combined with specific priors
to produce posterior distributions that belong to the same family as the prior.
For instance, a Gamma distributed prior for the mean of a Poisson distribution
results in a Gamma distributed posterior and in fact all models belonging to the
exponential family have natural priors (Gelman et al., 2004) (natural prior means
that the prior has the same functional form as the likelihood). For more complex
models it is sometimes possible to numerically solve Equation (A.1) or even better
to directly simulate samples from it and use those samples for (approximate) infer-
ence. Clearly though, direct sampling from a complicated arbitrary multivariate
distribution is not trivial, which is where MCMC methods can be used to simulate
samples asymptotically.

MCMC is a well-established technique for simulating samples from the posterior
to avoid the need to actually solve complicated integrals and has been used regu-
larly over the last few years mainly because it requires computational power. In
essence, as the name suggests, MCMC is Monte Carlo integration with the use of
Markov chains (Gilks et al., 1996). The idea is based on the fact that one can
draw samples from the posterior distribution $f(\theta|y)$, in appropriate proportions
from the whole range, which are not necessarily independent. One way of doing
that is by constructing a discrete homogeneous Markov chain in which $f(\theta|y)$ is
the stationary distribution of the chain. Under certain conditions (for details see
Gilks et al. (1996) Chapters 3 and 4), iterations from this chain will converge to
samples from the posterior, after an initial ‘burn-in’ period during which samples
are discarded. The method is quite straightforward and its major advantage is that
only the numerator of Equation (A.1) is required to utilise it. Two of the most widely used MCMC algorithms are discussed in reasonable detail below.

### A.1.1 Metropolis-Hastings Algorithm

There are many methods nowadays to construct MCMC chains but all these are based on the original work of Metropolis et al. (1953) which was later generalised by Hastings (1970) to establish MCMC as a powerful fitting mechanism in Bayesian modelling. Here we describe the Metropolis-Hastings algorithm since most of the other existing techniques are special cases to this one (Gilks et al., 1996). Consider then the posterior distribution \( f(\theta | y) \) which we need to sample from and let \( \theta^* \) denote a proposed move for the Markov chain in the next time step assuming the chain is at present in state \( \theta \). First, a transition kernel \( p(\theta, \theta^*) \) needs to be constructed in such a way ensuring that the posterior is the equilibrium distribution of the chain (Gamerman, 1997). Although it is not a required condition for convergence, if \( p(\theta, \theta^*) \) guarantees that the chain is reversible i.e.

\[
\begin{align*}
\int p(\theta, \theta^*) = f(\theta^* | y) p(\theta^*, \theta) \\
\int q(\theta^*, \theta) \alpha(\theta, \theta^*)
\end{align*}
\]

then this is sufficient to ensure that \( f(\theta | y) \) is the equilibrium distribution of the chain. The kernel here is composed of two parts, an arbitrary transition kernel \( q(\theta, \theta^*) \) and a probability \( \alpha(\theta, \theta^*) \) so that

\[
p(\theta, \theta^*) = \begin{cases} 
q(\theta, \theta^*)\alpha(\theta, \theta^*), & \text{if } \theta \neq \theta^* \\
1 - \int q(\theta, \theta^*)\alpha(\theta, \theta^*)d\theta^*, & \text{otherwise}
\end{cases}
\]

The acceptance probability proposed by Hastings (1970) is defined as

\[
\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{f(\theta^* | y)q(\theta^*, \theta)}{f(\theta | y)q(\theta, \theta^*)} \right\} \tag{A.2}
\]

which when combined with arbitrary \( q(\theta, \theta^*) \) it results in a reversible chain ensuring that the chain converges to the posterior. Notice how the posterior appears only as a ratio in Equation (A.2) so that the unknown constant of proportionality cancels.
A.1. Outline

The quantity \( q(\theta, \theta^*) \) is termed the proposal distribution and the algorithm works by sampling values from the proposal and then either accept or reject these values according to the acceptance probability. In practice, it is considerably easier to work with conditional proposal distributions \( q(\theta^*, \theta) \), instead of joint distributions \( q(\theta, \theta^*) \). In fact, much of the literature defines the acceptance probability as

\[
\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{f(\theta^* | y) q(\theta | \theta^*)}{f(\theta | y) q(\theta^* | \theta)} \right\}
\] (A.3)

for example see Gilks et al. (1996); Hitchcock (2003); Banerjee et al. (2004).

For clarity, the algorithm can be neatly described in the following steps:

1. Set \( i = 0 \) and choose arbitrary initial values \( \theta^{(0)} \) for \( \theta \)

2. Sample a candidate \( \theta^* \) for the next state of the chain from a pre-selected proposal distribution \( q(\theta^*, \theta^{(i)}) \)

3. Compute the acceptance probability \( \alpha(\theta^{(i)}, \theta^*) \) given by (A.3)

4. Sample \( u \) such that \( u \sim U(0, 1) \) and if \( u \leq \alpha(\theta^{(i)}, \theta^*) \) then set \( \theta^{(i+1)} = \theta^* \) 
   else set \( \theta^{(i+1)} = \theta^{(i)} \).

5. Set \( i = i + 1 \) and return to step 2 for a new candidate.

The choice of the proposal distribution in not restricted and it can have any form. More details of the algorithm are discussed in Chapter 7.

A.1.2 Gibbs Sampling

A particular variant of Metropolis-Hastings is Gibbs sampling. In the Metropolis-Hastings algorithm described earlier all components of the parameter vector \( \theta \) are updated at the same iteration in the chain. This is certainly not a restriction as \( \theta \) may be divided into sub-vectors or even single components, \( \theta_m = \{\theta_1, \theta_2, \ldots, \theta_m\} \), so that in each iteration there are at most \( m \) steps where in each step a value
A.1. Outline

is drawn from a proposal distribution \( q(\theta^{(i)}, \theta^*_m) \) which is a joint distribution of the current value \( \theta^{(i)} \) and the proposed block of components \( \theta^*_m \). Note that the vector \( \theta^*_{-m} \) containing all the elements of \( \theta^* \) not in \( \theta^*_m \) (i.e. the complement of \( \theta^*_m \)) is equal to \( \theta_{-m}^{(i)} \), the current value of the complement in the chain. Also, an acceptance probability needs to be calculated in each step:

\[
\alpha_m(\theta^{(i)}, \theta^*) = \min\left\{ 1, \frac{f(\theta^*|y)q(\theta^*, \theta^{(i)}_m)}{f(\theta^{(i)}|y)q(\theta^{(i)}, \theta^*_m)} \right\}
\]

(A.4)

where as before, \( \theta^{(i)} \) denotes the current value of \( \theta \) in iteration \( i \).

This multi-component Metropolis-Hastings is actually what was proposed by Metropolis et al. (1953) and is the idea behind Gibbs sampling. The difference though, is that in Gibbs sampling, the proposal distribution \( q(\theta^{(i)}, \theta_m) \) is the posterior conditional distribution of \( \theta_m \) given the current values of the other parameters \( \theta_{-m}^{(i)} \). Hence for any arbitrary value of \( \theta_m \) the proposal is:

\[
q(\theta^{(i)}, \theta_m) = f(\theta_m | \theta_{-m} = \theta_{-m}^{(i)}) = \frac{f(\theta_{-m}^{(i)} | \theta_m)}{f(\theta_{-m}^{(i)})}
\]

(A.5)

where \( f(\theta_{-m}^{(i)}) \) is the value of the marginal posterior density of \( \theta_{-m} \) at the current iteration \( i \). For convenience and clarity, the conditioning on the data \( y \) has been dropped from the posterior \( f(\theta|y) \). It should be clear from Equation (A.5) that \( \theta_{-m}^{(i)} = \theta_{-m}^* \) but also that \( q(\theta^*, \theta_m) = q(\theta^{(i)}, \theta_m) \) for any \( \theta_m \). The following equality then holds:

\[
\frac{f(\theta^{(i)} | y)}{q(\theta^{(i)} | \theta_m^*)} = \frac{f(\theta_{-m}^{(i)} | \theta_m)}{q(\theta_{-m}^{(i)} | \theta_m^*)} = \frac{f(\theta_{-m}^{(i)} | \theta_m)}{f(\theta_{-m}^{(i)})}
\]

so that when the right hand side and the left hand side of this equality are substituted in Equation A.4) the acceptance probability is always equal to 1 (O’Hagan and Foster, 2004). So in Gibbs sampling, proposed values are always accepted which is intuitively sensible since the proposal is the (conditional) posterior itself. In many applications of this sampler, a single component \( \theta_m \) from \( \theta \) is updated at a time so that it can be described in the following algorithm:

1. Set \( i = 0 \) and choose arbitrary initial values \( \{\theta^{(0)}_1, \theta^{(0)}_2, \ldots, \theta^{(0)}_p\} \)
A.1. Outline

2. Sample $\theta_{1}^{(i+1)}$ from $f(\theta_{1}|\theta_{2}^{(i)}, \ldots, \theta_{p}^{(i)}, y)$
   Sample $\theta_{2}^{(i+1)}$ from $f(\theta_{2}|\theta_{1}^{(i)}, \theta_{3}^{(i)}, \ldots, \theta_{p}^{(i)}, y)$
   \vdots
   Sample $\theta_{p}^{(i+1)}$ from $f(\theta_{p}|\theta_{1}^{(i)}, \ldots, \theta_{p-1}^{(i)}, y)$

3. Set $i = i + 1$ and repeat step 2

Although Gibbs sampling does not involve complicated proposals and calculations of acceptance probabilities, it does assume that one is able to sample for the full conditional posteriors $f(\theta_{m}|\theta_{m-1}^{(i)}, y)$. A particularly useful method of sampling from univariate unknown distributions is the so called acceptance rejection Metropolis sampling (ARMS) which is described in Gilks et al. (1995) and is successfully implemented in the software package WinBUGS (Spiegelhalter et al., 2003) which performs Gibbs sampling for standard likelihoods such as Poisson, Normal, Multinomial and so on.

A.1.3 Prior Distributions

Since MCMC is essentially a (numerical) sampling technique, there is no real restriction on the choice of prior distributions and in fact it is possible to use priors that are improper, i.e. that do not integrate to 1. In many situations, we will not actually have any prior information about the parameters in which case we need to use distributions which are vague, usually referred to as flat priors. Examples include the uniform distribution on the parameter support, or a Gaussian distribution with zero mean and huge variance for a parameter taking values on $(-\infty, \infty)$. A flat prior implies that the posterior distribution is approximately proportional to the likelihood. This is not to say that conjugate priors or even carefully chosen priors offer no advantage since in certain cases they may reduce numerical errors.
A.1. Outline

A.1.4 Convergence

Once the MCMC ‘machine’ is set up and running, it will start generating values from a chain which will eventually converge so that these can be considered samples from the posterior distribution $f(\theta|y)$. It is very important then to get a handle on when convergence has been reached and hence to decide what a sensible burn-in should be. Assuming that the chain has converged, it is also important to achieve a good rate of mixing, i.e. that the Markov chain moves rapidly throughout the support of the target distribution. Perhaps of secondary interest, is the so called notion of thinning where inference is based on every $k^{th}$ iteration of the chain and when $k$ is large enough, the consecutive draws from the posterior are roughly independent. This can be particularly useful when the number of simulated samples is large since their number is reduced by a factor of $k$ resulting in less computation time and storage. However Gilks et al. (1996) appropriately state that other than the reduction in computational and storage costs, there is no real advantage in thinning even when the samples are highly correlated, bearing in mind that even correlated simulations contain some information.

The simplest and most commonly used tool for assessing convergence is the visual inspection of time-series plots of the output which could consist of functions of the joint posterior or usually individual parameter samples. Formal methods for determining convergence exist, all of which make use of the actual MCMC output. Cowles and Carlin (1996) and Aldouni et al. (2006) provide very good reviews and comparisons of most of these convergence tools. In general, MCMC convergence diagnostics use output from a single or multiple parallel chains and they are either based on an arbitrary function of the results or they are based on the pure output itself. Here, we describe two of the most popular methods which are also built-in the R package coda.

Suppose that $\theta^{(i)}$, $i = 1, 2, \ldots$ are MCMC samples of a parameter $\theta$ and consider a real function $\phi = g(\theta)$ which will have a trajectory $\phi^{(i)} = g(\theta^{(i)})$. This trajectory is
ultimately a time series and Geweke (1992) proposes to use ergodic averages of this series to test convergence of the chain. Suppose now that we have \(m + n\) samples of the chain so we can calculate the following averages:

\[
\bar{\phi}_a = \frac{1}{n_a} \sum_{i=m+1}^{m+n} \phi^{(i)} \quad \text{and} \quad \bar{\phi}_b = \frac{1}{n_b} \sum_{i=m+n+1-n_b}^{m+n} \phi^{(i)}
\]

where \(n_a + n_b < n\). Assuming that \(m\) is the number of samples comprising the burn-in then \(\bar{\phi}_a\) and \(\bar{\phi}_b\) should represent ergodic averages at the beginning and end of the convergence period respectively and hence should behave similarly (Gamerman, 1997). As \(n\) gets larger and while \(n_a/n\) and \(n_b/n\) both remain constant then the test for convergence \(T_G\) is given by

\[
T_G = \frac{\bar{\phi}_a - \bar{\phi}_b}{\sqrt{\text{Var}(\hat{\phi}_a) + \text{Var}(\hat{\phi}_b)}} \sim N(0, 1)
\]

Geweke (1992) utilises spectral density estimation for the variances in the denominator and also recommends that \(n_a = 0.1n\) and \(n_b = 0.5n\). Note that this test is based on a single chain and in the case of parallel chains, it would have to be applied to each one independently.

One may wish to run more than one parallel Markov chain, ideally with dispersed initial values so that convergence to the posterior may be assessed with more certainty. Although multiple chains result in extra computation time, one must bear in mind that they are independent and as such, they do not need to be ran at the same time or even on the same processor. Furthermore, running more than one chain may pinpoint ‘bad’ starting values that result in slow convergence. Having many chains also means that one can compare these to monitor convergence since after convergence each chain should have the same trajectory and specifically, Gelman and Rubin (1992) propose a test that checks whether the variability between the chains is smaller than the dispersion within the chains. As before, consider a real function \(\phi = g(\theta)\) of the parameter \(\theta\) and assume that there are \(m\) parallel chains so that the possible trajectories of \(\phi\) are given by \(\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(m)}\}\) for \(i = 1, 2, \ldots, n\). The between chain variance \(B\) and the within chain variance \(W\)
A.1. Outline

are given by:

\[ B = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{\phi}_i - \bar{\phi})^2 \quad \text{and} \quad W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (\phi_{ij} - \bar{\phi}_i)^2 \]

where \( \bar{\phi}_i \) is the mean for chain \( i \) and \( \bar{\phi} \) is the mean of those means. If convergence has been reached, the whole of the \( mn \) values are draws from the posterior and the variance of \( \phi \), \( \sigma^2_\phi \) can be estimated as a weighted average of \( B \) and \( W \): \( \hat{\sigma}^2_\phi = \frac{n-1}{n} W + \frac{1}{n} B \). If the chains have not converged though, their starting overdispersed values will continue to affect the trajectories meaning that the variance of \( \phi \) will be overestimated by \( \hat{\sigma}^2_\phi \) but also it will be underestimated by \( W \). An indicator of convergence can then be given by \( \hat{R} = \sqrt{\hat{\sigma}^2_\phi/W} \) (Gamerman, 1997) and as \( n \to \infty \), \( \hat{R} \to 1 \) by the ergodic theorem. Convergence can therefore be assessed by how close \( \hat{R} \) is to 1.

A frequently used third method for monitoring convergence, which also helps to establish the length of the chain required to achieve precise estimates as well as a reasonable thinning number is the one by Raftery and Lewis (1992) which is not going to be described here. In general though, one of the most important aspects in MCMC is to reach converge reasonably quickly to make sure that the samples collected come from the posterior and then to collect enough of those to ensure that most, if not all, of the support of the posterior is covered. Note that the words ‘quick’ (convergence) and ‘enough’ (samples) are used loosely here and are very much dependent on the application and the experience of the MCMC user.

The methods described above apply equally to single parameters or to functions of many parameters. For instance, one could use some function \( g(\theta) \) which is a global summary of all model parameters \( \theta \) and perhaps the most natural choice for \( g(\theta) \) would be the posterior distribution \( p(\theta | y) \) itself. Samples of values from the posterior can be readily obtained from Metropolis-Hastings but also from Gibbs sampling with minimal effort which makes this a useful convergence diagnostic.
A.1.5 Posterior Predictive Diagnostics

A very important issue that arises after fitting a model is that of checking the goodness of fit. This is particularly important when fitting models through MCMC, since it allows application of complex models that are based on a set of assumptions which need to be verified. The concept behind methods that will be discussed here is perhaps old and simple: simulate data from a fitted model and then compare them with the actual data. In the Bayesian way of thinking, the data can be simulated from their posterior predictive distributions. It should be clear that the point of this is not to check whether the observed data were generated by the assumed model, but to quantify any discrepancies between data and model and assess whether they are due to chance or due to ‘bad’ fit (Gelman et al., 2004).

To put things in context, assume that we have observed data $y$ and that we consider a model with parameters $\theta$. Now suppose that the model is used to predict $y$ itself (i.e. generate the fitted values) to get $\tilde{y}$. In the Bayesian sense, this means we need to work with the posterior predictive distribution of $\tilde{y}$ given the data $y$, using the posterior distribution of the parameters $\theta$:

$$p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}|\theta)p(\theta|y)d\theta$$

When using MCMC, one can produce samples from $p(\tilde{y}|y)$ by using the samples from the posterior distribution of $\theta$ to generate response values from the model using simulation.

The natural problem that arises when applying posterior predictive checking is how to identify whether the discrepancy between model and data is significant in the statistical sense. In most cases, it should be clear which aspects in the data one wishes to capture best so the discrepancy may be described by test quantities $T(y, \theta)$, which sometimes are also called discrepancy variables. These variables are the analogue of test statistics in the classical sense with the main difference being that $T(y, \theta)$ may depend on the unknown parameters $\theta$ since the test will be evaluated using samples from $p(\theta|y)$ whereas classical test statistics cannot depend
on unknown quantities, meaning that most times one needs to restrict to pivotal test statistics. Clearly, when $T(y)$ only depends on the data, then it is actually a test statistic.

To formally assess the inability of the model to fit the data according to the posterior predictive distribution, a $p$-value may be obtained from the discrepancy variable to indicate whether $T(y, \theta)$ has an extreme value. The $p$-value may be formally defined (Geman et al., 1996) as

$$p\text{-value} = \Pr(T(\tilde{y}, \theta) \geq T(y, \theta) | y)$$

which is the probability that the generated data $\tilde{y}$ are extreme in comparison to the actual data, according to the discrepancy test. Note that $T(y, \theta)$ is not actually observed since $\theta$ is unknown yet the $p$-value is well defined, although here the reference distribution that this probability is based upon is the joint distribution of the replicated data $\tilde{y}$ and parameters $\theta$ i.e. $p(\tilde{y}, \theta | y)$. Again, when using MCMC, we can obtain samples from this joint distribution by taking the sample from $p(\theta | y)$ and for each of those draws we generate a value from the posterior predictive distribution $p(\tilde{y} | y)$ (i.e. simulate from the model). The sample from the posterior of the parameters and the sample from $p(\tilde{y} | y)$ can be considered as samples from the joint distribution $p(\tilde{y}, \theta | y)$. Then, these samples can be used to obtain samples from $T(\tilde{y}, \theta)$ and $T(y, \theta)$ so that the $p$-value may be estimated by the proportion of times that $T(\tilde{y}, \theta) \geq T(y, \theta)$.

When the test quantity only depends on the data, it is very useful to view things graphically, for instance by plotting the histogram of the simulated $T(\tilde{y})$ and adding on it the observed value of $T(y)$ to check whether it lies in the tails of the histogram. Graphical representation may in fact also be used when the discrepancy variable also depends on $\theta$ where a scatterplot of $T(\tilde{y}, \theta)$ vs $T(y, \theta)$ can be produced and one expects to see the points to be symmetrical around the 45° line if there is little discrepancy.

A potentially more useful definition of the test quantity $T(y, \theta)$ would be some
kind of summary which measures model fit such as the deviance

$$D(y, \theta) = -2 \log(p(y|\theta))$$  \hspace{1cm} (A.6)

where $p(y|\theta)$ is the likelihood. Strictly speaking the deviance is defined as minus twice the difference of the log-likelihood of the fitted model and the log-likelihood of the the saturated model. So here, $D(y, \theta)$ represents how ‘different’ the model it to the saturated model within the family of the model we are working with. In contrast to the classical $\chi^2$ test, the reference distribution is not based on large sample approximations, but is given by the samples from the posterior predictive distribution and a $p$-value can be estimated as described above.

### A.1.6 Model Comparison

Another important concept in statistical modelling is that of model comparison, which is indeed related to model fit but the main idea is that of comparing the fit of different (and perhaps nested) models. The main aspect when comparing models, is to asses whether the difference in fit between models is statistically significant and is not in fact occurring by chance (Gelman et al., 2004). There are of course many well established methods for model comparison, but here we concentrate on methods which quantify the difference between the data and each of the assumed models.

An important measure of model fit which is frequently used in model comparison, is the deviance which is given in Equation (A.6). The main reason for that is because the deviance $D(y, \theta)$ is closely related to the Kullback-Leibler distance (Kullback and Leibler, 1951), which is a measure for assessing the discrepancy between two distributions and in particular when one distribution is the true distribution of the data and the other is the model distribution. In brief, this measure which is also sometimes called relative entropy, represents the expected information lost when using the model distribution to approximate the true distribution of the data (Burnham and Anderson, 2002). In fact, the expected deviance is equal to twice
the Kullback-Leibler measure up to a constant which does not depend on model parameters $\theta$.

The expected deviance is thus a sensible measure to work with and a simple way to estimate it would be by defining

$$D(\bar{\theta}, y) = D(\hat{\theta}, y)$$

which is obtained by using a point estimate of $\theta$ (i.e. the mean or median of the posterior) to calculate the deviance. A more ‘Bayesian’ way of estimating the expected deviance would be by taking the mean of the posterior distribution of the deviance $D(\theta, y)$ which can be calculated for each MCMC sample so that its expectation may be estimated by

$$\overline{D}(\theta, y) = \frac{1}{N} \sum_{i=1}^{N} D(\theta^{(i)}, y)$$ (A.7)

where $N$ is the number of samples collected. Note that $\overline{D}(\theta, y)$ is closely related to what is suggested in the posterior predictive model checking in Section A.1.5 with the difference being that the measure is not defined over predicted data. This is more like a measure of adequacy and the intention is to use these measures to compare models. Spiegelhalter et al. (2002) point out that $\overline{D}(\theta, y)$ has in fact been used to informally compare between models.

The main idea behind many useful model comparison measures is to try and find a reasonable balance between model complexity and model fit in order to have a more sensible comparison between models. General measures such as the Akaike information criterion or AIC (Akaike, 1974) and the Bayesian information criterion or BIC (Schwarz, 1978) penalise the deviance for the number of parameters in the model to achieve that balance. In the Bayesian framework, finding the effective number of parameters $\theta$ in a model is not a trivial thing because the prior distributions $p(\theta)$ introduce a dependence structure between the parameters resulting in a reduction of the effective dimensionality.

The effective number of parameters in a Bayesian model should be a value which accounts for the number of unconstrained parameters, where a parameter that
A.1. Outline

‘draws’ more information from the prior than the data contributes less than a parameter that has very little prior information or constraints (Gelman et al., 2004). The contribution of each parameter to the effective dimensionality depends strongly of course on the data $\mathbf{y}$ so that a proposed measure for the effective number of parameters in a model $p_D$ is given by

$$p_D = \bar{D}(\theta, \mathbf{y}) - D(\bar{\theta}, \mathbf{y})$$

i.e. the difference between the mean posterior deviance and the deviance at the posterior estimates of the model parameters. The main advantage of $p_D$ described here is that it can be easily calculated using MCMC samples of the posterior $p(\theta | \mathbf{y})$ by basically using the estimate given by Equation (A.7) to get $\bar{D}(\theta, \mathbf{y})$ and by calculating the deviance at the posterior mean of $\theta$ to get $D(\bar{\theta}, \mathbf{y})$.

Measures such as the AIC and the BIC may be misleading in complex or hierarchical Bayesian models due to the fact that the effective number of parameters may be less than it seems which is why Spiegelhalter et al. (2002) introduced the deviance information criterion or DIC:

$$\text{DIC} = \bar{D}(\theta, \mathbf{y}) + p_D = 2\bar{D}(\theta, \mathbf{y}) - D(\bar{\theta}, \mathbf{y})$$

In essence, the DIC follows the same concept of classical measures which is to penalise a measure of fit with the number of parameters in the model. A different way of perceiving the DIC is by considering a measure which was suggested in Section A.1.5, namely the deviance for fitted values (replicated data) $\tilde{\mathbf{y}}$, i.e. $D(\theta, \tilde{\mathbf{y}})$. The expected value of this measure or the expected predictive deviance, i.e. $E[D(\theta, \tilde{\mathbf{y}})]$, is a better measure than (A.7) because it compares predicted data $\tilde{\mathbf{y}}$ to a model which was fitted using data $\mathbf{y}$ (Gelman et al., 2004). The DIC is then an approximate estimate of $E[D(\theta, \tilde{\mathbf{y}})]$.

Other than the fact that the DIC may be viewed as the Bayesian counterpart of classical measures such as the AIC, its main advantage over such measures is that it can be easily computed using MCMC samples. Furthermore, it is a general tool which can be theoretically used to compare between any model.
Appendix B

R and WinBUGS code

Here we present relevant WinBUGS and R code for fitting the models discussed in Chapters 4-7.

B.1 WinBUGS Code

As was discussed in Chapter 4, the likelihood of the NHPP is non-standard with regards to WinBUGS so that one needs to find an alternative way to formulate such a model. Here, we consider a method based on the so called ‘zeros trick’ described in Spiegelhalter et al. (2003). Briefly, the idea is to explicitly calculate the log-likelihood $L(y; \theta)$ (data $y$, parameters $\theta$) and then take $-1/L(y; \theta)$ and $1/L(y; \theta)$ as the lower and upper limit of a uniform distribution which models an artificial response of value zero.

The likelihood of the model is the the likelihood of the uniform distribution which is given by $(1/2)L(y; \theta)$ which is the desired likelihood. Note that by definition of MCMC the constant 1/2 will not affect sampling from the posterior.

For the mixed effects NHPP model in Chapter 4 the model for the North American data is formulated as:
B.1. WinBUGS Code

```bulgarian
model{
  for(j in 1:1349){
    zeros[j] <- 0
    zeros[j] ~ dunif(r[j],q[j])
    r[j] <- -1/p[j]
    q[j] <- 1/p[j]
    log(lc[j]) <- beta0+beta1*length[j]
    ### log-likelihood ###
    log(L[j]) <- delta[j]*fail[j]*log(theta[j])+(pow(t.start[j],theta[j])\n                  - pow(t.end[j],theta[j]))*lc[j]
                 - delta[j]*(1-theta[j])*sum.log.times[j] + delta[j]*fail[j]*log(lc[j])
    p[j] <- L[j]
    N.pred[j] ~ dpois(lambda[j]) ### Predictions
    lambda[j] <- (pow(Tp.end[j],theta[j]) - pow(Tp.start[j],theta[j]))*lc[j]
  }
  for(k in 1:1349){theta[k]~dgamma(a,b)}
  beta0~dnorm(0,0.001)
  beta1~dnorm(0,0.001)
  a~dgamma(0.5,0.005)
  b~dgamma(0.5,0.005)
}

For the zero-inflated NHPP model in Chapter 5:

```
B.1. WinBUGS Code

\[
\begin{align*}
\text{b}\sim \text{dgamma}(0.5, 0.005) \\
\text{gamma1}\sim \text{dnorm}(0, 0.001) \\
\text{gamma.age}\sim \text{dnorm}(0, 0.001)
\}
\]

For the aggregated NHPP model in Chapter 6:

\[
\begin{align*}
\text{model} &\{ \\
\quad \text{for}(j\text{ in } 1:1349)\{ \\
\quad \quad \text{fail}[j] \sim \text{dpois}(\mu[j]) \\
\quad \quad \mu[j] \Leftarrow (\text{pow}(t.\text{end}[j],\text{theta}[j]) - \text{pow}(t.\text{start}[j],\text{theta}[j]))*\text{l}c[j] \\
\quad \quad \log(\text{l}c[j]) \Leftarrow \text{beta0}+\text{beta1}*\text{length}[j] \\
\quad \quad \text{N.pred}[j] \sim \text{dpois}(\lambda[j]) \\
\quad \quad \lambda[j] \Leftarrow (\text{pow}(Tp.\text{end}[j],\text{theta}[j]) - \text{pow}(Tp.\text{start}[j],\text{theta}[j]))*\text{l}c[j] \\
\quad \}\}
\quad \text{for}(k\text{ in } 1:1349)\{\text{theta}[k]\sim \text{dgamma}(a, b)\}
\quad \text{beta0}\sim \text{dnorm}(0, 0.001) \\
\quad \text{beta1}\sim \text{dnorm}(0, 0.001) \\
\quad \text{a}\sim \text{dgamma}(0.5, 0.005) \\
\quad \text{b}\sim \text{dgamma}(0.5, 0.005)
\}\}
\]

For the aggregated zero-inflated NHPP model in Chapter 6:

\[
\begin{align*}
\text{model} &\{ \\
\quad \text{for}(j\text{ in } 1:1349)\{ \\
\quad \quad \text{fail}[j] \sim \text{dpois}(\phi[j]) \\
\quad \quad \text{u}[j] \sim \text{dbern}(p[j]) \\
\quad \quad \phi[j] \Leftarrow \mu[j]*u[j] \\
\quad \quad \mu[j] \Leftarrow (\text{pow}(t.\text{end}[j],\text{theta}[j]) - \text{pow}(t.\text{start}[j],\text{theta}[j]))*\text{l}c[j] \\
\quad \quad \log(\text{l}c[j]) \Leftarrow \text{beta0}+\text{beta1}*\text{length}[j] \\
\quad \quad \log(\text{B}[j]) \Leftarrow \text{gamma0}[j]+\text{gamma1}*\text{length}[j]+\text{gamma.age}*t.\text{end}[j] \\
\quad \quad p[j] \Leftarrow \text{B}[j]/(1+\text{B}[j]) \\
\quad \}\}
\quad \text{for}(k\text{ in } 1:1349)\{\text{theta}[k]\sim \text{dgamma}(a, b)\}
\quad \text{gamma0}[k]\sim \text{dnorm}(0, 0.01) \\
\quad \text{beta0}\sim \text{dnorm}(0, 0.001) \\
\quad \text{beta1}\sim \text{dnorm}(0, 0.001) \\
\quad \text{a}\sim \text{dgamma}(0.5, 0.005) \\
\quad \text{b}\sim \text{dgamma}(0.5, 0.005) \\
\quad \text{gamma1}\sim \text{dnorm}(0, 0.001) \\
\quad \text{gamma.age}\sim \text{dnorm}(0, 0.001)
\}\}
\]

182
B.2 R Code

The following represents code for an R function named `foralg` written to implement the forward algorithm used for the simulation experiments in Chapter 7:

```r
function (p.init,P,theta,phi1,beta0,beta1,n,y,x)
  p.init <- log(p.init)
  Z <- matrix(rep(1,9),3,3) - diag(3)

  ### CALCULATE VALUES OF H ###
  log.P <- log(P)
  log.h1 <- dpois(1:(n-1),phi1[1],log=T) - log(1-dpois(0,phi1[1])) # Calculate
  log.h2 <- dpois(1:(n-1),phi1[2],log=T) - log(1-dpois(0,phi1[2])) # values of
  log.h3 <- dpois(1:(n-1),phi1[3],log=T) - log(1-dpois(0,phi1[3])) # h upto n-1
  log.H1 <- c(log.h1[1],log.h2[1],log.h3[1]) # time distribution of 1 time step

  log.PH1 <- log.P + Z%*%diag(log.H1) # transition matrix*time distribution of 1 time step

  Ratios <- cbind(log.h1[2:(n-1)]-log.h1[1:(n-2)], # Matrix with the ratios of h(t) and
                  log.h2[2:(n-1)]-log.h2[1:(n-2)], # h(t-1) in each row starting at t=2
                  log.h3[2:(n-1)]-log.h3[1:(n-2)]) # time distribution of 1 time step

  log.s1 <- ppois(0:(n-1),phi1[1],lower.tail=F,log.p=T) - log(1-dpois(0,phi1[1])) # Calculate
  log.s2 <- ppois(0:(n-1),phi1[2],lower.tail=F,log.p=T) - log(1-dpois(0,phi1[2])) # values of
  log.s3 <- ppois(0:(n-1),phi1[3],lower.tail=F,log.p=T) - log(1-dpois(0,phi1[3])) # s upto n
  log.S1 <- c(log.s1[1],log.s2[1],log.s3[1]) # Survivals of 1 time step

  log.PS1 <- log.P + Z%*%diag(log.S1) # transition matrix*survivals of 1 time step

  S.Ratios <- cbind(log.s1[2:n]-log.h1[1:(n-1)], # Matrix with the ratios of s(t) and
                   log.s2[2:n]-log.h2[1:(n-1)], # h(t-1) in each row starting at t=2
                   log.s3[2:n]-log.h3[1:(n-1)]) # time distribution of 1 time step

  ### CALCULATE VALUES OF THE NHPP (LOG-LINEAR MODEL) ###
  mean1 <- ((1:n)^theta[1]-(0:(n-1))^theta[1])*exp(beta0+beta1*x)
  mean2 <- ((1:n)^theta[2]-(0:(n-1))^theta[2])*exp(beta0+beta1*x)
  mean3 <- ((1:n)^theta[3]-(0:(n-1))^theta[3])*exp(beta0+beta1*x)
  log.F <- cbind(dpois(y,mean1,log=T),dpois(y,mean2,log=T),
                  dpois(y,mean3,log=T)) # Calculating values of the NHPP

  ### FUNCTION TO CALCULATE THE LOG OF THE SUM OF LOGS ###
  sumlog2 <- function(vec)
  { M <- max(vec)
    if(M==Inf) {return(M)}
    result <- log(sum(exp(vec - M))) + M
    return(result)
  }

  ### START THE RECURSION ###
  alpha <- list() ## to store vectors alpha
  beta <- matrix(rep(0,n*3),n,3) ## to store vectors beta
  log.lik <- 1:n ## to store the log.likelihood
  dummy <- 0

  ### t=1 ###
```

---

183
B.2. R Code

```
gamma <- p.init + log.H1 + log.F[1,]
log.lik[1] <- sumlog2(gamma)

### t=2 ###
gamma <- gamma + Ratios[1,] + log.F[2,] # Diagonals
A <- diag(alpha[1])%*%Z + log.PH1 + Z%*%diag(log.F[2,]) # Off diagonals
alpha[2] <- c(sumlog2(c(gamma[1],beta[1,1])),sumlog2(c(gamma[2],beta[2,2])),sumlog2(c(gamma[3],beta[2,3])))
log.lik[2] <- sumlog2(alpha[2])

### t=3,...,n-1 ###
for(i in 3:(n-1)){
gamma <- gamma + Ratios[(i-1),] + log.F[i,] # Part of diagonals
beta[2:(i-1),] <- t(t(beta[2:(i-1),]) + Ratios[(i-2):1,]) + log.F[i,] # Part of diagonals
A <- diag(alpha[i-1])%*%Z + log.PH1 + Z%*%diag(log.F[i,]) # Off diagonals
beta[i,] <- c(sumlog2(A[2:3,1]),sumlog2(A[c(1,3),2]),sumlog2(A[1:2,3])) + log.lik[i-1] # Scaling back
dummy <- c(sumlog2(c(gamma[1],beta[2:(i-1),1])),sumlog2(c(gamma[2],beta[2:(i-1),2])),
         sumlog2(c(gamma[3],beta[2:(i-1),3])))
alpha[i] <- c(sumlog2(c(dummy[1],beta[i,1])),sumlog2(c(dummy[2],beta[i,2])),
             sumlog2(c(dummy[3],beta[i,3])))
log.lik[i] <- sumlog2(alpha[i])
alpha[i] <- alpha[i] - log.lik[i]
}

### t=n ###
gamma <- gamma + S.Ratios[(n-1),] + log.F[n,] # Part of diagonals
beta[2:(n-1),] <- t(t(beta[2:(n-1),]) + Ratios[(n-2):1,]) + log.F[n,] # Part of diagonals
A <- diag(alpha[n-1])%*%Z + log.PS1 + Z%*%diag(log.F[n,]) # Off diagonals
dummy <- c(sumlog2(c(gamma[1],beta[2:(n-1),1])),sumlog2(c(gamma[2],beta[2:(n-1),2])),
         sumlog2(c(gamma[3],beta[2:(n-1),3])))
alpha[n] <- c(sumlog2(c(dummy[1],beta[n,1])),sumlog2(c(dummy[2],beta[n,2])),
             sumlog2(c(dummy[3],beta[n,3])))
log.lik[n] <- sumlog2(alpha[n])
alpha[n] <- alpha[n] - log.lik[n]
return(log.lik[n])
}
```

The following represents an R function written to implement the Metropolis-Hastings sampler (one chain) for the simulations in Chapter 7 and note that use of the function `for.back` given earlier:

```
function (Y,x){
  library(MCMCpack)
  N <- length(Y)
  x <- 0
  for(i in 1:N){n[i] <- length(Y[i,])}
  ### FUNCTION TO EVALUATE THE LOG-POSTERIOR ###
  return(log.lik[n])
}
```
B.2. R Code

```r
post <- function(log.lhood, theta, beta, phi1)
{
        return(-Inf) # label switching
    else
    {
        dirichPrior <- sum(log(ddirichlet(c(0.7, 0.2, 0.1), c(1, 1, 1)))) # gives the same value always
        return(log.lhood + sum(dgamma(c(phi1), 0.001, 0.001, log=T)) +
               sum(dnorm(c(beta), 0, sqrt(1000), log=T)) * dirichPrior) # Dirichlet for P is 1 so not included
    } # end else
} # end post()

### INITIAL VALUES ###
paramA <- c(c(-1, -0.5, -0.1), 0.01, c(0.3, 0.5, 0.4)) # theta(1:3), beta1(4), phi(5:7)
beta0A <- -0.01 # intercept
parA <- c(0.5, 0.5, 0.5) # P(1:3)
thetaA <- exp(paramA[1:3])
phi1A <- exp(paramA[5:7])
P1 <- rbind(c(0, parA[1], 1-parA[1]), c(parA[2], 0, 1-parA[2]), c(parA[3], 1-parA[3], 0))
p.initA <- c(0.33, 0.33, 0.34)

### SETUP THINGS FOR SAMPLING ###
n.sims <- 60000
thin <- 5
sims.matrix <- matrix(nrow=n.sims/thin, ncol=7+1+6+3)
log.likelihoodA <- 0
for(j in 1:N)
    { FB <- for.alg(p.initA, P1, thetaA, phi1A, beta0A, beta1A, n[j], Y[j], x[j])
        log.likelihoodA <- log.likelihoodA + FB } # end for
log.postA <- post(log.likelihoodA, thetaA, c(beta0A, beta1A), phi1A)
accept1 <- accept2 <- accept3 <- accept4 <- accept5 <- accept6 <- 0
k <- kk <- 1
U1 <- runif(n.sims)
U2 <- runif(n.sims)
U3 <- runif(n.sims)
U4 <- runif(n.sims)
U5 <- runif(n.sims)
U6 <- runif(n.sims)
paramB <- paramA
ss1 <- 1e-05 # theta
ss2 <- 1e-07 # beta0
ss4 <- 1e-06 # beta1
ss3 <- 1e-03 # phi
VAR <- c(rep(ss1, 3), ss4, rep(ss3, 3), ss2)
Sigma <- diag(VAR)
Q <- mvnorm(n.sims, rep(0, 8), Sigma)
Z <- 150
Z2 <- 75

### COLLECT THE SAMPLES ###
for(i in 1:(n.sims-burn.in))
{...
```

---

185
### UPDATE FOR theta ###

```
paramB <- paramA + Q[k,]
thetaB <- exp(paramB[1:3])
log.likelihoodB <- 0
for(j in 1:N){
    FB <- for.alg(p.initA,P1,thetaB,phi1A,beta0A,beta1A,n[j],Y[j],x[j])
    log.likelihoodB <- log.likelihoodB + FB # end for
log.postB <- post(log.likelihoodB,thetaB,c(beta0A,beta1A),phi1A)
numer <- log.postB + sum(paramB[c(1:3)])
denom <- log.postA + sum(paramA[c(1:3)])
acceptance <- min(1,exp(numer - denom))
if(log.postB==-Inf){acceptance<-0} ### For label switching
accept1[kk] <- 0
if(U1[k] <= acceptance){
    thetaA <- thetaB
    accept1[kk] <- 1
    log.postA <- log.postB}
```
B.2. R Code

phi1A <- phi1B
accept5[kk] <- 1
log.postA <- log.postB
### Update for beta0 ###
beta0B <- beta0A + paramB[8]
log.likelihoodB <- 0
for(j in 1:N){
  FB <- for.alg(p.initA,P1,thetaA,phi1A,beta0B,beta1A,n[j],Y[j],x[j])
  log.likelihoodB <- log.likelihoodB + FB # end for
log.postB <- post(log.likelihoodB,thetaA,c(beta0B,beta1A),phi1A)
numer <- log.postB
 denom <- log.postA
acceptance <- min(1,exp(numer - denom))
accept2[kk] <- 0
if(U2[k] <= acceptance){
  beta0A <- beta0B
  accept2[kk] <- 1
  log.postA <- log.postB
### UPDATE FOR P ###
QP1 <- rdirichlet.theo(1,Z*P1[1,2:3])
QP2 <- rdirichlet.theo(1,Z*P1[2, c(1,3)])
QP3 <- rdirichlet.theo(1,Z*P1[3,1:2])
P2 <- rbind(c(0,QP1),c(QP2[1],0,QP2[2]),c(QP3,0))
log.likelihoodB <- 0
for(j in 1:N){
  FB <- for.alg(p.initA,P2,thetaA,phi1A,beta0A,beta1A,n[j],Y[j],x[j])
  log.likelihoodB <- log.likelihoodB + FB # end for
log.postB <- post(log.likelihoodB,thetaA,c(beta0A,beta1A),phi1A)
numer <- log.postB + log(ddirichlet(P1[1,2:3],Z*P2[1,2:3])) + log(ddirichlet(P1[2,c(1,3)],Z*P2[2,c(1,3)])) + log(ddirichlet(P1[3,1:2],Z*P2[3,1:2]))
denom <- log.postA + log(ddirichlet(P2[1,2:3],Z*P1[1,2:3])) + log(ddirichlet(P2[2,c(1,3)],Z*P1[2,c(1,3)])) + log(ddirichlet(P2[3,1:2],Z*P1[3,1:2]))
acceptance <- min(1,exp(numer - denom))
accept3[kk] <- 0
if(U3[k] <= acceptance){
P1 <- P2
accept3[kk] <- 1
log.postA <- log.postB
### UPDATE FOR p.init ###
p.initB <- as.vector(rdirichlet.theo(1,Z2*p.initA))
log.likelihoodB <- 0
for(j in 1:N){
  FB <- for.alg(p.initB,P1,thetaA,phi1A,beta0A,beta1A,n[j],Y[j],x[j])
  log.likelihoodB <- log.likelihoodB + FB # end for
B.2. R Code

```r
log.postB <- post(log.likelihoodB, thetaA, c(beta0A, beta1A), phi1A)
numer <- log.postB + log(ddirichlet(p.initA, Z2*p.initB))
denom <- log.postA + log(ddirichlet(p.initB, Z2*p.initA))
acceptance <- min(1, exp(numer - denom))
accept4[kk] <- 0
if(U4[k] <= acceptance){
  p.initA <- p.initB
  accept4[kk] <- 1
  log.postA <- log.postB
  if(jk%%thin == 0){
    sims.matrix[(kk/thin),] <- c(thetaA, beta0A, beta1A, phi1A, P1[1,2:3], P1[2,c(1,3)], P1[3,1:2], p.initA)
    kk <- kk+1
    k <- k+1
  }
}
}
```
References


References


Mailhot, A., G. Pelletier, J. F. Noel, and J. P. Villeneuve (2000). Modeling the evolution of the structural state of water pipe networks with brief recorded pipe
References

break histories: Methodology and Application. *Water resources research* 36(10), 3053-3062.


References


References


