2018

BAYESIAN HIERARCHICAL MODELS FOR LINEAR NETWORKS

Al-kaabawi, Zainab A A

UNIVERSITY OF
PLYMOUTH

BAYESIAN HIERARCHICAL MODELS FOR LINEAR NETWORKS

by

ZAINAB ABDULHUSSEIN ABDULLAH AL-KAABAWI

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing, Electronics and Mathematics

July 2018

Acknowledgements

I would like to thank my first supervisor (Director of Studies) Dr Yinghui Wei for her time and her guiding me through regular meetings and extensive feedback which have helped me to improve my work. I was happy because Yinghui is my main supervisor. She guided me in right way.

I would also like to thank my second supervisor Dr Rana Moyeed and the third supervisor Dr Malgorzata Wojtys for time and feedback. Finally, I would like to thank my brothers and sisters to their moral support.

AUTHOR'S DECLARATION

At no time during the registration for the degree of *Doctor of Philosophy* has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

A programme of advanced study was undertaken, which included *taught modules and attendance at training and study workshops*.

Taught modules:

Stochastic Calculus with Financial Applications, University of Plymouth, UK, 2014/2015.

Time Series Analysis and Modelling, University of Plymouth, UK, 2014/2015.

Data Modelling, University of Plymouth, UK, 2016/2017.

Multivariate Statistics and Experimental Design, University of Plymouth, UK, 2016/2017.

Academy for PhD training in statistics:

Statistical Computing and Statistical Inference, University of Cambridge, Cambridge, UK, 2016.

Applied Stochastic Process and Statistical Modelling, University of Oxford, Oxford, UK, 2017.

Posters at conferences:

Spatial Point Process on a Network: Parametric and Non-parametric Estimation of the Intensity of a Spatial Point Processes on a Network. *RESEARCH STUDENTS CONFERENCE*. University College Dublin, Dublin, Ireland, 14-17 June 2016.

Hierarchical Bayesian Models for Road Traffic Accidents. *International Conference of the Royal Statistical Society*. University of Strathclyde, Glasgow, UK, 4-7 September 2017.

Analysis of Road Traffic Accidents. *University of Plymouth Research Festival*. University of Plymouth, Plymouth, UK, 2018.

Word count of main body of thesis: 58,814

Signed Zainab Al-kaabawi

Date 8/10/2018

BAYESIAN HIERARCHICAL MODELS FOR LINEAR NETWORKS

ZAINAB ABDULHUSSEIN ABDULLAH AL-KAABAWI

Abstract

A motorway network is handled as a linear network. The purpose of this study is to highlight dangerous motorways via estimating the intensity of accidents and study its pattern across the UK motorway network. Two mechanisms have been adopted to achieve this aim. The first, the motorway-specific intensity is estimated by modelling the point pattern of the accident data using a homogeneous Poisson process. The homogeneous Poisson process is used to model all intensities but heterogeneity across motorways is incorporated using two-level hierarchical models. The data structure is multilevel since each motorway consists of junctions that are joined by grouped segments. In the second mechanism, the segment-specific intensity is estimated by modelling the point pattern of the accident data. The homogeneous Poisson process is used to model accident data within segments but heterogeneity across segments is incorporated using three-level hierarchical models. A Bayesian method via Markov Chain Monte Carlo simulation algorithms is used in order to estimate the unknown parameters in the models and a sensitivity analysis to the prior choice is assessed. The performance of the proposed models is checked through a simulation study and an application to traffic accidents in 2016 on the UK motorway network. The performance of the three-level frequentist model was poor. The deviance information criterion (DIC) and the widely applicable information criterion (WAIC) are employed to choose between the two-level Bayesian hierarchical model and the three-level Bayesian hierarchical model, where the results showed that the best fitting model was the three-level Bayesian hierarchical model.

Contents

1 Introduction **19**

1.1 Introduction . . . 19

1.1.1 Background . . . 19

1.1.2 Contributions . . . 20

1.2 Bayesian Inference . . . 20

1.2.1 Prior Distribution . . . 21

1.2.1.1 Conjugate Prior . . . 21

1.2.1.2 Non-informative Prior . . . 21

1.2.1.3 Jeffreys Prior . . . 23

1.2.2 Monte Carlo Integration . . . 24

1.2.3 Markov Chain Monte Carlo Methods . . . 25

1.2.3.1 The Metropolis-Hastings Algorithms . . . 26

1.2.3.2 Independence Sampler . . . 27

1.2.3.3 Gibbs Sampler . . . 27

1.2.3.4 Random Walk Metropolis . . . 28

1.3 Outline of the Thesis . . . 29

2 Point Process on the Line Segment **30**

2.1 Introduction . . . 30

2.2 Definitions . . . 30

2.3 Simulation of Inhomogeneous Process on the Line Segment . . . 31

2.4 Estimation . . . 32

2.4.1 Maximum Likelihood Estimation . . . 32

2.4.2 Simulated Example . . . 34

2.4.3 Bayesian Estimation . . . 36

2.4.4 Simulated Example . . . 37

2.5 Discussion . . . 40

3 Two-Level Hierarchical Models **41**

3.1 Introduction . . . 41

3.2 One-Stage Fully Bayesian Hierarchical Method (Model 1) . . . 43

3.2.1 Model Definition . . . 43

3.2.2 Likelihood Function . . . 43

3.2.3 Prior Distribution . . . 43

3.2.4 Posterior Distribution . . . 44

3.2.5 Estimation . . . 45

3.3 Two-Stage Semi-Bayesian Hierarchical Method (Model 2) . . . 48

3.3.1 Model Definition . . . 48

3.3.2 Likelihood Function . . . 49

3.3.3 Posterior Distribution . . . 49

3.3.4 Estimation . . . 51

3.4 Two-Stage Frequentist Hierarchical Method (Model 3) . . . 52

3.5 Non-hierarchical Bayesian and Frequentist Methods (Model 4 and 5) . . . 53

3.6 Estimation Results for Motorway Data . . . 53

3.7 Simulation Study . . . 68

3.7.1 Simulation Design . . . 68

3.7.2 Simulation Results . . . 69

3.8 Discussion . . . 76

4 Three-Level Hierarchical Models **79**

4.1 Introduction . . . 79

4.2 Three-Level Hierarchical Model . . . 80

4.2.1 Model Definition . . . 80

4.2.2 Likelihood Function . . . 81

4.2.3 Prior Distribution . . . 81

4.2.4 Posterior Distribution . . . 82

4.3 Bayesian Estimation . . . 84

4.4 Frequentist Estimation . . . 86

4.5 Estimation Results for Motorway Data . . . 88

4.6 Simulation Study . . . 100

4.6.1 Simulation Design . . . 100
4.6.2 Simulation Results . . . 100
4.7 Models Comparison . . . 103
4.7.1 Models Comparison using Information Criteria . . . 103
4.7.2 Models Comparison using Simulation Study . . . 105
4.8 Discussion . . . 107

5 Conclusion **109**
5.1 Summary . . . 109
5.2 Future work . . . 111

A Derivations and Plots of Chapter 3 **112**
A.1 One-Stage Fully Bayesian Hierarchical Method (Model 1) . . . 112
A.1.1 Likelihood Function . . . 112
A.1.2 Full Conditional Posterior Distributions . . . 112
A.1.2.1 Conditional Posterior Distribution of α_i . . . 112
A.1.2.2 Conditional Posterior Distribution of α . . . 113
A.1.2.3 Conditional Posterior Distribution of τ^2 . . . 114
A.1.2.4 Conditional Posterior Distribution of τ . . . 114
A.2 Two-Stage Semi-Bayesian Hierarchical Method (Model 2) . . . 115
A.2.1 Likelihood Function . . . 115
A.2.2 Joint Posterior Distribution . . . 115
A.2.3 Full Conditional Posterior Distributions . . . 115
A.2.3.1 Conditional Posterior Distribution of α_i . . . 115
A.2.3.2 Conditional Posterior Distribution of α . . . 117
A.2.3.3 Conditional Posterior Distribution of τ^2 . . . 118
A.3 Plots . . . 119

B Derivations and Plots of Chapter 4 **133**
B.1 Likelihood Function . . . 133
B.2 Joint Posterior Distribution . . . 134
B.3 Full Conditional Posterior Distributions . . . 134
B.3.1 Conditional Posterior Distribution of α_i . . . 134
B.3.2 Conditional Posterior Distribution of τ_i^2 . . . 135
B.3.3 Conditional Posterior Distribution of α . . . 136

B.3.4 Conditional Posterior Distribution of τ^2 137
B.4 Frequentist Estimation 137
B.5 Plots 141

Bibliography **152**

List of Tables

2.1 The maximum likelihood estimates, standard errors and 95% confidence intervals of parameters α_0 and α_1. The data are simulated from a nonhomogeneous spatial point process with intensity $\exp(3 + X(s))$ in the line segment $[0, 1]$. 35

2.2 Summary of the posterior sample of α_0 and α_1. 38

3.1 Results from fully Bayesian hierarchical model (Model 1) and semi-Bayesian hierarchical model (Model 2) as well as maximum likelihood parameter estimation results from frequenitist hierarchical model (Model 3). PM: Posterior Mean. PSD: Posterior Standard Deviation. CI: Credible Interval or confidence interval. 56

3.2 Results from the Bayesian non-hierarchical model(Model 4) and Maximum likelihood parameter estimation results from frequenitist non-hierarchical model (Model 5). CI: Credible Interval or confidence interval. 56

3.3 Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim \mathrm{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 71

3.4 Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim \mathrm{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 71

3.5 Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim \mathrm{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 72

3.6 Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim \mathrm{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 72

3.7 Simulation results under prior distributions $\tau \sim \text{unif}(0, 10^2)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 73

3.8 Simulation results under prior distributions $\tau \sim \text{unif}(0, 10^2)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 73

3.9 Simulation results under prior distributions $\tau \sim \text{HN}(0, 0.02)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 74

3.10 Simulation results under prior distributions $\tau \sim \text{HN}(0, 0.02)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 74

3.11 Simulation results using Bayesian method (Model 4) under prior distribution $\alpha \sim \text{N}(0, 10^2)$ and maximum likelihood method (Model 5) with true value $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 75

3.12 Simulation results using Bayesian method (Model 4) under prior distribution $\alpha \sim \text{N}(0, 10^2)$ and maximum likelihood method (Model 5) with true value $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability. 75

4.1 Posterior summary and frequentist estimates of parameters α, τ and λ of traffic accidents for 2016 year. $\lambda = \exp(\alpha)$ is the intensity of accidents per one kilometer. The prior of α is $\text{N}(0, 100)$. SD: standard deviation and CI: credible interval or confidence interval. HN represents the half-normal distribution. 90

4.2 Z-score (Geweke Statistic) resulting from fitting three-level hierarchical Bayesian model to traffic accident data for 2016. 98

4.3 Simulation results of frequentist method and Bayesian method under four prior distributions of τ^2 and the prior distribution $\alpha \sim \text{N}(0, 10^2)$ with true value of $\alpha = -5$ and -7. Time is recorded in seconds. MSE represents mean square error and CP the coverage probability. 102

4.4 *DIC* and *WAIC* criteria. 3LBHM represents the three-level Bayesian hierarchical model and 2LBHM represents the two-level Bayesian hierarchical model. 105

4.5 Simulation results from two-level Bayesian hierarchical model under four prior distributions of τ^2 and the prior distribution $\alpha \sim \text{N}(0, 10^2)$ with true value of $\alpha = -5$ and -7. Time is recorded in seconds. MSE represents mean square error and CP the coverage probability. 106

B.1 Simulation results of frequentist method. Time is recorded in second. Note: MSE represents mean square error and CP represents the coverage probability. 151
B.2 Simulation results of frequentist method. Time is recorded in second. Note: MSE represents mean square error and CP represents the coverage probability. 151

List of Figures

2.1 (a) plot of an intensity function $\lambda(s) = \exp(3 + 2s)$ and (b) simulated points from the inhomogeneous Poisson process. 32

2.2 Plots of maximised likelihood function in equation (2.4) of parameters α_0 and α_1 for the intensity function in equation (2.1). 35

2.3 Trace plots, autocorrelation function, histgram and density plots of MCMC chain for α_0 and α_1. Dashed red and black lines, respectively, represent medians of simulated chains and true values of α_0 and α_1. 39

3.1 Results from one-stage fully Bayesian hierarchical model (Model 1), two-stage semi-Bayesian hierarchical model (Model 2) and frequentist hierarchical model (Model 3) analysis of observed accident data on the 49 motorways in the UK for year 2016. Prior distributions are $\alpha \sim \mathrm{N}\left(0, 10^2\right)$ and $\tau^2 \sim$ Inv-Gamma $(0.001, 0.001)$. Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote the corresponding credible intervals and the solid vertical line represents the estimate of the overall log accident intensity α. 58

3.2 The posterior mean and 95% credible interval for the overall log-intensity α using Model 1, Model 2, Model 3. The Bayesian method for estimating the log-intensity of non-hierarchical model (Model 4) and the maximum likelihood estimation of the log-intensity of non-hierarchical model (Model 5). 59

3.3 Residuals plots. The predicted value of the number of accidents is calculated using the two-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. 60

3.4 An intensity of traffic accidents (λ_i) per one kilometer on the UK motorway network including 49 motorways. This plot is produced using the traffic accidents data for year 2016. Prior distributions $\alpha \sim \mathrm{N}(0, 100)$ and $\tau^2 \sim$ Inv-Gamma $(0.001, 0.001)$. 61

3.5 Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 63

3.6 Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 64

3.7 The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ± 1.96. 66

3.8 The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 67

4.1 Results from three-level hierarchical Bayesian model for accident data on the UK motorways for 2016 year. Prior distributions $\alpha \sim \text{N}(-6.65, 0.09^2)$ and $\tau^2 \sim \text{Inv-Gamma}(18.36, 58.06)$. Results include the posterior mean and the corresponding 95% credible interval for the log-intensity of accidents α_i on each motorway and the overall log-intensity of accidents α in (a) and the intensity of accidents $\lambda_i = 1000 \times \exp(\alpha_i)$ per one kilometer on each motorway and the overall intensity of accidents λ per one kilometer in (b). Square shapes represent posterior means of $\alpha_i,\ i = 1, ..., m$ in (a) and posterior means of $\lambda_i,\ i = 1, ..., m$ in (b). The diamond shape is used to represent the posterior mean of the overall log -intensity of accident α in (a) and the posterior mean of the overall intensity of accident λ in (b). Horizontal lines denote 95% credible intervals and the sold vertical line represents the posterior mean of the overall log-intensity of accidents α in (a) and the posterior mean of the overall intensity λ in (b). 91

4.2 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments. 92

4.3 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS denotes grouped segment. 93

4.4 Estimated intensities of traffic accidents (λ_i), $i = 1, ..., 49$ per one kilometer on the UK motorway network including 49 motorways. This plot is produced using the traffic accident data in year 2016. The intensity functions are estimated using Bayesian methods with prior distributions $\alpha \sim \text{N}(-6.65, 0.09^2)$ and $\tau^2 \sim \text{Inv-Gamma}(18.36, 58.06)$. 95

4.5 The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated using initial values for $\alpha = 0$ and $\tau = 0.1$ with a burn-in of 50,000 samples and a thinning interval of 100 samples. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean. 97

4.6 The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim$ N$(0, 100)$ and $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ± 1.96. 99

A.1 Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim$ N$(0, 100)$ and a weakly-informative $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 119

A.2 Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and non-informative $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 120

A.3 Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and non-informative $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 121

A.4 Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 122

A.5 Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 123

A.6 Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean. 124

A.7 The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 125

A.8 The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of th Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 126

A.9 The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 127

A.10 The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 128

A.11 The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the potential scale reduction factor (PSRF) of the Gelman-Rubin diagnostic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 129

A.12 The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 130

A.13 Results from one-stage fully Bayesian hierarchical model analysis of observed accidents data on the 49 motorways in the UK for year 2016. Prior distributions are $\text{Norm}(0, 10^2)$ of α and various prior distributions of the heterogeneity that are $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$, $\tau \sim \text{HN}(0, 0.02)$ and $\tau \sim \text{Unif}(0, 10^2)$. Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote corresponding credible intervals and the solid vertical line represent the estimate of the overall log accident intensity α. 131

A.14 Results from two-stage semi-Bayesian hierarchical model analysis of observed accidents data on the 49 motorways in the UK for year 2016. Prior distributions are Norm$(0, 10^2)$ of α and various prior distributions of the heterogeneity that are $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$, $\tau \sim$ HN$(0, 0.02)$ and $\tau \sim$ Unif$(0, 10^2)$. Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote corresponding credible intervals and the solid vertical line represent the estimate of the overall log accident intensity α. 132

B.1 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments. 141

B.2 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments. 142

B.3 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments. 143

B.4 Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments. 144

B.5 The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a weakly-informative prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean. 145

B.6 The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau \sim$ unif$(0, 100)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean. 146

B.7 The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau \sim$ HN$(0, 0.02)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean. 147

B.8 The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim$ N$(0, 100)$ and $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 148

B.9 The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim$ N$(0, 100)$ and $\tau \sim$ unif$(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin statistic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 149

B.10 The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96. 150

Chapter 1

Introduction

1.1 Introduction

1.1.1 Background

Traffic crashes have considerable impacts on human, economics and the society. To improve road safety, traffic accidents research often seek to determine prediction methods of traffic accidents. Traditional crash prediction models, such as generalized linear model, are widely used in traffic safety studies. However, these models are not able to consider multilevel data structure that is extensively existed due to technique used to collect or cluster traffic data (Huang and Abdel-Aty, 2010). Ignoring hierarchical nature of data may produce unreliable estimates of model parameters and statistical inference. This issue can be overcome by using hierarchical models. Hierarchical modelling is a statistical approach that is used to properly take account of multilevel data structure (Gelman and Hill, 2007; Huang and Abdel-Aty, 2010). Currently, hierarchical modelling has been employed in many research fields such as sociology, education, political science and public health. Shankar et al. (1998) are the first who employed hierarchical modelling in a traffic crash study. They showed that the explanatory power of crash models had been improved when site-specific random effects and time indicator were incorporated into the negative binomial regression model. Jones and Jørgensen (2003) expounded and discussed possible applications of hierarchical models in road traffic accidents in Norway. Then, the use of hierarchical modelling technique to represent multilevel data structure in crash prediction has been growing. In some research, hierarchical models were used to predict crash frequency (Mitra and Washington, 2007; Chin and Quddus, 2003; MacNab, 2003; Kim et al., 2007; Li et al., 2008; Quddus, 2008; Huang et al., 2009; Haque et al., 2010) and in other research, hierarchical models were developed to identify factors affecting crash severity (Jones and Jørgensen, 2003; Lenguerrand et al., 2006; Huang et al., 2008).

Huang and Abdel-Aty (2010) proposed five-level hierarchical approach to represent a framework of multilevel data structure in traffic safety. A five-level hierarchy represents traffic entity levels which are geographical region level, traffic site level, traffic crash level, driver vehicle unit level and occupant level. The geographical region level could represent a number of regions, countries, states or cities. Traffic site level could be road segments (link) or road junctions (node). Traffic crash level is characterized by crash severity, collision type or possible crash causes. Driver vehicle unit level is related to driver behaviour and vehicle maneuver. Different involved units in this level could be various drivers and characteristics of vehicle. Occupant level represents drivers and passengers involved in vehicle crash. Spatiotemporal level includes the geographic distribution of the regions or traffic sites and a number of time periods for pre-selected for a sites within regions.

1.1.2 Contributions

In traffic safety, there are no studies take account of hierarchical nature of traffic accident data on the UK motorway network. Instead, traditional crash prediction models such as generalized linear regression model are used to analyse traffic accidents on motorways. The UK motorway network is a linear network. However, current research on estimating an intensity on linear networks are limited to the maximum likelihood method. The main aim of this thesis is to develop a methodology for analysing traffic accident data on the UK motorway network. The contributions of this work are the development of Bayesian hierarchical models for estimating the intensity of traffic accidents and determination of dangerous motorways that have the highest intensity of accidents. These models are able to capture the heterogeneity in the intensity of accidents across grouped segments within a motorway and across motorways. The proposed models are evaluated through a simulation study.

1.2 Bayesian Inference

Bayesian inference is a major approach to statistical inference. Generally, one of tasks of Bayesian inference is to estimate unknown parameters or missing data (Congdon, 2007). Bayesian inference provides tools to create knowledge from data to update beliefs about parameters and missing data (Congdon, 2007). In Bayesian approach, parameters and missing observations are considered random variables. Let $P(\theta)$ denote the prior beliefs about a parameter θ, and $P(x|\theta)$ represents the probability or likelihood of the data, conditional on the prior beliefs about θ (Congdon, 2007). Using Bayes theorem, the posterior density function is

$$\pi(\theta|x) = \frac{P(x,\theta)}{P(x)} = \frac{P(x|\theta)P(\theta)}{P(x)}, \tag{1.1}$$

where $P(x) = \int P(\theta)P(x|\theta)d\theta$ or $\sum P(\theta)P(x|\theta)$ is a normalizing constant (sometimes named the marginal likelihood), and the probability $\pi(\theta|x)$ represents the updated or posterior probability beliefs about θ given the data (Congdon, 2007). The posterior density up to a normalising constant is

$$\pi(\theta|x) \propto P(x|\theta)P(\theta). \tag{1.2}$$

The posterior distribution is a product of likelihood and the prior distribution. The posterior distribution is the updated information about the parameter θ after having observed data (Gelman et al., 2003).

1.2.1 Prior Distribution

Prior distribution gives a summary of the prior information on θ. In other words, the information that is ready-made on parameter θ prior to the notice of an independent and identical random variables $x_1, x_2, ..., x_n$ (Marin and Robert, 2014). In Bayesian statistics, the choice of the prior distribution is a main matter since inference can be affected by the selection of the prior distribution (Marin and Robert, 2014). The decision to choose the prior distribution does not depend on powerful individual beliefs or crushing prior information but it relies on practical reasons (Marin and Robert, 2014). There are many types of prior distribution as described below.

1.2.1.1 Conjugate Prior

Conjugate prior distribution means that prior and posterior distributions have the same parametric family. In this case, likelihood structure is harmonious with prior. Parameters associated with prior distribution are called hyper parameters. For example, $x_1, x_2, ..., x_n$ are independent and identically distributed sample from the exponentially distribution with unknown mean λ (parametric model) and Gamma (α_0, β_0) as prior distribution of λ, where α_0 and β_0 are hyper parameters. Using Bayes' theorem the posterior distribution is Gamma $(n + \alpha_0, \beta_0 + \sum x_i)$. As you can see the prior distribution and posterior distribution have the same parametric family, so the prior distribution is conjugate prior distribution.

1.2.1.2 Non-informative Prior

The simplicity is the main cause to choose conjugate prior as our prior. However, the fixing of hyper parameter can cause difficulties in some settings and influence on the resultant inference (Marin and Robert, 2014). Therefore, one can use non-informative prior instead of conjugate prior. Non-

informative prior can be defined as cohesive extensions of the uniform distribution (Marin and Robert, 2014). Generally, the non-informative prior distribution expresses not having prior knowledge about model parameters before observing the data (Glickman and van Dyk, 2007) . In fact, a reference measure that is supplied by non-informative prior has the least amount of the possible impact on the resulting inference (Marin and Robert, 2014). However, the non-informative prior is suitable if the integral $\int P(\theta)d\theta$ is finite (Marin and Robert, 2014). For example, suppose that the parameter space is bounded and continuous that is $\Theta = [a, b],\ -\infty < a < b < \infty$, so the uniform distribution

$$P(\theta) = \frac{1}{b-a}, a < \theta < b, \tag{1.3}$$

is non-informative distribution for θ. The case be ambiguous when the parameter space is unbound $\Theta = (-\infty, \infty)$. In this case, the prior distribution takes the form

$$P(\theta) = k, \text{ any } k > 0, \tag{1.4}$$

and this distribution seems inappropriate as prior distribution because $\int P(\theta)d\theta = \infty$ (Carlin and Louis, 1997). This prior distribution is improper. But the posterior distribution can be found by using this improper distribution, if $\int P(x|\theta)\ d\theta = c$ where c is some finite value.

$$\pi(\theta|x) = \frac{P(x|\theta)}{\int P(x|\theta)\ d\theta} = \frac{P(x|\theta)}{c}. \tag{1.5}$$

This is called proper posterior distribution (Carlin and Louis, 1997). The proper posterior will not always arise, so the use of the improper prior should be done with caution (Carlin and Louis, 1997). The non-informative prior distribution is also called objective, vague, diffuse and sometimes a reference prior distribution (Glickman and van Dyk, 2007).

Actually, the non-informative prior distribution has some problems (Glickman and van Dyk, 2007). The numerous criteria to construct the non-informative prior distribution seldom give the same unique non-informative prior distribution (Glickman and van Dyk, 2007). In addition, some used methods to construct the non-informative prior distribution always assume that the uniform is the distribution for parameter model and this lead to a salient contradiction (Carlin and Louis, 1997). Indeed, the uniform prior distribution is not invariant under re-parametrization, so it is not a good non-informative (Carlin and Louis, 1997). For example, suppose that, one from the used method to construct non-informative prior distribution is performed on a data model with parameter θ, and then reparameterization is done to the same data model, where the parameter of this model is

$\gamma = \log(\theta)$. It would be desirable that the distributions on θ and γ were representing equivalent probabilistic information. It turns out that this is a difficult criterion to satisfy. The same used method to construct non-informative prior is applied to the reparameterized model (Carlin and Louis, 1997). If the prior of the reparameterized model is not uniform, then the uniformity cannot be considered as comprehensive definition of non-informative prior (Carlin and Louis, 1997).

1.2.1.3 Jeffreys Prior

Jeffreys prior is used in case that the prior distribution is not invariant under transformation (Carlin and Louis, 1997). Jeffreys prior relates with Fisher information matrix where there is one model parameter and it has the following form

$$I(\theta) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log P(X|\theta)}{\partial \theta^2} \right], \tag{1.6}$$

and Jeffreys prior for θ has the following form

$$P(\theta) = |I(\theta)|^{\frac{1}{2}}, \tag{1.7}$$

where $|I(\theta)|$ represents the determinant of the matrix $I(\theta)$ (Marin and Robert, 2014). Under transformation the Jeffreys prior for γ is

$$P(\gamma) = |I(\theta)|^{\frac{1}{2}} \left| \frac{d\theta}{d\gamma} \right|, \tag{1.8}$$

where $\left| \frac{d\theta}{d\gamma} \right|$ is the usual Jacobian transformation to the γ parameter (Carlin and Louis, 1997). In case that there were more than one parameter; Fisher information matrix takes the following form

$$I_{ij}(\boldsymbol{\theta}) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log P(X|\theta)}{\partial \theta_i \partial \theta_j} \right], \tag{1.9}$$

Equations (1.7) introduces the form to obtain non-informative prior, but in case of the high dimesions, this approach may not be suitable. When forming the non-informative prior under transformation, two important spacial cases appear (Carlin and Louis, 1997). First case, the density of X with parameter θ has the form $P(x|\theta) = P(x - \theta)$. In this case the parameter θ is named a location parameter and the density P is called a location parameter family (Berger, 1985). To find an invariant prior for θ under the location transformation of the form $Y = X + c$, the uniform along the θ domain is the invariant non-informative prior on θ . Therefore, $P(\theta) = k,\ \theta \in R,\ k > 0$ is the non-informative prior for a location parameter (Carlin and Louis, 1997).

In the second case, the density function of X has the form $P(x|\sigma) = \frac{1}{\sigma} P(\frac{x}{\sigma})$, then $\sigma > 0$ is named a

scale parameter and P is called a scale parameter family (Carlin and Louis, 1997). To obtain invariant prior for σ under scale transformation of the form $Y = cX$, for $c > 0$, So $P(\sigma) = \frac{k}{\sigma}$, $\sigma > 0$, $k > 0$ forms the non-informative prior for a scale parameter (Carlin and Louis, 1997). Our previous priors are improper prior because $\int_0^\infty P(\theta)d\theta = \infty$ (Carlin and Louis, 1997). If the density function of X has the form $P(x|\theta, \sigma) = \frac{1}{\sigma}P\left(\frac{x-\theta}{\sigma}\right)$, then P is called location-scale family (Carlin and Louis, 1997). In this case, the non-informative prior can be constructed using the previous non-informative priors and the independence concept, therefore, $P(\theta, \sigma) = \frac{k^2}{\sigma}$, $\theta \in R$, $\sigma > 0$ is the non-informative prior for location-scale parameters (Carlin and Louis, 1997).

1.2.2 Monte Carlo Integration

It is challenge to calculate the normalising constant explicitly, so we need MC integration. Monte Carlo integration uses simulation to solve integration problems. The first appearance of the idea of a MC integration was by Comte de Buffon in 1777, where random experiment was used to empirically examine Comte Buffon's probability calculation for the famous Buffon's needle experiment (Rizzo, 2008). Indeed, real development of Monte Carlo methods was after the second world war (Liu, 2001), when it was used in different scientific disciplines. One issue which arises in statistical inference and in other many branches of mathematics is integration problems, where in some cases, the integral cannot be evaluated analytically (Robert and Casella, 1999). In the Bayesian statistics, one can use Monte Carlo integration to find summary statistics of posterior distribution such as mean of posterior. The use of Monte Carlo integration is to evaluate definite integral $\int_D g(x)P(x)dx$, where $P(x)$ is a density function of a random variable x and $g(x)$ is a function in x. The mathematical expectation of $g(x)$ is $E(g(x)) = \int_D g(x)P(x)dx$. If D is an interval (a, b), then $P(x) = \frac{1}{b-a}$ is the probability density function of a uniform distribution. The basic idea to find this integral is generation of n random variables x_i from uniform distribution unif(a, b), so an unbiased estimator of $E(g(x))$ is a sample mean. In case that $\theta = \int_0^1 g(x)dx$, and $x_1, x_2, ..., x_n$ is a random sample from unif$(0, 1)$, so by using the Strong Law of Large Numbers, the Monte Carlo estimator of $E(g(x))$ is $\hat{\theta} = \frac{\sum g(x_i)}{n}$. In the different case, when limits of integral is from a to b, then we used change variables such that they are transformed to from 0 to 1. If $y = \frac{x-a}{b-a}$ and $dy = \left(\frac{1}{b-a}\right)dx$ are used as linear transformation, then

$$\int_a^b g(x)dx = \int_0^1 g(y(b-a)+a)(b-a)dy,$$

or for any uniform distribution unif(a, b), we can write integral as

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)\frac{1}{b-a}dx.$$

Algorithm 1.1 shows steps to find the Monte Carlo estimate of definite integral $\int_a^b g(x)dx$.

Algorithm 1.1 Algorithm to evaluate definite integral by Monte Carlo integration.

1- Generate $x_1, x_2, ..., x_n$ random sample from unif(a, b).
2- Compute $\hat{\theta} = (b-a)\frac{\sum g(x_i)}{n}$.

1.2.3 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods include a general framework to analyse numerous complex problems using simulation (Gilks et al., 1996). Precisely, MCMC can be defined as Monte Carlo integration using Markov Chains (Gilks et al., 1996). MCMC has been used in different statistical areas, but it has been largely used in Bayesian inference (Geyer, 1992). The main idea of Markov Chain Monte Carlo is drawing a sample from stationary distribution $P(.)$ to form irreducible ergodic Markov chain, where the chain is performed to enough time such that the chain becomes convergent to its stationary distribution (Rizzo, 2008). Markov chain is constructed by methods such as Metropolis-Hastings and Gibbs sampler (Rizzo, 2008). In Bayesian inference, observation, unknown parameters and missing data are considered random variables (Gilks et al., 1996). Suppose that $X = (X_1, X_2, ..., X_n)$ represents the observed data and θ represents parameters and missing data. The joint probability distribution $P(X, \theta)$ is the product of the prior distribution $P(\theta)$ and the likelihood $P(X|\theta)$, that means

$$P(X, \theta) = P(X|\theta)P(\theta). \tag{1.10}$$

Now using observed data X and Bayes Theorem, the distribution of θ given observed data X (the posterior distribution of θ conditional on X) can be formed as

$$\pi(\theta|X) = \frac{P(\theta)P(X|\theta)}{\int P(\theta)P(X|\theta)d\theta}. \tag{1.11}$$

Then the conditional expectation of a function $g(\theta)$ with respect to the posterior density is

$$\begin{aligned} E[g(\theta)] &= \int g(\theta)\pi(\theta|X)d\theta \\ &= \frac{\int g(\theta)P(\theta)P(X|\theta)d\theta}{\int P(\theta)P(X|\theta)d\theta}. \end{aligned} \tag{1.12}$$

To state the problem in more general terms,

$$E\left[g(\theta)\right] = \frac{\int g(u)\pi(u)du}{\int \pi(u)du}, \tag{1.13}$$

in Bayesian inference $\pi(.)$ denotes the posterior density. The expectation (1.13) can be evaluated even if $\pi(.)$ is known only up to a constant. This simplifies the problem because in practice the normalizing constant for a posterior density $\pi\left(\theta|X\right)$ is often difficult to evaluate. Indeed, integrations in equation (1.13) are impossible to evaluate using analytical approaches and for inaccuracies in case of high dimensions. It is also difficult to evaluate them by numerical approaches, but Markov Chain Monte Carlo can be used to evaluate such integrations (Gilks et al., 1996).

Markov Chain Monte Carlo methods use Monte Carlo integration to estimate the integral in equation (1.12) or equation (1.13) such that random observation $(X_1, X_2, ...)$ is simulated from target distribution $\pi(.)$ to be a realization of an irreducible ergodic Markov chain with stationary distribution $\pi(.)$ and by the generalized strong law of large numbers, so $\overline{g(X)}_n = \frac{1}{n}\sum_0^n g(X_i)$ converges with probability one to $E[(g(X))]$ as $n \longrightarrow \infty$. The chain needs to be generated for a period of time before it reaches a stationary behaviour. The period before stationarity for simulated chain is called the burn-in period or the initial transient phase of the Markov chain. This period is discarded, since the chain effect by initial values. Knowing chain validity to be a good approximation of the target distribution is through knowing the convergence of its distribution to the target distribution, where some form of statistical analysis is carried out to assess the convergence. This procedure is called convergence diagnostics (Brooks and Roberts, 1998).

1.2.3.1 The Metropolis-Hastings Algorithms

The Metropolis-Hastings algorithm is a technique for sampling from posterior distributions (target distribution) $P(x)$ using Markov Chain Monte Carlo method (Gelman et al., 2003). The idea behind The Metropolis-Hastings approaches is to construct the Markov Chain $\{X_s; s = 0, 1, 2, ...\}$ from proposal distribution q(.|X) that can be interpreted, that if a process is at the state X_s, then a candidate state Y is simulated from the proposal density (Chib and Greenberg, 1995). In case of the acceptance of candidate state Y the process will move from state X_s to state X_{s+1} and $X_{s+1} = Y$, or that the process remains at state X_s and $X_{s+1} = X_s$ (Rizzo, 2008). The choice of proposal distribution must lead to obtain the Markov chain such that it is irreducibility, positive recurrence, and aperiodicity, and it should have stationary distribution such that its stationary distribution must converge to the target distribution (Rizzo, 2008). Note that these conditions are called regularity conditions. The proposal

distribution may rely on the previous value X_s of the chain, and should have the same support set of the target distribution (Rizzo, 2008). Algorithm 1.2 is intended from Rizzo (2008) to illustrate how to simulate Markov chain from proposal distribution q(.|X_s) using Monte Carlo integration.

Algorithm 1.2 Metropolis-Hastings algorithm.

1- Set a proposal distribution q(.|X_s).
2- Draw X_1 from a proposal distribution q.
for all s from 1 to m **do**
  (a) Draw candidate value Y from q(.|X_s).
  (b) Draw u from unif(0, 1).
  if $u \leq \frac{P(Y)q(X_s|Y)}{P(X_s)q(Y|X_s)}$ **then**
    Accept Y and deliver $X_{s+1} = Y$
  else if $u > \frac{P(Y)q(X_s|Y)}{P(X_s)q(Y|X_s)}$ **then**
    Deliver $X_{s+1} = X_s$
  end if
end for

1.2.3.2 Independence Sampler

In independence sampler, a transition of a next position of chain does not rely on a previous position, so the proposal distribution q($X_s|Y$) takes the form q(X_s) and q($Y|X_s$) the form q(Y)(Rizzo, 2008). Independence sampler is used to simulate independent samples from proposal distribution which should be very close to the posterior distribution (Tierney, 1994). The acceptance probability of candidate point Y is $\alpha(X_s, Y) = \min\left(1, \frac{P(Y)q(X_s)}{P(X_s)q(Y)}\right)$. Algorithm 1.3 is presented from (Robert and Casella, 1999) to generate Markov chain with stationary distribution, which should be very close to the posterior distribution.

Algorithm 1.3 Independence sampler.

1- Define the proposal distribution q(x).
2- Initialize X_1
for all s in 2:m **do**
  (a) Generate u from unif(0,1).
  (b) Generate Y from proposal distribution q(x).
  if $u \leq \frac{P(Y)q(X_{s-1})}{P(X_{s-1})q(Y)}$ **then**
    $X_s = Y$
  else if $u > \frac{P(Y)q(X_{s-1})}{P(X_{s-1})q(Y)}$ **then**
    $X_s = X_{s-1}$
  end if
end for

1.2.3.3 Gibbs Sampler

The Gibbs sampler is a technique to simulate a chain from the target distribution $P(x)$. Gibbs sampler is considered a special case of Metropolis-Hastings sampler, and the first use of the term Gibbs sam-

pler was by Geman and Geman (1984), where they use Gibbs distribution to restore images. Indeed, Gibbs sampler is considerably used in classical statistics, but it has been widely used in Bayesian inference (Casella and George, 1992). In Bayesian analysis, Gibbs sampler generates a chain of a random variables from joint posterior distribution by sampling indirectly from marginal posterior distributions of joint posterior distribution (Casella and George, 1992). The target distribution is known up to the normalizing constant $\int P(x)dx$, and the prior distribution is chosen to be conjugate with likelihood (Gelfand, 2000). Applying Gibbs sampler needs using the multivariate target distribution, suppose that $X = (X_1, X_2, ..., X_n)$ is a vector, and $P(X)$ is the joint posterior distribution (target distribution) of X. Gibbs sampler generates a chain $(X^{(1)}, X^{(2)}, ..., X^{(m)})$, where every element in the chain is a vector and to simulate these vectors, the conditional densities are fully defined. Suppose that $X^{(0)} = (X_2^{(0)}, ..., X_n^{(0)})$ represents a starting point, then elements of vector X are simulated as following:

$X_1^{(1)}$ from conditional density $P(X_1|X_2^{(0)}, X_3^{(0)}, ..., X_n^{(0)})$.

$X_2^{(1)}$ from conditional density $P(X_2|X_1^{(1)}, X_3^{(0)}, ..., X_n^{(0)})$.

$X_3^{(1)}$ from conditional density $P(X_3|X_1^{(1)}, X_2^{(1)}, X_4^{(0)}, ..., X_n^{(0)})$.

and so on up to

$X_n^{(1)}$ from conditional density $P(X_n|X_1^{(1)}, X_2^{(1)}, ..., X_{n-1}^{(1)})$, and in the same way, it is used the vector $X^{(1)}$ for simulating the vector $X^{(2)}$, and so on up to the vector $X^{(m)}$ (Gelfand, 2000). The distribution of chain $(X^{(1)}, X^{(2)}, ..., X^{(m)})$ is stationary and converges to the target distribution, if it satisfies the regularity conditions of Markov chain (Gelman and Rubin, 1992b). Algorithm 1.4 is presented from (Marin and Robert, 2014, p. 90)to illustrate how to simulate a Markov chain using Gibbs sampler.

Algorithm 1.4 Gibbs sampler.

1- Initialize starting point $X^{(0)} = (X_1^{(0)}, X_2^{(0)}, ..., X_n^{(0)})$.
2- For iteration s from 1 to m
3- Generate $X_1^{(s)}$ from $P(X_1|X_2^{(s-1)}, X_3^{(s-1)}, ..., X_n^{(s-1)})$.
4- Generate $X_2^{(s)}$ from $P(X_2|X_1^{(s)}, X_3^{(s-1)}, ..., X_n^{(s-1)})$.
:
5- Generate $X_n^{(s)}$ from $P(X_n|X_1^{(s)}, X_2^{(s)}, ..., X_{n-1}^{(s)})$.

1.2.3.4 Random Walk Metropolis

In random walk metropolis, the proposal distribution is symmetric, that means $q(Y|X_s) = q(X_s|Y) = q(|X_s - Y|)$ so the proposal distribution is cancelled from the acceptance rate which becomes $\alpha(Y, X_s) = \frac{P(Y)}{P(X_s)}$. In addition, simulation of the next value of a chain relies on the current value of a chain. Algorithm 1.5 is presented from (Robert and Casella, 1999) to simulate posterior distribution using random walk metropolis sampler.

Algorithm 1.5 Random walk Metropolis sampler.

1- Given X_s.
2- Generate $Y \sim q(|X_s - Y|)$.
3- $X_{s+1} = Y$ with probability $\min(1, \frac{P(Y)}{P(X_s)})$.
4- $X_{s+1} = X_s$ otherwise.

1.3 Outline of the Thesis

The structure of this thesis is organized as follows:

In Chapter 2 we present the definition of a line segment. We introduce homogeneous Poisson processes and inhomogeneous Poisson processes on a line segment. We use maximum likelihood and Bayesian methods to estimate the intensity and illustrated the methods by using simulated data.

Chapter 3 presents two-level hierarchical models. The intensity of accidents is assumed homogeneous within motorway but heterogeneous across motorways. We introduce one-stage fully Bayesian hierarchical model, and two-stage semi-Bayesian hierarchical model. Two-stage frequentist hierarchical model is also presented. Bayesian and frequentist non-hierarchical models are compared. We conduct a simulation study to assess the performance of the proposed models. An application to the traffic accident data is presented.

In Chapter 4 we present three-level hierarchical models. We consider the intensity of accidents homogeneous within grouped segments whilst heterogeneous across grouped segments. We use a Bayesian method and frequentist approach for estimating the intensity of accidents. The performance of proposed models is assessed by a simulation study and application to traffic accidents data on the UK motorway network. In addition, we employ the deviance information criterion (DIC) and the widely applicable information criteria (WAIC) to choose between the two-level Bayesian hierarchical model and the three-level Bayesian hierarchical model. We classify the motorways into different risk categories according to the estimated accident intensity.

In chapter 5, we summarize the work in this thesis and introduce some proposed ideas for future research to extend Bayesian hierarchical models.

Chapter 2

Point Process on the Line Segment

2.1 Introduction

In this chapter, the focus is on a line segment as a study area of a spatial point pattern. Definitions of the line segment and related topics are introduced. We provide a description of models that are used to fit events on the line segment in particular homogeneous and inhomogeneous Poisson point processes. In this context, an intensity function of a spatial Poisson point process is defined on the line segment and realizations of the spatial Poisson point process on the line segment are generated. Statistical methods are considered for estimating the intensity function of the spatial point process on the line segment, including maximum likelihood estimation and Bayesian approach. This chapter aims to pave of the spatial point process on the linear network that will be introduced in the next chapter.

2.2 Definitions

As stated chapter 1, a point pattern is a collection of points or observed events over the study region. In this chapter, the line segment is considered as the study region. A line segment in the plane consists of two endpoints x and y and has a mathematical form $L = \left\{rx + (1-r)y : 0 \leq r \leq 1 \text{ and } x, y \in R^2\right\}$ where $|L|$ denotes its length which is the Euclidean distance between the endpoints (Ang et al., 2012). The definition of a Poisson point process on the line segment is the same definition of a Poisson point process on the plane, but the difference is that an intensity function of the Poisson point process represents the expected number of points per unit length instead of per unit area. The number of events (points) falling in the line segment L is denoted by $N(L)$ which is a random variable and has the Poisson distribution with mean $E(N(L)) = E(\mathbf{s} \cap L)$, where $\mathbf{s} = \{s_1, s_2, ..., s_n\}$ represents a set of

realizations of points on the line segment. Note that the locations of events are also random variables and the probability density function depends on the point process model.

Suppose a homogeneous Poisson point process on line segment with constant intensity function $\lambda > 0$. The number of points $N(L)$ within the study region L follows a Poisson distribution with mean $\lambda|L|$. Given the number of observed points in the line segment n, events locations are realisations from the uniform distribution over an interval $(0, |L|)$. An inhomogeneous Poisson point process on the line segment is one-dimensional point process, where the intensity function varies over the line segment. In this point process model, the intensity function is higher in some parts of the line segment than others. In this case, the number of points $N(L)$ has Poisson distribution with the mean $\Lambda(s) = E(N(L)) = \int_0^{|L|} \lambda(u) d_1 u$, we use the $d_1 u$ to indicate that the integration is done over one dimensional line segment where $\lambda(s)$ is the intensity function of events on the line segment L, namely, $N(L) \sim Poisson(\int_0^{|L|} \lambda(u) d_1 u)$. Given $N(L) = n$, the probability of every location event is $\frac{\lambda(s)}{E(N(L))}$, $0 \leq s \leq |L|$.

2.3 Simulation of Inhomogeneous Process on the Line Segment

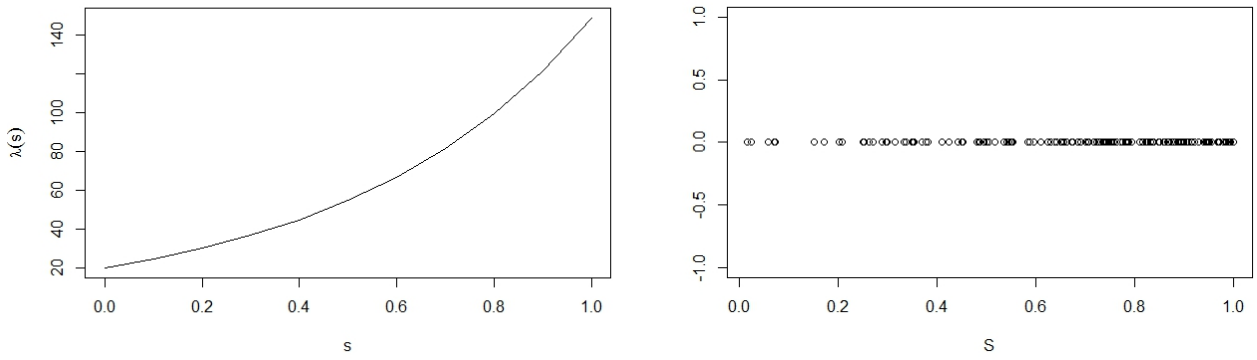
In this section, a simulation is done of an inhomogeneous Poisson point process with intensity function $\lambda(s)$ on a line segment L by rejection (Lewis and Shedler, 1978). Let s represent a distance from the beginning of the line segment. Following the method of Lewis and Schedler approach, $N^*(L)$ is generated with intensity function $\lambda^* \geq \max\{\lambda(s) : 0 \leq s \leq |L|\}$ such that the number of points $N^*(L)$ has a Poisson distribution with mean $\lambda^*|L|$. The points of the process $X_1^*, X_2^*, ..., X_{N^*}^*$ represent locations of events on the line segment L. Then, by thinning the points, the points are deleted with probability $1 - \frac{\lambda(X_i^*)}{\lambda^*}$. The remaining points with the number of points $N(L)$ represent an inhomogeneous Poisson point process with intensity function $\lambda(s)$ on line segment L and $f(s) = \lambda(s)/\lambda^*$ as the probability density function of retained points. In more detail, the inverse transform method is used for generating the points X_i^*, $i = 1, .., N^*$, on the line segment L (Rubinstein and Kroese, 2016). This method includes using the cumulative distribution function $F(s) = \int_0^s f(u) d_1 u$. The inverse of the cumulative distribution function is $F^{-1}(x) = s$, $0 \leq x \leq 1$, so that $x \sim$ unif$(0, 1)$ (Rubinstein and Kroese, 2016). Algorithm 2.1 shows the simulation of the inhomogeneous Poisson point process with intensity $\lambda(s) = \exp(\alpha_0 + \alpha_1 s)$ on the line segment L (Lewis and Shedler, 1976).

Algorithm 2.1 Simulation of an inhomogeneous Poisson process on line segment by rejection.

1. Generate $N^* \sim \text{Pois}(\lambda^*|L|)$.
2. Generate $x \sim \text{unif}(0, F(|L|))$.
3. Generate $u \sim \text{unif}(0, 1)$ independently of x.
4. Set candidate point $X^* = F^{-1}(x)$.

if $u \leq \lambda(X^*)/\lambda^*$ **then**
  Keep X^*.
else if $u > \lambda(X^*)/\lambda^*$ **then**
  Go to step 1.
end if

In the simulation, an assumption is made that the intensity function has a form $\lambda_0(s) = \exp(3 + 2s)$. Figure 2.1 shows the plots of this intensity function and the simulation of the points from the inhomogeneous Poisson point process on the line segment [0, 1] using the thinning method described in algorithm 2.1.



(a) (b)

Figure 2.1: (a) plot of an intensity function $\lambda(s) = \exp(3 + 2s)$ and (b) simulated points from the inhomogeneous Poisson process.

2.4 Estimation

2.4.1 Maximum Likelihood Estimation

In the spatial point processes models, spatial covariates may affect the intensity function of events. These covariates could be coordinates for events or covariate values on event locations (Baddeley et al., 2012). A model that allows us to reflect the relationship between spatial point process and spatial covariate is an inhomogeneous Poisson process with intensity function in the spatial covariates (Waagepetersen, 2008). In this section, a description is given of the maximum likelihood method for estimating the parameters of such models on a line segment. It is assumed that the intensity function $\lambda(s)$ of the inhomogeneous Poisson process $\{N(s), 0 < s < |L|\}$ with a sample of locations

$\mathbf{s} = \{s_1, ..., s_n\}$ of events in the line segment L depends on a spatial covariate $X(s)$ in the following way

$$\lambda(s; \alpha_0, \alpha_1) = \exp(\alpha_0 + \alpha_1 X(s)), \ 0 \leq s \leq |L|, \tag{2.1}$$

where $\Theta = \{\alpha_0, \alpha_1\}$ are unknown parameters to be estimated and $X(s)$ is a known covariate function at observed locations in the study region L. According to Cressie (1993), the likelihood of the inhomogeneous Poisson process is

$$\begin{aligned} L_L(\Theta|\mathbf{s}) &= \exp\left(-\int_0^{|L|} \lambda(u; \alpha_0, \alpha_1) d_1 u\right) \prod_{i=1}^{n} \lambda(s_i; \alpha_0, \alpha_1), \\ &= \exp\left(-\int_0^{|L|} \exp(\alpha_0 + \alpha_1 X(u)) \, d_1 u\right) \prod_{i=1}^{n} \exp(\alpha_0 + \alpha_1 X(s_i)), \\ &= \exp\left(-\int_0^{|L|} \exp(\alpha_0 + \alpha_1 X(u)) \, d_1 u\right) \exp\left(n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i)\right), \\ &= \exp\left(n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \int_0^{|L|} \exp(\alpha_0 + \alpha_1 X(u)) \, d_1 u\right), \end{aligned} \tag{2.2}$$

where the $d_1 u$ is one-dimensional integration over the line segment. The covariate $X(s)$ is only observed at locations of events, but not in the entire study region (Waagepetersen, 2008). Therefore, the integral in equation (2.2) cannot be calculated precisely. Berman and Turner (1992) overcame this problem by developing a numerical quadrature method in order to approximate likelihood function in equation (2.2). This method includes the approximation of the integral in equation (2.2) by a finite sum according to quadrature rule,

$$\int_0^{|L|} \exp(\alpha_0 + \alpha_1 X(u)) \, d_1 u = \sum_{j=1}^{m} w(u_j) \exp\left(\alpha_0 + \alpha_1 X(u_j)\right), \tag{2.3}$$

where $w(u_j), \ i = 1, ..., m$, are quadrature weights such that its sum is equal to the length of the line segment $|L|$. Let $\mathbf{q} = \{u_1, ..., u_m\}$ denote a set of quadrature points on the line segment L. The set of quadrature points is the union of the observed points $\mathbf{s}$ and a set of dummy points $\mathbf{d}$ which is a homogeneous dummy points process of constant intensity function. The choice of quadrature points should satisfy that observed points $\mathbf{s} = \{s_1, ..., s_n\} \subset \mathbf{q}$. The substitution of equation (2.3) into equation (2.2) gives the approximation of the likelihood function,

$$L_L(\Theta|\mathbf{s}) = \exp\left(n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} w(u_j) \exp\left(\alpha_0 + \alpha_1 X(u_j)\right)\right). \tag{2.4}$$

The log of the likelihood function in the equation (2.4) is:

$$\ell_L(\Theta) = n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j). \tag{2.5}$$

The **R** function **optim** is used to maximise the approximate log-likelihood in the equation (2.5). The function **optim** offers the approximate maximum log-likelihood estimates of α_0 and α_1.

2.4.2 Simulated Example

To demonstrate a maximisation of the likelihood function in equation (2.4), a simulated example is used. It is supposed that there is only one spatial covariate related with an intensity of events in equation (2.1) and the values of the covariate can be produced from the following function,

$$X(s) = \begin{cases} 2 & \text{if } 0 \leq s \leq 0.2, \\ 3 & \text{if } 0.2 < s \leq 0.7, \\ 4 & \text{if } 0.7 < s \leq 1. \end{cases} \tag{2.6}$$

In fact, the values of covariate must be known in all line segment (Rathbun et al., 2007). Therefore, dummy points are simulated in the line segment. The number of dummy points k is the product of p and the number of observed points, where p is a proportion of dummy points compared with observed points. In general, p can take values such that it leads to smaller standard error (Waagepetersen, 2008). In fact, when the number of dummy points is large, this gives accurate estimates. Here, p is chosen to be 0.25 and the dummy points are generated with probability density function unif(0,1). The number of observed points that are simulated on the line segment [0,1] is n. In the simulation of observed points $\mathbf{s}$, true values of model parameters in equation (2.1) are $\alpha_0 = 3$ and $\alpha_1 = 1$. Let $\mathbf{q} = \mathbf{s} \bigcup \mathbf{d}$ denote a set of quadrature points. To produce quadrature weights w_j, segment [0,1] is divided into k sub-segments $L_1, ..., L_k$ such that each sub-segment only includes one dummy point and it may or may not contain data points. All quadrature points $u_j \in \mathbf{q}$ within a given sub-segment L_j receive the same weight w_j. The quadrature weight w_j for a quadrature point $u_j \in \mathbf{q}$ falling in a sub-segment L_j is the length of L_j divided by the number of quadrature points u_j falling in L_j. Algorithm 2.2 illustrates the method to simulate the inhomogeneous point process with the intensity function depending on the covariate function and the simulation of the dummy points.

Algorithm 2.2 Simulation of the inhomogeneous Poisson process with intensity depending on covariate $X(s)$, simulation of dummy points and evaluation of log-likelihood.

1- Initialize $\alpha_0 = 3$ and $\alpha_1 = 1$.
2- Set $\Lambda_1 = 0.2\exp(\alpha_0 + 2\alpha_1)$, $\Lambda_2 = 0.5\exp(\alpha_0 + 3\alpha_1)$ and $\Lambda_3 = 0.3\exp(\alpha_0 + 4\alpha_1)$.
3- Simulate three homogeneous Poisson processes N_1, N_2 and N_3 with means Λ_1, Λ_2 and Λ_3 respectively.
4- Simulate locations of events as $\mathbf{s}_1 \sim \text{unif}(N_1, 0, 0.2)$, $\mathbf{s}_2 \sim \text{unif}(N_1, 0.2, 0.7)$ and $\mathbf{s}_3 \sim \text{unif}(N_3, 0.7, 1)$.
5- Set $\mathbf{s} = \mathbf{s}_1 \cup \mathbf{s}_2 \cup \mathbf{s}_3$.
6- Simulate dummy points $\mathbf{d}$ from runif(0, 1), where p=0.25.
7- Divide the line segment [0,1] into k sub-segments such that every sub-segment receives one dummy point.
8- Let $\mathbf{q} = \mathbf{s} \cup \mathbf{d}$.
9- Weight for each point in sub-segment L_j is $W_j = \frac{\text{the length of } L_j}{\text{the number of points in } L_j}$.
10- Use the function **optim** to maximise log-likelihood in equation (2.5).

The MLE estimates are $\hat{\alpha_0} = 2.9901$ and $\hat{\alpha_1} = 1.0058$. Figure 2.2 shows the sketch of log-likelihood profiles. Table 2.1 displays estimated values and 95% confidence intervals of parameters α_0 and α_1.

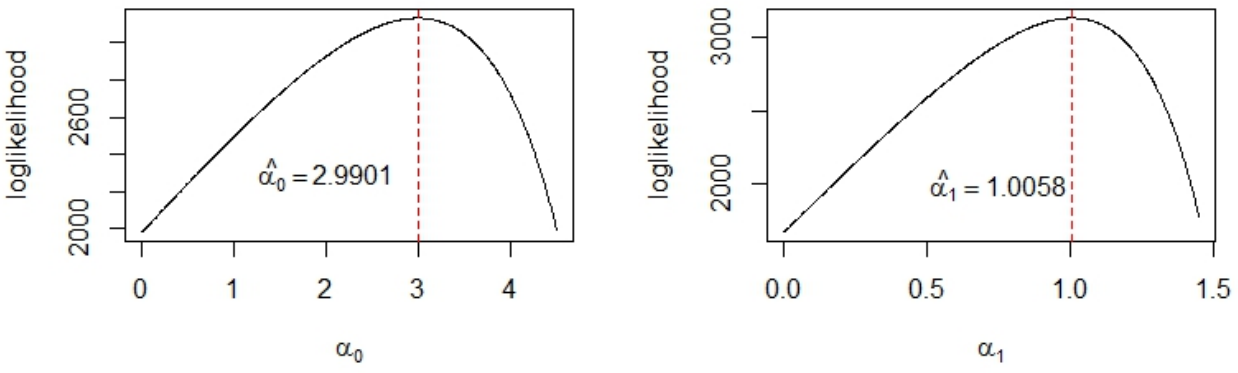


Figure 2.2: Plots of maximised likelihood function in equation (2.4) of parameters α_0 and α_1 for the intensity function in equation (2.1).

Parameter	True value	Estimated value	Standard Error	95% CI
α_0	3	2.9901	0.2532	(2.4938, 3.4863)
α_1	1	1.0058	0.0706	(0.8674, 1.1442)

Table 2.1: The maximum likelihood estimates, standard errors and 95% confidence intervals of parameters α_0 and α_1. The data are simulated from a nonhomogeneous spatial point process with intensity $\exp(3 + X(s))$ in the line segment $[0, 1]$.

2.4.3 Bayesian Estimation

In this section, Bayesian methods will be used for estimating α_0 and α_1 in the likelihood specified in equation (2.4). The basic idea of the Bayesian estimation is to update the belief about model parameters Θ by combining the prior belief about parameters with the observed data $\mathbf{s} = \{s_1, ..., s_n\}$. The independent prior distributions of the model parameters α_0 and α_1 and updated belief are defined as the joint prior distribution $P(\Theta) = P(\alpha_0)P(\alpha_1)$ and the joint posterior distribution $\pi(\Theta|s)$ respectively and the observed data are represented by the likelihood function $L(\Theta|\mathbf{s})$. Using Bayes theorem, the posterior distribution of Θ is given by

$$\pi(\Theta|\mathbf{s}) = \frac{L(\Theta|\mathbf{s})\ P(\Theta)}{\int_0^{|L|} L(\Theta|\mathbf{s})\ P(\Theta) d_1\Theta}, \tag{2.7}$$

where the integral in the denominator is called a normalising constant. Since the output of the integral is a function in the observed data $\mathbf{s}$, so the posterior distribution can be written as

$$\pi(\Theta|\mathbf{s}) \propto L(\Theta|\mathbf{s})\ P(\Theta). \tag{2.8}$$

From equation (2.4), the likelihood function is

$$L(\Theta|\mathbf{s}) = \exp\left(n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j)\right). \tag{2.9}$$

We specify Gamma(a, b) and Gamma(c, d) as independent prior distributions for the unknown parameters α_0 and α_1. The joint posterior probability density for unknown parameters α_0 and α_1 is given by

$$\begin{aligned}
\pi(\Theta|\mathbf{s}) &= L\left(\Theta|\mathbf{s}\right) P\left(\Theta\right) \\
&= P(\mathbf{s}|\alpha_0, \alpha_1)P(\alpha_0)P(\alpha_1) \\
&= \exp\left(n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j)\right) \alpha_0^{a-1} \exp\left(-b\alpha_0\right)\ \alpha_1^{c-1} \exp\left(-d\alpha_1\right) \\
&= \alpha_0^{a-1}\alpha_1^{c-1} \exp\left(-b\alpha_0 - d\alpha_1 + n\alpha_0 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j)\right).
\end{aligned} \tag{2.10}$$

Then, the conditional posterior distributions are derived as,

$$\pi(\alpha_0|\alpha_1, \mathbf{s}) = \alpha_0^{a-1} \exp\left(-b\alpha_0 + n\alpha_0 - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j)\right), \tag{2.11}$$

and

$$\pi(\alpha_1|\alpha_0, \mathbf{s}) = \alpha_1^{c-1} \exp\left(-d\alpha_1 + \alpha_1 \sum_{i=1}^{n} X(s_i) - \sum_{j=1}^{m} \exp\left(\alpha_0 + \alpha_1 X(u_j)\right) w(u_j)\right). \tag{2.12}$$

Both conditional posterior distributions in equations (2.11) and (2.12) do not have a known form. Hence, it is not possible to use the Gibss sampler. Instead, Metropolis-Hastings within Gibbs sampler is used to simulate chains from proposal distributions q_1 and q_2 of parameters α_0 and α_1, respectively. Algorithm 2.3 shows that Metropolis-Hastings within Gibbs sampler is implemented such that candidate values of both model parameters are accepted according to the following probabilities:

$$r_1\left(\alpha_0^{(t-1)}, \acute{\alpha}_0\right) = \min\left(\frac{\pi\left(\acute{\alpha}_0|\alpha_1^{(t-1)}\right) q_1\left(\alpha_0^{(t-1)}, \acute{\alpha}_0\right)}{\pi\left(\alpha_0^{(t-1)}|\alpha_1^{(t-1)}\right) q_1\left(\acute{\alpha}_0, \alpha_0^{(t-1)}\right)}, 1\right). \tag{2.13}$$

$$r_2\left(\alpha_1^{(t-1)}, \acute{\alpha}_1\right) = \min\left(\frac{\pi\left(\acute{\alpha}_1|\alpha_0^{(t)}\right) q_2\left(\alpha_1^{(t-1)}, \acute{\alpha}_1\right)}{\pi\left(\alpha_1^{(t-1)}|\alpha_0^{(t)}\right) q_2\left(\acute{\alpha}_1, \alpha_1^{(t-1)}\right)}, 1\right). \tag{2.14}$$

Algorithm 2.3 Metropolis-Hastings within Gibbs sampler.

Set initial values, $\alpha_0^{(0)}$ and $\alpha_1^{(0)}$.
For each iteration t, $t = 1, ..., T$
Step 1. Update α_0.

- Generate a proposed value $\acute{\alpha}_0 \sim q_1(., \alpha_0^{(t-1)})$.
- Calculate the probability $r_1\left(\alpha_0^{(t-1)}, \acute{\alpha}_0\right)$ specified in equation (2.13).
- With probability r_1, set $\alpha_0^{(t)} = \acute{\alpha}_0$, otherwise set $\alpha_0^{(t)} = \alpha_0^{(t-1)}$.

Step 2. Update α_1.

- Generate a proposed value $\acute{\alpha}_1 \sim q_2(., \alpha_1^{(t-1)})$.
- Calculate the probability $r_2\left(\alpha_1^{(t-1)}, \acute{\alpha}_1\right)$ specified in equation (2.14).
- With probability r_2, set $\alpha_1^{(t)} = \acute{\alpha}_1$, otherwise set $\alpha_1^{(t)} = \alpha_1^{(t-1)}$.

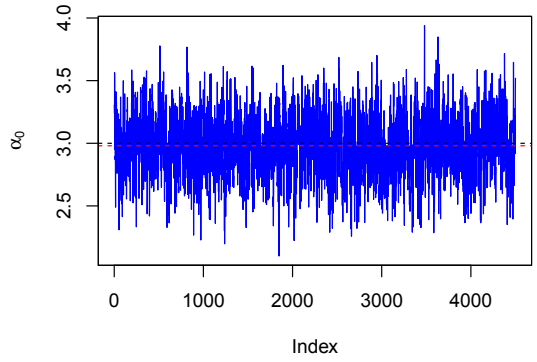
2.4.4 Simulated Example

In this section, algorithm 2.3 is applied to the simulated data of section 2.4.2. The simulated data represent points of the inhomogeneous spatial Poisson process $\{N(s); 0 < s < |L|\}$ with the intensity function $\lambda(s) = \exp(\alpha_0 + \alpha_1 X(s))$ on the line segment $L = [0, 1]$. Here, $X(s)$ is a covariate function that is defined in equation (2.6). True values of $\alpha_0 = 3$ and $\alpha_1 = 1$ were chosen. The proposal distributions q_1 of α_0 and q_2 of α_1 are respectively $N(\alpha_0^{(t-1)}, 0.04)$ and $N(\alpha_1^{(t-1)}, 0.003)$, where $\alpha_0^{(t-1)}$ and $\alpha_1^{(t-1)}$ represent the current values of the simulated chains for α_0 and α_1. The values of parameters

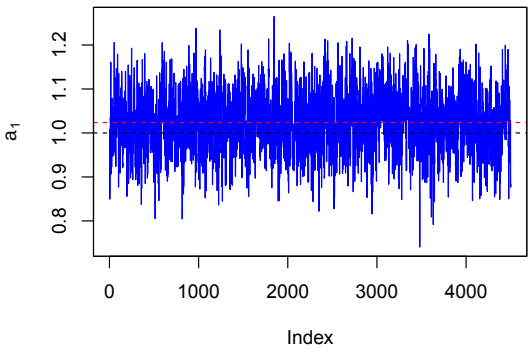
of prior distributions are $a = b = c = d = 0.01$. Algorithm 2.3 is run for 500,000 iterations with discarding 50,000 iterations as a burn-in period as well as a thinning interval of 100. The sample size was sufficient to ensure that the chain had converged and that there was enough samples after the burn-in to ensure reasonable estimates. Figure 2.3 shows a trace plot, autocorrelation function and histogram as well as imposed posterior density for the parameters of α_0 and α_1. In this figure, the first row displays the trace plots of parameters after thinning. The second row offers the autocorrelation function plots of parameters after thinning. The final row presents the histograms and imposed posterior density plots of parameters after thinning. Table 2.2 also shows an actual input value, posterior mean and median, standard error and 95% credible interval (CI) after thinning for each parameter. Simulated chains are well mixed and autocorrelation plots of thinning chains show that correlation within produced samples decays fast at lag 6 as it can be seen in Figures 2.3(c) and 2.3(d). The posterior means and medians are comparable to the input values and the 95% credible interval are reasonably tight so that no problem is apparent in the MCMC implementation of the model. The acceptance rates of α_0 and α_1 chains are 0.32 and 0.24 and these acceptance rates are within the range of (0.24, 0.40) (Gelman et al., 1996).

Parameter	True value	Posterior mean	Posterior median	Standard deviation	95% CI.
α_0	3	2.9795	2.9799	0.2457	(2.5015, 3.4503)
α_1	1	1.0229	1.0236	0.0681	(0.8908 , 1.1560)

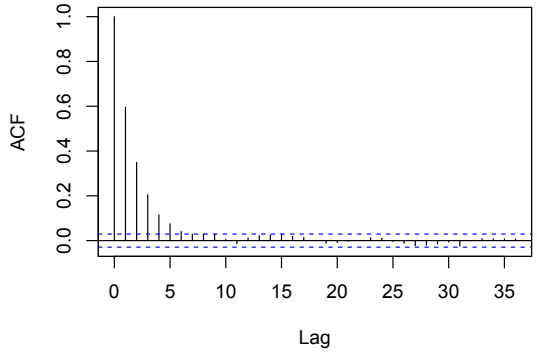
Table 2.2: Summary of the posterior sample of α_0 and α_1.



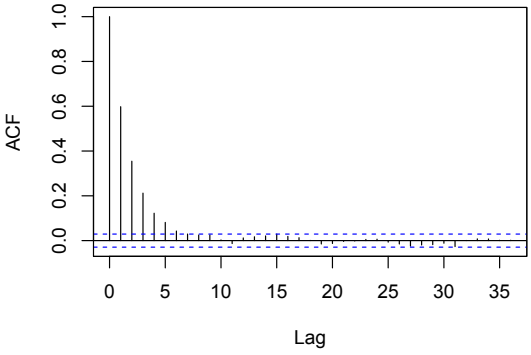
(a) Trace of α_0



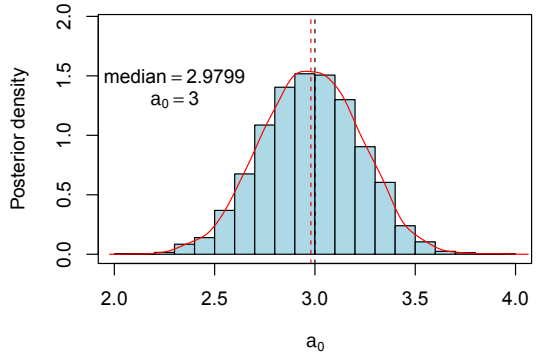
(b) Trace of α_1



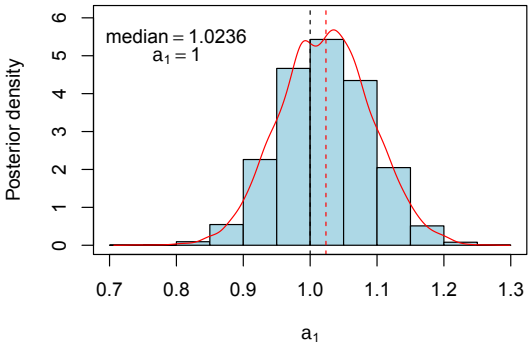
(c) Autocorrelation of α_0



(d) Autocorrelation of α_1



(e) Histgram and density of α_0



(f) Histgram and densityof α_1

Figure 2.3: Trace plots, autocorrelation function, histgram and density plots of MCMC chain for α_0 and α_1. Dashed red and black lines, respectively, represent medians of simulated chains and true values of α_0 and α_1.

2.5 Discussion

The main object of this chapter is preparation for the spatial point process on a linear network that consists of a set of line segments in the plane linked to each other by nodes. Therefore, definitions were provided of a line segment in the plane, spatial Poisson point processes in the line segment involving homogeneous and inhomogeneous point processes. Methods that are parametric for estimating an intensity of events on the line segment were displayed. The parametric method includes the maximum likelihood and the Bayesian estimations. In the Bayesian framework, Metropolis-Hastings within Gibbs sampler is used. In implementation, simulation studies are fitted to spatial point process models on the line segment. In these studies, algorithms were introduced to simulate inhomogeneous spatial point processes with intensity depending on coordinates of events as well as with intensity depending on a spatial covariate in the line segment. The sampler performed well, and has converged to the target distribution (the posterior distribution of α_0 and α_1). When the findings from Bayesian and maximum likelihood estimations were compared, it could be seen that the performance for both methods is well and similar. Therefore, the question is which method should be used? The choice between two methods depends on the inferential framework and computational issues (Farrell and Ludwig, 2008). The main difference between approaches lies in the their philosophical idea. Bayesian modelling is interested in presenting the posterior distribution by multiplying prior and likelihood. While maximum likelihood approach is interested in point estimates of parameters. In addition, in the Bayesian analysis, we need to set the prior distributions of parameters that require knowledge in the models. These priors allow for the addition of related information into our models. The other difference includes the ease application and implementation of these methods. The maximum likelihood estimation only require the likelihood function and minimising routine that is available and straightforward for models fitting of data on the line segment.

Chapter 3

Two-Level Hierarchical Models

3.1 Introduction

The objective of traffic safety studies is to determine danger spots on road networks that involve a high density of traffic accidents (Okabe and Sugihara, 2012). In the early studies of traffic accidents, count data were used to identify the distribution of the danger spots. For this type of road accidents data, observations are non-negative integer values (Ahmed et al., 2014). For example, the number of accidents is count data such that it is calculated with respect to road segments and then the number of traffic accidents is used to produce the density of accidents on each road segment in order to investigate the risk spots (Okabe and Sugihara, 2012). In this approach, a road network should be divided into road segments. Some studies considered different road segment lengths (Ceder and Livneh, 1978; Ng and Hauer, 1989; Stern and Zehavi, 1990; Miaou, 1994). Other studies considered the same road segment lengths (Golob et al., 1990; Thomas, 1996; Black, 1991; Erdogan et al., 2008; Yamada and Thill, 2010). In spatial analysis, there is a problem called modifiable areal unit (Openshaw, 1979; Thomas, 1996). This term means that statistical results could be affected by the scale of spatial unit, namely, the lengths of road segments may lead to different results (Okabe and Sugihara, 2012). To avoid the modifiable areal unit problem resulting from the use of count data in traditional statistical analysis, the individual data of accidents on the road network can be used (Okabe and Sugihara, 2012).

In this chapter, the motorway network is considered as a linear network and road accidents as a spatial point pattern involving the spatial locations of accidents. Baddeley et al. (2015) studied point processes on the linear network where they defined the linear network as vertices that are joined by straight line segments in two dimensions. A point process on a linear network has the same properties as the point process in two dimensions except for an intensity of points along the network

where it represents the expected number of points per unit length of network (Baddeley et al., 2015). For this study, the unit length is a meter. Let $\mathbf{X}$ denote the point process on the linear network L with the intensity λ and $\mathbf{B}$ is a subset of L. The parameter λ is called a homogeneous intensity if the expected number of points falling in $\mathbf{B}$ is $E(\mathbf{X} \cap \mathbf{B}) = \lambda|\mathbf{B}|$ where $|\mathbf{B}|$ is the length of $\mathbf{B}$ (Baddeley et al., 2015). An intensity function $\lambda(s)$ at all locations s on L is called an inhomogeneous intensity if the expected number of points falling in the subset $\mathbf{B}$ of L is $E(\mathbf{X} \cap \mathbf{B}) = \int_L \lambda(u) d_1 u$ where the $d_1 u$ is one-dimensional integration over the line segment (Baddeley et al., 2015). Methods used in order to estimate the intensity function on a linear network include point process models that are used to fit the point pattern dataset. This requires specifying the form of the intensity function where the parameters of the model are estimated using the maximum likelihood method (Baddeley et al., 2015). Currently, there are no published papers which use Bayesian inference to analyse a spatial point pattern on the linear network. Therefore, in this chapter, the aim is to estimate the intensity function of accidents and study its pattern across the UK motorway network using a Bayesian approach. The motorway-specific intensity function is estimated by modelling the point pattern of the accident data using a homogeneous Poisson process. The homogeneous Poisson process is used to model all intensity functions but heterogeneity is incorporated across motorways using a hierarchical approach. The parameters in the hierarchical models are estimated by one-stage fully Bayesian, two-stage semi-Bayesian and frequentist approach. The non-hierarchical model involves both Bayesian and frequentist approaches. In the Bayesian approach, a sensitivity analysis is conducted by using different priors. The performance of the proposed models is evaluated using a simulation study. The dataset used in this chapter is obtained from the website of the Department for Transport in Great Britain. The data include locations of accidents on 49 motorways in the UK for 2016. The intensity is defined as the expected number of traffic accidents per unit length (meter).

3.2 One-Stage Fully Bayesian Hierarchical Method (Model 1)

3.2.1 Model Definition

The analysis of traffic accident data from all motorways in a single step is called a one-stage approach. Let m denote the total number of motorways. The number of accidents n_i on the motorway i ($i = 1, ..., m$) follows a Poisson distribution with mean $\lambda_i L_i$ where L_i represents the length of motorway i and λ_i is the accidents intensity on the motorway i per unit length. Here $\lambda_i L_i$ is the expected number of accidents on the motorway i and it can vary from motorway to motorway because each motorway could have different conditions and features. Let $\alpha_i = \log \lambda_i$ denote the log-intensity function and assume it follows a normal distribution $\mathrm{N}\left(\alpha, \tau^2\right)$. Thus, the model for traffic accidents on the motorway is:

$$\begin{aligned} n_i &\sim \text{Pois}(\lambda_i L_i),\ i = 1, ..., m, \\ \alpha_i &\sim \mathrm{N}(\alpha, \tau^2). \end{aligned} \tag{3.1}$$

Here α is the overall log-intensity and τ^2 is the between-motorway variance. In this model, each accident's location follows uniform distribution on interval $(0, L_i)$.

3.2.2 Likelihood Function

Let $\mathbf{N} = \{n_i,\ i = 1, ..., m\}$ represent the accident count and $\Theta = \left\{\alpha_1, \alpha_2, ..., \alpha_m, \alpha, \tau^2\right\}$ the model parameters. The likelihood for model (3.1) is given by,

$$L\left(\mathbf{N}|\Theta\right) \propto \prod_{i=1}^{m} \exp\left(n_i \alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi\tau^2}}\right) \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right). \tag{3.2}$$

For details about deriving the likelihood function see Appendix A.1.1.

3.2.3 Prior Distribution

A prior distribution is an important part in the Bayesian approach. The specification of the prior distribution depends on available information about unknown parameters. If the prior information is not enough or unavailable, then the useful choice of the prior distribution is a non-informative prior. Another option is a vague prior with a large variance. The vague prior distribution is selected when the amount of data is not small (Stojanovski et al., 2011). On the other hand, an informative prior takes into account any belief or knowledge about unknown parameters. Furthermore, the prior distribution leads to a posterior distribution which has the same distribution family as the prior (Gel-

man et al., 2014). This type of prior is called conjugate prior. A conjugate prior with a large variance leads to the vague prior. The strategy for specifying prior distributions for the parameters in the hierarchical model includes conjugate, vague and weakly-informative priors.

For α, conjugate normal prior $\mathrm{N}\left(\mu_0, \sigma_0^2\right)$ is assigned. We consider a conjugate inverse gamma prior with shape α_0 and rate β_0 for τ^2. We specify a uniform prior $\mathrm{unif}\,(0, a)\,,\ a > 0$ as prior distribution on the between-motorway standard deviation (τ) (Lambert et al., 2005). The half-normal prior is specified as $\tau \sim \mathrm{HN}\left(0, \theta^2\right)$, where $\theta^2 = \dfrac{\pi}{2\sigma^2}$ and $\sigma > 0$ as detailed in Klaus et al. (2015).

3.2.4 Posterior Distribution

Posterior if the prior distribution on τ^2 is an inverse gamma distribution

The posterior distribution is the product of the likelihood and the prior distribution. Therefore, the joint posterior density function for parameters given data is

$$\pi(\Theta|\mathbf{N}) = \prod_{i=1}^{m} \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi\tau^2}}\right) \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\alpha - \mu_0)^2}{2\sigma_0^2}\right) \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\tau^2\right)^{-\alpha_0 - 1} \exp\left(-\beta_0/\tau^2\right). \tag{3.3}$$

The conditional posterior distribution of α_i is given by,

$$\pi\left(\alpha_i|\alpha, \tau^2, \mathbf{N}\right) \propto \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right). \tag{3.4}$$

For details about deriving the conditional posterior distribution of α_i see Appendix A.1.2.1.

The conditional posterior distribution of α is a normal distribution $\mathrm{N}\left(\mu_\alpha, \sigma_\alpha^2\right)$ with mean and variance:

$$\mu_\alpha = \frac{\dfrac{\sum_{i=1}^{m} \alpha_i}{\tau^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{m}{\tau^2} + \dfrac{1}{\sigma_0^2}},\ \sigma_\alpha^2 = \frac{1}{\dfrac{m}{\tau^2} + \dfrac{1}{\sigma_0^2}}. \tag{3.5}$$

For details about deriving the conditional posterior distribution of α see Appendix A.1.2.2.

The conditional posterior distribution of τ^2 is given by,

$$\pi(\tau^2|\alpha, \alpha_1, ..., \alpha_m, \mathbf{N}) \propto \left(\tau^2\right)^{-(\alpha_0 + \frac{m}{2}) - 1} \exp\left(-\frac{\beta_0 + \dfrac{\sum_{i=1}^{2} (\alpha_i - \alpha)^2}{2}}{\tau^2}\right). \tag{3.6}$$

Hence, τ^2 has an inverse gamma distribution with shape $a = \alpha_0 + \frac{m}{2}$ and rate $b = \beta_0 + \frac{\sum_{i=1}^{m}(\alpha_i - \alpha)^2}{2}$. For details about deriving the conditional posterior distribution of τ^2 see Appendix A.1.2.3.

Posterior if the prior distribution on τ is a uniform distribution

The joint posterior density of parameters given data is

$$\pi(\Theta|\mathbf{N}) \propto \prod_{i=1}^{m} \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{m}\left(\frac{1}{\sqrt{2\pi\tau^2}}\right)\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right). \tag{3.7}$$

The conditional posterior distributions of $\alpha_i, i = 1, ..., m$ and α are as in equations (3.4) and (3.5). The conditional posterior density of τ is given by,

$$\pi(\tau|\alpha_1, ..., \alpha_m, \alpha, \mathbf{Y}) \propto \left(\frac{1}{\sqrt{2\pi\tau^2}}\right)^m \exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right). \tag{3.8}$$

Posterior if the prior distribution on τ is a half-normal distribution

The joint posterior distribution of parameters given data is

$$\pi(\Theta|\mathbf{N}) = \prod_{i=1}^{m} \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{m}\left(\frac{1}{\sqrt{2\pi\tau^2}}\right)\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right)\times\frac{2\theta}{\pi}\exp\left(-\frac{\tau^2\theta^2}{\pi}\right). \tag{3.9}$$

The conditional posterior distributions of $\alpha_i,\ i = 1, ..., m$ and α are the same as in equations (3.4) and (3.5). The conditional posterior distribution of τ is

$$\pi(\tau|\alpha_1, ..., \alpha_m, \alpha, \mathbf{N}) \propto \tau^{-m}\exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2} - \frac{\tau^2\theta^2}{\pi}\right),\ \tau > 0. \tag{3.10}$$

For details on the derivation of the conditional posterior distribution of τ, see Appendix A.1.2.4.

3.2.5 Estimation

In equations (3.5) and (3.6), the conditional posterior distributions of α and τ^2 given other parameters have a known form, but the conditional posterior distributions of $\alpha_i, i = 1, ..., m$ given other parameters in equation (3.4) do not have known forms. Therefore, Metropolis-Hastings within Gibbs sampler

is used to generate samples (Markov chain) of $\alpha_i, i = 1, ..., m$, α and τ^2. The Metropolis-Hastings sampler does not directly generate samples from the full conditional distribution. Instead, a proposal distribution is chosen given the current value of the parameter, $\boldsymbol{\alpha}_i^{(t-1)}$, where t is iteration index. The proposal distribution $\mathrm{q}_1(., \alpha_i^{(t-1)})$ for the proposed value $\acute{\alpha}_i$ is the normal distribution with mean equalling to current value $\alpha_i^{(t-1)}$ and variance is chosen such that an acceptance rate of $\acute{\alpha}_i$ is between 0.24 and 0.40 (Gelman et al., 1996). Then the Metropolis-Hastings is defined by two steps: firstly, generate a proposed value, $\acute{\alpha}_i$, from the proposal distribution, $\mathrm{q}_1(., \alpha_i^{(t-1)})$; secondly, the proposal value is accepted as the next value with the probability

$$r_1(\alpha_i^{(t-1)}, \acute{\alpha}_i) = \min\left\{\frac{\pi\left(\acute{\alpha}_i|\alpha^{(t-1)}, \tau^{2(t-1)}\right)\mathrm{q}_1\left(\alpha_i^{(t-1)}, \acute{\alpha}_i\right)}{\pi\left(\alpha_i^{(t-1)}|\alpha^{(t-1)}, \tau^{2(t-1)}\right)\mathrm{q}_1\left(\acute{\alpha}_i, \alpha_i^{(t-1)}\right)}, 1\right\}. \tag{3.11}$$

If the proposed value is rejected, then the current value is taken as the next value in the Markov chain. The uniform prior distribution on τ leads to the posterior distribution on τ that is given in equation (3.8). This posterior distribution does not have a closed form, therefore, the Metropolis-Hastings sampler is used. In order to move into the next state, the following two steps are defined: firstly, draw a proposal value, $\acute{\tau}$, from the proposal distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$. The proposal distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$ is a normal distribution with current state $\tau^{(t-1)}$ as mean and variance 0.9. Secondly the proposed value is accepted with the probability

$$r_2(\tau^{(t-1)}, \acute{\tau}) = \min\left\{\frac{\pi\left(\acute{\tau}|\alpha^{(t)}, \alpha_1^{(t)}, ..., \alpha_m^{(t)}\right)\mathrm{q}_2\left(\tau^{(t-1)}, \acute{\tau}\right)}{\pi\left(\tau^{(t-1)}|\alpha^{(t)}, \alpha_1^{(t)}, ..., \alpha_m^{(t)}\right)\mathrm{q}_2\left(\acute{\tau}, \tau^{(t-1)}\right)}, 1\right\}. \tag{3.12}$$

The conditional posterior in equation (3.10) is produced by using half-normal prior distribution and it does not have a closed form. Therefore, the Metropolis-Hastings sampler is utilized to simulate Markov chain of τ. This sampler includes generating the proposed value $\acute{\tau}$ from the proposal distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$ and accepting this value with the probability $r_2\left(\tau^{(t-1)}, \acute{\tau}\right)$ which is described in equation (3.12). The proposal distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$ is a normal distribution with current state $\tau^{(t-1)}$ as mean and variance 0.09. Hence, the algorithm for estimating parameters of Model 1 with the three prior distributions is given in Algorithm 3.1.

Algorithm 3.1 Sampling from the full conditional posterior distributions of parameters for the two-level Bayesian hierarchical model (Model 1) using Metropolis-Hastings within Gibbs sampling.

Set initial values, $\boldsymbol{\alpha}^{(0)} = (\alpha_1^{(0)}, ..., \alpha_m^{(0)})$, $\alpha^{(0)}$ and $\tau^{2(0)}$.
For each iteration t.

Step 1: Update $\boldsymbol{\alpha}$ one by one.

[1.1] Generate a proposed value, $\acute{\alpha}_i \sim \mathrm{q}_1(., \alpha_i^{(t-1)})$.

[1.2] Calculate the probability $r_1(\alpha_i^{(t-1)}, \acute{\alpha}_i)$ specified in equation (3.11).

[1.3] With probability r_1, set $\alpha_i^{(t)} = \acute{\alpha}_i$, otherwise set $\alpha_i^{(t)} = \alpha_i^{(t-1)}$.

[1.4] Repeat steps 1.1 to 1.3 for all $\alpha_i, i = 1, ..., m$.

Step 2: Update full conditional posterior density $\pi\left(\alpha^{(t)} | \alpha_i^{(t)}, \tau^{2^{(t-1)}}, \mathbf{N}\right),\ i = 1, ..., m$, specified in equation (3.5).

Step 3: Update τ.

[3.1] Generate a proposed value, $\acute{\tau} \sim \mathrm{q}_2(., \tau^{(t-1)})$.

[3.2] Calculate the probability $r_2(\tau^{(t-1)}, \acute{\tau})$ specified in equation (3.12).

[3.3] With probability r_2, set $\tau^{(t)} = \acute{\tau}$, otherwise set $\tau^{(t)} = \tau^{(t-1)}$.

3.3 Two-Stage Semi-Bayesian Hierarchical Method (Model 2)

In the two-stage method, traffic accidents from each motorway are analysed separately in order to obtain summary statistics (such as point estimates and their standard deviations), then, they are combined by hierarchical models (Burke et al., 2017).

3.3.1 Model Definition

In a two-stage approach, the log-intensity will be estimated for each motorway separately in stage one using a frequentist approach. The resulting estimates from stage one are then used as data in stage two, where log intensities are assumed to arise from a common population distribution with an unknown mean and variance. At stage one, firstly all motorways are analysed independently to estimate the log-intensity of accidents α_i $(i = 1, ..., m)$ for the motorway i. The accident location s_{ij} is assumed to follow a uniform distribution on interval $(0, L_i)$,

$$s_{ij} \sim \text{unif}\,(0, L_i)\,,\ i = 1, ..., m, j = 1, ..., n_i \tag{3.13}$$

where L_i represents the length of motorway i and n_i is the total number of accidents on the motorway i. The number of accidents n_i follows the Poisson distribution with mean $\lambda_i L_i$. Let $\mathbf{N} = \{n_i, i = 1, ..., m\}$. The likelihood function is:

$$\begin{aligned} L\,(\mathbf{N}|\alpha_i) &= P\,(\mathbf{s}_i) \times P\,(n_i) \\ &= \frac{1}{L_i^{n_i}} \times \frac{(\lambda_i L_i)^{n_i} \exp\,(-\lambda_i L_i)}{n_i!} \\ &\propto \exp\,(n_i \alpha_i - L_i \exp\,(\alpha_i))\,, \end{aligned} \tag{3.14}$$

where $s_i = \left(s_{ij}, i = 1, ..., m, j = 1, ..., n_i\right)$. The estimated log-intensity of accidents and its standard deviation for each motorway i are obtained using maximum likelihood estimation. The **MaxLik** function in the MaxLik package is used (Chandler et al., 2013).

In the second stage, the Bayesian approach is used to estimate the overall log-intensity. The second stage of the model can be formulated as:

$$\begin{aligned} \mathrm{y}_i &\sim \mathrm{N}\left(\alpha_i, \sigma_i^2\right), \\ \alpha_i &\sim \mathrm{N}\left(\alpha, \tau^2\right), \end{aligned} \tag{3.15}$$

where $y_i = \hat{\alpha}_i$ is the maximum likelihood estimate of the log accidents intensity of motorway i and σ_i^2 its corresponding variance. In this model, y_i and σ_i^2 for all $i = 1, ..., m$ are known from stage one. The parameters α_i for all $i = 1, ..., m$ are unknown nuisance parameters. α represents an overall mean. τ^2 represents the variability between motorways (heterogeneity). This model is termed a random effects model because it allows the log-intensity function to vary from one motorway to the other.

3.3.2 Likelihood Function

Setting $\mathbf{y} = (y_1, ..., y_m)$ and $\Theta = (\alpha_1, ..., \alpha_m, \alpha, \tau^2)$, the likelihood function is given by:

$$L(\mathbf{y}|\Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right). \tag{3.16}$$

For details about deriving the likelihood function see Appendix A.2.1.

3.3.3 Posterior Distribution

The same prior distributions for α and τ^2 as the ones described in section 3.2.3 are considered.

Posterior if the prior distribution on τ^2 is an inverse gamma distribution

To produce the joint posterior density function, the likelihood in equation (3.16) is combined with the prior density function of α and the prior density function of τ^2. The joint posterior distribution of the parameters given data is given by:

$$\begin{aligned}\pi(\Theta|\mathbf{y}) = &\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right) \\ &\times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\alpha - \mu_0)^2}{2\sigma_0^2}\right) \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\tau^2\right)^{-\alpha_0 - 1} \exp\left(-\beta_0/\tau^2\right).\end{aligned} \tag{3.17}$$

For details about deriving the joint posterior distribution see Appendix A.2.2.

Using equation (3.17), the conditional posterior density of α_i given other parameters is obtained:

$$\pi\left(\alpha_i|\alpha, \tau^2, \mathbf{y}\right) \propto \exp\left(-\frac{(\alpha_i - \mu_{\alpha_i})^2}{2\sigma_{\alpha_i}^2}\right), \tag{3.18}$$

and this is a normal distribution with mean $\mu_{\alpha_i} = \dfrac{\dfrac{y_i}{\sigma_i^2} + \dfrac{\alpha}{\tau^2}}{\dfrac{1}{\sigma_i^2} + \dfrac{1}{\tau^2}}$ and variance $\sigma_{\alpha_i}^2 = \dfrac{1}{\dfrac{1}{\sigma_i^2} + \dfrac{1}{\tau^2}}$.

For details about deriving the conditional posterior density of α_i see Appendix A.2.3.1.

Using equation (3.17), the conditional posterior density of α is obtained by considering α as a random variable and α_i, τ^2 as known. Hence,

$$\pi(\alpha|\alpha_1, ..., \alpha_m, \tau^2, \mathbf{y}) \propto \exp\left(-\frac{(\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2}\right), \tag{3.19}$$

where this represents the normal distribution for α with mean $\mu_\alpha = \dfrac{\dfrac{\sum_{i=1}^m \alpha_i}{\tau^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{m}{\tau^2} + \dfrac{1}{\sigma_0^2}}$ and variance $\sigma_\alpha^2 = \dfrac{1}{\dfrac{m}{\tau^2} + \dfrac{1}{\sigma_0^2}}$.

For details about deriving the conditional posterior density of α see Appendix A.2.3.2.

Using equation (3.17), the conditional posterior density of τ^2 given other parameters is derived. Hence:

$$\pi\left(\tau^2|\alpha_1, ..., \alpha_m, \alpha, \mathbf{y}\right) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-(\alpha_0 + \frac{m}{2}) - 1} \exp\left(-\frac{\beta_0 + \frac{\sum_{i=1}^m (\alpha_i - \alpha)^2}{2}}{\tau^2}\right). \tag{3.20}$$

The above density represents the density function of an inverse gamma distribution with $a = \alpha_0 + \dfrac{m}{2}$ and $b = \beta_0 + \dfrac{\sum_{i=1}^m (\alpha_i - \alpha)^2}{2}$. For more details about deriving the conditional posterior distribution of τ^2 see Appendix A.2.3.3.

Posterior if the prior distribution on τ is a uniform distribution

The joint posterior distribution of parameters given data can be written as follows:

$$\pi\left(\Theta|\mathbf{y}\right) \propto \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \prod_{i=1}^m \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right)$$
$$\times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\alpha - \mu_0)^2}{2\sigma_0^2}\right). \tag{3.21}$$

The conditional posterior distributions of α_i and α are the same as the distributions in equations (3.18) and (3.19), while the conditional posterior distribution of τ is :

$$\pi\left(\tau|\alpha_1,...,\alpha_m,\alpha,\mathbf{y}\right) \propto \left(\frac{1}{\tau^2}\right)^{\frac{m}{2}} \exp\left(-\frac{\sum_{i=1}^{m}(\alpha_i-\alpha)^2}{2\tau^2}\right). \tag{3.22}$$

Posterior if the prior distribution on τ is a half-normal distribution

The joint posterior distribution of parameters given data is

$$\begin{aligned}\pi\left(\Theta|\mathbf{y}\right) \propto \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{\left(\mathrm{y}_i-\alpha_i\right)^2}{2\sigma_i^2}\right)\times\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\\ \times\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right)\times\frac{2\theta}{\pi}\exp\left(-\frac{\tau^2\theta^2}{\pi}\right).\end{aligned} \tag{3.23}$$

The conditional posterior distribution of α_i and α are the same as in equations (3.18) and (3.19). The conditional posterior distribution of τ is

$$\pi\left(\tau|\alpha_1,...,\alpha_m,\alpha,\mathbf{y}\right) = \left(\frac{1}{\tau^2}\right)^{\frac{m}{2}} \exp\left(-\frac{\sum_{i=1}^{m}(\alpha_i-\alpha)^2}{2\tau^2}-\frac{\tau^2\theta^2}{\pi}\right). \tag{3.24}$$

3.3.4 Estimation

When the conditional posterior distribution is in a closed form, Gibbs sampler is used to generate Markov chain of the parameter. To simulate from the conditional posterior on τ in (3.22), the Metropolis-Hastings within Gibbs sampler is used. A symmetric proposal distribution q(., $\tau^{(t-1)}$)is chosen given the current value of the parameter, $\tau^{(t-1)}$, where t is an iteration index. We chose $\mathrm{N}\left(\tau^{(t-1)},\sigma^2\right)$ with variance $\sigma^2 = 0.09$ as proposal distribution of proposed value $\acute{\tau}$. To move from the current state to the next state, we define the following two steps. Firstly, generate a proposed value, $\acute{\tau}$, from the proposal distribution ,q(., $\tau^{(t-1)}$); secondly, the proposed value is accepted with the probability

$$r(\tau^{(t-1)},\acute{\tau}) = \min\left\{\frac{\pi\left(\acute{\tau}|\alpha^{(t)},\alpha_1^{(t)},...,\alpha_m^{(t)}\right)\mathrm{q}\left(\tau^{(t-1)},\acute{\tau}\right)}{\pi\left(\tau^{(t-1)}|\alpha^{(t)},\alpha_1^{(t)},...,\alpha_m^{(t)}\right)\mathrm{q}\left(\acute{\tau},\tau^{(t-1)}\right)},1\right\}. \tag{3.25}$$

If the proposed value is rejected, then the current value is accepted. The form of the conditional posterior on τ in equation (3.24) is unknown, so the Metropolis-Hastings within Gibbs sampler is used to generate the Markov chain from the conditional posterior distributions of α_i, α and τ. Algorithm 3.2 shows the Metropolis-Hastings within Gibbs sampler to generate from conditional posterior distributions in equations (3.18), (3.19), (3.22) and (3.24).

Algorithm 3.2 Sampling from the full conditional posterior distributions of parameters for the two-level semi-Bayesian hierarchical model (Model 2) using Metropolis-Hastings within Gibbs sampling.

Initialise $\alpha^{(0)}$ and $\tau^{2^{(0)}}$

For each iteration t,

1. Update α_i by Gibbs sampler, $\alpha_i^{(t)} \sim \mathrm{N}(\mu_{\alpha_i}^{(t-1)}, \sigma_{\alpha_i}^{2(t-1)}), i = 1, .., m$ that is defined in (3.18).
2. Update α by Gibbs sampler, $\alpha^{(t)} \sim \mathrm{N}(\mu_{\alpha}^{(t-1)}, \sigma_{\alpha}^{2(t-1)}),\ i = 1, ..., m$ that is defined in (3.19).
3. Update τ.
   - Generate a proposed value, $\acute{\tau} \sim \mathrm{q}(., \tau^{(t-1)})$.
   - Calculate the probability $r(\tau^{(t-1)}, \acute{\tau})$ specified in equation (3.25).
   - With probability r, set $\tau^{(t)} = \acute{\tau}$, otherwise set $\tau^{(t)} = \tau^{(t-1)}$.

3.4 Two-Stage Frequentist Hierarchical Method (Model 3)

This section describes a two-stage approach to fit the intensity of accidents on the UK motorway network.

In stage one, the intensity function for each motorway is estimated using the maximum likelihood method to obtain the estimated log-intensity function y_i and the corresponding standard deviation σ_i.

In stage two, the log-intensity function across motorways is combined to produce an overall log-intensity estimate. The model can be formulated as:

$$\begin{aligned} \mathrm{y}_i &\sim \mathrm{N}(\alpha_i, \sigma_i^2), \\ \alpha_i &\sim \mathrm{N}(\alpha, \tau^2). \end{aligned} \tag{3.26}$$

Here y_i represents the estimated intensity on log scale for motorway i, α_i represents the true log-intensity and σ_i^2 is the within-motorway variance corresponding to y_i; α is the overall intensity on log scale and τ^2 represents the between-motorway heterogeneity.

The model was set up as in (3.26) above with distributional assumptions of normality for y_i and α_i. The marginal distribution of each estimated log-intensity y_i is therefore normal with mean α and variance $\left(\sigma_i^2 + \tau^2\right)^{-1}$ (Hardy and Thompson, 1996). Hence the contribution of motorway i to the likelihood for α and τ^2 is given by,

$$L_i\left(\alpha, \tau^2\right) = \frac{1}{\sqrt{2\pi\left(\sigma_i^2 + \tau^2\right)}} \exp\left(-\frac{\left(\mathrm{y}_i - \alpha\right)^2}{2\left(\sigma_i^2 + \tau^2\right)}\right). \tag{3.27}$$

For m independent motorways, the likelihood is given by the product of the individual motorway likelihoods, so the likelihood on all motorways is

$$L\left(\alpha, \tau^2\right) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\left(\sigma_i^2 + \tau^2\right)}} \exp\left(-\frac{\left(\mathrm{y}_i - \alpha\right)^2}{2\left(\sigma_i^2 + \tau^2\right)}\right). \tag{3.28}$$

In order to estimate α and τ, we use the R function **rma** that employs the maximum likelihood method (ML).

3.5 Non-hierarchical Bayesian and Frequentist Methods (Model 4 and 5)

In this section, a non-hierarchical Bayesian model is considered for the accidents on the whole UK motorway network for comparison to hierarchical models. In this model, the number of accidents on the whole network is considered as a homogeneous process. The total number of accidents N has a Poisson distribution with a mean $|L| \exp(\alpha)$, where $\lambda = \exp(\alpha)$ represents the intensity of accidents on the network and $|L|$ is the network length:

$$N \sim \text{Pois}(\lambda |L|). \tag{3.29}$$

The likelihood function is given as:

$$L(N|\alpha) \propto \exp\left(N\alpha - |L| \exp(\alpha)\right). \tag{3.30}$$

$N(\mu_0, \sigma_0^2)$ is chosen to be a non-informative prior distribution for the parameter α which is the log-intensity of accidents on the motorway network. So the posterior density function of α is

$$\pi(\alpha|N) = \exp\left(N\alpha - |L| \exp(\alpha)\right) \exp\left(\frac{(\alpha - \alpha_0)^2}{2\sigma_0^2}\right). \tag{3.31}$$

The posterior density of α does not have a closed form, so the Metropolis-Hastings sampler is used to make inference about the posterior distribution of α. Also the maximum likelihood method is used to analyse the non-hierarchical model in equation (3.30).

3.6 Estimation Results for Motorway Data

Non-informative and weakly-informative prior distribution

One could choose a conjugate normal prior $\mathrm{N}(0, 100)$ of α. We consider a conjugate inverse gamma

prior for τ^2 with shape and rate $\alpha_0 = \beta_0 = 0.001$. The weakly-informative prior is $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$. This prior is chosen to assess the sensitivity to the choice of prior parameters. Another option is a non-informative prior distribution for τ which is a uniform prior unif $(0, 100)$ (Lambert et al., 2005). Finally, the non-informative half-normal distribution HN$(0, 0.02)$ may be specified as prior on τ (Lambert et al., 2005).

Results

In this section, results from analysis the observed accident data described in section 3.1 are provided. For models 1, 2 and 3, two parameters which are the overall log-intensity (α) of accidents per meter and the heterogeneity between motorway (τ) are estimated. The MCMC simulation process requires specifying starting points for the parameters. Therefore, the initial values $\alpha = 0$ and $\tau = 0.1$ are specified. The MCMC algorithm was run of Model 1 for 100,000 iterations with burn-in 10,000 and thinning interval 10 and of Model 2 for 50,000 iterations with burn-in 5,000 and thinning interval 10. The first 10% of the iterations are discard in order to minimize the effect of the initial values on the posterior inference.

Table 3.1 displays estimation results for three hierarchical models given various prior distributions. In the frequentist hierarchical method (Model 3), overall log accidents intensity (α) across all motorways is estimated to be -6.811 with a standard deviation of 0.099 and a 95% confidence interval $(-7.004, -6.618)$. The heterogeneity between motorway (τ) was estimated to be 0.641 with a standard deviation of 0.096 and a 95% confidence interval $(0.535, 0.885)$. For all prior distributions specified for Model 2, results from hierarchical models 2 and 3 are similar. In more detail, estimates of the overall log-intensity of accidents of Model 2 including the mean posterior and its standard deviation as well as 95% credible interval are similar to overall log-intensity of accidents estimated from Model 3. In the same way, estimates of the heterogeneity between motorway are similar for both models 2 and 3. Regarding Model 1, the posterior mean and its standard deviation are slightly different from models 2 and 3 where the difference is obvious in estimates of α and τ with respect to all specified prior distributions. Results from Model 1 show that the prior distribution has a slight influence on the posterior.

On the other hand, Table 3.2 shows that non-hierarchical Bayesian and maximum likelihood methods gave different estimates for parameters of models 4 and 5 where the posterior mean of Model 4 parameter is -6.489 and a point estimate of Model 5 parameter is -6.289. For non-hierarchical models, there was no difference between the range of Bayesian credible intervals and the likelihood confidence interval. A comparison of Tables 3.1 and 3.2 shows the disagreement in results of hierarchical and non-hierarchical models, where there is a variability in the posterior mean estimates and

corresponding standard deviations of models 4 and 5 versus models 1, 2 and 3.

Prior distribution	Parameters	Model 1			Model 2			Model 3		
		PM	PSD	95% CI	PM	PSD	95% CI	PM	PSD	95% CI
$\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$	α	-6.820	0.094	(-7.008, -6.637)	-6.810	0.100	(-7.014, -6.618)	-6.811	0.099	(-7.004, -6.618)
	τ	0.601	0.078	(0.468, 0.770)	0.650	0.081	(0.507, 0.832)	0.641	0.096	(0.535, 0.885)
$\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$	α	-6.856	0.106	(-7.065, -6.651)	-6.810	0.100	(-7.016, -6.618)	-6.811	0.099	(-7.004, -6.618)
	τ	0.691	0.086	(0.542, 0.881)	0.653	0.081	(0.510, 0.834)	0.641	0.096	(0.535, 0.885)
$\tau \sim$ HN $(0, 0.02)$	α	-6.857	0.106	(-7.071, -6.654)	-6.813	0.102	(-7.021, -6.615)	-6.811	0.099	(-7.004, -6.618)
	τ	0.700	0.088	(0.549, 0.890)	0.665	0.085	(0.522, 0.852)	0.641	0.096	(0.535, 0.885)
$\tau \sim$ unif $(0, 100)$	α	-6.858	0.108	(-7.072, -6.644)	-6.811	0.102	(-7.015, -6.612)	-6.811	0.099	(-7.004, -6.618)
	τ	0.699	0.088	(0.550, 0.899)	0.662	0.085	(0.517, 0.847)	0.641	0.096	(0.535, 0.885)

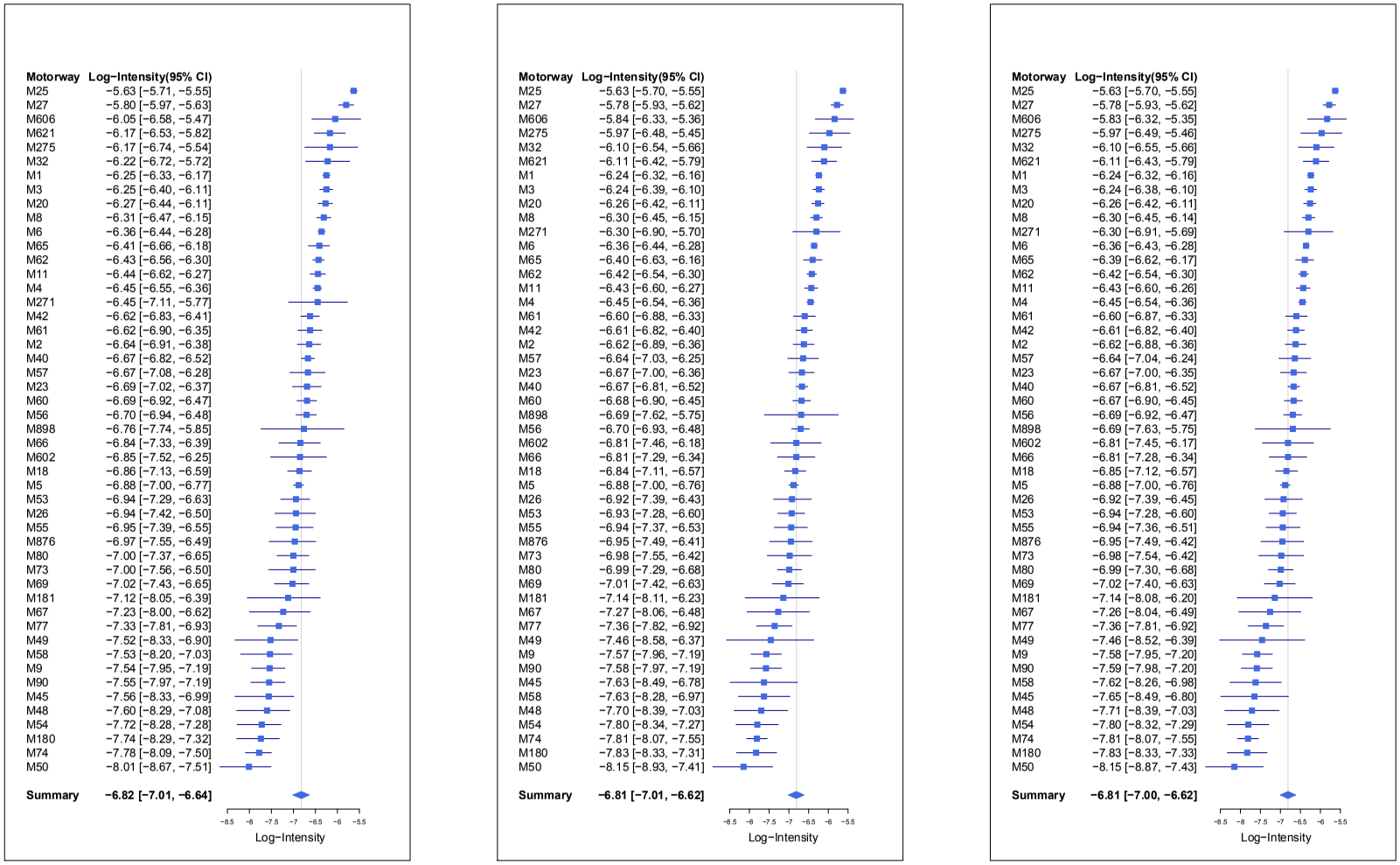
Table 3.1: Results from fully Bayesian hierarchical model (Model 1) and semi-Bayesian hierarchical model (Model 2) as well as maximum likelihood parameter estimation results from frequenitist hierarchical model (Model 3). PM: Posterior Mean. PSD: Posterior Standard Deviation. CI: Credible Interval or confidence interval.

Model 4			Model 5		
Posterior mean	Posterior standard deviation	95%CI	Point estimate	Standard deviation	95%CI
-6.489	0.014	(-6.517, -6.461)	-6.289	0.013	(-6.314, -6.263)

Table 3.2: Results from the Bayesian non-hierarchical model(Model 4) and Maximum likelihood parameter estimation results from frequenitist non-hierarchical model (Model 5). CI: Credible Interval or confidence interval.

Figure 3.1 illustrates the posterior means (point estimates) and corresponding 95% credible intervals for the estimates of the log-intensity of accidents α_i on each motorway i and the overall log-intensity of accidents α from the analysis of hierarchical models 1, 2 and 3 using prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim \text{N}(0, 10^2)$. From Figure 3.1, it is clear that the posterior mean and the point estimate of the overall log-intensity α of accidents are similar for models 2 and 3 and the width of the 95% credible interval is similar. The estimates of the posterior mean and the 95% credible interval of Model 1 are slightly different. The posterior mean of the log-intensity of accidents on each motorway is comparable between models 2 and 3, but there is a variability in the posterior means and the width of 95% credible intervals of the log-intensity of accidents on the motorways M45, M56, M58 and M898. Model 1 appears to have a different performance in the inference regarding the estimation of width of 95% credible interval for some α_i. Findings from Model 1 are similar to those from models 2 and 3 in point and credible interval estimates of α_i for motorways M25, M27, M606, M60, M1, M3, M20, M8, M18, M5, M55, M876, M69, M181, M67, M77, M49, M48, M54 and M50.

Figure 3.1 shows that three motorways with the highest intensity of traffic accidents are M25, M27 and M606. The M25 has the highest intensity of traffic accidents where the expected number of accidents is 3.59 per one kilometer. The M27 has the second highest intensity of traffic accidents with the expected number of accidents 3.03 per one kilometer. The M606 completes the top three with the expected number of accidents 2.36 per one kilometer. In addition, Figure 3.1 shows that three motorways with the lowest intensity of accidents are M50, M74 and M180. The lowest intensity of accidents is on the M50 with the expected number of accidents 3.32 per 10 kilometers of the M50. In terms of the lowest for intensity of traffic accidents, the M74 occupies the second rank where the expected number of accidents is 4.18 per 10 kilometers. The third rank for the lowest accidents intensity is for the M180 with the expected number of accidents 4.35 per 10 kilometers.



(a) Model 1 (b) Model 2 (c) Model 3

Figure 3.1: Results from one-stage fully Bayesian hierarchical model (Model 1), two-stage semi-Bayesian hierarchical model (Model 2) and frequentist hierarchical model (Model 3) analysis of observed accident data on the 49 motorways in the UK for year 2016. Prior distributions are $\alpha \sim \mathrm{N}\left(0, 10^2\right)$ and $\tau^2 \sim$ Inv-Gamma $(0.001, 0.001)$. Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote the corresponding credible intervals and the solid vertical line represents the estimate of the overall log accident intensity α.

Similarly, Figure A.14 in Appendix shows that a prior distribution has no effect on the findings from Model 2 where point and credible interval estimates of overall log-intensity are similar for all specified prior distributions. Whereas, the effects of the prior distribution on estimates of Model 1 parameters are not present (very weak) as a non-informative gamma prior gave a very similar point estimate of the overall log-intensity (Figure 3.1 and Appendix Figure A.13).

In Figure 3.2, a comparison is made among the posterior mean, point estimate and 95% credible interval of the parameter α from models 1, 2 and 3 using a hierarchical model with those obtained from models 4 and 5 using a non-hierarchical model. This figure shows the discrepancies between hierarchical models and non-hierarchical models

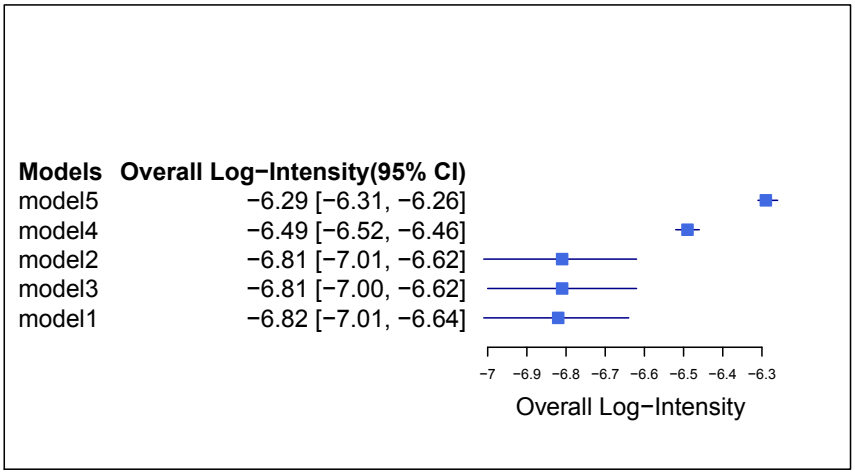


Figure 3.2: The posterior mean and 95% credible interval for the overall log-intensity α using Model 1, Model 2, Model 3. The Bayesian method for estimating the log-intensity of non-hierarchical model (Model 4) and the maximum likelihood estimation of the log-intensity of non-hierarchical model (Model 5).

Plotting residuals considered the common diagnostic technique to assess the appropriateness of the model. Let N_i represent the observed number of accidents on motorway i and $\hat{N}_i$ denote the predicted value of the number of accidents on motorway i. Let $\hat{\lambda}_i$ represent Bayesian estimate of the accidents intensity $\hat{N}_i = \hat{\lambda}_i L_i$ (Baddeley et al., 2015). The residual (R_i) of the observed data can be defined as $R_i = N_i - \hat{N}_i$. Here, $\hat{\lambda}_i (i = 1, ..., m)$ are the estimated intensities of motorways in Figure 3.1(a).

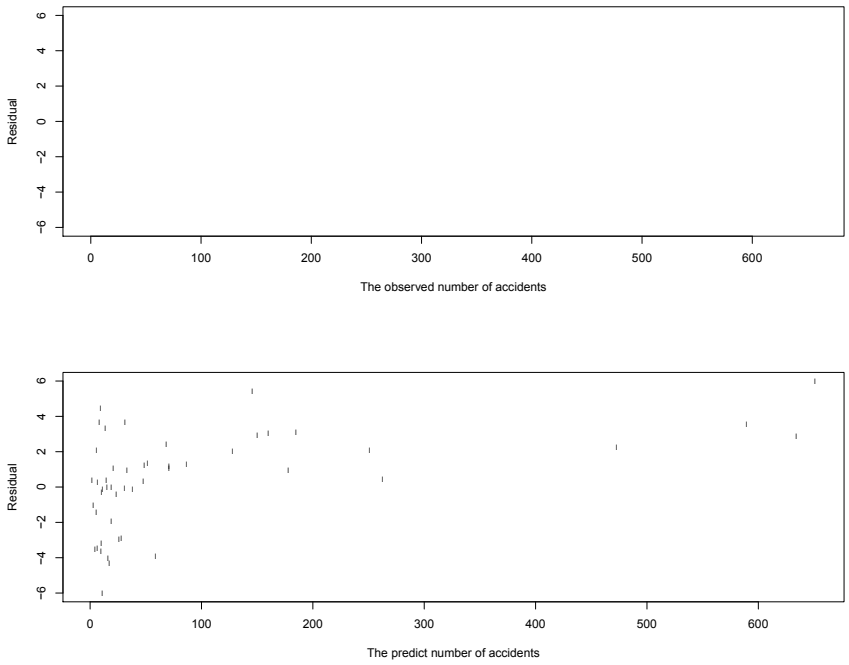


Figure 3.3: Residuals plots. The predicted value of the number of accidents is calculated using the two-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016.

The predicted intensities of accidents, $\hat{\lambda}_i$ $(i = 1, ..., m)$, after fitting Model 1 for the UK motorway network data were divided into five levels. Level one includes $\hat{\lambda}_i < 0.5$ pointing out a very low-risk. Level two includes $0.5 \leq \hat{\lambda}_i < 1$ pointing out a low-risk. Level three includes $1 \leq \hat{\lambda}_i < 2$ pointing out a moderate-risk. Level four includes $2 \leq \hat{\lambda}_i < 3$ pointing out a high-risk. Level five includes $\hat{\lambda}_i \geq 3$ pointing out a very high-risk. Figure 3.4 shows that a general level of the intensity of accidents on the UK motorway network is the moderate-risk where the moderate-risk motorways are M32, M1, M3, M20, M8, M6, M65, M62, M11, M4, M271, M42, M61, M2, M40, M57, M23, M60, M56, M898, M66, M602, M18 and M5. Motorways M25 surrounding almost all of Greater London, England, except North Ockendon, in the United Kingdom and M27 in Hampshire, England, starting west-east from Cadnam to Portsmouth, have a very high-risk level. The expected numbers of accidents are 3.59 per one kilometer of M25 and 3.03 per one kilometer of M27. The motorways M54, M180, M74 and M50 form the lowest risk motorways and their estimated intensities are 4.4, 4.4, 4.2, 3.3 per 10 kilometers. Figure 3.4, moreover, illustrates that the risk intensity level for motorways M606, M621 and M275 is high and the expected number of accidents is 2.36 per one kilometer of M606 and 2.09 per one kilometer of both M621 and M275.

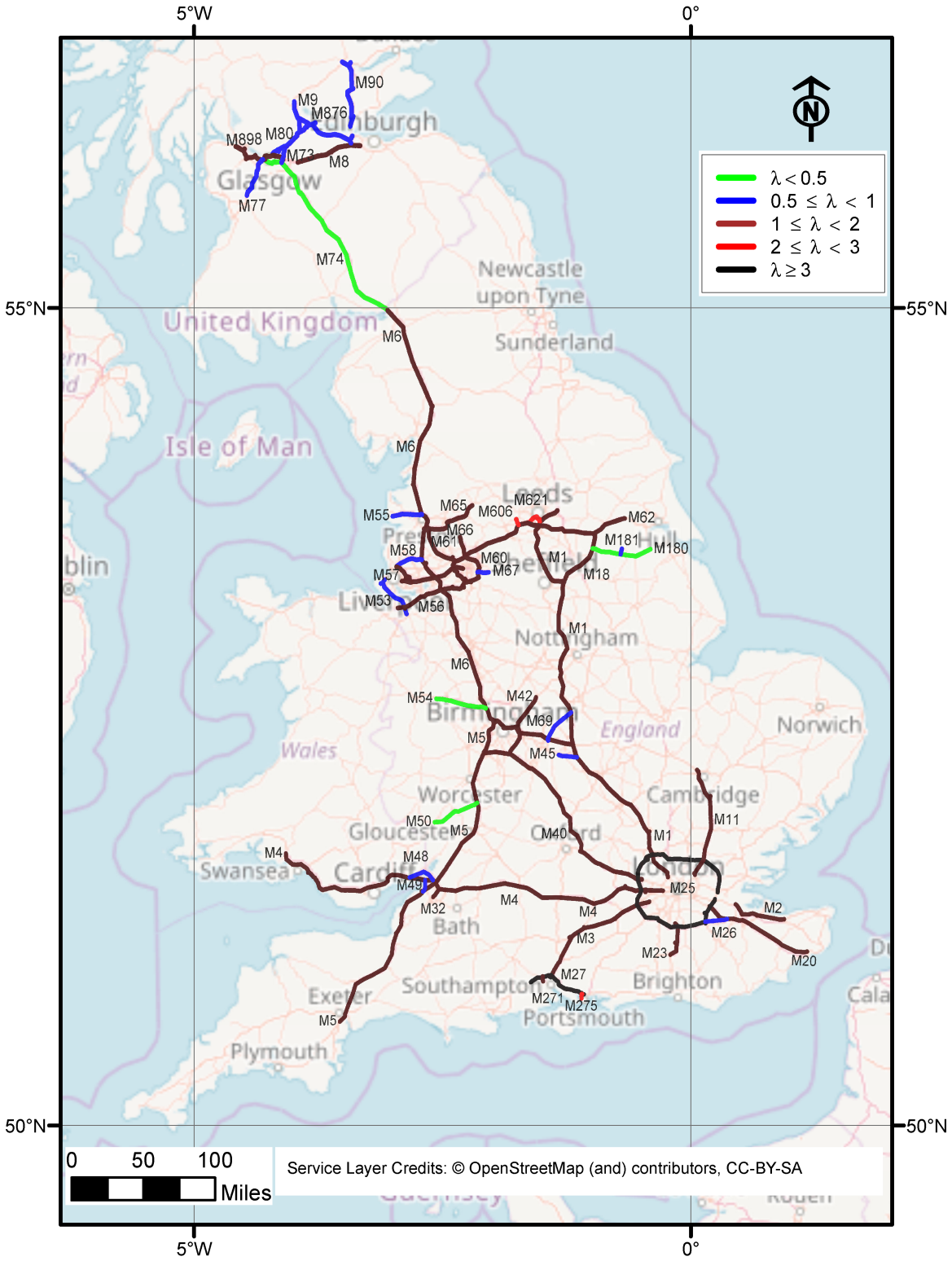


Figure 3.4: An intensity of traffic accidents (λ_i) per one kilometer on the UK motorway network including 49 motorways. This plot is produced using the traffic accidents data for year 2016. Prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}\,(0.001, 0.001)$.

The important issue in the MCMC analysis is an assessment of convergence. There are several diagnostic tools to investigate whether the simulated Markov chain converges to the stationary distribution. The most popular diagnostic tool is a visual inspection involving trace, autocorrelation and density plots (van de Schoot and Depaoli, 2014). Figures 3.5 and 3.6 show the trace plots, autocorrelation functions (ACF) and histograms for the parameters α and τ of both models 1 and 2 with prior distribution $\tau^2 \sim$ Inv-Gamma(0.001, 0.001). For each of these figures, the first row includes the trace plots of all model parameters, where the blue colour represents the generated samples and the dashed red line is the posterior mean. The second row represents ACF plots of all model parameters. The third row represents the marginal density histograms of all model parameters. The dashed red line is the posterior mean. As can be seen from the results in Table 3.1 and Figures 3.5 and 3.6, the samplers perform well in estimating the true plots of parameter α suggesting that there is no correlation between the samples produced by the samplers; with regard to ACF plots of τ, autocorrelation decays quickly in all figures across both models. More specifically, both the α and τ parameters appear to have a slight correlation at the first lags and then begin to fade quickly. This was easily treated by thinning. Figures A.1, A.2, A.3, A.4, A.5 and A.6 in Appendix show that the trace plots, autocorrelation functions and histograms for the parameters α and τ of models 1 and 2 for the three remaining prior distributions are insensitive to priors.

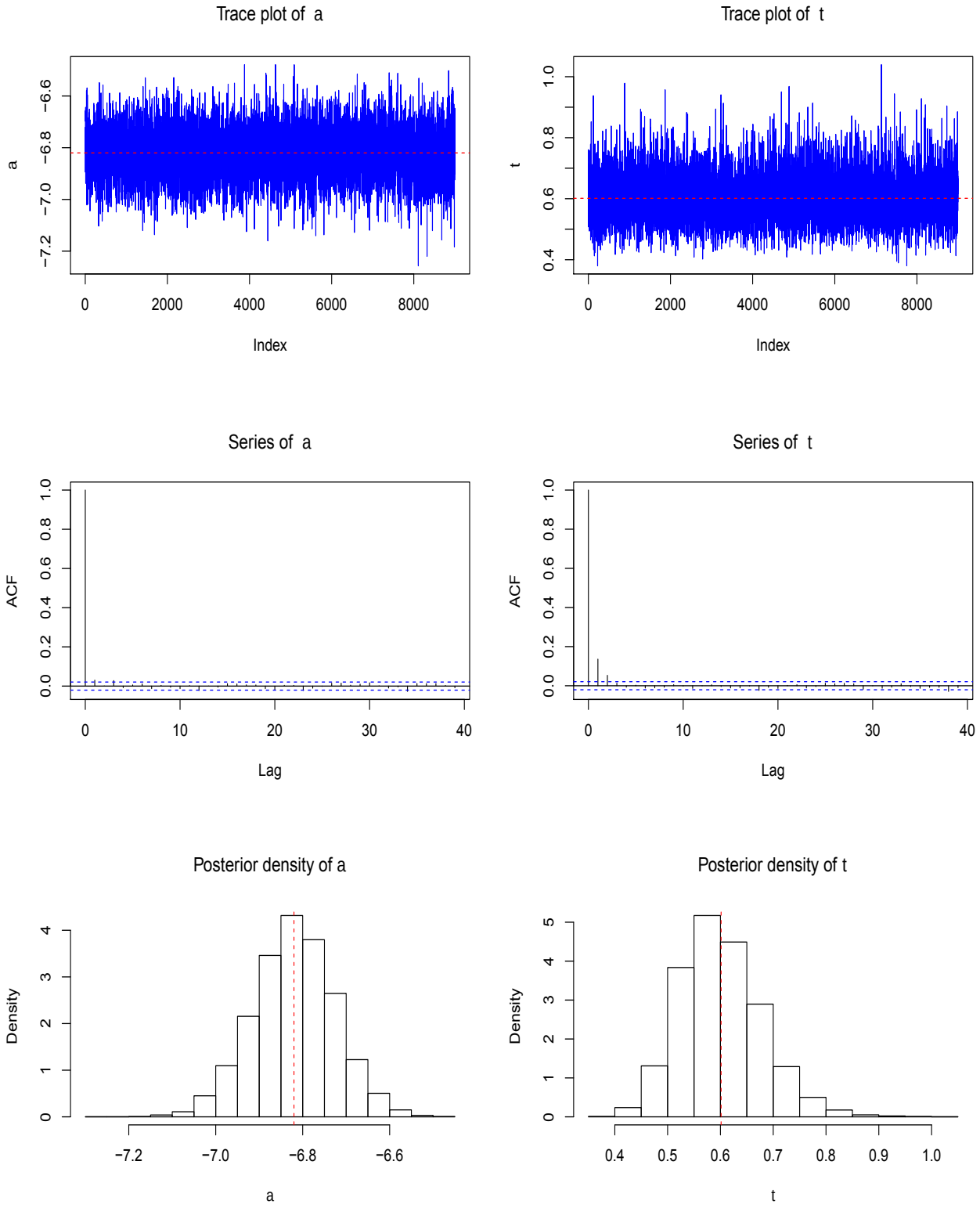


Figure 3.5: Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

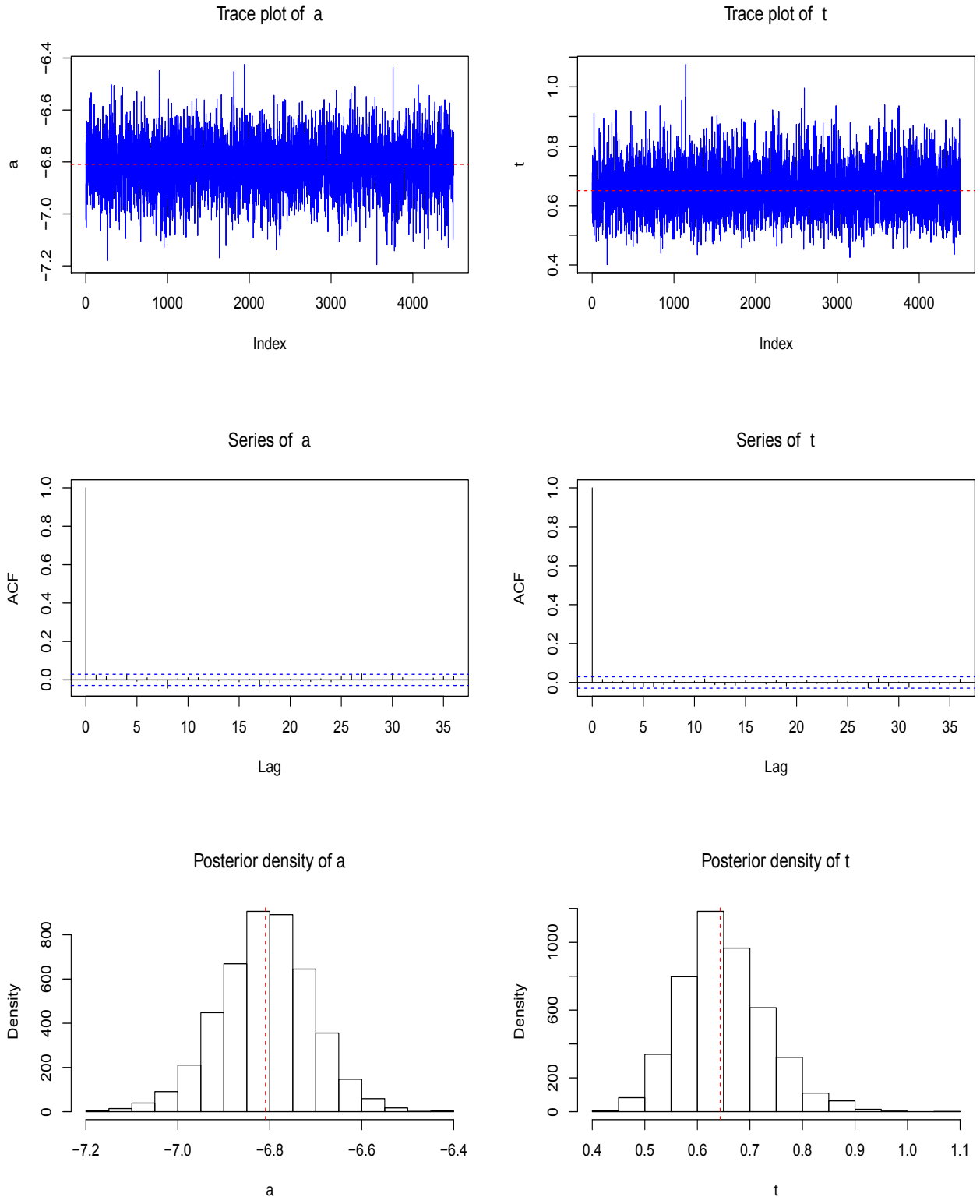


Figure 3.6: Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

However, the visual inspection to assess the convergence does not guarantee that the chain has a stationary distribution (Hamra et al., 2013). Therefore, Gelman-Rubin or Geweke diagnostic inspections are used. The Gelman-Rubin statistic evaluates a difference between the variance within multiple chains and the variance between multiple chains through calculating the Gelman-Rubin statistic $\hat{R}$ (Gelman and Rubin, 1992a). To implement Gelman-Rubin diagnostic, two MCMC simulations are run with two different overdispersed starting values $(\alpha, \tau) = (-10, 0.25)$ and $(10, 3)$ using algorithms 3.1 and 3.2 . The first row in Figures 3.7 and 3.8 shows that the trace plots of the two chains with different starting values of parameters α and τ are stationary. The second row in Figures 3.7 and 3.8 shows that the Gelman-Rubin statistic is less than 1.2. In the same context, from graphs in the first row of Figures A.7, A.8, A.9, A.10, A.11 and A.12 in Appendix, it is clear that the Gelman-Rubin statistics of the MCMC chains of the parameters α and τ of both models 1 and 2 across the other prior distributions are less than 1.2.

The convergence is also investigated using the Geweke diagnostic for each parameter across Model 1 and Model 2. The Geweke diagnostic splits a chain into two parts of iterations and measures the similarity between the mean of the first part of the iterations and the mean of the last part of the iterations by standard normal statistic Z (Geweke, 1991; Cowles and Carlin, 1996). The third row in Figures 3.7 and 3.8 shows the Geweke statistic Z versus the first part of the iteration. The Z score is within the interval $(-1.96, 1.96)$ (Best et al., 1995). From all diagnostic results, it is concluded that the MCMC chain of parameters α and τ across both models have the stationary distribution.

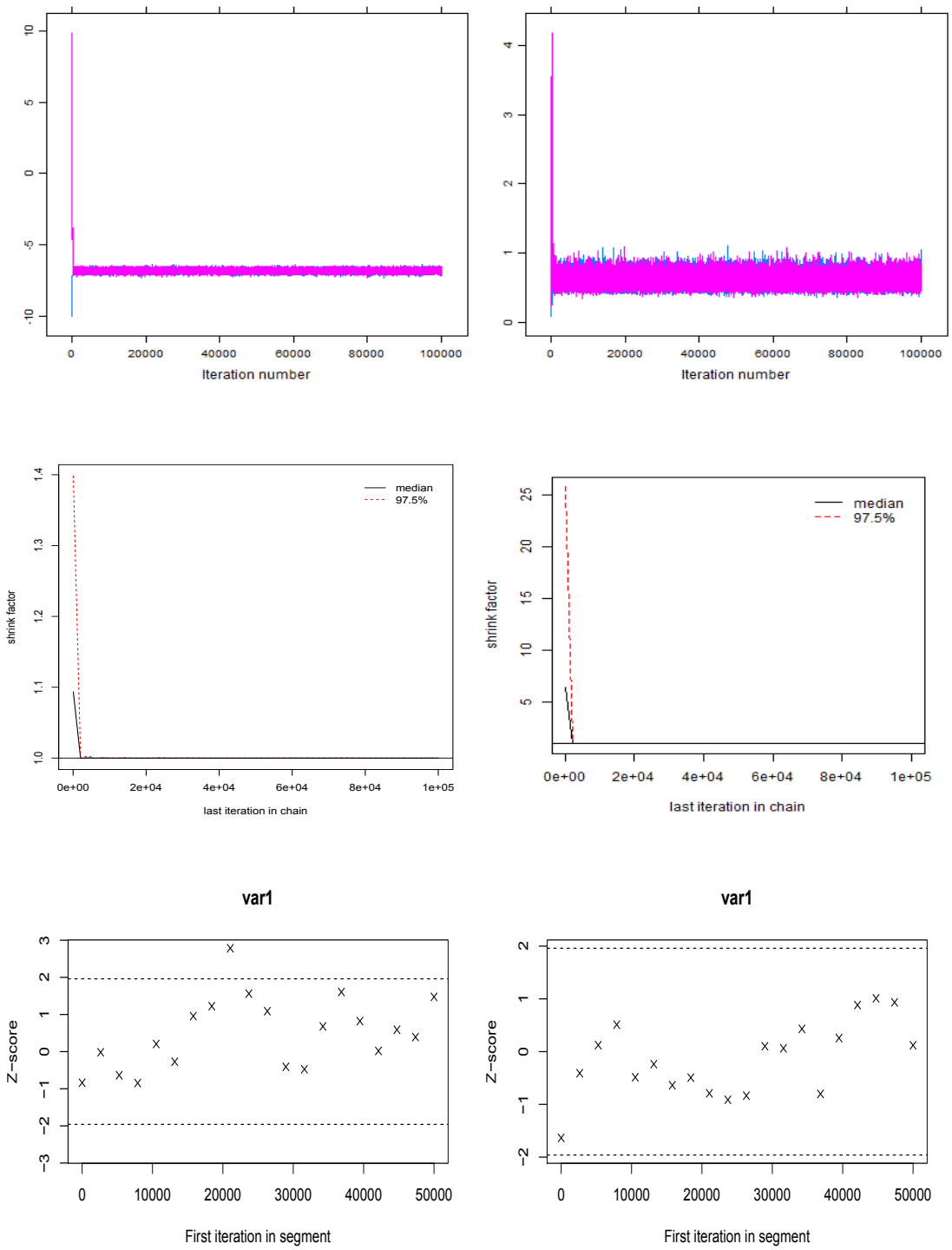


Figure 3.7: The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

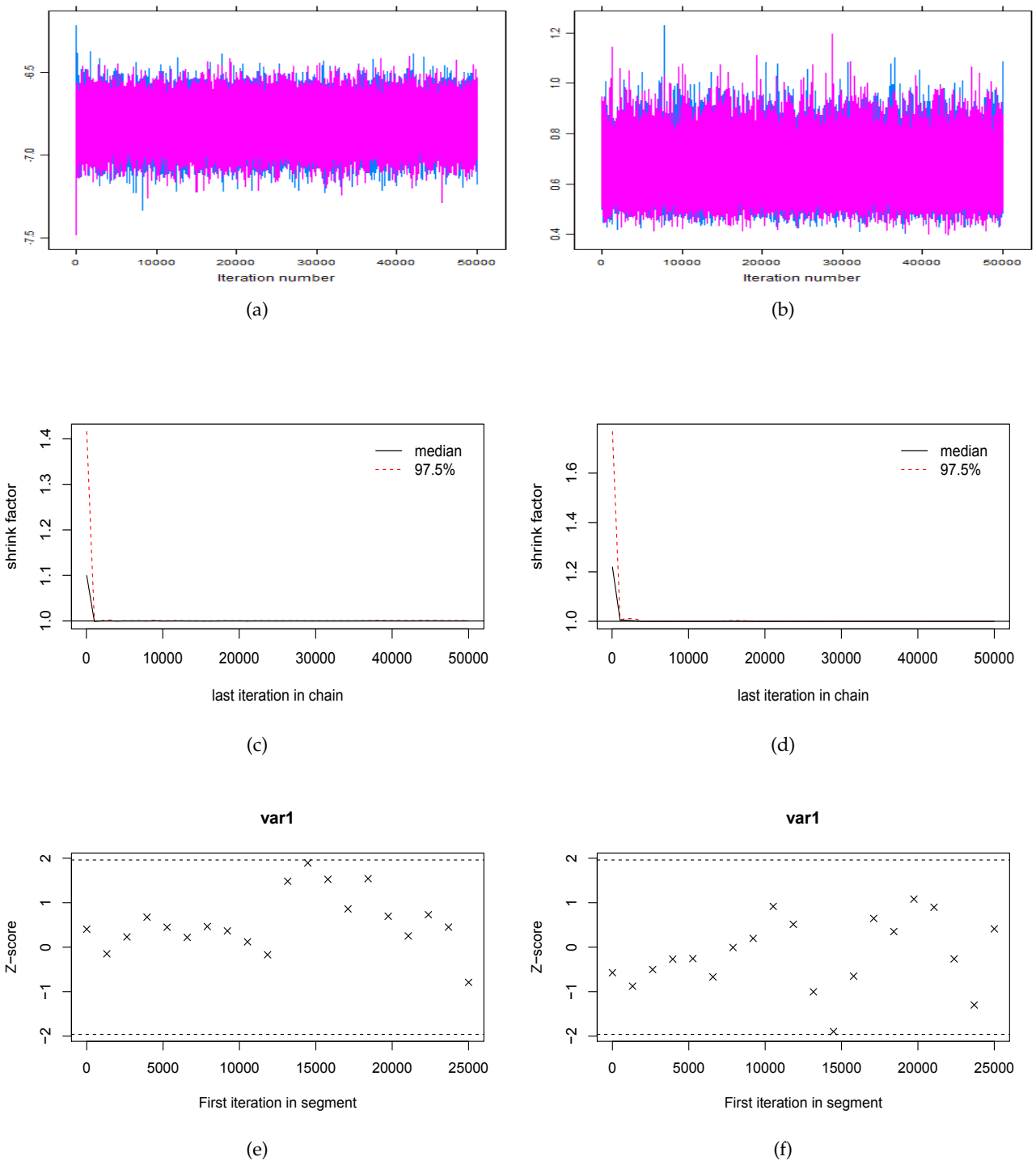


Figure 3.8: The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

3.7 Simulation Study

3.7.1 Simulation Design

In this section, a simulation study is conducted to assess the performance of the models described in sections 3.2-3.5. The design of this simulation study includes three steps. Firstly, the true values of model parameters α and τ should be specified, where α is the overall log-intensity of accidents per meter and τ is the between-motorway standard deviation. For this simulation, six scenarios are considered for the true values of parameters α and τ. The true values of α are taken to be -7 and -1. If the overall log-intensity is chosen to be $\alpha < -9$, then the number of accidents on the motorway will be equal to zero. The between motorways standard deviation τ is set to be 0.3, 0.8 and 1.5 to reflect the variation between motorways. A magnitude of 0.3 would indicate that there is not much variation in the motorway specific log-intensity while a magnitude of 1.5 would result in much more variation between motorways. These true values of parameters α and τ are chosen to be close to the results for the observed data set. Secondly, the log-intensity α_i on motorway i ($i = 1, ..., m$) is drawn from a normal distribution with mean α and standard deviation τ. Thirdly, the data set which represents the number of accidents n_i ($i = 1, ..., m$) on the motorway i is generated from a Poisson distribution with mean $L_i \exp(\alpha_i)$, where L_i is the length of the motorway i. The second and third steps were repeated 1,000 times for each scenario.

The performance and precision of the simulation are measured by comparing the simulated results with the true values that were used to produce the simulated data (Burton et al., 2006). Because of the potential variation of results across criteria, three performance criteria are tested: bias, *mean square error* (MSE) and *coverage probability* (CP) of parameter estimates (Collins et al., 2001). Here, an explanation of criteria is briefly presented. The average of the estimates over all simulations is utilized to calculate the bias in the parameter estimate which represents the difference between the average of the estimates (mean) over all simulations and the true value of parameter used to produce simulated data (Collins et al., 2001). The second criterion, the *mean square error* of the parameter estimate is a useful tool to measure the overall accuracy and it is equal to the squared bias of estimate plus its variance (Collins et al., 2001); $MSE = \text{Bias}\,(\hat{\alpha}, \alpha)^2 + \text{Var}\,(\hat{\alpha})$. The *coverage probability* is the percentage of 95% credible intervals that contain the true value of parameter (Burton et al., 2006). The *coverage probability* should be close to 95% (Kontopantelis and Reeves, 2012).

3.7.2 Simulation Results

This section describes and discuss simulation results. Tables 3.4, 3.6, 3.8, 3.10 and 3.12 show that for scenarios with true value of $\alpha = -7$, the performance of means for each of the overall log-intensity and of the between motorway standard deviation for the fully Bayesian hierarchical model is better than those for the other models across all prior distributions. For scenarios with true value $\alpha = -1$, means values for each of the three hierarchical models are similar and accurate across all prior distributions except the uniform prior distribution within range (0, 100). For this prior with true value $\tau = 1.5$, the mean of τ obtained for fully Bayesian hierarchical model is better than those obtained for semi-Bayesian and frequentist hierarchical models. Findings for Bayesian and frequentist non-hierarchical models are similar for all scenarios and the mean of the log-intensity is poor in some scenarios. For example, for scenario with $\alpha = -1$ and $\tau = 1.5$ the mean of the overall log-intensity α is very inaccurate and it is not close to the true value of α, also for scenario with true values of $\alpha = -1$ and $\tau = 0.8$ the mean of α is far from the true value of α. Means of α and τ indicate that the fully Bayesian approach performs slightly better than the semi-Bayesian and frequentist approaches. The performance of non-hierarchical models is poor compared with the performance of hierarchical models because of ignoring heterogeneity in non-hierarchy structure.

For the scenario with true value $\alpha = -1$, the bias in α for the fully Bayesian hierarchical model is similar to those for semi-Bayesian hierarchical model across all prior distributions. In Tables 3.3 and 3.5, the bias in τ for the frquentist hierarchical model is slightly larger than the bias obtained for the fully Bayesian and semi-Bayesian hierarchical models for the prior distributions of τ^2, Inv-Gamma(0,001, 0.001) and Inv-Gamma(0.1, 0.1). Tables 3.7 and 3.9 show that the bias in τ obtained for the fully Bayesian and semi-Bayesian models using unif$(0, 10^2)$ and HN$(0, 0.02)$ as prior distributions of τ is larger than the bias obtained for the same models but using Inv-Gamma$(0.001, 0.001)$ and Inv-Gamma$(0.1, 0.1)$ as prior distributions of τ^2. This indicates a sensitivity of τ to the prior specifications.

For the true value of $\alpha = -7$, Tables 3.4, 3.6, 3.8 and 3.10 show that estimates of α and τ for the fully Bayesian hierarchical model have less bias compared with those for other hierarchical models used. With regards to the non-hierarchical models, the magnitude of the bias in α for both non-hierarchical models is larger than the magnitude of the bias in α for all hierarchical models, as can be seen in Tables 3.11 and 3.12.

Tables 3.3-3.10 show that the MSE of α and τ for the three hierarchical models are similar. The MSE of the non-hierarchical models for α is larger than the MSE obtained for the hierarchical models as can be seen in Tables 3.11 and 3.12.

A 95% credible interval is calculated for each simulated data set for each scenario, then the coverage probability is estimated. For the true value $\alpha = -1$, the performance of the coverage probability of α and τ for the hierarchical models is similar, where in some cases coverage probabilities are just slightly below 0.95. In this scenario there is no general direction of the coverage probability of both parameters α and τ through all used prior distributions. However, the coverage probability of α for the hierarchical models across all the prior distributions used, increases as the heterogeneity between motorways increases.

For scenario with true value $\alpha = -7$, the coverage probability of both parameters α and τ for the fully Bayesian hierarchical model is better than those of the semi-Bayesian and frequentist hierarchical models and it is closest to 0.95, as can be seen in Tables 3.4, 3.6, 3.8 and 3.10. In addition, across prior distributions Inv-Gamma$(0.1, 0.1)$, unif$(0, 10^2)$ and $HN(0, 0.02)$, the fully Bayesian hierarchical model gave coverage that does not exceed 0.96 or drop under 0.94.

Tables 3.11 and 3.12 show that the performance of the parameter α for the non-hierarchical model is very poor and is not comparable with those for the hierarchical models. In general, the fully Bayesian hierarchical model performed better than other hierarchical models in terms of bias and coverage. The performance of the mean and MSE for the fully Bayesian hierarchical model was similar to other hierarchical models. Finally, the performance of the non-hierarchical models was extremely poor in terms of mean, bias, MSE and coverage probability.

Regarding the computing time, for all scenarios and across all prior distributions, Model 1 took from 11593 seconds to 18516 seconds; Model 2 took from 806 seconds to 1469 seconds and Model 3 took from 34 seconds to 44 seconds. Model 4 took from 317 seconds to 325 seconds, while Model 5 took few seconds. Thus, it is clear that Model 1 takes the longest computing time.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	time
0.3	α	-1.00	0.0018	0.0018	94.7%	11747	-1.00	0.0019	0.0018	94.4%	841	-1.00	0.0018	0.0018	94.1%	42
	τ	0.30	0.0008	0.0009	95.4%		0.30	0.0008	0.0009	95.4%		0.30	-0.0044	0.0009	95.6%	
0.8	α	-0.99	0.0071	0.0134	95.3%	11679	-0.99	0.0073	0.0134	95.6%	806	-0.99	0.0072	0.0134	94.2%	40
	τ	0.80	0.0041	0.0066	95.7%		0.80	0.0039	0.0066	96.1%		0.79	-0.0098	0.0065	96.0%	
	α	-0.99	0.0077	0.0426	96.4%	18516	-0.99	0.0079	0.0427	96.7%	827	-0.99	0.0075	0.0426	95.9%	35
1.5	τ	1.50	0.0019	0.0238	95.5%		1.50	0.0018	0.0237	95.6%		1.49	-0.0238	0.0234	95.4%	

Table 3.3: Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-7.00	0.0033	0.0028	94.8%	11667	-6.97	0.0264	0.0034	92.7%	909	-6.97	0.0268	0.0034	90.4%	42
	τ	0.30	-0.0037	0.0021	93.8%		0.29	-0.0097	0.0020	94.1%		0.29	-0.0150	0.0021	94.7%	
0.8	α	-6.98	0.0194	0.0142	94.7%	11664	-6.94	0.0592	0.0166	91.6%	939	-6.94	0.0595	0.0166	90.5%	39
	τ	0.79	-0.0104	0.0083	95.2%		0.79	-0.0331	0.0088	94.8%		0.75	-0.0465	0.0096	94.9%	
1.5	α	-6.94	0.0558	0.0478	95.0%	11609	-6.88	0.1170	0.0554	91.1%	955	-6.88	0.1157	0.0551	89.5%	37
	τ	1.45	-0.0530	0.0270	94.7%		1.40	-0.1036	0.0338	92.3%		1.37	-0.1277	0.0386	93.8%	

Table 3.4: Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-1.00	0.0018	0.0018	95.0%	11660	-1.00	0.0018	0.0018	95.3%	814	-1.00	0.0018	0.0018	94.1%	43
	τ	0.31	0.0072	0.0009	95.5%		0.31	0.0072	0.0009	94.7%		0.30	-0.0044	0.0009	95.6%	
0.8	α	-0.99	0.0073	0.0134	95.2%	11593	-0.99	0.0072	0.0134	95.3%	829	-0.99	0.0072	0.0134	94.2%	40
	τ	0.81	0.0051	0.0066	96.0%		0.81	0.0050	0.0066	96.0%		0.79	-0.0098	0.0065	96.0%	
1.5	α	-0.99	0.0080	0.0426	96.1%	11903	-0.99	0.0078	0.0427	96.3%	842	-0.99	0.0075	0.0426	95.9%	44
	τ	1.50	0.0005	0.0235	95.4%		1.50	0.0004	0.0235	95.5%		1.48	-0.0238	0.0234	95.4%	

Table 3.5: Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-7.00	0.0020	0.0028	95.6%	11896	-6.98	0.0251	0.0033	93.7%	916	-6.97	0.0268	0.0034	90.4%	43
	τ	0.31	0.0097	0.0018	95.5%		0.30	0.0041	0.0016	96.3%		0.29	-0.0150	0.0021	94.7%	
0.8	α	-6.98	0.0193	0.0141	95.0%	11691	-6.94	0.0590	0.0166	91.2%	904	-6.94	0.0595	0.0166	90.5%	39
	τ	0.79	-0.0090	0.0082	95.6%		0.77	-0.0315	0.0086	95.1%		0.75	-0.0465	0.0096	94.9%	
1.5	α	-6.94	0.0559	0.0478	94.9%	11683	-6.88	0.1171	0.0555	90.5%	904	-6.88	0.1157	0.0551	89.5%	37
	τ	1.45	-0.0546	0.0272	94.7%		1.40	-0.1050	0.0338	92.2%		1.40	-0.1277	0.0386	93.8%	

Table 3.6: Simulation results under prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-1.00	0.0017	0.0018	94.8%	12492	-1.00	0.0019	0.0018	95.5%	1301	-1.00	0.0018	0.0018	94.1%	40
	τ	0.31	0.0050	0.0010	95.0%		0.31	0.0078	0.0010	94.2%		0.30	-0.0044	0.0009	95.6%	
0.8	α	-0.99	0.0073	0.0134	95.5%	12707	-0.99	0.0072	0.0134	95.7%	1286	-0.99	0.0072	0.0134	94.2%	42
	τ	0.81	0.0131	0.0069	95.6%		0.83	0.0265	0.0077	94.9%		0.79	-0.0098	0.0065	96.0%	
1.5	α	-0.99	0.0077	0.0426	96.6%	12580	-0.99	0.0077	0.0426	96.6%	1329	-0.99	0.0075	0.0426	95.9%	39
	τ	1.52	0.0181	0.0246	95.2%		1.54	0.0380	0.0257	95.1%		1.48	-0.0238	0.0234	95.4%	

Table 3.7: Simulation results under prior distributions $\tau \sim \text{unif}(0, 10^2)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-7.00	0.0025	0.0028	95.8%	12486	-6.98	0.0251	0.0033	93.3%	1364	-6.97	0.0268	0.0034	90.4%	43
	τ	0.30	0.0044	0.0021	94.5%		0.30	0.0041	0.0020	94.9%		0.29	-0.015	0.0021	94.7%	
0.8	α	-6.98	0.0184	0.0141	95.5%	12432	-6.94	0.0566	0.0163	92.5%	1469	-6.94	0.0595	0.0166	90.5%	38
	τ	0.80	0.0007	0.0083	95.4%		0.79	-0.0060	0.0083	96.4%		0.75	-0.0465	0.0096	94.9%	
1.5	α	-6.95	0.0548	0.0478	94.9%	12465	-6.89	0.1134	0.0546	91.6%	1448	-6.88	0.1157	0.0551	89.5%	37
	τ	1.47	-0.0354	0.0261	95.3%		1.44	-0.0635	0.0275	94.5%		1.37	-0.1277	0.0386	93.8%	

Table 3.8: Simulation results under prior distributions $\tau \sim \text{unif}(0, 10^2)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-1.00	0.0017	0.0018	94.9%	17933	-1.00	0.0019	0.0018	95.6%	1005	-1.00	0.0018	0.0018	94.1%	40
	τ	0.31	0.0050	0.0010	94.9%		0.31	0.0049	0.0010	95.1%		0.30	-0.0044	0.0009	95.6 %	
0.8	α	-0.99	0.0071	0.0134	95.3%	17724	-0.99	0.0073	0.0134	95.4%	993	-0.99	0.0072	0.0134	94.2%	37
	τ	0.81	0.0129	0.0069	95.4%		0.81	0.0127	0.0069	95.6%		0.79	-0.0098	0.0065	96.0%	
1.5	α	-0.99	0.0076	0.0426	96.7%	17767	-0.99	0.0080	0.0427	96.6%	963	-0.99	0.0075	0.0426	95.9%	34
	τ	1.52	0.0179	0.0245	95.5%		1.52	0.0175	0.0245	95.3%		1.48	-0.0238	0.0234	95.4%	

Table 3.9: Simulation results under prior distributions $\tau \sim \text{HN}(0, 0.02)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 1					Model 2					Model 3				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-7.00	0.0026	0.0028	95.5%	18073	-6.97	0.0256	0.0034	93.1%	1027	-6.97	0.0268	0.0034	90.4%	40
	τ	0.31	0.0046	0.0021	94.7%		0.30	-0.0018	0.0020	94.2%		0.29	-0.0150	0.0021	94.7%	
0.8	α	-6.98	0.0184	0.0141	95.2%	17807	-6.94	0.0582	0.0165	91.8%	1025	-6.94	0.0595	0.0166	90.5%	37
	τ	0.80	0.0006	0.0083	95.4%		0.78	-0.0223	0.0083	95.5%		0.75	-0.0465	0.0096	94.9%	
1.5	α	-6.95	0.0546	0.0477	94.5%	17828	-6.88	0.1156	0.0551	91.2%	1057	-6.88	0.1157	0.0551	89.5%	36
	τ	1.47	-0.0355	0.0261	95.2%		1.41	-0.0866	0.0310	93.2%		1.37	-0.1277	0.0386	93.8%	

Table 3.10: Simulation results under prior distributions $\tau \sim \text{HN}(0, 0.02)$ and $\alpha \sim \text{N}(0, 10^2)$ and with true value of $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 4					Model 5				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-0.96	0.0444	0.0075	1.5%	319	-0.96	0.0444	0.0075	1.5%	3
0.8	α	-0.68	0.3235	0.15	0.4%	323	-0.68	0.3235	0.15	0.4%	3
1.5	α	0.02	1.0235	1.2569	0%	325	0.02	1.0235	1.2569	0%	2

Table 3.11: Simulation results using Bayesian method (Model 4) under prior distribution $\alpha \sim \text{N}(0, 10^2)$ and maximum likelihood method (Model 5) with true value $\alpha = -1$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

		Model 4					Model 5				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-6.96	0.042	0.0071	34.1%	319	-6.96	0.042	0.0071	34.1%	5
0.8	α	-6.69	0.307	0.1362	4.6%	317	-6.69	0.307	0.1362	4.6%	4
1.5	α	-6.01	0.9887	1.1828	0.2%	321	-6.01	0.9887	1.1828	0.2%	4

Table 3.12: Simulation results using Bayesian method (Model 4) under prior distribution $\alpha \sim \text{N}(0, 10^2)$ and maximum likelihood method (Model 5) with true value $\alpha = -7$ where time is recorded in seconds. Note: MSE represents mean square error and CP represents the coverage probability.

3.8 Discussion

In this chapter, five models are proposed for the traffic accident data on the UK motorway network involving 49 motorways. These models are one-stage fully Bayesian hierarchical model (Model 1), two-stage semi-Bayesian hierarchical model (Model 2), two-stage frequentist hierarchical model (Model 3), non-hierarchical Bayesian model (Model 4) and non-hierarchical frequentist model (Model 5). The use of the hierarchical model allows the following parameters to be estimated: the intensity of accidents for each motorway, the overall intensity of accidents on the whole motorway network and the between-motorway standard deviation. For the fully Bayesian and semi-Bayesian models, Model 1 and Model 2, various prior distributions have been specified for the between-motorway variance to investigate the sensitivity for the choice of the prior. These prior distributions are either non-informative or weakly-informative. The Metropolis Hastings within Gibbs sampler is introduced to estimate the Model 1 and Model 2 parameters, and two sampling algorithms have been provided. The performance of the proposed models has been assessed and compared using a real application and simulated data. The real application includes fitting the five described models to traffic accident data for the 49 motorways in the UK for 2016. Results of the real application have been summarised in the form of the posterior mean/ point estimate, standard deviation/standard error and 95% credible/confidence interval for the log-intensity of accidents for each motorway, the overall log-intensity of accidents and the between-motorway standard deviation. In the hierarchical models context, from the forest plots of the posterior mean and 95% credible interval for each motorway, it is clear that the motorway M25 has the highest log-intensity of accidents (see Figure 3.1 for more details). The parameters estimates $\hat{\alpha}_i,\ i = 1, ..., m$ that were obtained using models 2 and 3 were not exactly consistent with those obtained using Model 1. For example, the forest plot produced from using the frequentist model illustrates that the M180 has the second lowest log-intensity of accidents, whereas the forest plot of Model 1 shows that M180 has the third lowest log-intensity of accidents (see Figure 3.1 for more details). As for the non-hierarchical models, their performance was slightly different. In general, the hierarchical models indicate a much better performance compared with the non-hierarchical models. This is because the hierarchical structure takes into account the heterogeneity in the intensity of accidents across the motorway network, while the non-hierarchy models ignore this. In addition, a simulation study was conducted to assess the performance of all the models. In this simulation study, two true values of the log-intensity α of accidents and three different true values of the heterogeneity between motorway τ were chosen in order to ensure the simulation robustness. These values are $\alpha = -7$ and -1 and $\tau = 0.3, 0.8$ and 1.5 that lead to 6 scenarios of simulated data. The simulation study compares the different specific models using four criteria

that include the average of the posterior mean/point estimate over all simulations where this average represents a parameter estimator $\hat{\alpha}$. The second criterion is the bias in the parameter estimator. The third criterion is the mean square error of the parameter estimator. The last criterion is the coverage probability (CP) which is the actual probability that the 95% credible interval contains the true value of the parameter. The results of the simulation study show that for some scenarios Model 1 gave accurate mean estimates for α and τ that are less biased compared with those obtained from Models 2 and 3. The results of the simulation study also show that Models 2 and 3 have a lower coverage probability compared to Model 1 for all the parameters when the data are simulated with a large true value of the intensity and the between-motorway standard deviation. The performance of Model 1 is better than Model 2 and Model 3 for data with a high intensity of accidents. However, all three models give similar results for MSE and the mean for all the parameters. It can be seen that the Bayesian hierarchical model has a smaller bias, a lower MSE and a better coverage probability compared to a frequentist hierarchical model. The results of the simulation study show that the performance of non-hierarchical models is not comparable with that of hierarchical models since the bias and MSE of the log-intensity of accidents are quite large. In addition, the coverage probability is less than the accepted ratios and at times it is close to zero.

Based on results for simulated data sets, the hierarchical models (Models 1, 2 and 3) perform better than the non-hierarchical models (Models 4 and 5). The performance of the hierarchical models is good in terms of the point estimate, bias of point estimate, MSE and actual coverage of interval estimates (see Tables 3.3-3.10 for more details), but the performance of the non-hierarchical models was poor with respect to these evaluation criteria (see Tables 3.11 and 3.12 for more details). In the Bayesian methods context, hierarchical Bayesian approaches (Models 1 and 2) perform well and they have produced accurate results compared with the non-hierarchical Bayesian model (Model 4) when estimating the log-intensity of accidents α. The advantage of using Models 1 and 2 is represented by the decreasing bias and MSE of the overall log-intensity of accidents estimator of α and acceptable levels of the coverage probability. The non-hierarchical Bayesian method faced difficulty in attaining the required level of actual coverage. In addition, this model produced a biased estimator of the overall log-intensity of accidents with large MSE. Practically, when the interest is estimation and the number of observations is large then there seems not to be a great difference in selecting hierarchical over non-hierarchical models (Farrell and Ludwig, 2008). In the context of frequentist methods, the non-hierarchical frequentist model performed poorly in terms of the bias, MSE and actual coverage (see Tables 3.11-3.12 for more details). Thus, when comparing Bayesian and frequentist approaches, the Bayesian hierarchical models appeared better than the frequentist hierarchical model in terms

of the coverage probability and bias (see Table 4.1, 3.6 and 3.10). From the above discussion, we can conclude that the best models for the UK motorway accidents data are the Bayesian hierarchical models (models 1 and 2).

The current findings are noteworthy and enable the highlighting of dangerous motorways in the UK network, in terms of intensity of traffic accidents. The findings from Model 1 analysis suggest that motorways M25 and M27 have the highest accident intensity on the UK motorway network for 2016. The estimated intensity values are 3.59 and 3.03 per one kilometer of M25 and M27. Findings also revealed that motorways M54, M180, M74, M50 appear to have the lowest intensity of accidents with estimated values of intensity 4.4, 4.4, 4.2, 3.3 per 10 kilometers. Results from the analysis of models 2 and 3, however, indicate that the expected numbers of accidents, $\hat{\lambda}$, are 2.89, 3.98, 4.06 and 4.10 per 10 kilometers of M50, M180, M74 and M54. Estimated intensities for M25 and M27 are 3.59 and 3.09 per one kilometer.

Chapter 4

Three-Level Hierarchical Models

4.1 Introduction

One the most common approaches to study crash data is the crash prediction model. Classical prediction models for crashes (e.g. generalized linear model) do not take into account a multilevel structure of data leading to a limitation of models (Huang and Abdel-Aty, 2010). According to Huang and Abdel-Aty (2010), this limitation is in the estimated method, each crash or vehicle involvement harmonizes with individual situation resulting in independent residuals, but the assumption of residuals independence may be invalid. This is due to traffic data collection and the clustering process leading to a multilevel structure of data. Consequently, it may produce inaccurate estimates of model parameters and statistical inference. Hierarchical models are used to address the multilevel data structure. Actually, hierarchical modelling has been employed in many research fields such as sociology, education, political science, and public health, but the first employment of hierarchical models in a traffic safety field was by Shankar et al. (1998). They show that the explanatory power of crash models had been improved when site-specific random effects and time indicators were incorporated into the negative binomial regression model. Jones and Jørgensen (2003) explained and discussed possible applications of hierarchical models in road traffic accidents in Norway. Researchers have shown an increased interest in the hierarchical modelling approach to account for the multilevel data structure in crash prediction (Huang and Abdel-Aty, 2010). Some researchers used hierarchical models to predict crash frequency, whereas other researchers presented hierarchical modelling to recognize factors affecting crash severity.

In the previous chapter, the two-level hierarchical model has been proposed to analyse traffic accident data on the UK motorway network. In this model, a heterogeneity across motorways was only considered. This means, the intensity of accidents was considered constant within each motorway, but

varied among motorways. However, each motorway consists of junctions that are joined by grouped segments. These grouped segments are called links. Each link consists of start and end points. Each link has a uniquely referenced *Count Point* (*CP*) that is called "Mark". Some heterogeneities across-link may exist due to multilevel data structure. That is, the intensity of accidents may be inhomogeneous across-link and homogeneous within-link. Ignoring these heterogeneities adds a variance to the accident data and the case over-dispersion. Thus, without convenient methods to calculate the cross-grouped segments, heterogeneities may produce the underestimated estimates of the standard error in the intensity of accidents. The heterogeneity across-link can be taken into account using the three-level hierarchical model.

Bayesian inference via Markov Chain Monte Carlo methods are used to estimate the unknown parameters in the proposed model. We conducted a sensitivity analysis to different priors choices. The frequentist approach is also used for estimating model parameters. The proposed model was fitted to accident data on the UK motorway network for 2016. We evaluated the performance of the proposed models using a simulation study. We compare between the two-level Bayesian hierarchical model described in the previous chapter and the three-level Bayesian hierarchical model using information criteria and simulation study.

4.2 Three-Level Hierarchical Model

4.2.1 Model Definition

The number of accidents in each grouped segments is assumed as a homogeneous process and is assumed a non-homogeneous process across segments. Let m denote the total number of motorways and $n_i\,(i = 1, ..., m)$ the number of grouped segments for each motorway i. Suppose that the intensity of accidents per meter is $\lambda_{ij},\ i = 1, ..., m$ and $j = 1, ..., n_i$, where i is the index of motorway and j is the index of grouped segments. The number of accidents n_{ij} on each grouped segments follows a Poisson distribution with mean $\lambda_{ij}L_{ij}$, where L_{ij} represents the length (in meter) of the grouped segments j for motorway i. Let $\alpha_{ij} = \log \lambda_{ij}$ denote the log-intensity function. The three-level hierarchical model is written as follows,

$$
\begin{aligned}
n_{ij} &\sim \text{Poisson}\left(\lambda_{ij}L_{ij}\right), i = 1, ..., m; j = 1, .., n_i, \\
\alpha_{ij} &\sim \text{N}\left(\alpha_i, \tau_i^2\right), \\
\alpha_i &\sim \text{N}(\alpha, \tau^2).
\end{aligned} \tag{4.1}
$$

The second level includes the log-intensity of accidents, α_{ij} on each grouped segments and the log-intensity of accidents, α_i on each motorway as well as the between grouped segments heterogeneity, τ_i^2. The third level includes the overall log-intensity of accidents, α and the between motorway heterogeneity, τ^2. The intensity of accidents is constant on grouped segments that have the same mark, but it varies across grouped segments and motorways.

4.2.2 Likelihood Function

Let $\mathbf{N} = \left\{n_{ij},\ i = 1,..,m \text{ and } j = 1,...,n_i\right\}$ be the accident count. Let Θ denote model parameters $\left\{\alpha_{11},...,\alpha_{mn_m},\alpha_1,...,\alpha_m,\tau_1^2,...,\tau_m^2,\alpha,\tau^2\right\}$. Let $\boldsymbol{\gamma} = \{\alpha_{11},...,\alpha_{mn_m}\}$, $\boldsymbol{\alpha} = \{\alpha_1,...,\alpha_m\}$ and $\boldsymbol{\tau}^2 = \left\{\tau_1^2,...,\tau_m^2\right\}$. It is assumed that the accidents are uniformly distributed within each groped segments.
The likelihood function for the proposed model (4.1) is given by,

$$\begin{aligned} L(\mathbf{N}|\Theta) &= P(\mathbf{N}|\boldsymbol{\gamma}) \times P\left(\boldsymbol{\gamma}|\boldsymbol{\alpha},\boldsymbol{\tau}^2\right) \times P\left(\boldsymbol{\alpha}|\alpha,\tau^2\right) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{n_i} \exp\left(n_{ij}\alpha_{ij} - L_{ij}\exp\left(\alpha_{ij}\right)\right) \\ &\times \prod_{i=1}^{m}\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}}\ \exp\left(-\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}}\ \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right). \end{aligned} \tag{4.2}$$

For details about deriving the likelihood function see Appendix B.1.

4.2.3 Prior Distribution

The prior distribution reflects the available information about model parameters (Lesaffre and Lawson, 2012). In the Bayesian analysis of hierarchical models, the choice of the prior distribution for variance parameters is an important issue (Daniels, 1999). Different prior distributions were used for τ^2. A common choice of prior for variance parameters is an inverse gamma distribution and it is a conjugate distribution of the normal distribution (Gelman et al., 2013). An inverse gamma prior was assigned for τ^2 with shape α_0 and rate β_0 where parameters α_0 and β_0 of the prior distribution are hyper-parameters. The alternative priors are uniform prior $\text{unif}(0,a)$, $a > 0$ for τ and half-normal prior for τ with mean 0 and variance $\theta^2 = \frac{\pi}{2\sigma^2}$, $\sigma^2 > 0$.
For the τ_i^2 $(i = 1,...,m)$, we use Inv-Gamma(a_0,b_0) as prior distribution with shape a_0 and rate b_0. As the prior distribution for α, $\text{N}(\mu_0,\sigma_0^2)$ was used. A good choice of prior should minimize standard errors of the parameter estimates.

4.2.4 Posterior Distribution

Posterior if the prior distribution on τ^2 is an inverse gamma distribution

A joint posterior distribution for parameters Θ is given by,

$$\begin{aligned}\pi\left(\Theta|\mathbf{N}\right) \propto & \prod_{i=1}^{m}\prod_{j=1}^{n_i}\exp\left(n_{ij}\alpha_{ij}-L_{ij}\exp\left(\alpha_{ij}\right)\right)\times\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\tau_i^2}}\exp\left(-\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right)\\ & \times\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\left(\alpha_i-\alpha\right)^2}{2\tau^2}\right)\times\prod_{i=1}^{m}\frac{b_0^{a_0}}{\Gamma(a_0)}\left(\tau_i^2\right)^{-a_0-1}\exp\left(\frac{-b_0}{\tau_i^2}\right)\\ & \times\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(\frac{-\left(\alpha-\mu_0\right)^2}{2\sigma_0^2}\right)\times\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-\alpha_0-1}\exp\left(\frac{-\beta_0}{\tau^2}\right).\end{aligned} \tag{4.3}$$

For details about joint posterior distribution of parameters Θ see Appendix B.2. The conditional posterior distribution of α_{ij} is given by,

$$\pi\left(\alpha_{ij}|\boldsymbol{\alpha},\boldsymbol{\tau}^2,\mathbf{N}\right)\propto\exp\left(n_{ij}\alpha_{ij}-L_{ij}\exp\left(\alpha_{ij}\right)\right)\exp\left(-\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right). \tag{4.4}$$

The conditional posterior distribution of α_i is a normal distribution $\mathrm{N}(\mu_{\alpha_i},\sigma_{\alpha_i}^2)$ with mean and variance:

$$\mu_{\alpha_i}=\frac{\frac{\sum_{j=1}^{n_i}\alpha_{ij}}{\tau_i^2}+\frac{\alpha}{\tau^2}}{\frac{n_i}{\tau_i^2}+\frac{1}{\tau^2}}\text{ and }\sigma_{\alpha_i}^2=\frac{1}{\frac{n_i}{\tau_i^2}+\frac{1}{\tau^2}}. \tag{4.5}$$

For details about deriving the conditional posterior distribution of α_i see Appendix B.3.1. The conditional posterior distribution of τ_i^2 is given by,

$$\tau_i^2\sim\text{Inv-Gamma}\left(\frac{n_i}{2}+a_0,\sum_{j=1}^{n_i}\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2}+b_0\right). \tag{4.6}$$

For details about deriving the conditional posterior distribution of τ_i^2 see Appendix B.3.2.

The conditional posterior distribution of α is a $\mathrm{N}(\mu_\alpha,\sigma_\alpha^2)$ with mean and variance:

$$\mu_\alpha=\frac{\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}}\text{ and }\sigma_\alpha^2=\frac{1}{\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}}. \tag{4.7}$$

For details about deriving the conditional posterior distribution of α see Appendix B.3.3. The conditional posterior distribution of τ^2 is given by,

$$\tau^2 \sim \text{Inv-Gamma}\left(\frac{m}{2} + \alpha_0, \frac{\sum_{i=1}^{m}(\alpha_i - \alpha)^2}{2} + \beta_0\right), \tag{4.8}$$

which is an inverse gamma distribution. For details about deriving the conditional posterior distribution of τ^2 see Appendix B.3.4.

Posterior if the prior distribution on τ is a Uniform distribution

The probability density function of the uniform prior distribution on τ is constant, so it does not appear in the joint posterior distribution. Hence, the joint posterior distribution is given by,

$$\begin{aligned}\pi(\Theta|\mathbf{N}) \propto & \prod_{i=1}^{m}\prod_{j=1}^{n_i}\exp\left(n_{ij}\alpha_{ij} - L_{ij}\exp\left(\alpha_{ij}\right)\right) \times \prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\tau_i^2}}\exp\left(-\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right) \\ & \times \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right) \times \prod_{i=1}^{m}\frac{b_0^{a_0}}{\Gamma(a_0)}\left(\tau_i^2\right)^{-a_0-1}\exp\left(\frac{-b_0}{\tau_i^2}\right) \\ & \times \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(\frac{-(\alpha-\mu_0)^2}{2\sigma_0^2}\right).\end{aligned} \tag{4.9}$$

The conditional posterior distributions of α_{ij}, α_i, τ_i^2, $i = 1, ..., m$; $j = 1, ..., n_i$ and α are the same in equations (4.4)-(4.7). The conditional posterior distribution of τ is given by,

$$\pi(\tau|\boldsymbol{\alpha}, \alpha, \mathbf{N}) \propto \left(\frac{1}{\sqrt{2\pi\tau^2}}\right)^m \exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right). \tag{4.10}$$

Posterior if the prior distribution on τ is a half-normal distribution

The joint posterior distribution is given by,

$$\begin{aligned}\pi(\Theta|\mathbf{N}) \propto & \prod_{i=1}^{m}\prod_{j=1}^{n_i}\exp\left(n_{ij}\alpha_{ij} - L_{ij}\exp\left(\alpha_{ij}\right)\right) \\ & \times \prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\tau_i^2}}\exp\left(-\frac{\left(\alpha_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right) \times \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right) \\ & \times \prod_{i=1}^{m}\frac{b_0^{a_0}}{\Gamma(a_0)}\left(\tau_i^2\right)^{-a_0-1}\exp\left(\frac{-b_0}{\tau_i^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(\frac{-(\alpha-\mu_0)^2}{2\sigma_0^2}\right) \\ & \times \frac{2\theta}{\pi}\exp\left(-\frac{\tau^2\theta^2}{\pi}\right), \tau > 0.\end{aligned} \tag{4.11}$$

The conditional posterior distributions of α_{ij}, α_i, τ_i^2, $i = 1, ..., m; j = 1, ..., n_i$ and α are the same in equations (4.4)-(4.7). The conditional posterior distribution of τ is given by,

$$\pi(\tau|\boldsymbol{\alpha}, \alpha, \mathbf{N}) \propto \tau^{-m} \exp\left(-\sum_{i=1}^{m} \frac{(\alpha_i - \alpha)^2}{2\tau^2} - \frac{\tau^2\theta^2}{\pi}\right), \ \tau > 0. \tag{4.12}$$

4.3 Bayesian Estimation

In this section, Bayesian estimation of model (4.1) is performed using Metropolis-Hastings within Gibbs sampler. We generate random samples from conditional posterior distributions of α, α_i, τ_i^2, $i = 1, ..., m$ which are closed forms. Conditional posterior distributions of α_{ij}, $i = 1, ..., m; j = 1, ..., n_i$ are not closed forms. In this case, MCMC is used. Normal proposal distributions are specified for α_{ij}, $i = 1, ..., m; j = 1, ..., n_i$ with mean $\alpha_{ij}^{(t-1)}$ and variance σ_{ij}^2, where t ($t = 1, ..., M$) is iteration index. The variance σ_{ij}^2 is chosen such that an acceptance rate is within the range of $(0.24, 0.40)$ (Gelman et al., 1996). Suppose that $\mathrm{q}_1(.)$ denotes the proposal distribution. A value $\acute{\alpha}_{ij}$ generated from the proposal distribution $\mathrm{q}_1\left(., \alpha_{ij}^{(t-1)}\right)$ is accepted or rejected with probability

$$r_1(\alpha_{ij}^{(t-1)}, \acute{\alpha}_{ij}) = \min\left\{\frac{\pi\left(\acute{\alpha}_{ij}|\alpha_i^{(t-1)}, \tau_i^{2(t-1)}\right)\mathrm{q}_1\left(\alpha_{ij}^{(t-1)}, \acute{\alpha}_{ij}\right)}{\pi\left(\alpha_{ij}^{(t-1)}|\alpha_i^{(t-1)}, \tau_i^{2(t-1)}\right)\mathrm{q}_1\left(\acute{\alpha}_{ij}, \alpha_{ij}^{(t-1)}\right)}, 1\right\}, \tag{4.13}$$

where the conditional posterior distributions (4.10) and (4.12) of τ are not available as closed forms, we simulate τ using MCMC. This requires the specification of a proposal probability distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$ of $\acute{\tau}$. A normal distribution was chosen with mean equalling to current value $\tau^{(t-1)}$ and variance 0.09. A new value $\acute{\tau}$ is generated from the proposal distribution $\mathrm{q}_2\left(., \tau^{(t-1)}\right)$ with acceptance probability

$$r_2(\tau^{(t-1)}, \acute{\tau}) = \min\left\{\frac{\pi\left(\acute{\tau}|\alpha^{(t)}, \alpha_1^{(t)}, ..., \alpha_m^{(t)}\right)\mathrm{q}_2\left(\tau^{(t-1)}, \acute{\tau}\right)}{\pi\left(\tau^{(t-1)}|\alpha^{(t)}, \alpha_1^{(t)}, ..., \alpha_m^{(t)}\right)\mathrm{q}_2\left(\acute{\tau}, \tau^{(t-1)}\right)}, 1\right\}. \tag{4.14}$$

We present the Metropolis-Hastings within Gibbs algorithm (4.1) in a general formulation for sampling from the posterior distributions defined in section 4.2.4.

Algorithm 4.1 Sampling from the full conditional posterior distributions of parameters for the three-level Bayesian hierarchical model using Metropolis-Hastings within Gibbs sampling.

Set initial value, $\alpha_{ij}^{(0)}$, $\alpha_i^{(0)}$, $\tau_i^{2(0)}$, $\alpha^{(0)}$ and $\tau^{2(0)}$, $i = 1, ..., m,$ and $j = 1, ..., n_i$.
For each iteration t.

Step 1: update α_{ij} one by one.

1.1 Generate a proposed value, $\acute{\alpha}_{ij} \sim q_1(., \alpha_{ij}^{(t-1)})$.

1.2 Calculate the probability $r_1(\alpha_{ij}^{(t-1)}, \acute{\alpha}_{ij})$ specified in equation (4.13).

1.3 With probability $r_1(\alpha_{ij}^{(t-1)}, \acute{\alpha}_{ij})$, set $\alpha_{ij}^{(t)} = \acute{\alpha}_{ij}$, otherwise set $\alpha_{ij}^{(t)} = \alpha_{ij}^{(t-1)}$.

1.4 Repeat steps 1.1 to 1.3 for all $\alpha_{ij}, i = 1, ..., m$ and $j = 1, ..., n_i$.

Step 2: Update full conditional posterior density $\pi\left(\alpha_i^{(t)} | \alpha_{ij}^{(t)}, \tau_i^{2(t-1)}, \alpha^{(t-1)}, \tau^{2(t-1)}, \mathbf{N}\right)$, $i = 1, ..., m$ and $j = 1, ..., n_i$ specified in equation (4.5).

Step 3: Update full conditional posterior density $\pi\left(\tau_i^{2(t)} \mid \alpha_{ij}^{(t)}, \alpha_i^{(t)}, \mathbf{N}\right)$, $i = 1, ..., m$ and $j = 1, ..., n_i$ specified in equation (4.6).

Step 4: Update full conditional posterior density $\pi\left(\alpha^{(t)} | \alpha_i^{(t)}, \tau^{2(t-1)}, \mathbf{N}\right), i = 1, ..., \mathrm{m}$ specified in equation (4.7).

Step 5: Update τ.

5.1 Generate a proposed value, $\acute{\tau} \sim q_2(., \tau^{(t-1)})$.

5.2 Calculate the probability $r_2(\tau^{(t-1)}, \acute{\tau})$ specified in equation (4.14).

5.3 With probability $r_2(\tau^{(t-1)}, \acute{\tau})$, set $\tau^{(t)} = \acute{\tau}$, otherwise set $\tau^{(t)} = \tau^{(t-1)}$.

4.4 Frequentist Estimation

In this section, the maximum likelihood method is performed on three stages in order to estimate parameters of the hierarchical model (4.1).
In stage one, the log-intensity of accidents $\left(\alpha_{ij}\right)$ is estimated for each grouped segments where the first part of likelihood function in equation (4.2) is used

$$L_{ij}\left(\alpha_{ij};\mathbf{N}\right)=\exp\left(n_{ij}\alpha_{ij}-L_{ij}\exp\left(\alpha_{ij}\right)\right). \tag{4.15}$$

The log-likelihood function is

$$\ell_{ij}\left(\alpha_{ij};\mathbf{N}\right)=n_{ij}\alpha_{ij}-L_{ij}\exp\left(\alpha_{ij}\right). \tag{4.16}$$

The maximum likelihood estimate (M.L.E.) of α_{ij} is the value that maximises $\ell_{ij}\left(\alpha_{ij};\mathbf{N}\right)$. The M.L.E. can be found by differentiating ℓ_{ij} with respect to α_{ij} and equalling derivative of ℓ_{ij} to zero. Thus the point estimate of α_{ij} is given by,

$$\hat{\alpha}_{ij}=\log n_{ij}-\log L_{ij}. \tag{4.17}$$

To calculate the standard error of $\hat{\alpha}_{ij}$ we use the Fisher information matrix $I(\hat{\alpha}_{ij})$ that is a scalar containing the entry

$$I(\hat{\alpha}_{ij})=-E\left[H\left(\hat{\alpha}_{ij}\right)\right]=-E\left[\frac{\partial^2\ell_{ij}}{\partial\alpha_{ij}^2}\right]=L_{ij}\exp\left(\alpha_{ij}\right), \tag{4.18}$$

where $H\left(\hat{\alpha}_{ij}\right)$ represents the Hessian matrix. The square root of the inverse of the Fisher information scalar is an estimator of the standard error for $\hat{\alpha}_{ij}$

$$SE\left(\hat{\alpha}_{ij}\right)=\sqrt{I^{-1}(\hat{\alpha}_{ij})}=\left(L_{ij}\exp\left(\alpha_{ij}\right)\right)^{-\frac{1}{2}}. \tag{4.19}$$

In stage two, the log-intensity of accidents α_i ($i=1,...,m$) is estimated for each motorway. Here, the point estimates $\hat{\alpha}_{ij},\ i=1,...,m$ and $j=1,...,n_i$ are substituted in the likelihood function in equation (4.2). The relevant part of the likelihood function is given by,

$$L_i\left(\boldsymbol{\alpha},\boldsymbol{\tau}^2;\hat{\boldsymbol{\gamma}}\right)=\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\tau_i^2}}\ \exp\left(-\frac{\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right). \tag{4.20}$$

The log-likelihood function is

$$\ell_i\left(\boldsymbol{\alpha}, \boldsymbol{\tau}^2; \hat{\boldsymbol{\gamma}}\right) = \sum_{i=1}^{m}\left[-\frac{n_i}{2}\log\tau_i^2 - \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right]. \tag{4.21}$$

The maximum likelihood estimates of α_i and τ_i^2 are values that maximise $\ell_i\left(\boldsymbol{\alpha}, \boldsymbol{\tau}^2; \hat{\boldsymbol{\gamma}}\right)$. The maximum likelihood estimates $\hat{\alpha}_i$ and $\hat{\tau}_i^2$ can be found by partial derivative with respect to α_i and τ_i^2 respectively and equalling the partial derivative of ℓ_i to zero. Hence, the point estimates of α_i and τ_i^2 are given by,

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n_i}\hat{\alpha}_{ij}}{n_i} \text{ and } \hat{\tau}_i^2 = \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\hat{\alpha}_i\right)^2}{n_i}. \tag{4.22}$$

To obtain the standard errors of $\hat{\alpha}_i$ and $\hat{\tau}_i^2$, we use the Fisher information matrix that is (2×2) matrix containing entries

$$I\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) = -E\left[H\left(\hat{\alpha}_i, \hat{\tau}_i^2\right)\right] = -E\begin{bmatrix}\dfrac{\partial^2\ell_i}{\partial\alpha_i^2} & \dfrac{\partial^2\ell_i}{\partial\alpha_i\partial\tau_i^2}\\ \dfrac{\partial^2\ell_i}{\partial\tau_i^2\partial\alpha_i} & \dfrac{\partial^2\ell_i}{\partial(\tau_i^2)^2}\end{bmatrix} = \begin{bmatrix}\dfrac{n_i}{\tau_i^2} & 0\\ 0 & \dfrac{n_i}{2\tau_i^4}\end{bmatrix}, \tag{4.23}$$

where $H\left(\hat{\alpha}_i, \hat{\tau}_i^2\right)$ represents the Hessian matrix. For more details about obtaining the elements of the Hessian matrix and finding the elements of the Fisher information matrix see Appendix B.4. The inverse of the Fisher information matrix is given by,

$$I^{-1}\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) = \begin{bmatrix}\dfrac{\tau_i^2}{n_i} & 0\\ 0 & \dfrac{2\tau_i^4}{n_i}\end{bmatrix}. \tag{4.24}$$

The standard errors of $\hat{\alpha}_i$ and $\hat{\tau}_i^2$ are the square root of diagonal elements, and hence are given by,

$$SE\left(\hat{\alpha}_i\right) = \sqrt{\frac{\tau_i^2}{n_i}}. \tag{4.25}$$

$$SE\left(\hat{\tau}_i^2\right) = \sqrt{\frac{2\tau_i^4}{n_i}}. \tag{4.26}$$

In stage three, α and τ^2 are estimated. The relevant part of likelihood is

$$L(\alpha, \tau^2; \hat{\alpha}_i) = \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\ \exp\left(-\frac{(\hat{\alpha}_i-\alpha)^2}{2\tau^2}\right). \tag{4.27}$$

The estimated values $\hat{\alpha}_i$ and $SE(\hat{\alpha}_i)$ are used as data. The M.L.E. is obtained by **rma**.

4.5 Estimation Results for Motorway Data

Non-informative and weakly informative prior distributions

A non-informative prior distribution reflects the lack of prior information about parameter (Lesaffre and Lawson, 2012). In this case, the prior distribution is dominated by the likelihood function. As a result, the prior distribution has negligible to influence the posterior distribution of parameter. The choice of the non-informative prior requires selecting parameters of prior so that the variance of prior is large enough. A conjugate prior could be non-informative, or weakly-informative, such as Inv-Gamma$(0.001, 0.001)$ (non-informative prior) and Inv-Gamma$(0.1, 0.1)$ (weakly-informative prior) for τ^2. As a sensitivity analysis, we use the uniform prior unif$(0, 100)$ and a half-normal prior distribution HN$(0, 0.02)$ for τ, both are non-informative priors on τ (Thompson et al., 1997; Lambert et al., 2005). As for Inv-Gamma(a_0, b_0) prior on τ_i^2 $(i = 1, ..., m)$, we place $a_0 = b_0 = 0.001$. Finally, a non-informative normal prior distribution was used with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 100$ for α.

Informative prior distributions

An informative prior describes specific pre-existing information about parameter (Lesaffre and Lawson, 2012). Consequently, the prior distribution has impacts on the posterior distribution of parameter. The informative Inv-Gamma (α_0, β_0) prior is specified for τ^2 and $\mathrm{N}\left(\mu_0, \sigma_0^2\right)$ for α. This requires specifying parameters of priors. The maximum likelihood estimates of τ^2 and α and their standard errors of traffic accident data from an earlier year (e.g. 2015) will be used for specifying the informative priors in 2016 data. More specifically, the parameters for gamma prior are calculated from solving the following equation:

$$\begin{aligned} E\left(\tau^2\right) &= \frac{\alpha_0}{\beta_0} = \hat{\tau}^2_{ML} \\ \mathrm{var}\left(\tau^2\right) &= \frac{\alpha_0}{\beta_0^2} = SE(\hat{\tau}^2_{ML}), \end{aligned} \tag{4.28}$$

where $\hat{\tau}^2_{ML}$ is the maximum likelihood estimate, 0.3162 and $SE(\hat{\tau}^2_{ML})$ is the standard error of $\hat{\tau}^2_{ML}$, 0.0738, both are obtained from analysing the accident data in 2015. Solving the equations in (4.28), we obtain $\alpha_0 = 18.36$ and $\beta_0 = 58.06$. Thus, the informative prior for τ^2 is Inv-Gamma$(18.36, 58.06)$. Similarly, $\mu_0 = -6.65$ and $\sigma_0^2 = 0.09^2$.

Results

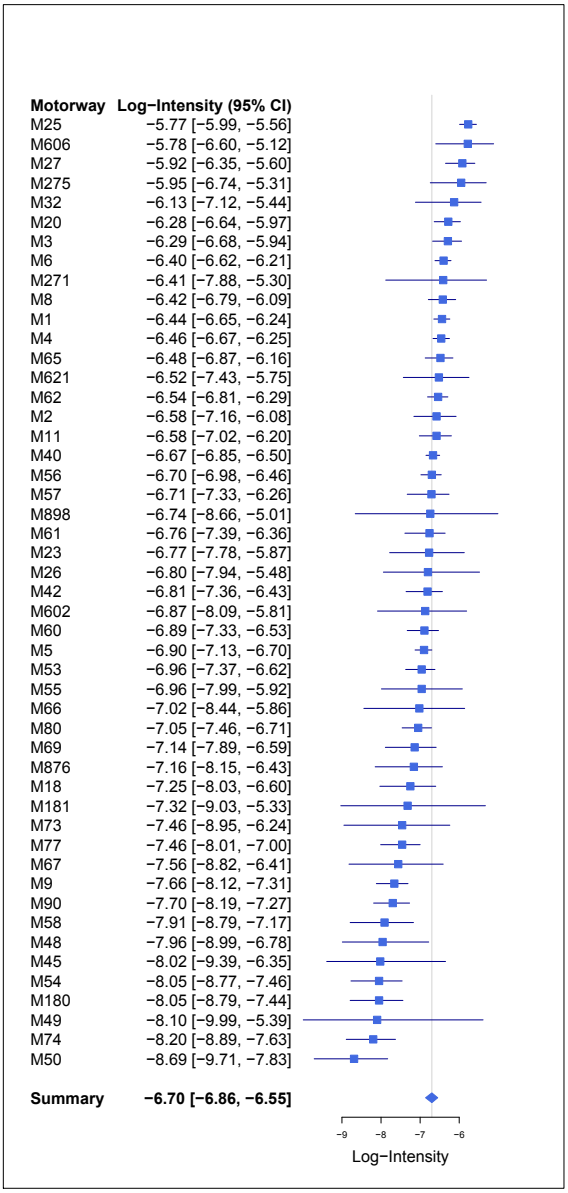
The three-level hierarchical model is used to analyse the observed accident data for 2016. The model parameters include the overall log-intensity of traffic accidents, α, on the UK motorways and the standard deviation between-motorway, τ. The MCMC is run for 500,000 iterations with discarding 50,000 iterations as a burn-in period and a thinning interval of 100.

Table 4.1 shows that in a Bayesian framework, the posterior mean, median, standard deviation and 95% credible interval for α and τ are similar between non-informative and weakly-informative priors of τ where informative prior distribution of τ is used, the length of the credible interval of α is shorter than the corresponding credible intervals of α based on the non and weakly-informative priors. In addition, the standard deviation for $\hat{\alpha}$ based on informative priors is smaller than that based on another priors. Table 4.1 displays the overall intensity of accidents (λ) per one kilometer. It is clear that results for λ from Bayesian methods are similar, except for the informative prior where λ is greater, but close to the estimate from the frequentist method.

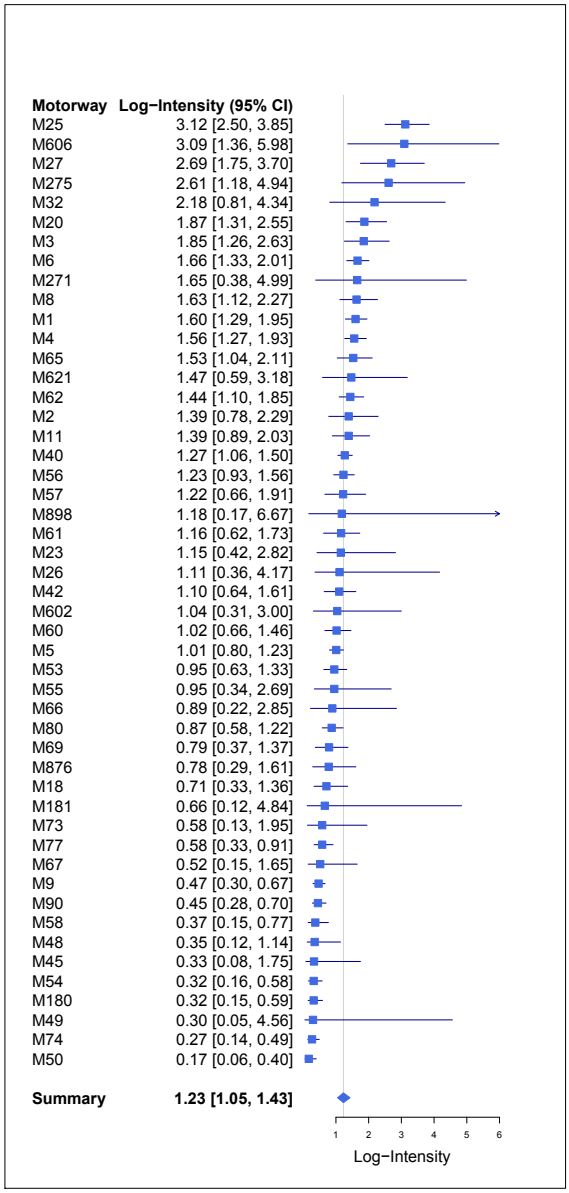
Methods							λ	
	Prior distribution	Parameter	Mean	Median	SD	95% CI	Mean	95% CI
Bayesian	$\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$	α	-6.93	-6.93	0.11	(-7.15, -6.71)	0.98	(0.79, 1.22)
		τ	0.65	0.65	0.09	(0.49, 0.85)		
	$\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$	α	-6.93	-6.93	0.11	(-7.15, -6.71)	0.98	(0.79, 1.22)
		τ	0.66	0.65	0.09	(0.50, 0.85)		
	$\tau \sim$ unif$(0, 100)$	α	-6.93	-6.93	0.11	(-7.16, -6.71)	0.98	(0.78, 1.22)
		τ	0.67	0.66	0.09	(0.51, 0.87)		
	$\tau \sim$ HN$(0, 0.02)$	α	-6.93	-6.93	0.11	(-7.16, -6.71)	0.98	(0.78, 1.22)
		τ	0.67	0.66	0.09	(0.51, 0.87)		
	$\alpha \sim$ N$(-6.65, 0.09^2)$	α	-6.70	-6.70	0.08	(-6.86, -6.55)	1.23	(1.05, 1.43)
	$\tau^2 \sim$ Inv-Gamma$(18.36, 58.06)$	τ	1.34	1.33	0.11	(1.15, 1.57)		
Frequentist		α	-6.64	-	0.08	(-6.80, -6.49)	1.31	(1.12, 1.53)
		τ	0.51	-	0.06	(0.41, 0.66)		

Table 4.1: Posterior summary and frequentist estimates of parameters α, τ and λ of traffic accidents for 2016 year. $\lambda = \exp(\alpha)$ is the intensity of accidents per one kilometer. The prior of α is N$(0, 100)$. SD: standard deviation and CI: credible interval or confidence interval. HN represents the half-normal distribution.

Figure 4.1 shows that the highest intensity of accidents is on M25 where the expected number of accidents (λ) is 3.12 per one kilometer. The second highest intensity is on M606 with $\lambda = 3.09$ per one kilometer. The third highest intensity is on M27 where it equals to 2.69 per one kilometer. However, M50, M74 and M49 have the lowest intensity of accidents such that the expected numbers of accidents are respectively 1.7, 2.7 and 3.0 per 10 kilometers. Some motorways have the similar value of the intensity of accidents such as M2 and M11 both with $\lambda = 1.39$ per one kilmeter, M53 and M55 both with $\lambda = 9.5$ per 10 kilometers, M73 and M77 both with $\lambda = 5.8$ per 10 kilometers and M54 and M180 both with $\lambda = 3.2$ per 10 kilometers.



(a) log-intensity of accidents α and α_i $(i = 1, ..., m)$



(b) intensity of accidents λ and λ_i $(i = 1, ..., m)$

Figure 4.1: Results from three-level hierarchical Bayesian model for accident data on the UK motorways for 2016 year. Prior distributions $\alpha \sim \text{N}(-6.65, 0.09^2)$ and $\tau^2 \sim \text{Inv-Gamma}(18.36, 58.06)$. Results include the posterior mean and the corresponding 95% credible interval for the log-intensity of accidents α_i on each motorway and the overall log-intensity of accidents α in (a) and the intensity of accidents $\lambda_i = 1000 \times \exp(\alpha_i)$ per one kilometer on each motorway and the overall intensity of accidents λ per one kilometer in (b). Square shapes represent posterior means of α_i, $i = 1, ..., m$ in (a) and posterior means of λ_i, $i = 1, ..., m$ in (b). The diamond shape is used to represent the posterior mean of the overall log -intensity of accident α in (a) and the posterior mean of the overall intensity of accident λ in (b). Horizontal lines denote 95% credible intervals and the sold vertical line represents the posterior mean of the overall log-intensity of accidents α in (a) and the posterior mean of the overall intensity λ in (b).

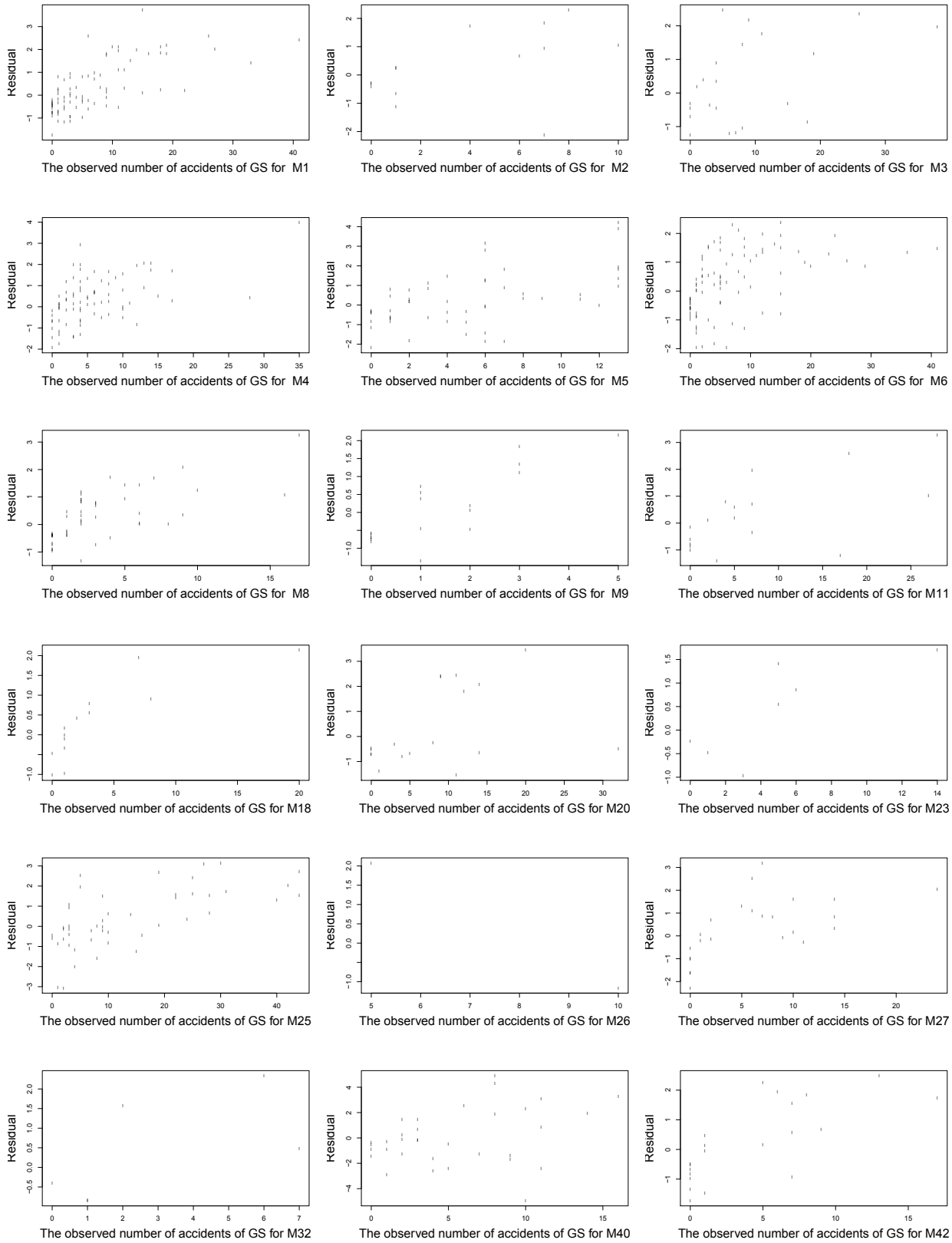


Figure 4.2: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments.

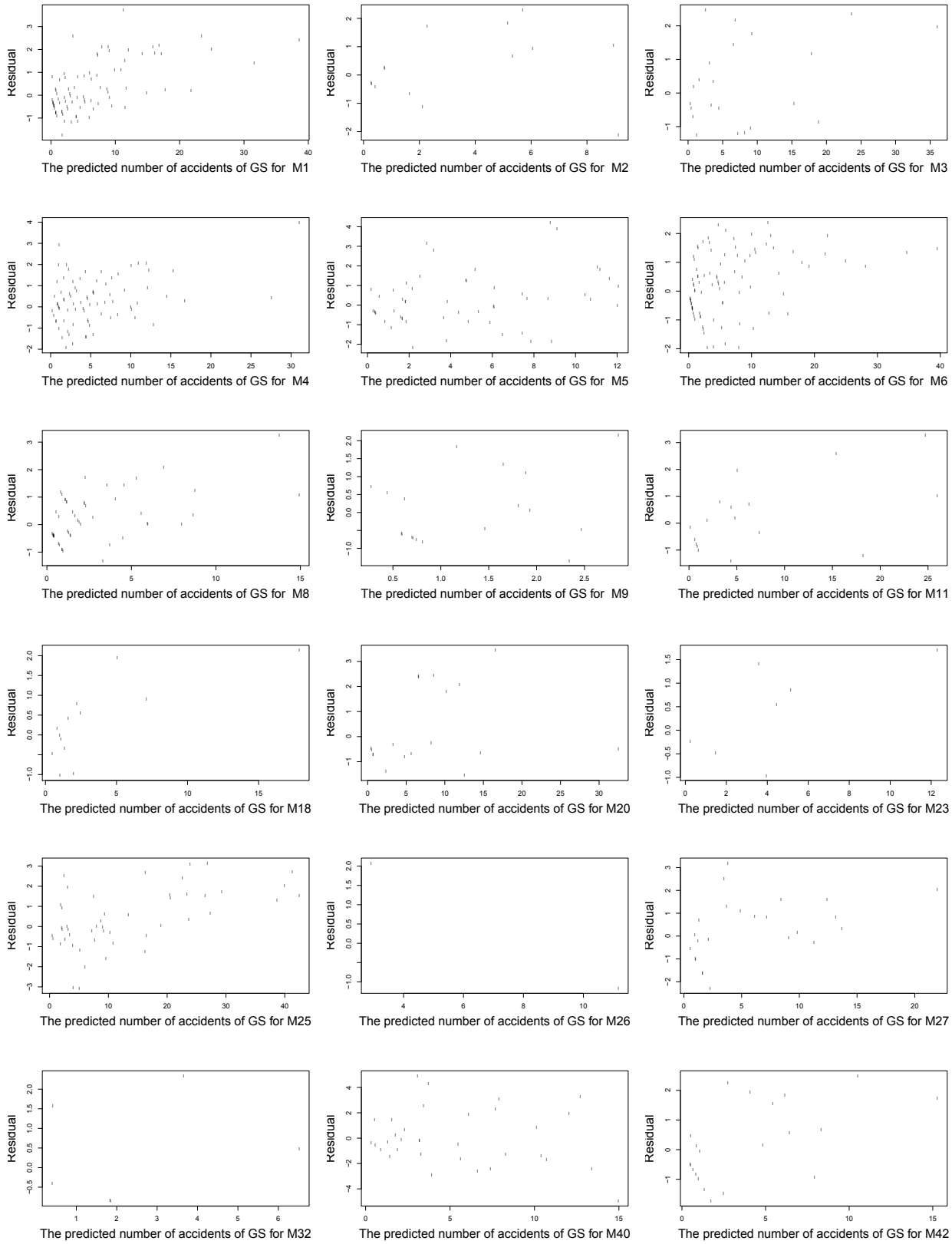


Figure 4.3: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS denotes grouped segment.

Based on estimation results of three-level Bayesian hierarchical model for the UK motorway data, the estimate of the intensity of accidents on the UK motorway network is classified into five categories.

Category one ($\lambda < 0.5$) is referred to a very low risk; Category two ($0.5 \leq \lambda < 1$) is referred to a low risk; Category three ($1 \leq \lambda < 2$) is referred to a moderate risk; Category four ($2 \leq \lambda < 3$) is referred to a high risk. Finally, category five ($\lambda \geq 3$) is referred to a very high risk. The moderate-risk level represents the general intensity of accidents level of the UK motorway network. Based on the results in Figures 4.4, motorways: M27, M275 and M32 are at high risk, whereas motorways: M25 and M606 form the highest risk motorways, where the expected number of accidents is above 3 per one kilometer for both motorways. On the other hand, motorways: M9, M90, M58, M45, M48, M49, M180, M54, M74 and M50 have the lowest risk such that the expected number of accidents is lower than 0.5 for these motorways.

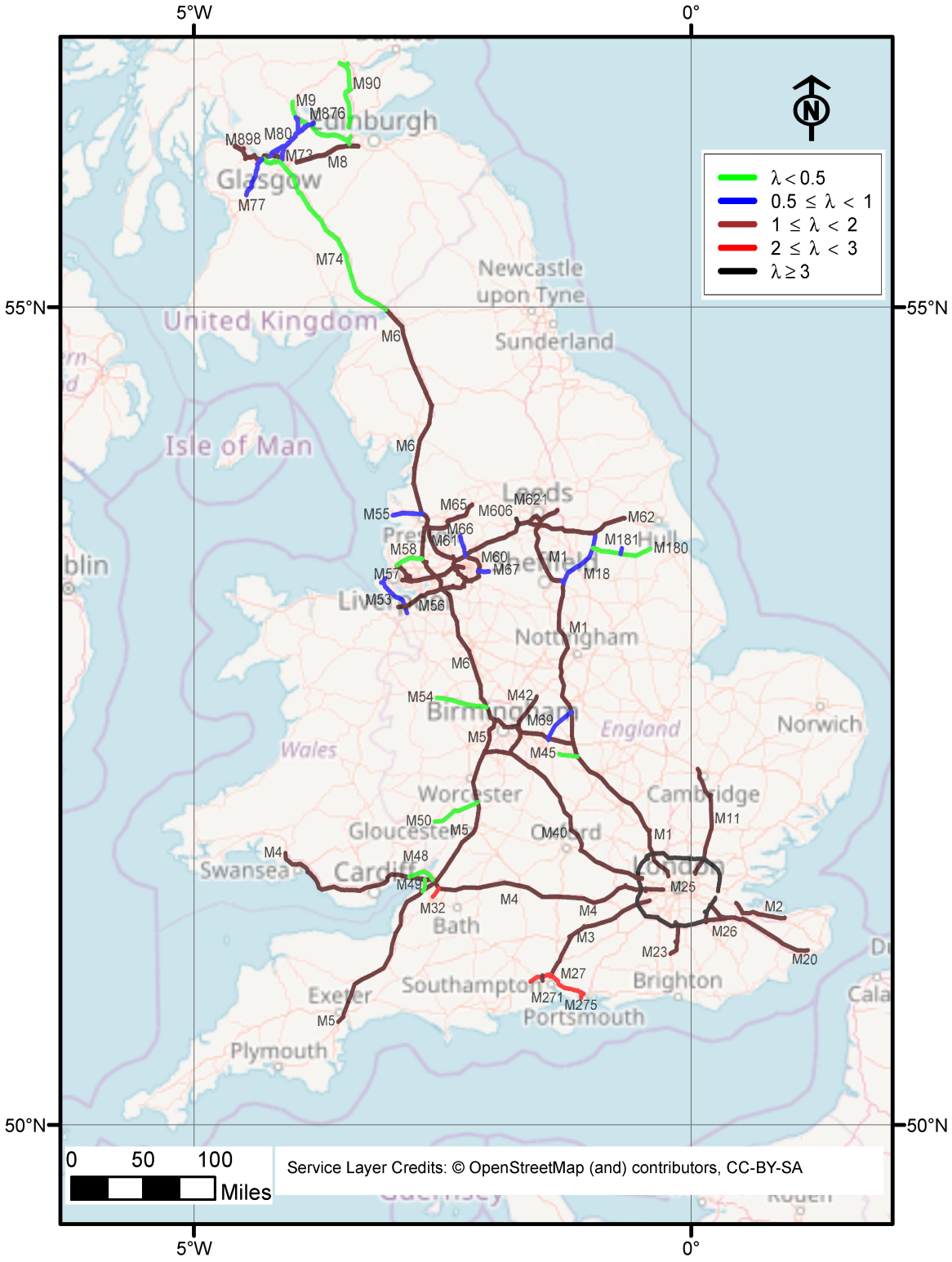


Figure 4.4: Estimated intensities of traffic accidents (λ_i), $i = 1, ..., 49$ per one kilometer on the UK motorway network including 49 motorways. This plot is produced using the traffic accident data in year 2016. The intensity functions are estimated using Bayesian methods with prior distributions $\alpha \sim \text{N}(-6.65, 0.09^2)$ and $\tau^2 \sim \text{Inv-Gamma}(18.36, 58.06)$.

In this section, two approaches are used to assess the convergence of MCMC. In the first approach, mixing within a single chain was examined. This approach involves inspecting trace plots and autocorrelation function (ACF) plots of MCMC samples for the parameter. The trace plots are used for the visual diagnostics of convergence. Trace plots for two parameters α and τ with four proposed prior distributions (the non-informative and weakly-informative priors) for τ, each has a good mixing as shown in the first row of Figure 4.5 and in the first row of Figures B.5, B.6 and B.7 in Appendix B.5. It is known that MCMC methods produce correlated samples, but samples with smaller correlation indicate that algorithm is more efficient. The second row of Figure 4.5 and the second row of Figures B.5, B.6 and B.7 in Appendix B.5 show that the autocorrelation (ACF) for both parameters α and τ is low across four proposed prior distributions for τ. The third row in Figure 4.5 and the third row in Figures B.5, B.6 and B.7 in Appendix B.5 show histograms of both parameters under four different prior distributions for τ. There is no big effect for different prior distributions on the correlation of the simulated chains.

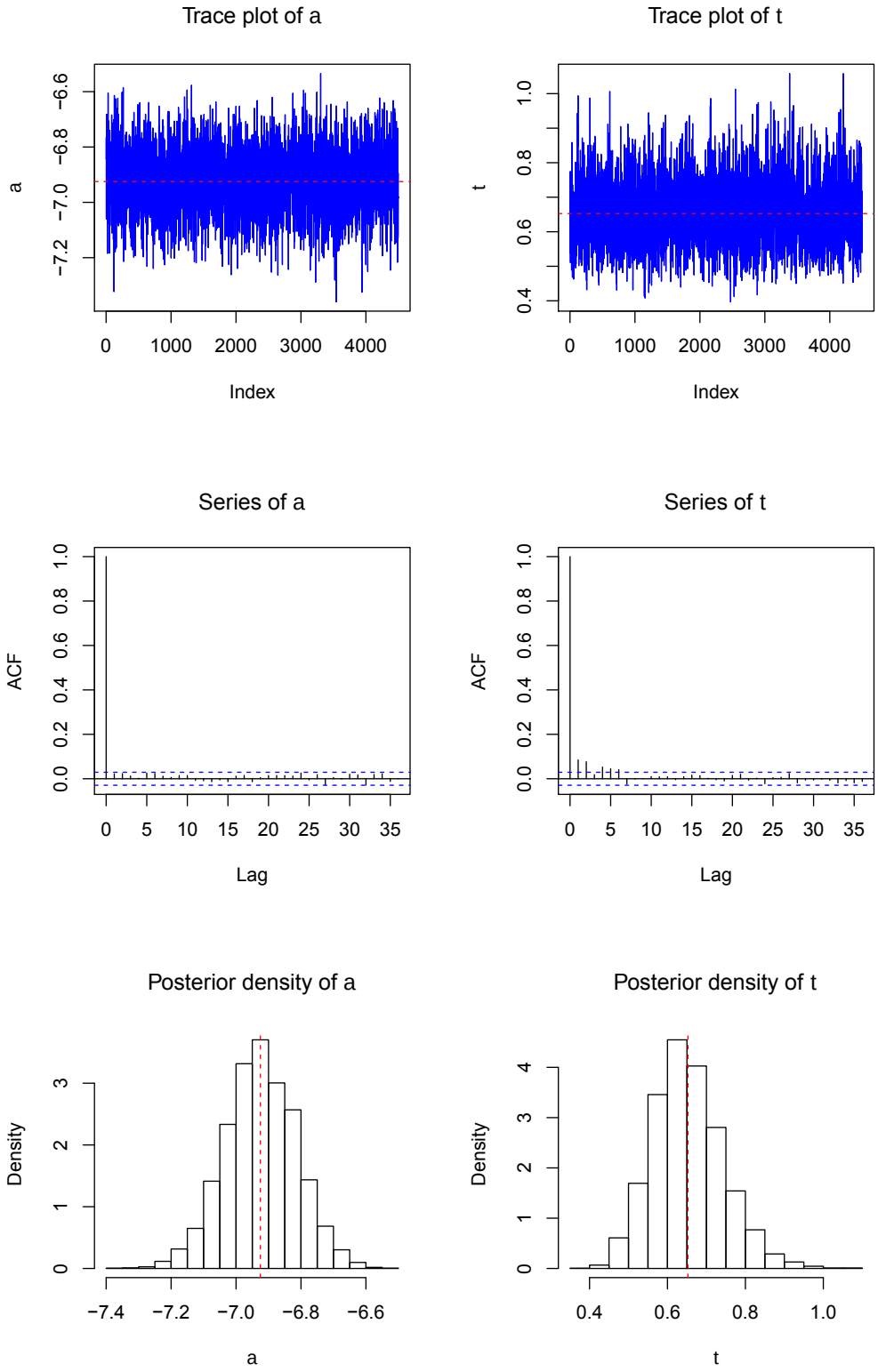


Figure 4.5: The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated using initial values for $\alpha = 0$ and $\tau = 0.1$ with a burn-in of 50,000 samples and a thinning interval of 100 samples. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean.

The second approach for checking the convergence of the MCMC sampling involves checking multiple

chains for each parameter (Gelman and Rubin, 1992a). Two chains were simulated for each parameter with over-dispersed starting points which are $\alpha = -10, 10$ and $\tau = 0.25, 3$. The number of iterations for each chain is $500,000$. The Gelman-Rubin statistic ($\hat{R}$) is calculated. This diagnostic compares between the within-chain variance and the between-chain variance. As discussed in Chapter 5, if $\hat{R} < 1.2$, then the MCMC algorithm is converged. In three-level Bayesian hierarchical model, the Gelman-Rubin statistics are less than 1.2 across four different prior distributions and this is confirmed by plots in the second row in Figure 4.6 and in the second row in Figures B.8, B.9 and B.10 in Appendix B.5 that show median and 97.5% quantile of the sampling distribution of the $\hat{R}$. The first row in Figure 4.6 shows the trace plots for parameters α and τ. Each trace plot illustrates that two chains started from different positions and they converged to the same posterior distribution.

Prior distributions	Parameter	Z-score
$\tau^2 \sim$ Inv-Gamma$(0.001, 0.001)$	α	0.6333
	τ	1.9710
$\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$	α	-0.2436
	τ	1.0050
$\tau \sim$ unif$(0, 100)$	α	1.3680
	τ	-0.3274
$\tau \sim$ HN$(0, 100)$	α	1.5450
	τ	0.2452
$\tau^2 \sim$ Inv-Gamma$(18.3574, 58.0563)$	α	0.8362
	τ	0.03797

Table 4.2: Z-score (Geweke Statistic) resulting from fitting three-level hierarchical Bayesian model to traffic accident data for 2016.

Geweke (1991) presented the other convergence diagnostic method. A single chain is divided into two parts such that the first part represents 10% of iterations and the second part represents 50% iterations. The sample mean and variance were calculated for each part. The Geweke convergence diagnostic (Z-score) is the difference between these two means divided by the standard error of their difference. When the length of the chain is large, the sampling distribution of the diagnostic statistic will be the standard normal distribution. The simulated chain converges to the posterior distribution if Z-score is between ±1.96 (Sahlin, 2011). Table 4.2 shows the Geweke convergence statistic of simulated chain for both parameters model α and τ using Algorithm 4.1 with four different prior distributions. The values of Z do not give evidence that simulated chains do not converge to the posterior distribution. The third row in Figure 4.6 shows the plots of Z-scores versus the first 50% of iterations. There are little values of Z-scores lie outside th 95% confidence interval. This indicates that samples of Z-scores follow a standard-normal distribution. Hence, there is evidence indicating the convergence of MCMC to the posterior distribution.

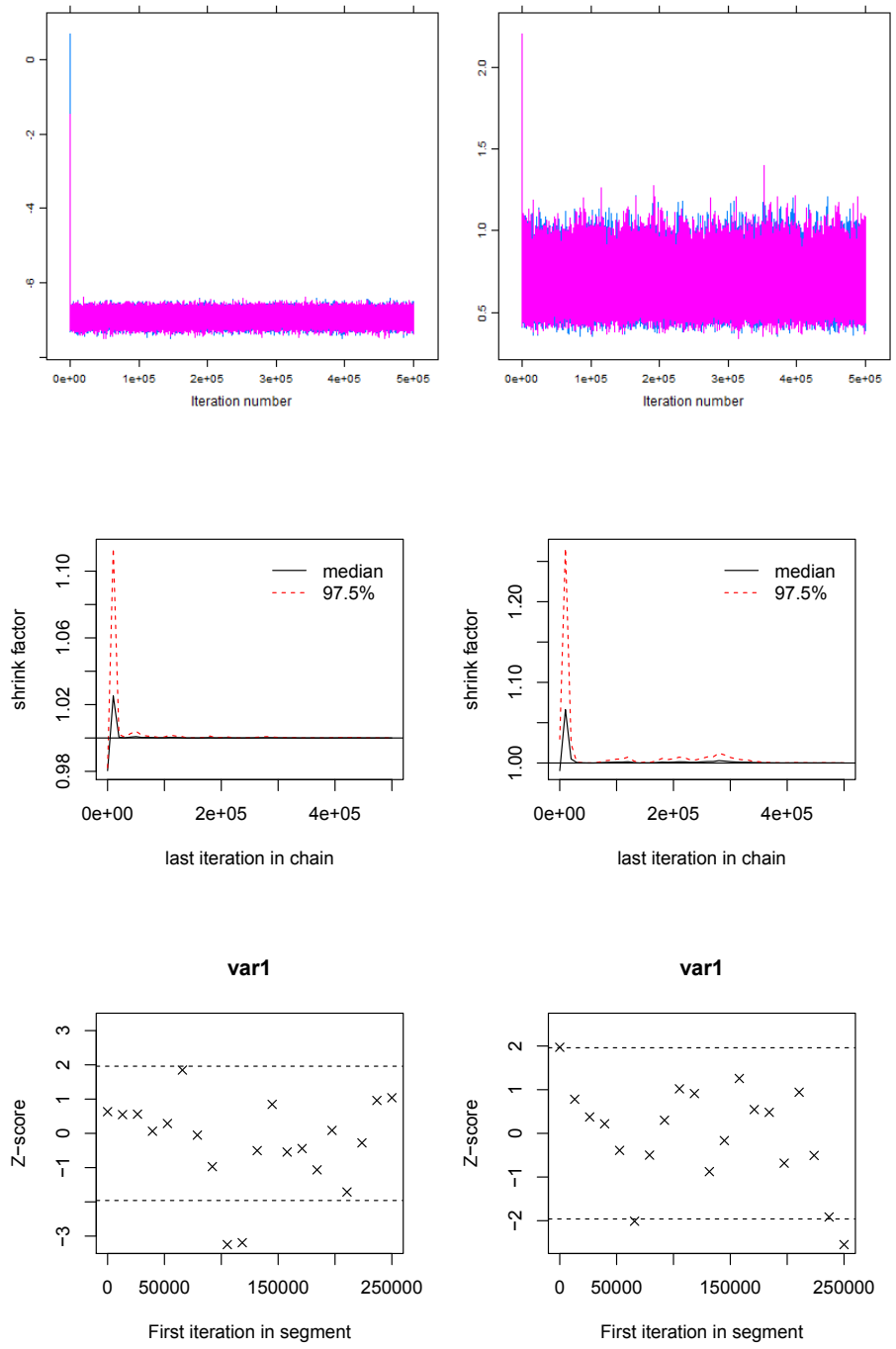


Figure 4.6: The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.001, 0.001)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

4.6 Simulation Study

4.6.1 Simulation Design

To evaluate the performance of methods, a simulation study was conducted. One thousand data sets are generated and each data set was simulated according to the following model:

$$\begin{aligned}
&\alpha_i \sim \mathrm{N}(\alpha, \tau^2), \\
&\alpha_{ij} \sim \mathrm{N}\left(\alpha_i, \tau_i^2\right), \\
&n_{ij} \sim \mathrm{Pois}\left(L_{ij} \exp\left(\alpha_{ij}\right)\right), i = 1, ..., m; j = 1, .., n_i.
\end{aligned} \tag{4.29}$$

Six different scenarios of simulation are considered with $\alpha = -5$ and -7, $\tau = 0.3, 0.7$ and 1.5. The three-level hierarchical model given in (4.1) is fitted for each generated dataset and both Bayesian and frequentist approaches described in sections 4.3 and 4.4 are respectively used to estimate model parameters α and τ. The three-level Bayesian and frequentist hierarchical models are evaluated and compared based on *mean square error* (MSE) and *coverage probability* (CP).

4.6.2 Simulation Results

In Table 4.3, posterior mean for both α and τ is very close to the true value of parameter across all prior distributions for τ and six scenarios. Estimates from frequentist method for both α and τ are close to the true values of parameter across scenarios $(\alpha, \tau) = (-5, 0.3), (-5, 0.7)\text{and}(-5, 1.5)$. Thus, the Bayesian approach performs better than frequentist approach in the terms of parameter estimates.
In the Bayesian method, the absolute value of Bias is small in general. The absolute value of the magnitude of bias of estimate of α slightly increases as the true value of τ increases from 0.3 to 1.5 as in Table 4.3. The absolute value of the magnitude of bias of estimate τ slightly increases as the true value of τ increases for scenarios with true values $\alpha = -5$ and -7 and across two priors uniform and half normal distributions for τ. For non- and weakly-informative Inv-Gamma priors for τ^2, the absolute value of the amount of bias of τ estimate decreases when the true value of τ increases for the true value -7. This means that τ slightly affect the amount of bias. The frequentiat approach produced high biases of point estimates of α and τ across six scenarios, but it produced bigger bias with the true value $\alpha = -7$. In summary, Bayesian method seems to give less biased results than the frequentist method.
Table 4.3 shows that the MSE for both model parameters slightly increases when τ increases for all prior distributions of τ and all simulation scenarios. Magnitudes of the MSE for both α and τ are

similar across scenarios with true value $\alpha = -5$ and $\alpha = -7$ as shown in Table 4.3.

Note that for the true value $\alpha = -7$, the frequentist approach produced larger MSE of α and τ compared with those for the true value $\alpha = -5$. In addition, the MSE of α and τ obtained from the frequentist method is larger than those obtained from Bayesian method for the true value $\alpha = -7$. It can be concluded that the performance of the Bayesian method in terms of MSE is better than the performance of the frequentist method. In general, the frequentist method was performing poorly in terms of MSE.

Bayesian coverage probability values with all proposed prior distributions of τ were close to nominal 95% credible interval for both parameters. The frequentist method produced poor coverage probability values for the true value $\alpha = -7$ and for both parameters, where it was 0 for α. Henderson et al. (2000) shows that the separate analysis using the two-stage method is not performing well compared with the one-stage method. Browne et al. (2006) showed that marginal quasi-likelihood method produced zero value of the coverage probability for a random effect variance parameter (σ_u^2) in random-effects logistic regression (RELR) model. Marginal quasi-likelihood estimation method also yielded very undercoverage probability of random-effect variance ($\sigma_V^2 = 2.4$) and the fixed effect parameter ($\beta_2 = 17.6$) in the RELR model. In the simulation study, for the true value $\alpha = -5$, the coverage probability values are better than those for $\alpha = -7$, but they still under-coverage for both parameters. In our Frequentist approach, the coverage probability of α decreases as the true value of α decreases from -3 to -7 as shown in Tables B.1 and B.2 in Appendix B.5. In general, the performance of the coverage probability in Bayesian method for the both model parameters is better than the frequentist method. The inferior performance of frequentist method is because of biased estimates and large standard deviations of estimates.

			$\alpha = -5$					$\alpha = -7$				
	True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
Inv-Gamma(0.001, 0.001)	0.3	α	-5.0032	-0.0032	0.0026	94.7%	559,000	-6.9940	0.0060	0.0037	93.8%	258,760
		τ	0.3001	0.0001	0.0019	95.0%		0.2937	-0.0063	0.0034	93.5%	
	0.7	α	-5.0035	-0.0035	0.0116	95.6%	549,926	-7.0003	-0.0003	0.0132	95.0%	436,154
		τ	0.7004	0.0004	0.0066	96.2%		0.6941	-0.0059	0.0085	94.5%	
	1.5	α	-5.0052	-0.0052	0.0500	94.8%	542,032	-7.0198	-0.0198	0.0592	93.1%	263,827
		τ	1.4937	-0.0063	0.0262	95.4%		1.5048	0.0048	0.0336	94.9%	
Inv-Gamma(0.1, 0.1)	0.3	α	-5.0037	-0.0037	0.0026	95.2%	584,324	-7.0014	-0.0014	0.0037	95.9%	545,479
		τ	0.3131	0.0131	0.0018	95.8%		0.3164	0.0164	0.0025	96.3%	
	0.7	α	-5.0035	-0.0035	0.0116	95.6%	577,621	-6.9944	0.0056	0.0136	94.5%	507,931
		τ	0.7025	0.0025	0.0065	96.1%		0.6946	-0.0054	0.0085	94.2%	
	1.5	α	-5.0052	-0.0052	0.0501	94.5%	571,022	-7.0187	-0.0187	0.0536	95.2%	569,115
		τ	1.4920	-0.0080	0.0260	95.2%		1.5032	0.0032	0.0340	94.2%	
unif(0, 100)	0.3	α	-5.0035	-0.0035	0.0026	95.0%	545,730	-7.0007	-0.0007	0.0037	95.2%	736,853
		τ	0.3082	0.0082	0.0020	94.9%		0.3063	0.0063	0.0033	94.4%	
	0.7	α	-5.0041	-0.0041	0.0116	95.6%	533,863	-6.9957	0.0043	0.0137	94.8%	555,450
		τ	0.7111	0.0111	0.0068	96.4%		0.7048	0.0048	0.0089	94.3%	
	1.5	α	-5.0062	-0.0062	0.0504	95.5%	580,176	-7.0217	-0.0217	0.0535	95.5%	537,030
		τ	1.5133	0.0133	0.0271	95.8%		1.5287	0.0287	0.0361	93.9%	
HN(0, 0.02)	0.3	α	-5.0035	-0.0035	0.0026	95.2%	574,984	-7.0004	-0.0004	0.0037	95.2%	548,584
		τ	0.3082	0.0082	0.0020	95.0%		0.3064	0.0064	0.0033	94.1%	
	0.7	α	-5.0041	-0.0041	0.0116	95.6%	584,604	-6.9957	0.0043	0.0136	94.7%	497,680
		τ	0.7111	0.0111	0.0068	96.3%		0.7050	0.0050	0.0089	94.5%	
	1.5	α	-5.0061	-0.0061	0.0504	95.5%	546,481	-7.0216	-0.0216	0.0535	95.5%	583,913
		τ	1.5132	0.0132	0.0270	95.7%		1.5286	0.0286	0.0362	93.9%	
Frequentist method	0.3	α	-4.9305	0.0695	0.0073	72.4%	537	-6.5961	0.4039	0.1650	0%	479
		τ	0.3082	0.0082	0.0027	81.7%		0.2650	-0.0350	0.0037	89.8%	
	0.7	α	-4.9166	0.0834	0.0169	86.0%	528	-6.5297	0.4703	0.2271	0%	547
		τ	0.6602	-0.0398	0.0072	94.1%		0.4942	-0.2058	0.0472	27.1%	
	1.5	α	-4.8524	0.1476	0.0593	88.0%	514	-6.3587	0.6413	0.4315	0%	549
		τ	1.3102	-0.1898	0.0532	84.7%		0.9650	-0.5350	0.3012	4.2%	

Table 4.3: Simulation results of frequentist method and Bayesian method under four prior distributions of τ^2 and the prior distribution $\alpha \sim \mathrm{N}(0, 10^2)$ with true value of $\alpha = -5$ and -7. Time is recorded in seconds. MSE represents mean square error and CP the coverage probability.

4.7 Models Comparison

We present two methods to select the best fitting model from our two candidate models (two and three-level Bayesian hierarchical models). Here, two- and three-level Bayesian hierarchical models are only compared since simulation results in sections 3.7 and 4.6 show that the performance of Bayesian method is better than the frequentist method. Specially, the three-level frequentist hierarchical model performed poorly. Regarding Bayesian methods, the comparison of two- and three-levels hierarchical models is done across all proposed prior distributions.

4.7.1 Models Comparison using Information Criteria

In this section, the most commonly used two criteria are described and employed to compare Bayesian hierarchical models that are the deviance information criterion (DIC) developed by Spiegelhalter et al. (2002) and the Watanabe-Akaike or widely applicable information criterion (WAIC) proposed by Watanabe (2010).

Deviance Information Criterion (DIC)

The DIC is the sum of the two terms that are measures of a goodness of fit of the model (deviance statistic) and a complexity (the number of free parameters in the model). The effective number of parameters in the model is the posterior mean of the deviance minus the deviance at the posterior estimates of the parameters. Thus, the effective number of parameters is given by,

$$PD = \overline{D(\Theta)} - D(\overline{\Theta}) = E_{\Theta|\text{data}}\left[-2\log\left\{L\left(\Theta;\text{data}\right)\right\}\right] + 2\log\left[L\left(\hat{\Theta};\text{data}\right)\right], \tag{4.30}$$

where $D(\Theta) = -2\log\left\{L\left(\Theta;\text{data}\right)\right\}+2\log\left\{f(\text{data})\right\}$ is Bayesian deviance, $\overline{D(\Theta)} = E\left[-2\log\left\{L\left(\Theta;\text{data}\right)\right\}\right]+2\log\left\{f(\text{data})\right\}$ and $\hat{\Theta}$ is posterior mean of the parameter Θ. According to Spiegelhalter et al. (2002), the deviance information criterion is given by,

$$\text{DIC} = \overline{D\left(\Theta\right)} + PD = -4E\left[\log\left\{L\left(\Theta;\text{data}\right)\right\}\right] + 2\log\left[L\left(\hat{\Theta};\text{data}\right)\right]. \tag{4.31}$$

The $\overline{D\left(\Theta\right)}$ represents the expected deviance. The $L\left(\hat{\Theta};\text{data}\right)$ represents the likelihood function with $\hat{\Theta}$ posterior means. These posterior means are produced using MCMC.

Watanabe-Akaike or widely applicable information criterion (WAIC)

The WAIC is an improvement on the deviance information criterion. The WAIC avoids the existing problems in the DIC and it is fully Bayesian since posterior estimates contribute to the formulation of this criterion. Let $\text{Y} = \left(\text{y}_1, ..., \text{y}_n\right)$ represent observed data. Let $L\left(\text{Y}|\Theta\right)$ be a likelihood function and

$\pi(\Theta|\mathrm{Y})$ represent a posterior distribution. The WAIC also includes terms for the fit and the complexity of a model. The measure of fit is the log pointwise predictive density (lppd) that is given by,

$$\text{lppd} = \log \prod_{i=1}^{n} P\left(\mathrm{y}_i|\mathrm{Y}\right) = \sum_{i=1}^{n} \log E_{\Theta}\left[L\left(\mathrm{y}_i|\Theta\right)\right] = \sum_{i=1}^{n} \log \int L\left(\mathrm{y}_i|\Theta\right)\pi\left(\Theta|\mathrm{Y}\right)d\Theta. \tag{4.32}$$

In practice, lppd can be calculated using simulated samples $\Theta^{(t)}$, $(t = 1, ..., M)$ from the posterior distribution $\pi(\Theta|\mathrm{Y})$ that are generated using MCMC methods. So the log pointwise predictive density is given by,

$$\text{lppd} = \sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{t=1}^{M} L\left(\mathrm{y}_i|\Theta^{(t)}\right)\right). \tag{4.33}$$

The measure of complexity is an effective number of parameters that is also called bias correction (Gelman et al., 2014). According to Gelman et al. (2014) the bias correction of the WAIC is given by,

$$PWAIC = 2\sum_{i=1}^{n}\left[\log\left(E_{\text{post}}L\left(\mathrm{y}_i|\Theta\right)\right) - E_{\text{post}}\left(\log L\left(\mathrm{y}_i|\Theta\right)\right)\right]. \tag{4.34}$$

The above formula of *PWAIC* is rewritten where the expectations are replaced by the average over the M posterior simulations $\Theta^{(t)}$, that is

$$PWAIC = 2\sum_{i=1}^{n}\left[\log\left(\frac{1}{M}\sum_{t=1}^{M} L\left(\mathrm{y}_i|\Theta^{(t)}\right)\right) - \frac{1}{M}\sum_{t=1}^{M}\log L\left(\mathrm{y}_i|\Theta^{(t)}\right)\right]. \tag{4.35}$$

Using the log pointwise predictive density of data (lppd) in equation (4.33) and a bias correction (*PWAIC*) in equation (4.35), the expected log pointwise predictive density of data (elppd) is given by,

$$\widehat{\text{elppd}}_{WAIC} = \text{lppd} - PWAIC. \tag{4.36}$$

According to (Gelman et al., 2014), *WAIC* is given by,

$$WAIC = -2\left(\widehat{\text{elppd}}_{WAIC}\right) = 2\sum_{i=1}^{n}\log\left(\frac{1}{M}\sum_{t=1}^{M} L\left(\mathrm{y}_i|\Theta^{(t)}\right)\right) - 4\sum_{i=1}^{n}\frac{1}{M}\sum_{t=1}^{M}\log L\left(\mathrm{y}_i|\Theta^{(t)}\right). \tag{4.37}$$

Table 4.4 shows that DIC and WAIC for the three-level Bayesian hierarchical model with the non-informative prior Inv-Gamma(0.001, 0.001) are lower than those for two-level Bayesian hierarchical model. This indicates that the three-level Bayesian Hierarchical is appropriate and closest to the real model of observed data compared with the two-level Bayesian hierarchical model.

Model	2LBHM	3LBHM
DIC	100408.3	85204.3
WAIC	89412.6	1049.9

Table 4.4: *DIC* and *WAIC* criteria. 3LBHM represents the three-level Bayesian hierarchical model and 2LBHM represents the two-level Bayesian hierarchical model.

4.7.2 Models Comparison using Simulation Study

Simulation Design

The term "model misspecification" means the wrongly fitted model to data (Yoo and Slate, 2005). The model misspecification affects estimation results leading to the producing wrong or biased estimates. To investigate the effects of model misspecification, we recall the same data sets generated in section 4.6.1 that are simulated depending on three-level hierarchical model (4.1), and fit data sets generated using two-level Bayesian hierarchical model (3.1). We provide posterior mean, bias, mean square error and coverage probability criteria described in chapter 3 section 3.7.1 to investigate whether model (3.1) is able to analyse data when heterogeneity is incorporated across grouped segments of motorway. The same prior distribution in section 3.6 and the same initial values in section 3.6 are utilized in the Bayesian analysis, and 100,000 iterations were run with burn-in 10,000 and thinning interval of 10 to get posterior samples for α and τ using algorithm 3.1.

Simulation Results

Tables 4.5 shows that the two-level Bayesian hierarchical model produced biased estimates, large mean square errors and extremely bad coverage probability values for both model parameters. The coverage probability values were 0 or close to 0 for τ and exactly equal or close to 100% for α when the true value of the standard deviation between motorway $\tau = 0.3$ and 0.7. This indicates that the fitted model (two-level hierarchical Bayesian model) is incorrect.

			$\alpha = -5$					$\alpha = -7$				
	True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
Inv-Gamma(0.001, 0.001)	0.3	α	-4.8199	0.1801	0.0387	100%	20147	-6.8121	0.1879	0.0437	99.9%	19592
		τ	1.6270	1.3270	1.7673	0%		1.6670	1.3670	1.8768	0%	
	0.7	α	-4.8165	0.1835	0.0479	99.2%	17955	-6.8004	0.1996	0.0576	98.7%	21103
		τ	1.7362	1.0362	1.0868	0%		1.7856	1.0856	1.1946	0%	
	1.5	α	-4.8167	0.1833	0.0841	96.6%	19410	-6.822	0.178	0.0868	97.0%	19959
		τ	2.188	0.688	0.5142	0.37%		2.2409	0.7409	0.5989	0.34%	
Inv-Gamma(0.1, 0.1)	0.3	α	-4.8197	0.1803	0.0387	100%	19623	-6.8121	0.1879	0.0438	99.9%	20937
		τ	1.6245	1.3245	1.7607	0%		1.6645	1.3645	1.8699	0%	
	0.7	α	-4.8167	0.1833	0.0478	99.6%	18536	-6.8002	0.1998	0.0577	98.4%	20114
		τ	1.734	1.034	1.0821	0%		1.7829	1.0829	1.1887	0%	
	1.5	α	-4.8163	0.1837	0.084	96.2%	17885	-6.8219	0.1781	0.0865	97.1%	19965
		τ	2.1836	0.6836	0.5079	03.7%		2.2368	0.7368	0.5926	0.35%	
unif(0, 100)	0.3	α	-4.8200	0.1800	0.0388	100%	19398	-6.8129	0.1871	0.0435	100%	21003
		τ	1.6452	1.3452	1.816	0%		1.6869	1.3869	1.9318	0%	
	0.7	α	-4.8168	0.1832	0.0478	99.4%	19788	-6.8014	0.1986	0.0573	99.0%	22654
		τ	1.7562	1.0562	1.1289	0%		1.8075	1.1075	1.2430	0%	
	1.5	α	-4.8169	0.1831	0.0837	96.7%	22420	-6.8243	0.1757	0.086	97.4%	20769
		τ	2.2136	0.7136	0.5511	0.31%		2.2696	0.7696	0.6437	03.0%	
HN(0, 0.02)	0.3	α	-4.8195	0.1805	0.0389	100%	23911	-6.8129	0.1871	0.0435	100%	24285
		τ	1.6455	1.3455	1.8169	0%		1.6869	1.3869	1.9319	0%	
	0.7	α	-4.8167	0.1833	0.0478	99.5%	24054	-6.8014	0.1986	0.0573	99.0%	24808
		τ	1.7561	1.0561	1.1287	0%		1.8075	1.1075	1.2429	0%	
	1.5	α	-4.8167	0.1833	0.0839	96.5%	24214	-6.8243	0.1757	0.0860	97.4%	24460
		τ	2.2134	0.7134	0.5509	0.32%		2.2697	0.7697	0.6438	3.0%	

Table 4.5: Simulation results from two-level Bayesian hierarchical model under four prior distributions of τ^2 and the prior distribution $\alpha \sim \mathrm{N}(0, 10^2)$ with true value of $\alpha = -5$ and -7. Time is recorded in seconds. MSE represents mean square error and CP the coverage probability.

4.8 Discussion

A methodology for modelling accident data at the segment level of the UK motorway network is proposed. The model is built up by sub-dividing the UK motorway network into grouped segments, where the three-level hierarchical model was used to take into account the heterogeneity across segments and motorways. The model has been applied to traffic accident data on the UK motorway network in 2016 to classify the dangerous motorways. The Bayesian Markov Chain Monte Carlo methods and frequentist method are used to estimate model parameters. In the Bayesian method, a sensitivity analysis with different prior distributions specifications for τ^2 has been performed to investigate the effect of the prior choice on the resulting posterior distributions of α and τ. We have used a non-informative, weakly-informative and informative priors. The analysis revealed that Bayesian results are not sensitive to the choice of prior distributions. Gelman-Rubin statistic, Geweke statistic, ACF and trace plots have been used to monitor the convergence of posterior distributions of model parameters, α and τ. These convergence diagnostic methods have indicated the convergence of the MCMC chains. Regarding the frequentist approach, the maximum likelihood method has been separately used for each level of model. Information criteria (DIC and WAIC) and simulation study were used to compare between the two-level and three-level Bayesian hierarchical models.

In a simulation study and a real application, we have examined the performance of Bayesian and frequentist methods for fitting the three-level hierarchical model. The simulation results showed that the performance of the three-level Bayesian hierarchical model is better than the frequentist method in the most simulation scenarios proposed. Indeed, the frequentist method encounters difficulty in attaining the nominal 95% coverage probability in the scenarios with true value $\alpha = -7$ where it yielded very poor coverage probability for α (coverage probability = 0) and considerable undercoverage for τ (coverage probability = 4.2% and 27.1%). This is due to a large bias in the estimates of α and τ that also leaded to large MSE in the both parameters.

The Bayesian method produced mean estimates that were close to unbiased for both parameters, α and τ, with all priors and with coverage probability close to 95% for all scenarios. Results from simulation study demonstrated that the value of MSE and bias are sensitive to the choice of the τ value. This means that an increase in the τ value results in an increase in the MSE and bias of α and τ.

Based on DIC, WAIC and simulation study, the three-level Bayesian hierarchical model was chosen as the best fitting model to traffic accident data on the UK motorway network since it yielded the lowest DIC and WAIC. The three-level Bayesian hierarchical model provided good information about the intensity of accidents on the UK motorway network for 2016. According to levels of accident intensity risk (very low, low, moderate, high, very high), the moderate risk level was identified as the general intensity of accident level of the UK motorway network. The very high-risk level was observed on M25 and M606, whereas the very low-risk level on M9, M90, M58, M45, M48, M49, M180, M54, M74 and M50

Chapter 5

Conclusion

5.1 Summary

This dissertation focused on Bayesian hierarchical models for analysing road accidents on the UK motorway network. Work along this line has been gradated from a line segment to a linear network. This work has determined which the most dangerous motorways in the UK network based on the estimated intensity of traffic accidents.

Chapter 1 presented a general review of the Bayesian inference. In Bayesian inference, Bayes's theorem, prior distributions and MCMC methods were reviewed.

Chapter 2 defined a line segment and discussed homogeneous and inhomogeneous spatial Poisson processes on the line segment. The discussion includes using the maximum likelihood and Bayesian methods to produce useful estimation of the intensity function of simulated events from an inhomogeneous spatial Poisson process on the line segment. In the Bayesian approach, the Metropolis-Hastings within Gibbs sampling was used to provide posterior summaries of the intensity of a spatial point process. The two estimation methods gave similar summaries of an inhomogeneous Poisson model parameters. The aim of chapter 2 was to pave the way for defining spatial point process on the linear network.

Chapter 3 and 4 presented our main contributions where we proposed Bayesian hierarchical models to account for the multilevel nature of data on the UK motorway network. These models have not been used for the UK motorway network before. Using our proposed hierarchical models, we identified motorways with highest and lowest intensities of accidents, classified motorways into different risk categories, and estimated the overall intensity of accidents.

Chapter 3 demonstrated the use of one-stage fully Bayesian hierarchical model, two-stage semi-Bayesian hierarchical model, two-stage frequentist hierarchical model, Bayesian non-hierarchical

model and frequentist non-hierarchical model for analysing traffic accident data on the UK motorway network. In the Bayesian approach, we conducted a sensitivity analysis to evaluate the impact of the prior distributions. In general, our proposed hierarchical models were not sensitive to the prior specification of the between-motorway standard deviation. We assessed the performance of all proposed models using a simulation study and real application that includes traffic accident data on the UK motorway network for 2016.

In the simulation study, different scenarios were conducted. We examined three performance criteria, bias, *mean square error* (MSE) and *coverage probability* (CP) of parameter estimates. The simulation results showed that the performance of the fully Bayesian hierarchical model is better than those of the semi-Bayesian and frequentist hierarchical models in terms of bias and coverage probability for some simulation scenarios. The performance of all the three models is similar in terms of MSE. The results of the simulation study show that non-hierarchical models perform poorly in terms of all the evaluation criteria. In the real application, the analysis of the fully Bayesian hierarchical model showed that M25 and M27 seem to have the highest accident intensity on the UK motorway network for 2016. Results also showed that motorways M180, M74, M50 appear to have the lowest intensity of accidents.

Chapter 4 proposed using a three-level hierarchical model to incorporate a multilevel data structure. We assume accident intensity is homogeneous within grouped segments but heterogeneous across grouped segments. The three-level hierarchical model was evaluated using a simulation study and traffic accident data on the UK motorway network for 2016. The simulation results showed that the Bayesian three-level hierarchical model performed better than the frequentist model in the most simulation scenarios proposed. The frequentist method faced difficulty in attaining the required level of actual coverage in some scenarios because of a large bias in the estimates of the overall log-intensity of accidents and the between motorway standard deviation. The results of the analysis of the model showed that M25 and M606 have high intensity of accidents and M9, M90, M58, M45, M48, M49, M180, M54, M74 and M50 have low intensity of accidents.

Information criteria (DIC and WAIC) and simulation study were used to compare between the two-level and three-level Bayesian hierarchical models. The values of criteria DIC and WAIC for three-level Bayesian hierarchical model are less than those for two-level Bayesian hierarchical model. This means that the best fitting model is the three-level Bayesian hierarchical model.

5.2 Future work

In the published literature, point patterns were simulated on the two-dimensional space (Moller and Waagepetersen, 2003; Ripley, 1991; Cressie, 1992; Symanzik, 2005). Points represent the locations of events. The points are simulated either uniformly or non-uniformly (inhomogeneously) across the study region. In the inhomogeneous case, the adjustment on the locations is made so that the point pattern inhomogenously distributes across the study region in a two-dimensional space. However, the UK motorway network is a linear network consisting of line segments. The methods for simulating points in two-dimensional space cannot be directly applied to simulate locations of accidents on a linear network. A possible future extension is to simulate an inhomogeneous Poisson process on a linear network.

Spatial covariate on a linear network is a quantity that could imaginably be measured at any location on the network (Baddeley et al., 2015). In the analysis of road traffic accidents, spatial covariate may include road width, the distance to the nearest road intersection and the sighting distance along the road. Future extension may include studying the effect of spatial covariates on the pattern of accidents. Models may be extended to include accident data from different years to study how the intensity of accidents change over time.

Appendix A

Derivations and Plots of Chapter 3

A.1 One-Stage Fully Bayesian Hierarchical Method (Model 1)

A.1.1 Likelihood Function

The likelihood function is

$$\begin{aligned}
L(\mathbf{N}|\Theta) &= P\left(\mathbf{N}|\alpha_1,...,\alpha_m\right) P(\alpha_1,...,\alpha_m|\alpha,\tau^2) f\left(s_1,...,s_{n_i}\right), \\
&= \prod_{i=1}^{m} \frac{(\lambda_i L_i)^{n_i} \exp\left(-\lambda_i L_i\right)}{n_i!} \times \frac{1}{L_i^{n_i}} \times \prod_{i=1}^{\mathrm{m}} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right), \\
&\propto \prod_{i=1}^{\mathrm{m}} \lambda_i^{n_i} \exp\left(-L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{\mathrm{m}} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right), \\
&\propto \prod_{i=1}^{m} \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \prod_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi\tau^2}}\right) \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right). \qquad \text{(A.1)}
\end{aligned}$$

A.1.2 Full Conditional Posterior Distributions

A.1.2.1 Conditional Posterior Distribution of α_i

The conditional posterior distribution of α_i is

$$\begin{aligned}
\pi\left(\alpha_i|\alpha,\tau^2,\mathbf{N}\right) &= \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \left(\frac{1}{\sqrt{2\pi\tau^2}}\right) \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right), \\
&\propto \exp\left(n_i\alpha_i - L_i \exp\left(\alpha_i\right)\right) \times \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right). \qquad \text{(A.2)}
\end{aligned}$$

A.1.2.2 Conditional Posterior Distribution of α

The conditional posterior distribution of α is

$$\begin{aligned}
\pi\left(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{N}\right) &= \left(\frac{1}{\sqrt{2\pi\tau^2}}\right)^m \exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right),\\
&\propto \exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right)\times\exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right),\\
&\propto \exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}-\sum_{i=1}^{m}\frac{(\alpha-\alpha_i)^2}{2\tau^2}\right),\\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}(\alpha-\mu_0)^2-\frac{1}{2\tau^2}\sum_{i=1}^{m}(\alpha-\alpha_i)^2\right),\\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}\left(\alpha^2-2\alpha\mu_0+\mu_0^2\right)-\frac{1}{2\tau^2}\sum_{i=1}^{m}\left(\alpha^2-2\alpha_i\alpha+\alpha_i^2\right)\right),\\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}\left(\alpha^2-2\alpha\mu_0\right)-\frac{1}{2\tau^2}\left(m\alpha^2-2\alpha\sum_{i=1}^{m}\alpha_i\right)\right),\\
&\propto \exp\left(-\frac{1}{2}\left[\alpha^2\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)-2\alpha\left(\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2}+\frac{\mu_0}{\sigma_0^2}\right)\right]\right).
\end{aligned}\tag{A.3}$$

Multiplying and dividing by $\dfrac{m}{\tau^2}+\dfrac{1}{\sigma_0^2}$ and letting $\mu_\alpha = \dfrac{\dfrac{\sum_{i=1}^{m}\alpha_i}{\tau^2}+\dfrac{\mu_0}{\sigma_0^2}}{\dfrac{m}{\tau^2}+\dfrac{1}{\sigma_0^2}}$ produces

$$\pi(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{N}) \propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)\left(\alpha^2-2\alpha\mu_\alpha\right)\right).\tag{A.4}$$

Completing square to obtain the mean for the normal distribution is formed by adding and subtracting with μ_α^2

$$\begin{aligned}
\pi(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{N}) &\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)\left(\alpha^2-2\alpha\mu_\alpha+\mu_\alpha^2\right)\right),\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)(\alpha-\mu_\alpha)^2\right).
\end{aligned}\tag{A.5}$$

Therefore, $\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{N}$ follows the normal distribution with mean μ_α and corresponding variance σ_α^2

$$\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{N} \sim \mathrm{N}(\mu_\alpha,\sigma_\alpha^2),\tag{A.6}$$

where

$$\mu_\alpha = \frac{\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}} \text{ and } \sigma_\alpha^2 = \frac{1}{\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}}.$$

A.1.2.3 Conditional Posterior Distribution of τ^2

The conditional posterior distribution of τ^2 is

$$\begin{aligned}
\pi(\tau^2|\alpha_1,...,\alpha_m,\alpha,\mathbf{N}) &= \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-\alpha_0-1}\exp\left(-\beta_0/\tau^2\right), \\
&\propto \left(\tau^2\right)^{-\frac{m}{2}}\exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\left(\tau^2\right)^{-\alpha_0-1}\exp\left(-\beta_0/\tau^2\right), \\
&\propto \left(\tau^2\right)^{-(\alpha_0+\frac{m}{2})-1}\exp\left(-\frac{\beta_0+\frac{\sum_{i=1}^{2}(\alpha_i-\alpha)^2}{2}}{\tau^2}\right). \qquad \text{(A.7)}
\end{aligned}$$

Therefore,

$$\tau^2|\alpha_1,...,\alpha_m,\alpha,\mathbf{N} \sim \text{Inv-Gamma}\left(\alpha_0+\frac{m}{2},\beta_0+\frac{\sum_{i=1}^{m}(\alpha_i-\alpha)^2}{2}\right). \qquad \text{(A.8)}$$

A.1.2.4 Conditional Posterior Distribution of τ

The conditional posterior distribution on τ can be derived as

$$\begin{aligned}
\pi(\tau|\alpha_1,...,\alpha_m,\alpha,\mathbf{N}) &= \prod_{i=1}^{m}\left(\frac{1}{\sqrt{2\pi\tau^2}}\right)\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{2\theta}{\pi}\exp\left(-\frac{\tau^2\theta^2}{\pi}\right), \\
&\propto \left(\frac{1}{\sqrt{2\pi\tau^2}}\right)^m\exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{2\theta}{\pi}\exp\left(-\frac{\tau^2\theta^2}{\pi}\right), \\
&\propto \tau^{-m}\exp\left(-\sum_{i=1}^{m}\frac{(\alpha_i-\alpha)^2}{2\tau^2}-\frac{\tau^2\theta^2}{\pi}\right),\ \tau>0. \qquad \text{(A.9)}
\end{aligned}$$

A.2 Two-Stage Semi-Bayesian Hierarchical Method (Model 2)

A.2.1 Likelihood Function

Setting $\mathbf{y} = (y_1, ..., y_m)$, the likelihood function can be given as follows:

$$\begin{aligned}\pi\left(\mathbf{y}|\alpha_1, ..., \alpha_m, \alpha, \tau^2\right) &= P\left(\mathbf{y}|\alpha_1, ..., \alpha_m\right) P\left(\alpha_1, ..., \alpha_m|\alpha, \tau^2\right), \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right). \end{aligned} \tag{A.10}$$

A.2.2 Joint Posterior Distribution

The joint posterior distribution is

$$\pi\left(\alpha_1, ..., \alpha_m, \alpha, \tau^2|\mathbf{y}\right) = P\left(\mathbf{y}|\alpha_1, ..., \alpha_m\right) P\left(\alpha_1, ..., \alpha_m|\alpha, \tau^2\right) P(\alpha) P(\tau^2), \tag{A.11}$$

$$\begin{aligned}\pi\left(\alpha_1, ..., \alpha_m, \alpha, \tau^2|\mathbf{y}\right) = &\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right), \\ &\times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\alpha - \mu_0)^2}{2\sigma_0^2}\right) \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\tau^2\right)^{-\alpha_0 - 1} \exp\left(-\beta_0/\tau^2\right). \end{aligned} \tag{A.12}$$

A.2.3 Full Conditional Posterior Distributions

A.2.3.1 Conditional Posterior Distribution of α_i

The conditional posterior density of α_i is

$$\begin{aligned}\pi\left(\alpha_i|\alpha, \tau^2, \mathbf{y}\right) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \alpha_i\right)^2}{2\sigma_i^2}\right) \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right), \\ &\propto \exp\left(-\frac{\left(\alpha_i - y_i\right)^2}{2\sigma_i^2} - \frac{(\alpha_i - \alpha)^2}{2\tau^2}\right), \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2}\left(\alpha_i^2 - 2\alpha_i y_i + y_i^2\right) + \frac{1}{\tau^2}\left(\alpha_i^2 - 2\alpha_i\alpha + \alpha^2\right)\right)\right), \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2}\left(\alpha_i^2 - 2\alpha_i y_i\right) + \frac{1}{\tau^2}\left(\alpha_i^2 - 2\alpha_i\alpha\right)\right)\right), \\ &\propto \exp\left(-\frac{1}{2}\left(\alpha_i^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right) - 2\alpha_i\left(\frac{y_i}{\sigma_i^2} + \frac{\alpha}{\tau^2}\right)\right)\right). \end{aligned}$$

Multiplying and dividing by $\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}$ and letting $\mu_{\alpha_i} = \frac{\frac{y_i}{\sigma_i^2} + \frac{\alpha}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}$ produces

$$\pi\left(\alpha_i|\alpha, \tau^2, \mathbf{y}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)\left(\alpha_i^2 - 2\alpha_i\mu_{\alpha_i}\right)\right). \tag{A.13}$$

Completing square to obtain the mean for the normal distribution is formed by adding and subtracting with μ_{α_i}

$$\begin{aligned}\pi\left(\alpha_i|\alpha, \tau^2, \mathbf{y}\right) &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)\left(\alpha_i^2 - 2\alpha_i\mu_{\alpha_i} + \mu_{\alpha_i}^2\right)\right), \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)(\alpha_i - \mu_{\alpha_i})^2\right). \end{aligned} \tag{A.14}$$

Therefore, $\alpha_i|\alpha, \tau^2, \mathbf{y}$ follows the normal distribution with mean μ_{α_i} and corresponding variance $\sigma_{\alpha_i}^2$ i.e.

$$\alpha_i \sim \mathrm{N}(\mu_{\alpha_i}, \sigma_{\alpha_i}^2), \tag{A.15}$$

where

$$\mu_{\alpha_i} = \frac{\frac{y_i}{\sigma_i^2} + \frac{\alpha}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \text{ and } \sigma_{\alpha_i}^2 = \frac{1}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}.$$

A.2.3.2 Conditional Posterior Distribution of α

Using equation (3.17), the conditional posterior density of α is calculated by considering α as a random variable and α_i, τ^2 as constants. Hence,

$$\begin{aligned}
\pi(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{y}) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\alpha-\mu_0)^2}{2\sigma_0^2}\right), \\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}(\alpha-\mu_0)^2 - \frac{1}{2\tau^2}\sum_{i=1}^{m}(\alpha_i-\alpha)^2\right), \\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}\left(\alpha^2-2\alpha\mu_0+\mu_0^2\right) - \frac{1}{2\tau^2}\sum_{i=1}^{m}\left(\alpha_i^2-2\alpha_i\alpha+\alpha^2\right)\right), \\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}\left(\alpha^2-2\alpha\mu_0\right) - \frac{1}{2\tau^2}\left(m\alpha^2-2\alpha\sum_{i=1}^{m}\alpha_i\right)\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\alpha^2\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right) - 2\alpha\left(\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2}+\frac{\mu_0}{\sigma_0^2}\right)\right]\right). \qquad \text{(A.16)}
\end{aligned}$$

Multiplying and dividing by $\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}$ and letting $\mu_\alpha = \dfrac{\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}}$ produces

$$\pi(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)\left(\alpha^2-2\alpha\mu_\alpha\right)\right). \qquad \text{(A.17)}$$

Completing square to obtain the mean for the normal distribution is formed by adding and subtracting with μ_α^2

$$\begin{aligned}
\pi(\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{y}) &\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)\left(\alpha^2-2\alpha\mu_\alpha+\mu_\alpha^2\right)\right), \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)(\alpha-\mu_\alpha)^2\right). \qquad \text{(A.18)}
\end{aligned}$$

Therefore, $\alpha|\alpha_1,...,\alpha_m,\tau^2,\mathbf{y}$ follows the normal distribution with mean μ_α and corresponding variance σ_α^2

$$\alpha \sim \mathrm{N}(\mu_\alpha,\sigma_\alpha^2), \qquad \text{(A.19)}$$

where

$$\mu_\alpha = \frac{\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}} \text{ and } \sigma_\alpha^2 = \frac{1}{\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}}.$$

A.2.3.3 Conditional Posterior Distribution of τ^2

Using equation (3.17), we derive the conditional posterior density of τ^2 given other parameters. Hence,

$$\begin{aligned}\pi(\tau^2|\alpha_1,...,\alpha_m,\alpha,\mathbf{y}) &= \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-\alpha_0-1}\exp\left(-\beta_0/\tau^2\right),\\ &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-(\alpha_0+\frac{m}{2})-1}\exp\left(-\frac{\beta_0+\frac{\sum_{i=1}^{m}(\alpha_i-\alpha)^2}{2}}{\tau^2}\right). \qquad \text{(A.20)}\end{aligned}$$

Therefore

$$\tau^2|\alpha_1,...,\alpha_m,\alpha,\mathbf{y} \sim \text{Inv-Gamma}\,(a,b)\,, \qquad \text{(A.21)}$$

where

$$a = \alpha_0 + \frac{m}{2} \text{ and } b = \beta_0 + \frac{\sum_{i=1}^{m}(\alpha_i-\alpha)^2}{2}.$$

A.3 Plots

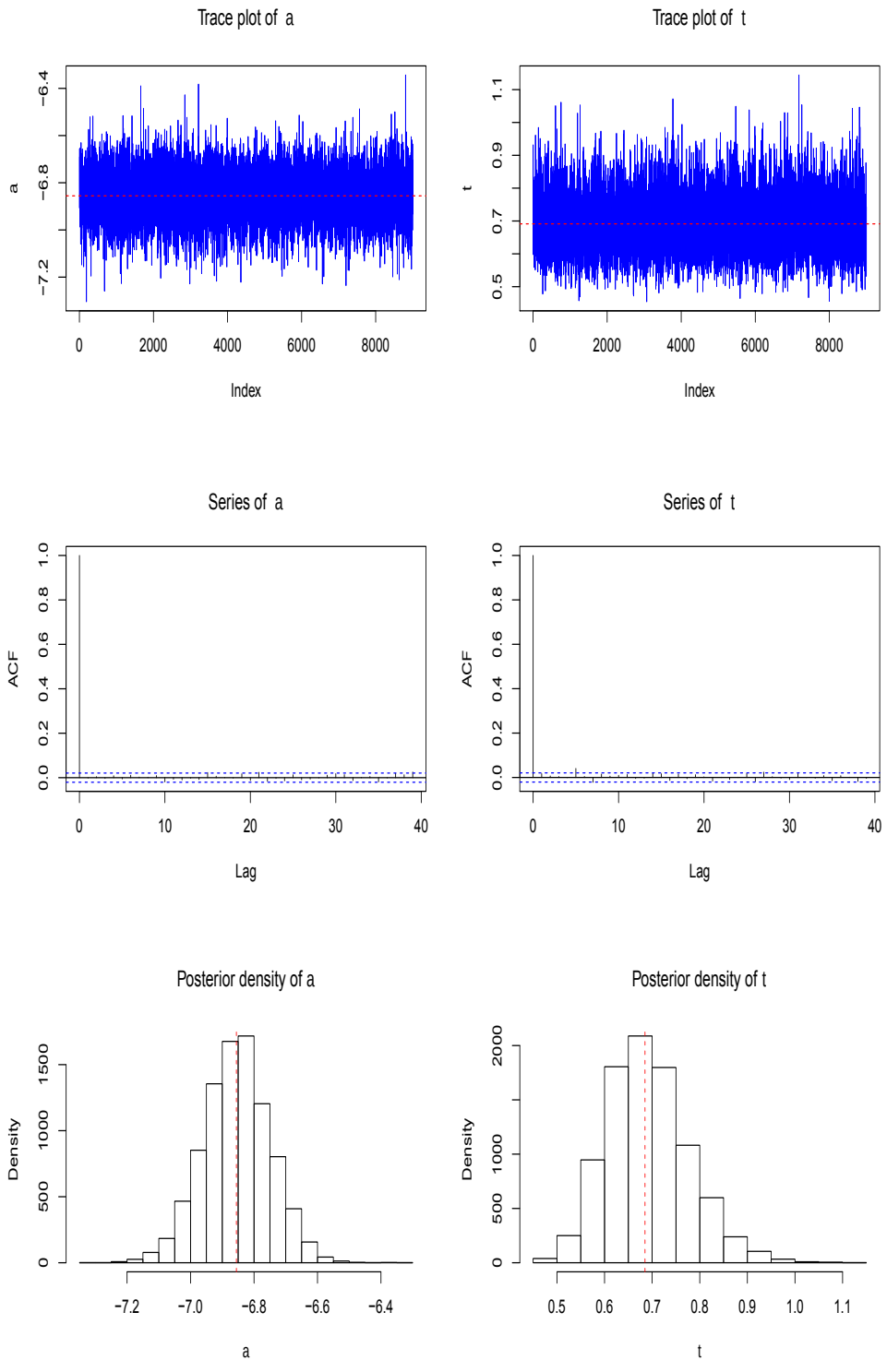


Figure A.1: Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and a weakly-informative $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

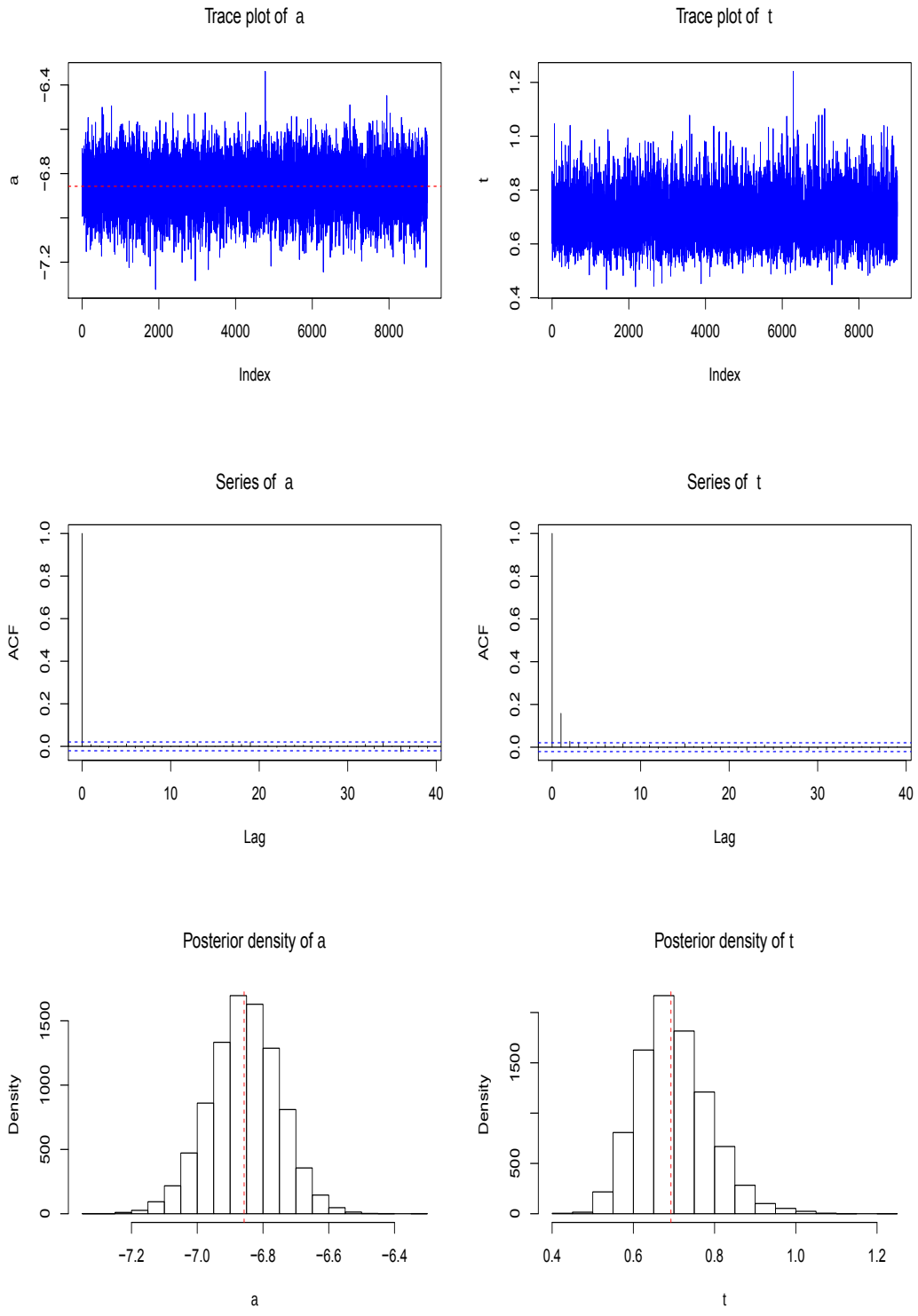


Figure A.2: Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \mathrm{N}(0, 100)$ and non-informative $\tau \sim \mathrm{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

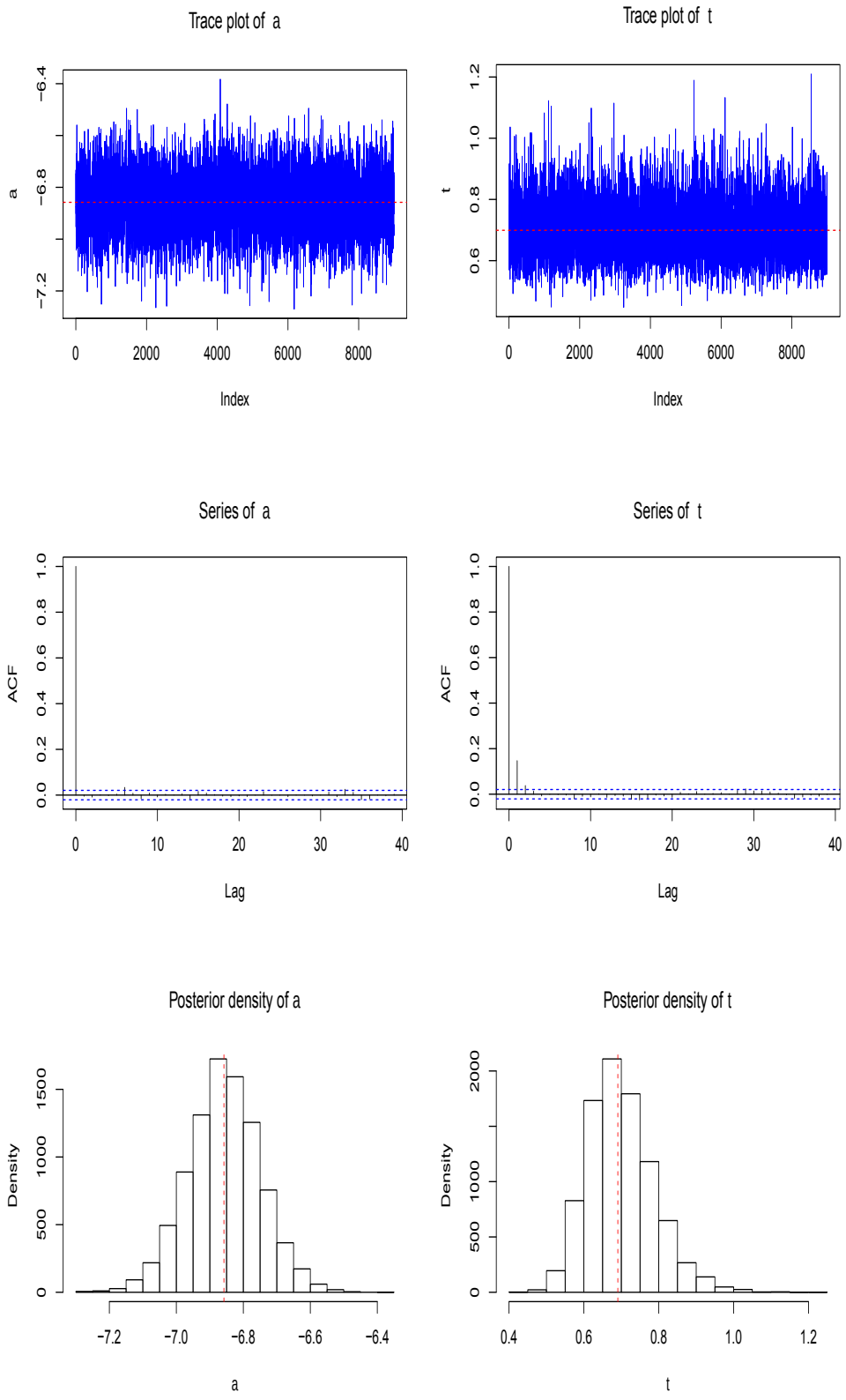


Figure A.3: Trace plots, ACF functions and histograms of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 with prior distributions $\alpha \sim \text{N}(0, 100)$ and non-informative $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ with 10,000 samples discarded burned-in from 100,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

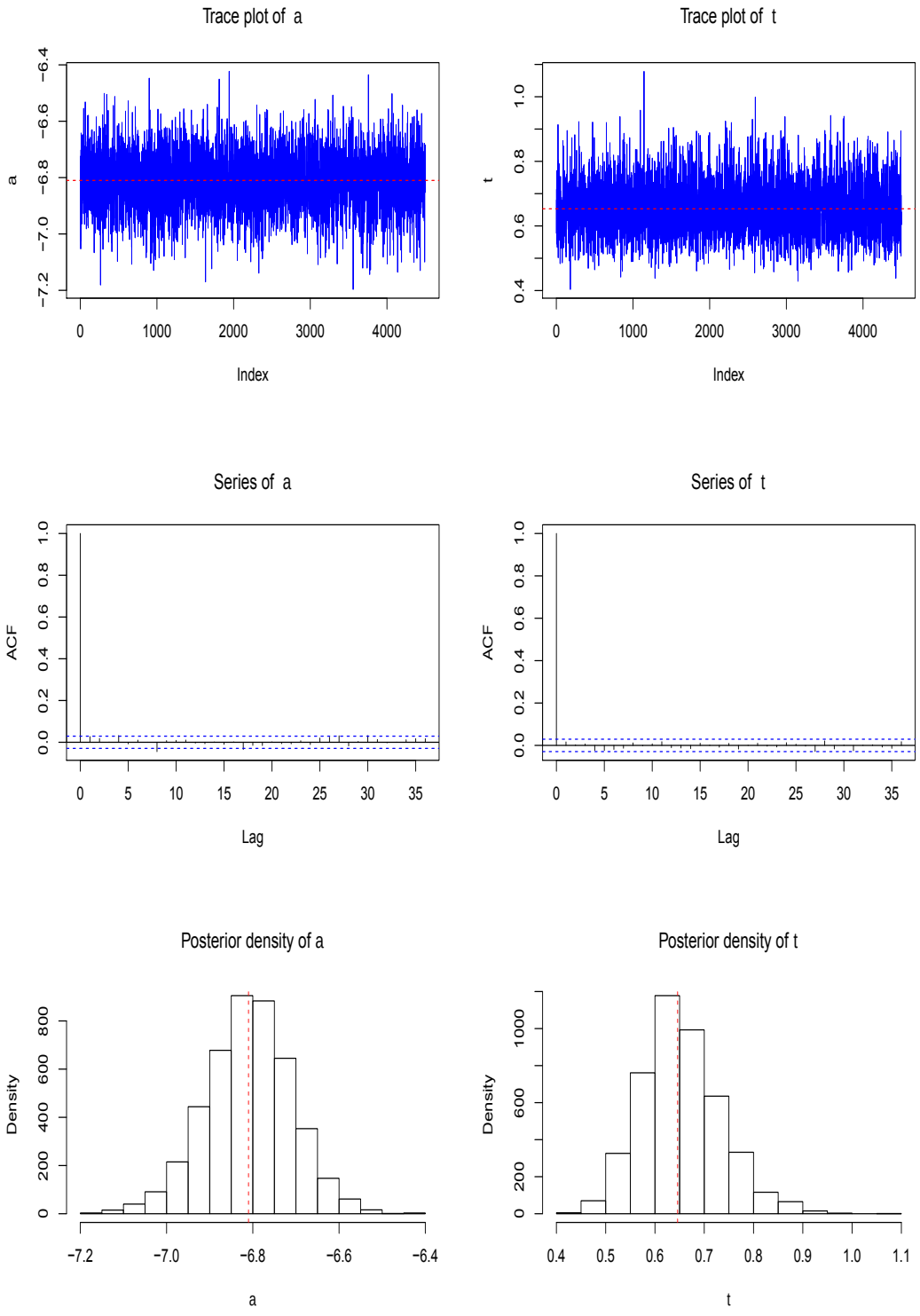


Figure A.4: Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

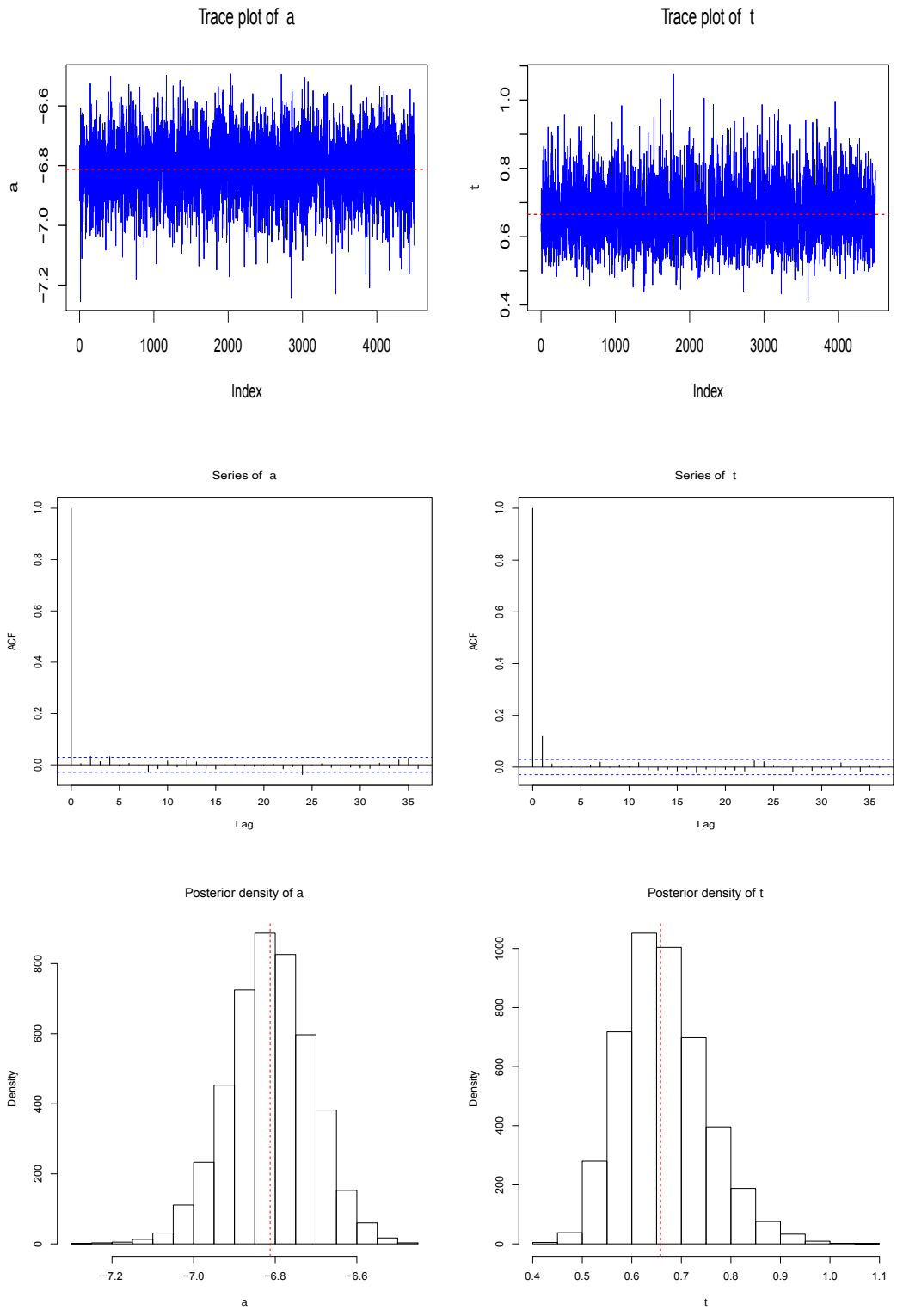


Figure A.5: Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

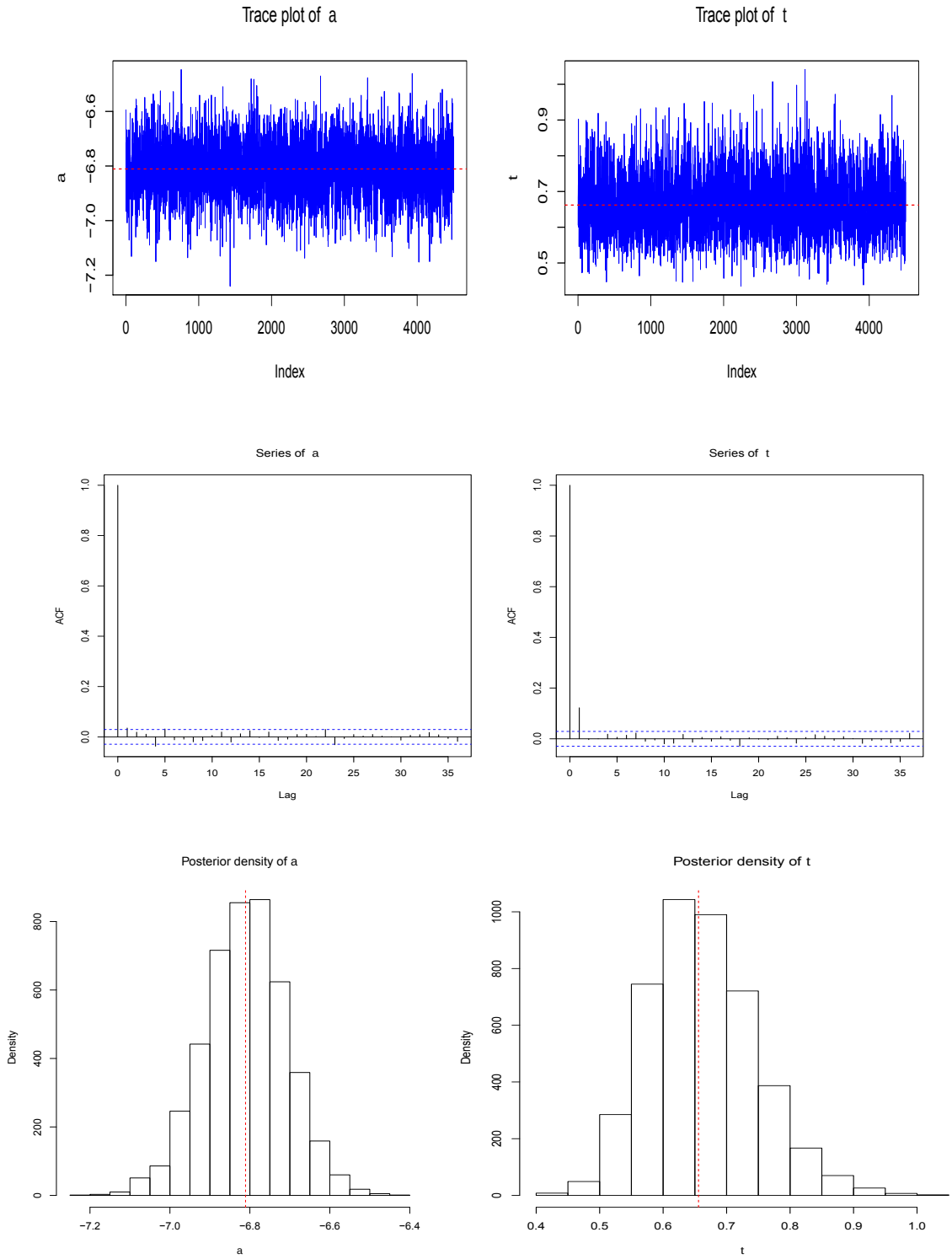


Figure A.6: Trace plots, ACF functions and histograms of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 with prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ with 5,000 samples discarded burned-in from 50,000 samples and the horizontal red dashed line in the trace plots shows the posterior mean. The graphs in the second row show the ACF functions of the parameters α and τ. The graphs in the third row show the histograms of the parameters α and τ. The vertical red dashed line shows the posterior mean.

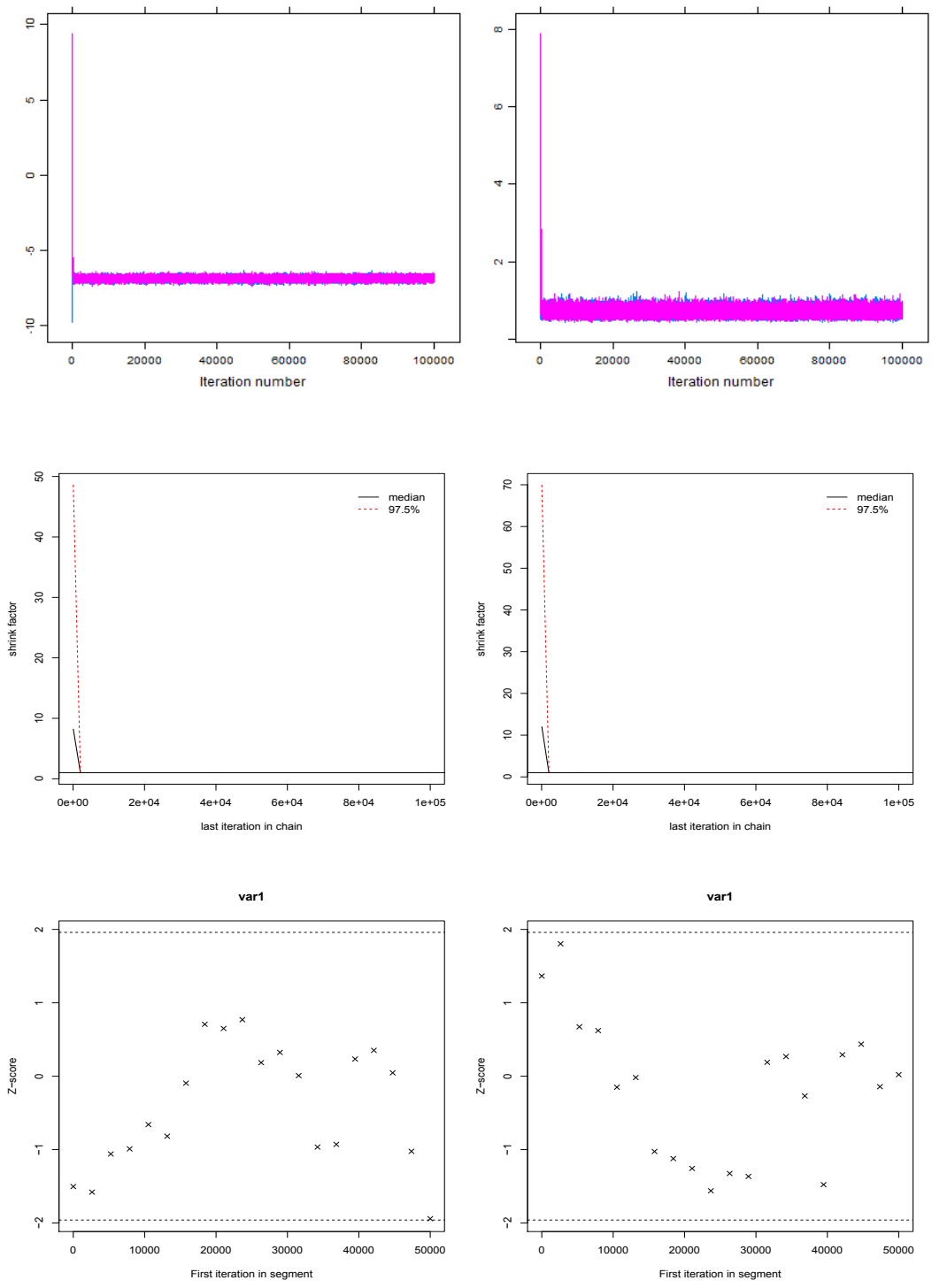


Figure A.7: The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

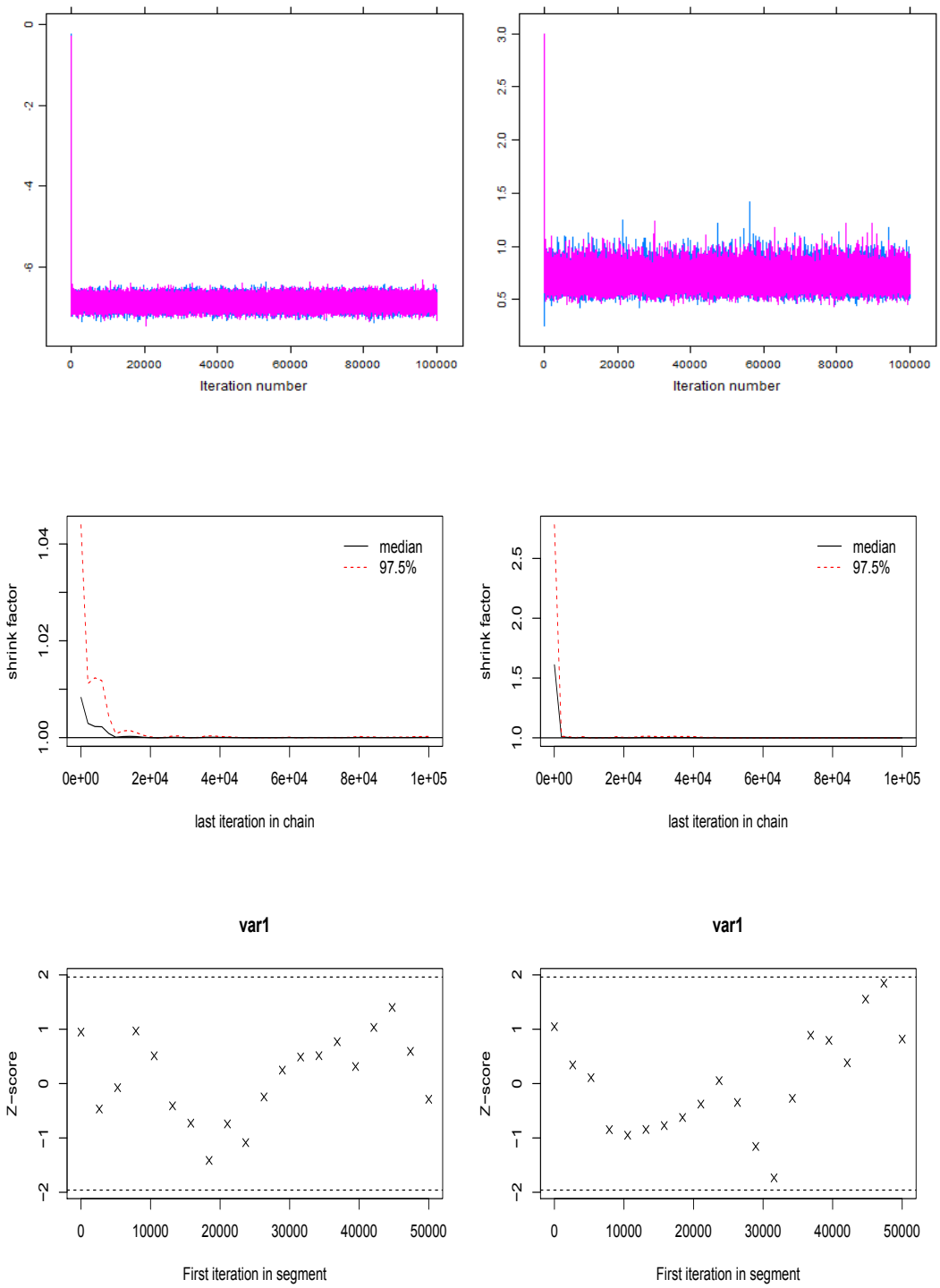


Figure A.8: The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \mathrm{N}(0, 100)$ and $\tau \sim \mathrm{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of th Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ± 1.96.

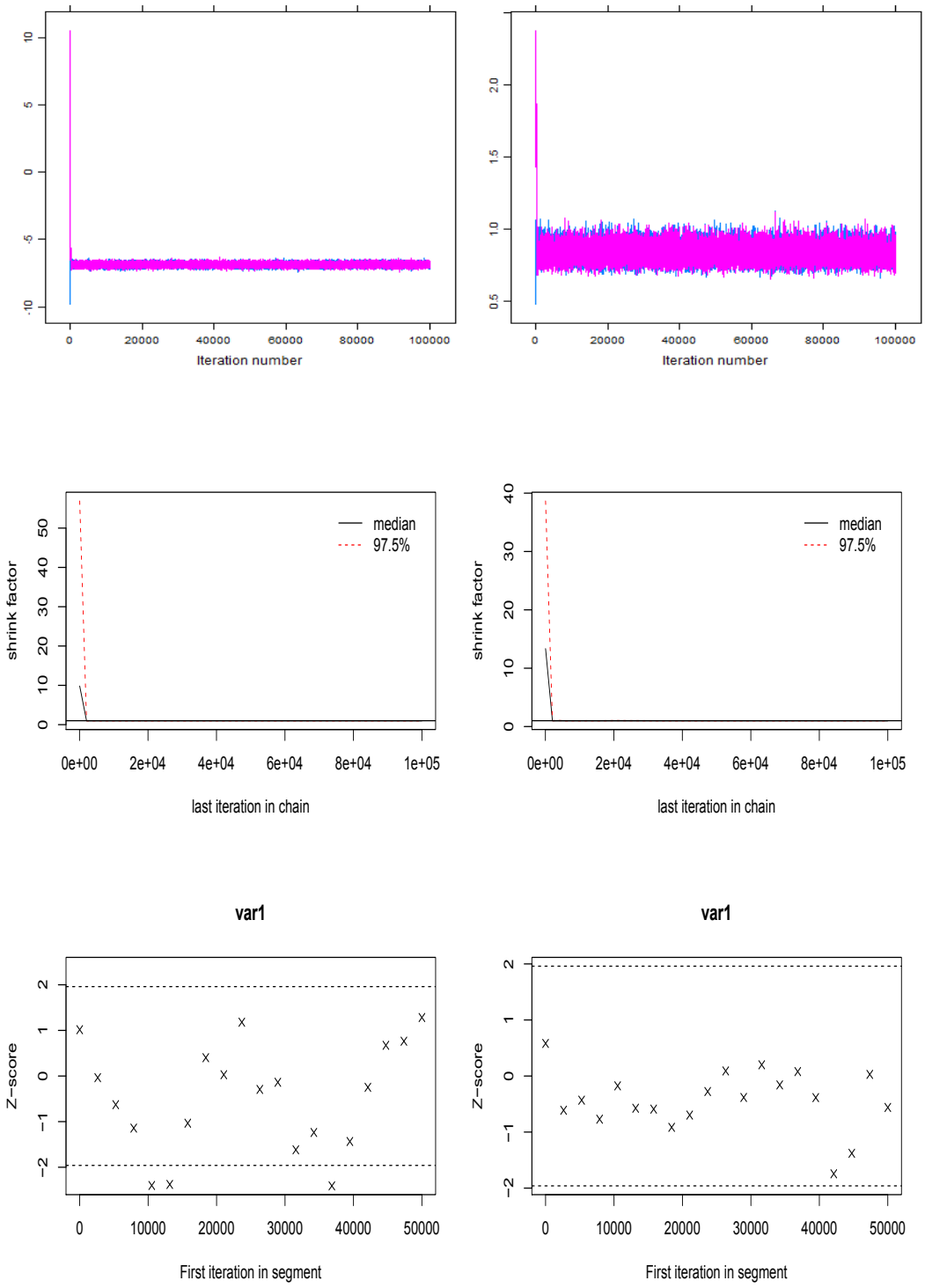
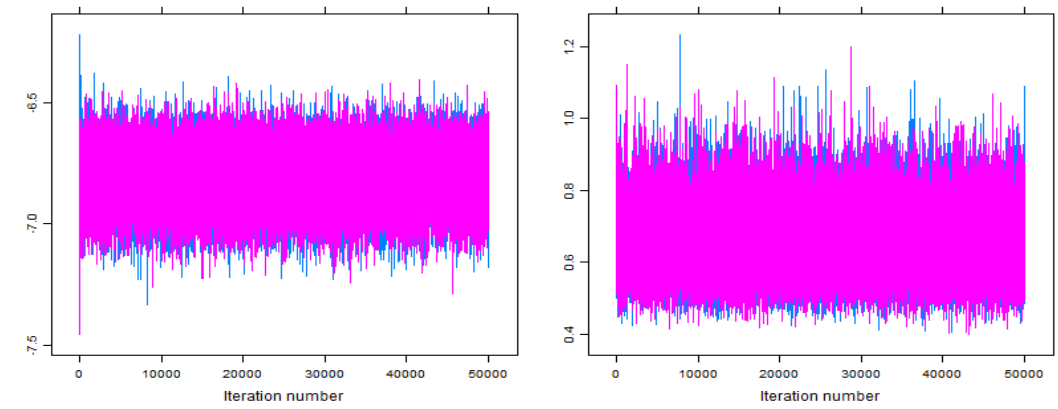
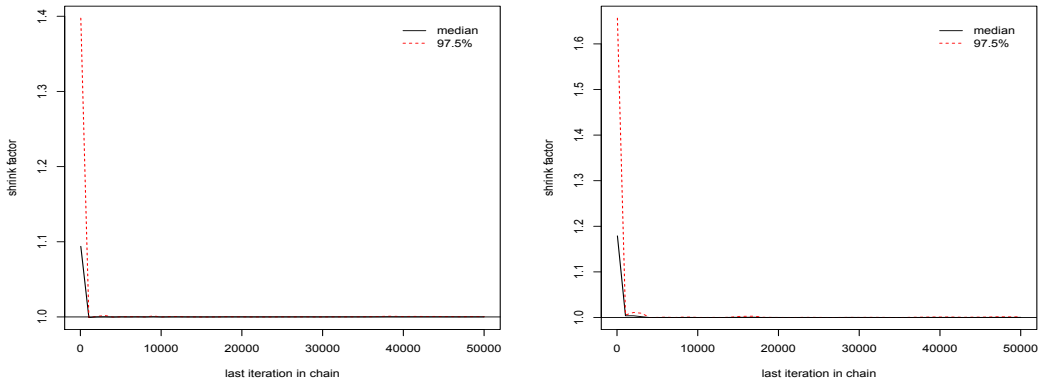


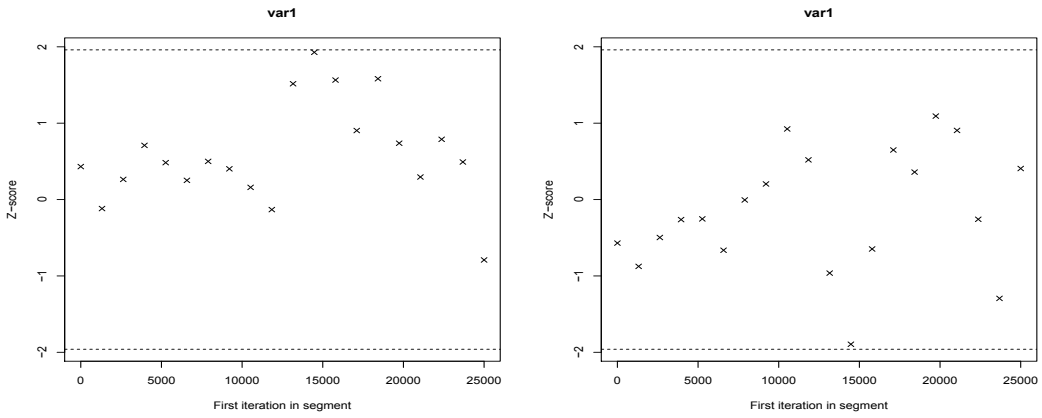
Figure A.9: The diagnostic convergence graphs of posterior parameters of the fully Bayesian hierarchical model. The model is fitted using algorithm 3.1 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 50,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.



(a) (b)



(c) (d)



(e) (f)

Figure A.10: The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

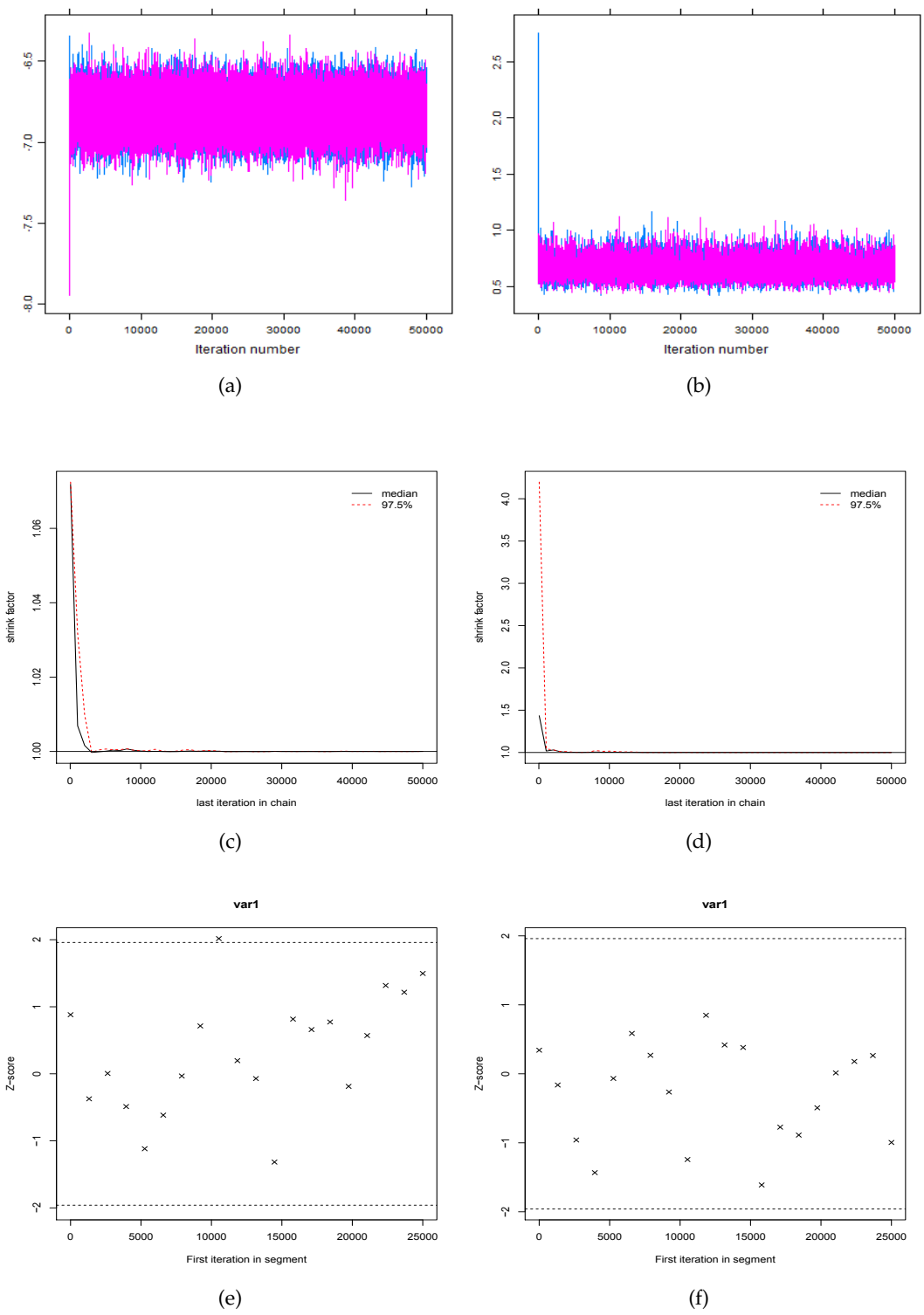
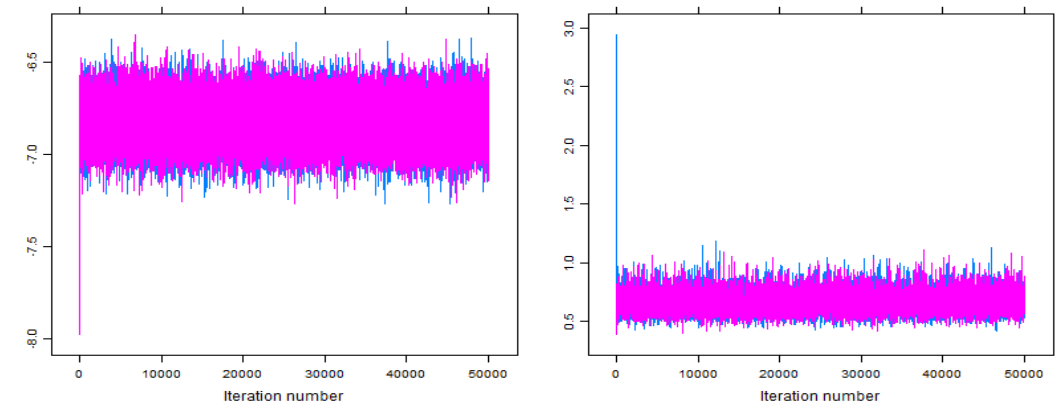
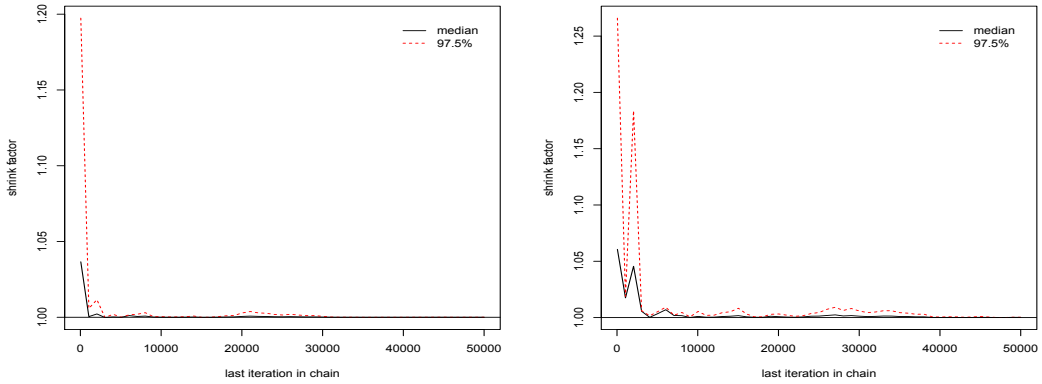


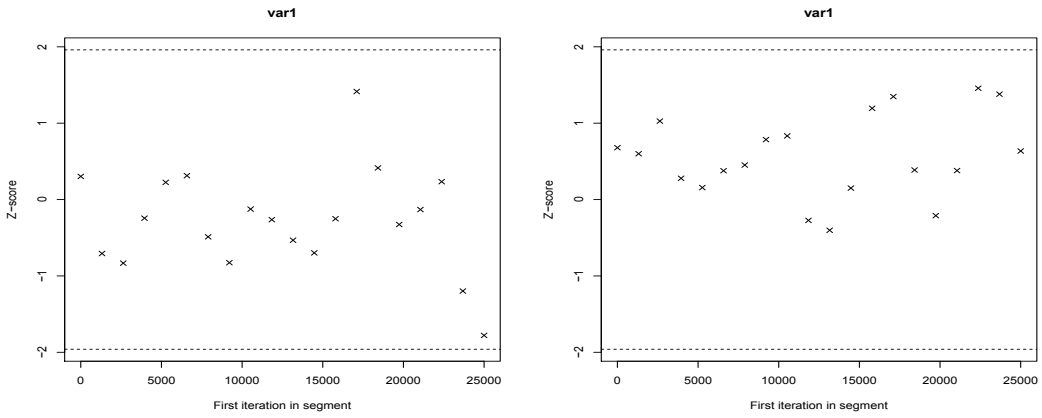
Figure A.11: The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the potential scale reduction factor (PSRF) of the Gelman-Rubin diagnostic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ± 1.96.



(a) (b)

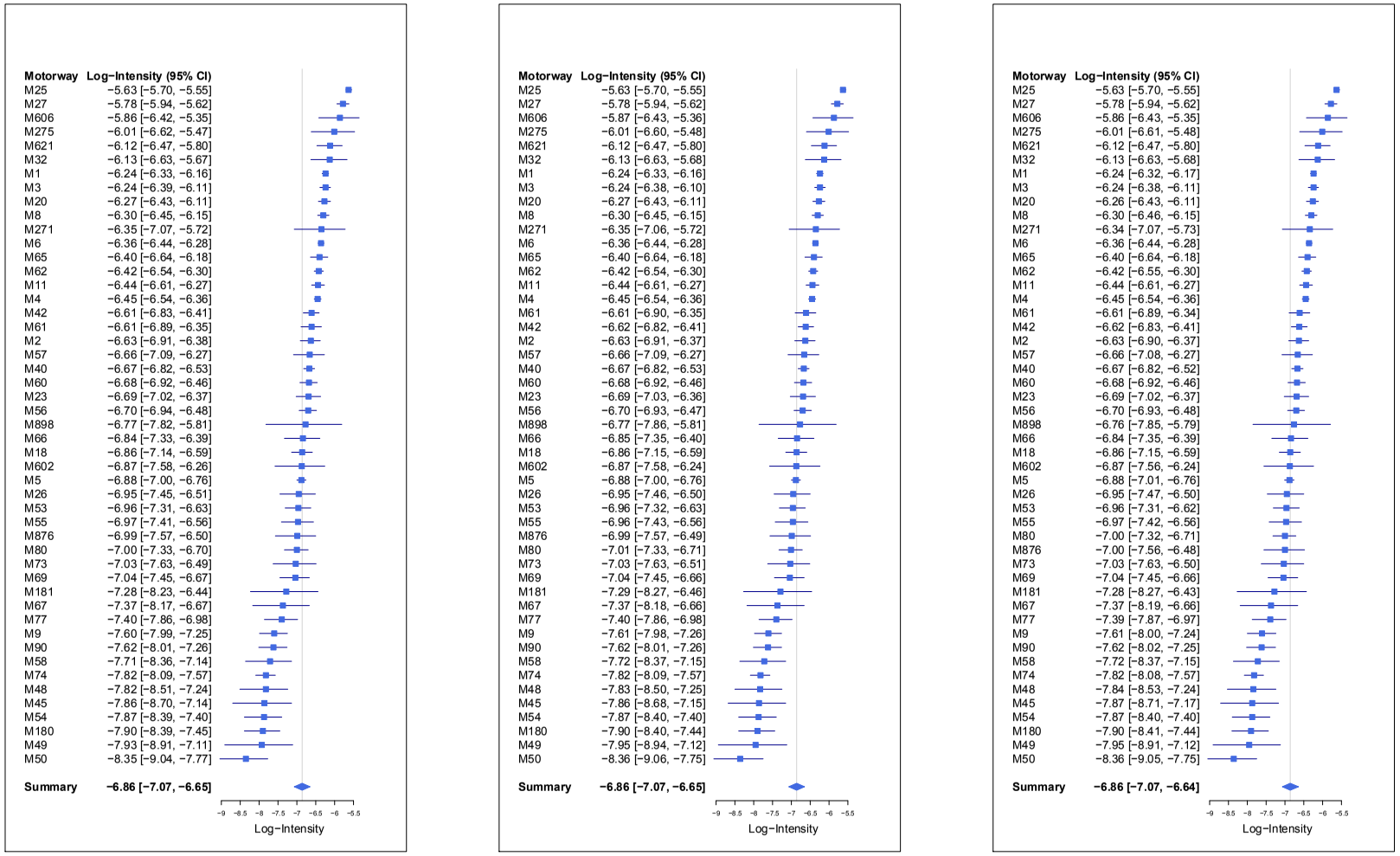


(c) (d)



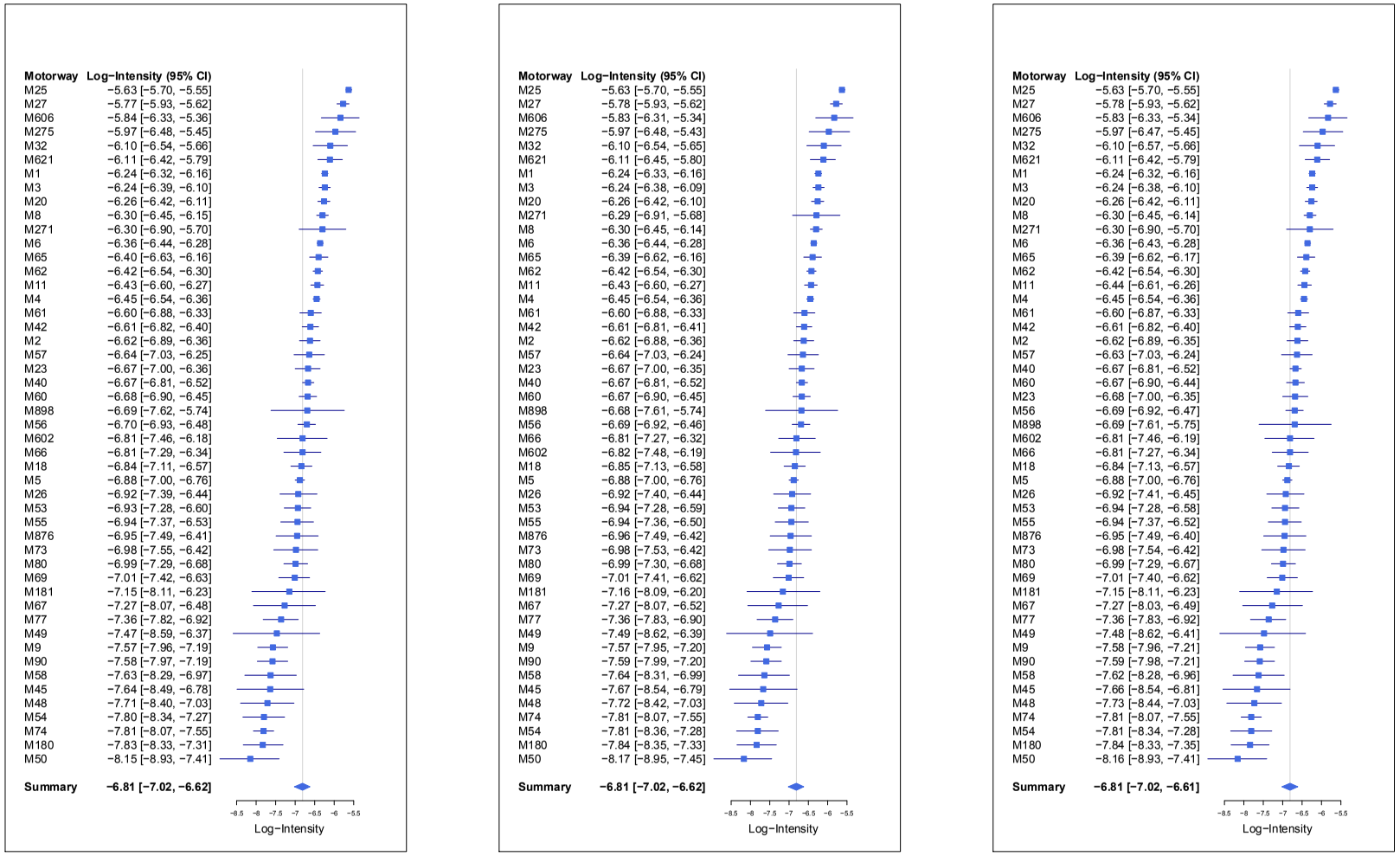
(e) (f)

Figure A.12: The diagnostic convergence graphs of posterior parameters of the semi-Bayesian hierarchical model. The model is fitted using algorithm 3.2 under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plot is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plot is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black solid and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor to stabilize around value of 1 for the last 25,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.



(a) Inv-Gamma(0.1, 0.1) (b) HN(0, 0.02) (c) Unif(0, 10^2)

Figure A.13: Results from one-stage fully Bayesian hierarchical model analysis of observed accidents data on the 49 motorways in the UK for year 2016. Prior distributions are Norm(0, 10^2) of α and various prior distributions of the heterogeneity that are $\tau^2 \sim$ Inv-Gamma(0.1, 0.1), $\tau \sim$ HN(0, 0.02) and $\tau \sim$ Unif(0, 10^2). Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote corresponding credible intervals and the solid vertical line represent the estimate of the overall log accident intensity α.



(a) Inv-Gamma(0.1, 0.1) (b) HN(0, 0.02) (c) Unif(0, 10^2)

Figure A.14: Results from two-stage semi-Bayesian hierarchical model analysis of observed accidents data on the 49 motorways in the UK for year 2016. Prior distributions are Norm(0, 10^2) of α and various prior distributions of the heterogeneity that are $\tau^2 \sim$ Inv-Gamma(0.1, 0.1), $\tau \sim$ HN(0, 0.02) and $\tau \sim$ Unif(0, 10^2). Square shapes represent means/point estimates of α_i, $i = 1, ..., m$ and the diamond shape is used to represent the mean/point estimate of the overall log accident intensity α. Horizontal lines denote corresponding credible intervals and the solid vertical line represent the estimate of the overall log accident intensity α.

Appendix B

Derivations and Plots of Chapter 4

B.1 Likelihood Function

$$
\begin{aligned}
L\left(\mathbf{N}|\Theta\right) &= P\left(\mathbf{N}|\boldsymbol{\gamma}\right) \times P\left(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \boldsymbol{\tau}^2\right) \times P\left(\boldsymbol{\alpha}|\alpha, \tau^2\right), \\
&= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{\left(\lambda_{ij} L_{ij}\right)^{n_{ij}} \exp\left(-\lambda_{ij} L_{ij}\right)}{n_{ij}!} \times \frac{1}{L_{ij}^{n_{ij}}} \\
&\times \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(-\frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \\
&\times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right), \\
&\propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{L_{ij}^{n_{ij}} \exp\left(n_{ij}\alpha_{ij}\right) \exp\left(-L_{ij} \exp\left(\alpha_{ij}\right)\right)}{n_{ij}!} \times \frac{1}{L_{ij}^{n_{ij}}} \\
&\times \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(-\frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right), \\
&\propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \exp\left(n_{ij}\alpha_{ij} - L_{ij} \exp\left(\alpha_{ij}\right)\right) \\
&\times \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(-\frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\alpha_i - \alpha)^2}{2\tau^2}\right). \qquad \text{(B.1)}
\end{aligned}
$$

B.2 Joint Posterior Distribution

$$
\begin{aligned}
\pi\left(\Theta|\mathbf{N}\right) &= P\left(\mathbf{N}|\gamma\right) P\left(\gamma|\boldsymbol{\alpha}, \boldsymbol{\tau}^2\right) P\left(\boldsymbol{\alpha}|\alpha, \tau^2\right) \prod_{i=1}^{m} P\left(\tau_i^2\right) P\left(\alpha\right) P\left(\tau^2\right), \\
&\propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \exp\left(n_{ij}\alpha_{ij} - L_{ij} \exp\left(\alpha_{ij}\right)\right) \\
&\times \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(-\frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right) \\
&\times \prod_{i=1}^{m} \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\tau_i^2\right)^{-a_0-1} \exp\left(\frac{-b_0}{\tau_i^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-\left(\alpha - \mu_0\right)^2}{2\sigma_0^2}\right) \\
&\times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\tau^2\right)^{-\alpha_0-1} \exp\left(\frac{-\beta_0}{\tau^2}\right). \qquad \text{(B.2)}
\end{aligned}
$$

B.3 Full Conditional Posterior Distributions

B.3.1 Conditional Posterior Distribution of α_i

Using equation (B.2), the conditional posterior density of α_i is calculated given other parameters:

$$
\begin{aligned}
\pi\left(\alpha_i|\boldsymbol{\gamma}, \boldsymbol{\tau}^2, \alpha, \tau, \mathbf{N}\right) &= \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(\frac{-\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{-\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right), \\
&\propto \exp\left(-\frac{\sum_{j=1}^{n_i}\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2} - \frac{\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right), \\
&\propto \exp\left(-\frac{\sum_{j=1}^{n_i}\left(\alpha_i - \alpha_{ij}\right)^2}{2\tau_i^2} - \frac{\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_i^2}\sum_{j=1}^{n_i}\left(\alpha_i - \alpha_{ij}\right)^2 + \frac{1}{\tau^2}\left(\alpha_i - \alpha\right)^2\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_i^2}\sum_{j=1}^{n_i}\left(\alpha_i^2 - 2\alpha_{ij}\alpha_i + \alpha_{ij}^2\right) + \frac{1}{\tau^2}\left(\alpha_i^2 - 2\alpha\alpha_i + \alpha^2\right)\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_i^2}\sum_{j=1}^{n_i}\left(\alpha_i^2 - 2\alpha_{ij}\alpha_i\right) + \frac{1}{\tau^2}\left(\alpha_i^2 - 2\alpha\alpha_i\right)\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left(\left(\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}\right)\alpha_i^2 - 2\alpha_i\left(\frac{\sum_{j=1}^{n_i}\alpha_{ij}}{\tau_i^2} + \frac{\alpha}{\tau^2}\right)\right)\right). \qquad \text{(B.3)}
\end{aligned}
$$

Multiplying and dividing by $\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}$ and letting $\mu_{\alpha_i} = \dfrac{\frac{\sum_{j=1}^{n_i} \alpha_{ij}}{\tau_i^2} + \frac{\alpha}{\tau^2}}{\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}}$ produces

$$\pi\left(\alpha_i|\gamma, \boldsymbol{\tau}^2, \alpha, \tau, \mathbf{N}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}\right)\left(\alpha_i^2 - 2\alpha_i\mu_{\alpha_i}\right)\right). \tag{B.4}$$

Completing square to obtain the mean for the normal distribution is formed by summing and subtracting with μ_{α_i}

$$\begin{aligned}\pi\left(\alpha_i|\gamma, \boldsymbol{\tau}^2, \alpha, \tau, \mathbf{N}\right) &\propto \exp\left(-\frac{1}{2}\left(\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}\right)\left(\alpha_i^2 - 2\alpha_i\mu_{\alpha_i} + \mu_{\alpha_i}^2\right)\right), \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}\right)(\alpha_i - \mu_{\alpha_i})^2\right).\end{aligned} \tag{B.5}$$

Therefore, $\alpha_i|\gamma, \boldsymbol{\tau}^2, \alpha, \tau, \mathbf{N}$ follows the normal distribution with mean μ_{α_i} and corresponding variance $\sigma_{\alpha_i}^2$

$$\alpha_i \sim \mathrm{N}(\mu_{\alpha_i}, \sigma_{\alpha_i}^2), \tag{B.6}$$

where

$$\mu_{\alpha_i} = \frac{\frac{\sum_{j=1}^{n_i} \alpha_{ij}}{\tau_i^2} + \frac{\alpha}{\tau^2}}{\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}} \text{ and } \sigma_{\alpha_i}^2 = \frac{1}{\frac{n_i}{\tau_i^2} + \frac{1}{\tau^2}}.$$

B.3.2 Conditional Posterior Distribution of τ_i^2

$$\begin{aligned}\pi\left(\tau_i^2|\gamma, \boldsymbol{\alpha}, \mathbf{N}\right) &= \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left(-\frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \frac{b_0^{a_0}}{\Gamma(a_0)}\left(\tau_i^2\right)^{-a_0-1} \exp\left(\frac{-b_0}{\tau_i^2}\right), \\ &\propto \left(\tau_i^2\right)^{-\frac{n_i}{2}} \exp\left(-\sum_{j=1}^{n_i} \frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2\tau_i^2}\right) \times \frac{b_0^{a_0}}{\Gamma(a_0)}\left(\tau_i^2\right)^{-a_0-1} \exp\left(\frac{-b_0}{\tau_i^2}\right), \\ &\propto \left(\tau_i^2\right)^{-(\frac{n_i}{2}+a_0)-1} \exp\left(-\frac{b_0 + \sum_{j=1}^{n_i} \frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2}}{\tau_i^2}\right).\end{aligned} \tag{B.7}$$

So the posterior distribution of τ_i^2 given other parameters is

$$\tau_i^2 \sim \text{Inv-Gamma}\left(a_{\tau_i^2} = \frac{n_i}{2} + a_0, b_{\tau_i^2} = \sum_{j=1}^{n_i} \frac{\left(\alpha_{ij} - \alpha_i\right)^2}{2} + b_0\right). \tag{B.8}$$

B.3.3 Conditional Posterior Distribution of α

Using equation (B.2), the conditional posterior density of α is calculated by considering α as a random variable and other parameters as constants. Hence:

$$\begin{aligned}
\pi\left(\alpha|\boldsymbol{\alpha}, \tau^2, \mathbf{N}\right) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{-\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-\left(\alpha - \mu_0\right)^2}{2\sigma_0^2}\right), \\
&\propto \exp\left(-\frac{\sum_{i=1}^{m}\left(\alpha_i - \alpha\right)^2}{2\tau^2}\right) \times \exp\left(\frac{-\left(\alpha - \mu_0\right)^2}{2\sigma_0^2}\right), \\
&\propto \exp\left(-\frac{\sum_{i=1}^{m}\left(\alpha_i - \alpha\right)^2}{2\tau^2} - \frac{\left(\alpha - \mu_0\right)^2}{2\sigma_0^2}\right), \\
&\propto \exp\left(-\frac{\sum_{i=1}^{m}\left(\alpha - \alpha_i\right)^2}{2\tau^2} - \frac{\left(\alpha - \mu_0\right)^2}{2\sigma_0^2}\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\sum_{i=1}^{m}\left(\alpha - \alpha_i\right)^2}{\tau^2} + \frac{\left(\alpha - \mu_0\right)^2}{\sigma_0^2}\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\sum_{i=1}^{m}\left(\alpha^2 - 2\alpha_i\alpha + \alpha_i^2\right)}{\tau^2} + \frac{\left(\alpha^2 - 2\mu_0\alpha + \mu_0^2\right)}{\sigma_0^2}\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\sum_{i=1}^{m}\left(\alpha^2 - 2\alpha_i\alpha\right)}{\tau^2} + \frac{\left(\alpha^2 - 2\mu_0\alpha\right)}{\sigma_0^2}\right]\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\alpha^2\left(\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}\right) - 2\alpha\left(\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2} + \frac{\mu_0}{\sigma_0^2}\right)\right]\right).
\end{aligned} \tag{B.9}$$

Multiplying and dividing by $\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}$ and letting $\mu_\alpha = \dfrac{\frac{\sum_{i=1}^{m}\alpha_i}{\tau^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}}$ produces

$$\pi\left(\alpha|\boldsymbol{\alpha}, \tau^2, \mathbf{N}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2} + \frac{1}{\sigma_0^2}\right)\left(\alpha^2 - 2\alpha\mu_\alpha\right)\right). \tag{B.10}$$

Completing square to obtain the mean for the normal distribution is formed by summing and subtracting with μ_α

$$\begin{aligned}\pi\left(\alpha|\boldsymbol{\alpha},\tau^2,\mathbf{N}\right) &\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)\left(\alpha^2-2\alpha\mu_\alpha+\mu_\alpha^2\right)\right),\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}\right)(\alpha-\mu_\alpha)^2\right).\end{aligned} \tag{B.11}$$

Therefore, $\alpha|\boldsymbol{\alpha},\tau^2,\mathbf{N}$ follows the normal distribution with mean μ_α and corresponding variance σ_α^2

$$\alpha \sim \mathrm{N}(\mu_\alpha,\sigma_\alpha^2), \tag{B.12}$$

where

$$\mu_\alpha = \frac{\frac{\sum_{i=1}^m \alpha_i}{\tau^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}} \text{ and } \sigma_\alpha^2 = \frac{1}{\frac{m}{\tau^2}+\frac{1}{\sigma_0^2}}.$$

B.3.4 Conditional Posterior Distribution of τ^2

Using equation (B.2), we derive the conditional posterior density of τ^2 given other parameters. Hence:

$$\begin{aligned}\pi\left(\tau^2|\boldsymbol{\alpha},\alpha,\mathbf{N}\right) &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\frac{-(\alpha_i-\alpha)^2}{2\tau^2}\right)\times\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\tau^2\right)^{-\alpha_0-1}\exp\left(\frac{-\beta_0}{\tau^2}\right),\\ &\propto \left(\tau^2\right)^{-\left(\frac{m}{2}+\alpha_0\right)-1}\exp\left(-\frac{\sum_{i=1}^m(\alpha_i-\alpha)^2}{2\tau^2}-\frac{\beta_0}{\tau^2}\right),\\ &\propto \left(\tau^2\right)^{-\left(\frac{m}{2}+\alpha_0\right)-1}\exp\left(-\frac{\beta_0+\frac{\sum_{i=1}^m(\alpha_i-\alpha)^2}{2}}{\tau^2}\right).\end{aligned} \tag{B.13}$$

Therefore, $\tau^2|\boldsymbol{\alpha},\alpha,\boldsymbol{\tau}^2,\mathbf{N}$ follows the inverse gamma distribution with shape $a_{\tau^2}=\frac{m}{2}+\alpha_0$ and scale $b_{\tau^2}=\frac{\sum_{i=1}^m(\alpha_i-\alpha)^2}{2}+\beta_0$.

B.4 Frequentist Estimation

The second part of the likelihood function is given by,

$$L_i\left(\boldsymbol{\alpha},\boldsymbol{\tau}^2;\hat{\boldsymbol{\gamma}}\right) = \prod_{i=1}^m\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\tau_i^2}}\exp\left(-\frac{\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right). \tag{B.14}$$

The first and the second derivatives of the log-likelihood function with respect to α_i and τ_i^2 are given by,

$$\ell_i\left(\boldsymbol{\alpha}, \boldsymbol{\tau}^2; \hat{\gamma}\right) = \sum_{i=1}^{m}\left[-\frac{n_i}{2}\log\tau_i^2 - \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{2\tau_i^2}\right]. \tag{B.15}$$

$$\frac{\partial \ell_i}{\partial \alpha_i} = \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)}{\tau_i^2}. \tag{B.16}$$

$$\frac{\partial^2 \ell_i}{\partial \alpha_i^2} = \sum_{j=1}^{n_i} -\frac{1}{\tau_i^2}. \tag{B.17}$$

$$\frac{\partial \ell_i}{\partial \tau_i^2} = -\frac{n_i}{2\tau_i^2} + \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{2\tau_i^4}. \tag{B.18}$$

$$\frac{\partial^2 \ell_i}{\partial (\tau_i^2)^2} = \frac{n_i}{2\tau_i^4} - \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{\tau_i^6}. \tag{B.19}$$

$$\frac{\partial^2 \ell_i}{\partial \tau_i^2 \partial \alpha_i} = -\frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)}{\tau_i^4}. \tag{B.20}$$

$$\frac{\partial^2 \ell_i}{\partial \alpha_i \partial \tau_i^2} = -\frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)}{\tau_i^4}. \tag{B.21}$$

By equalling both $\frac{\partial \ell_i}{\partial \alpha_i}$ and $\frac{\partial \ell_i}{\partial \tau_i^2}$ to zero, maximum estimators of α_i and τ_i^2 are given by,

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n_i}\hat{\alpha}_{ij}}{n_i} \text{ and } \hat{\tau}_i^2 = \frac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\hat{\alpha}_i\right)^2}{n_i}, \tag{B.22}$$

and the Hessian matrix is given by,

$$H\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) = \begin{bmatrix} \frac{\partial^2 \ell_i}{\partial \alpha_i^2} & \frac{\partial^2 \ell_i}{\partial \alpha_i \partial \tau_i^2} \\ \frac{\partial^2 \ell_i}{\partial \tau_i^2 \partial \alpha_i} & \frac{\partial^2 \ell_i}{\partial (\tau_i^2)^2} \end{bmatrix}. \tag{B.23}$$

The Fisher information matrix is given by,

$$
\begin{aligned}
I\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) &= -E\left[H\left(\hat{\alpha}_i, \hat{\tau}_i^2\right)\right] \\
&= -\begin{bmatrix} E\left(-\dfrac{n_i}{\tau_i^2}\right) & E\left(-\dfrac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)}{\tau_i^4}\right) \\ E\left(-\dfrac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)}{\tau_i^4}\right) & E\left(\dfrac{n_i}{2\tau_i^4}-\dfrac{\sum_{j=1}^{n_i}\left(\hat{\alpha}_{ij}-\alpha_i\right)^2}{\tau_i^6}\right) \end{bmatrix},
\end{aligned} \tag{B.24}
$$

since $\alpha_{ij} \sim \mathrm{N}\left(\alpha_i, \tau_i^2\right)$. Then

$$
\begin{aligned}
\frac{\alpha_{ij}-\alpha_i}{\tau_i} &\sim \mathrm{N}(0,1). \\
\left(\frac{\alpha_{ij}-\alpha_i}{\tau_i}\right)^2 &\sim \mathrm{X}^2(1). \\
\sum_{j=1}^{n_i}\left(\frac{\alpha_{ij}-\alpha_i}{\tau_i}\right)^2 &\sim \mathrm{X}^2(n_i). \\
\sum_{j=1}^{n_i}\left(\alpha_{ij}-\alpha_i\right)^2 &\sim \tau_i^2\mathrm{X}^2(n_i),
\end{aligned} \tag{B.25}
$$

hence

$$
E\left(\sum_{j=1}^{n_i}\left(\alpha_{ij}-\alpha_i\right)^2\right) = \tau_i^2 n_i. \tag{B.26}
$$

$$
E\left(\sum_{j=1}^{n_i}\left(\alpha_{ij}-\alpha_i\right)\right) = \sum_{j=1}^{n_i}\left[E\left(\alpha_{ij}\right)-E\left(\alpha_i\right)\right] = \sum_{j=1}^{n_i}\left[\alpha_i-\alpha_i\right] = 0. \tag{B.27}
$$

Therefore the Fisher information matrix is given by,

$$
I\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) = \begin{bmatrix} \dfrac{n_i}{\tau_i^2} & 0 \\ 0 & \dfrac{n_i}{2\tau_i^4} \end{bmatrix}, \tag{B.28}
$$

and the inverse of the Fisher information matrix is given by,

$$
I^{-1}\left(\hat{\alpha}_i, \hat{\tau}_i^2\right) = \begin{bmatrix} \dfrac{\tau_i^2}{n_i} & 0 \\ 0 & \dfrac{2\tau_i^4}{n_i} \end{bmatrix}. \tag{B.29}
$$

The standard errors of $\hat{\alpha}_i$ and $\hat{\tau}_i^2$ are

$$SE\left(\hat{\alpha}_i\right) = \sqrt{\frac{\tau_i^2}{n_i}}. \tag{B.30}$$

$$SE\left(\hat{\tau}_i^2\right) = \sqrt{\frac{2\tau_i^4}{n_i}}. \tag{B.31}$$

B.5 Plots

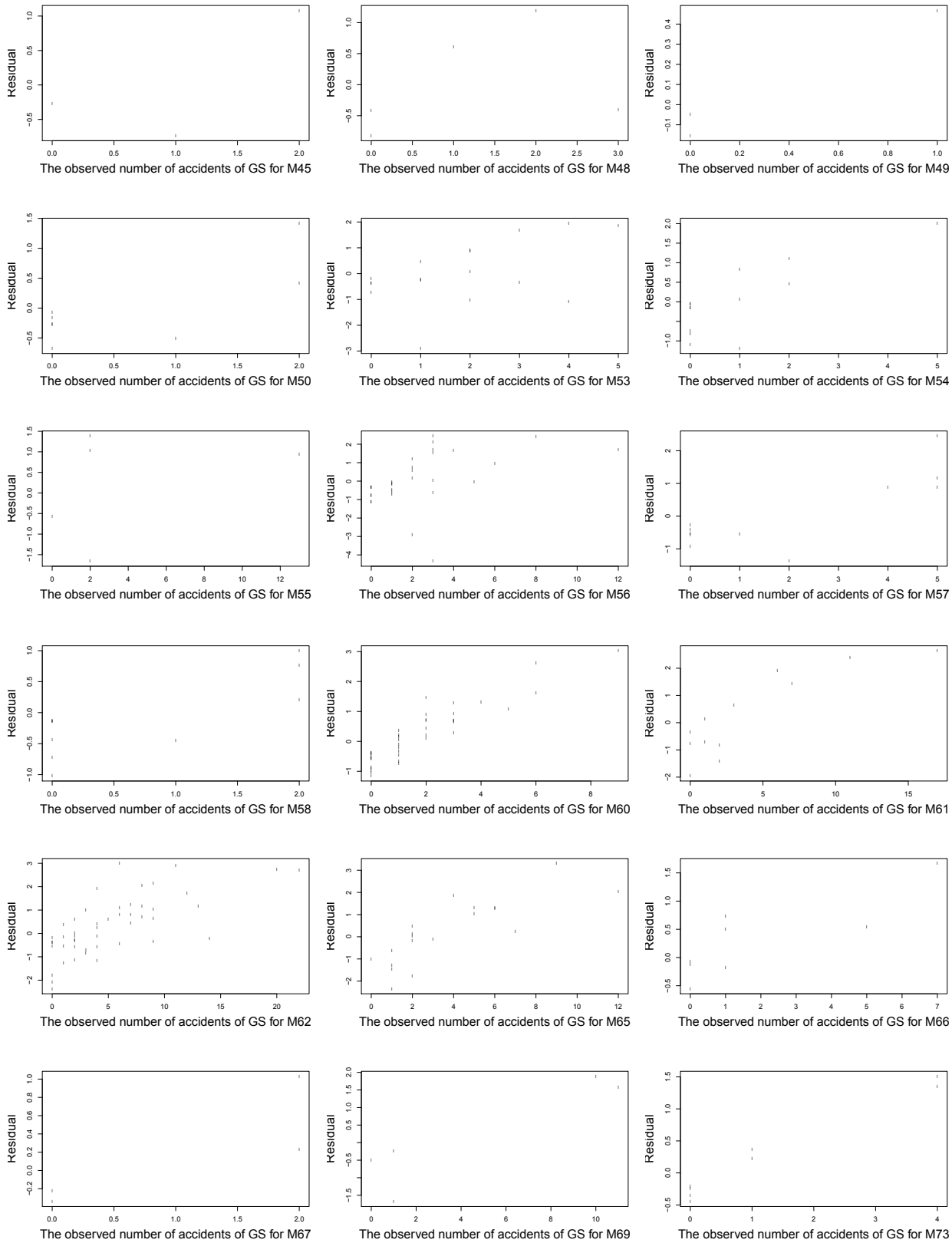


Figure B.1: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments.

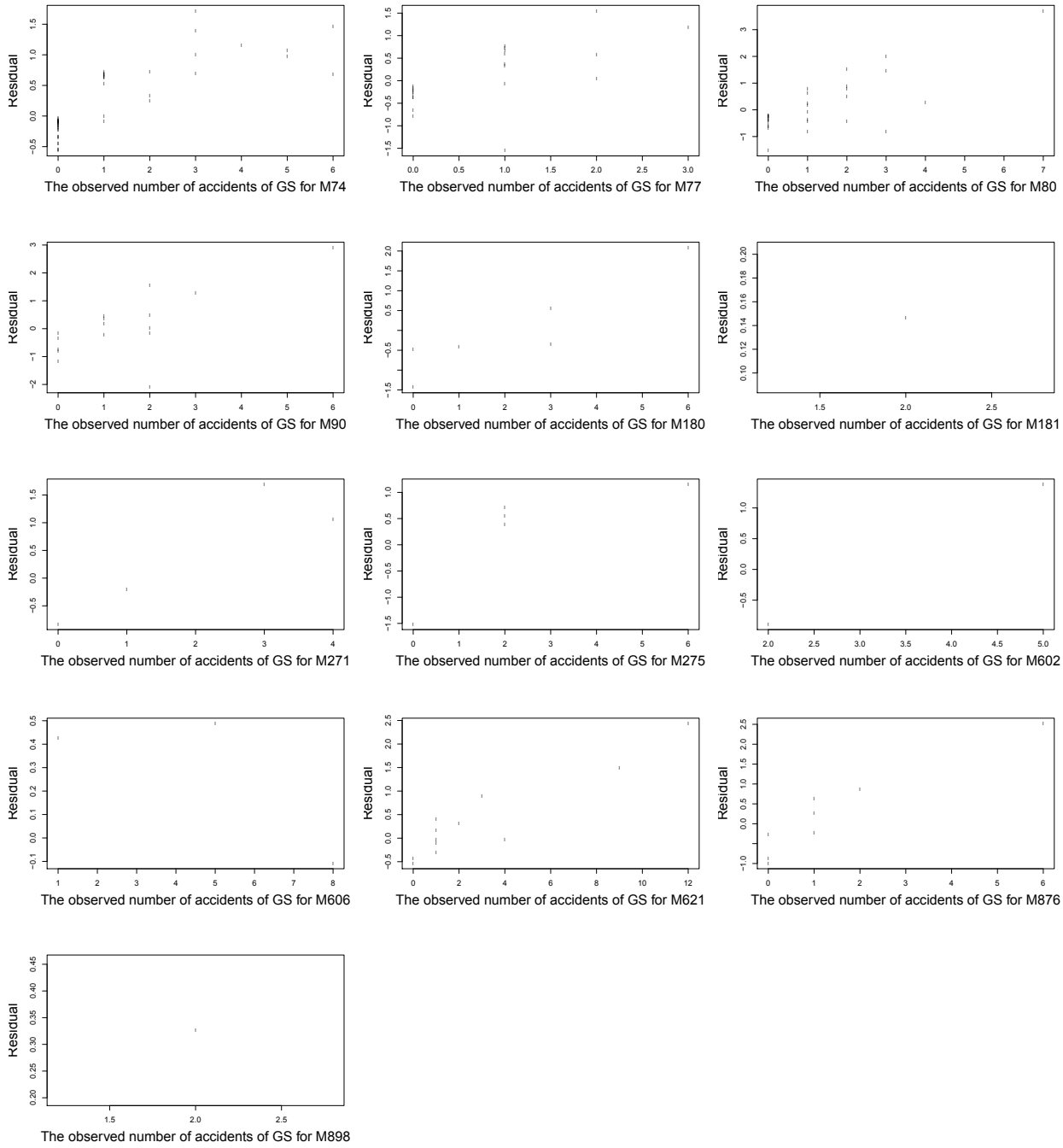


Figure B.2: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments.

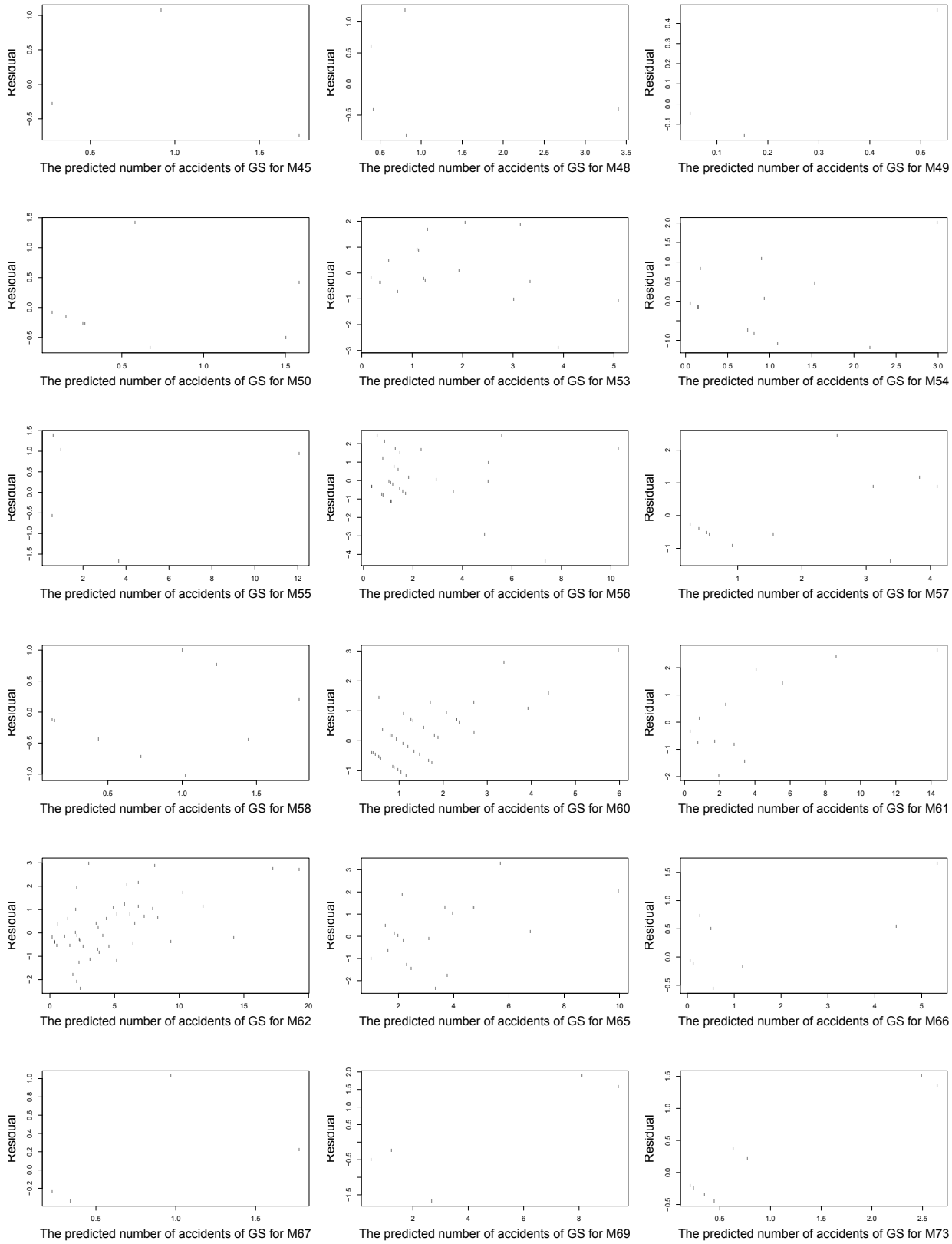


Figure B.3: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments.

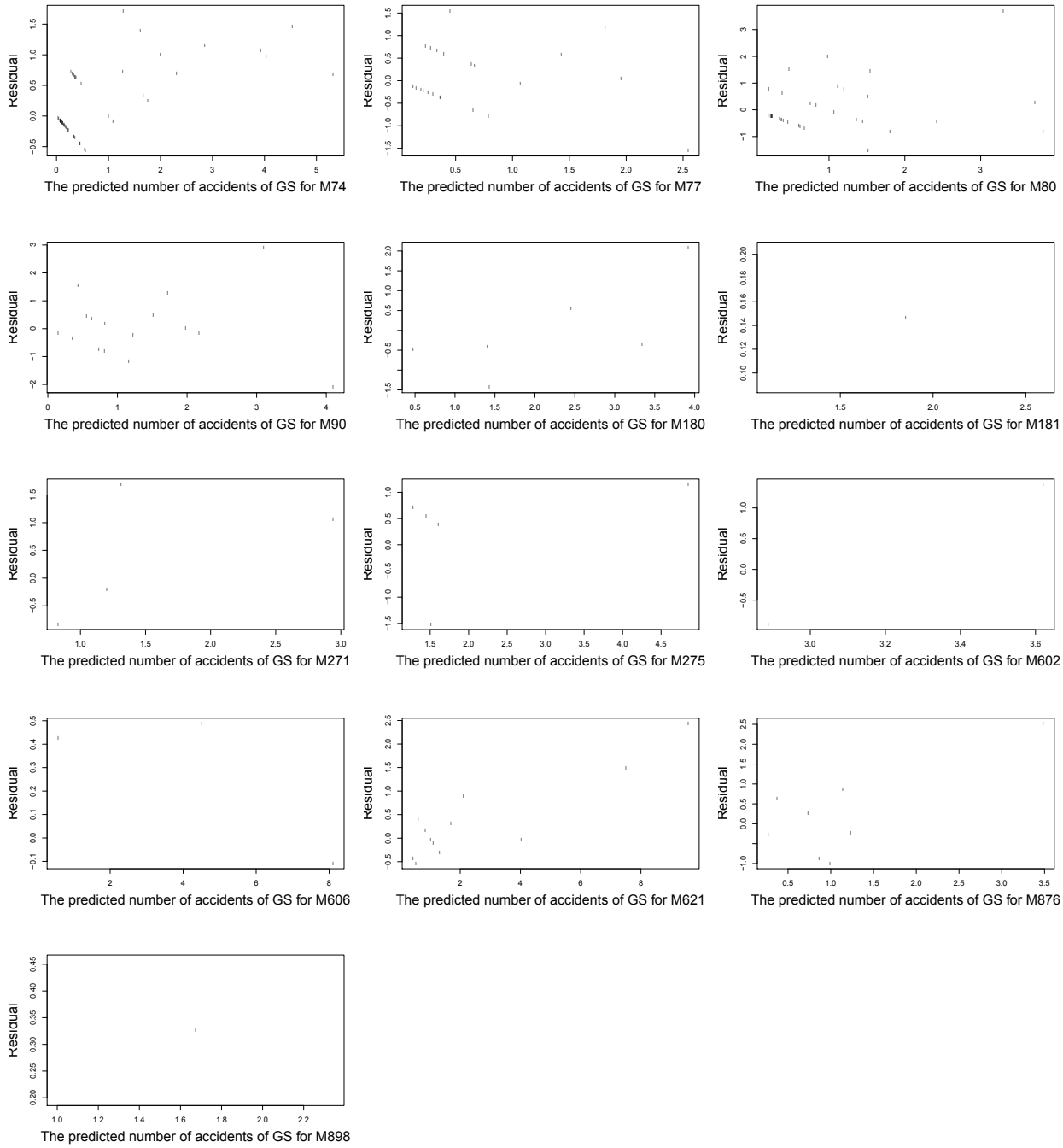


Figure B.4: Residuals plots. The predicted value of the number of accidents is calculated using the three-level Bayesian hierarchical model fitted to the traffic accidents on the UK motorway network for 2016. GS represents grouped segments.

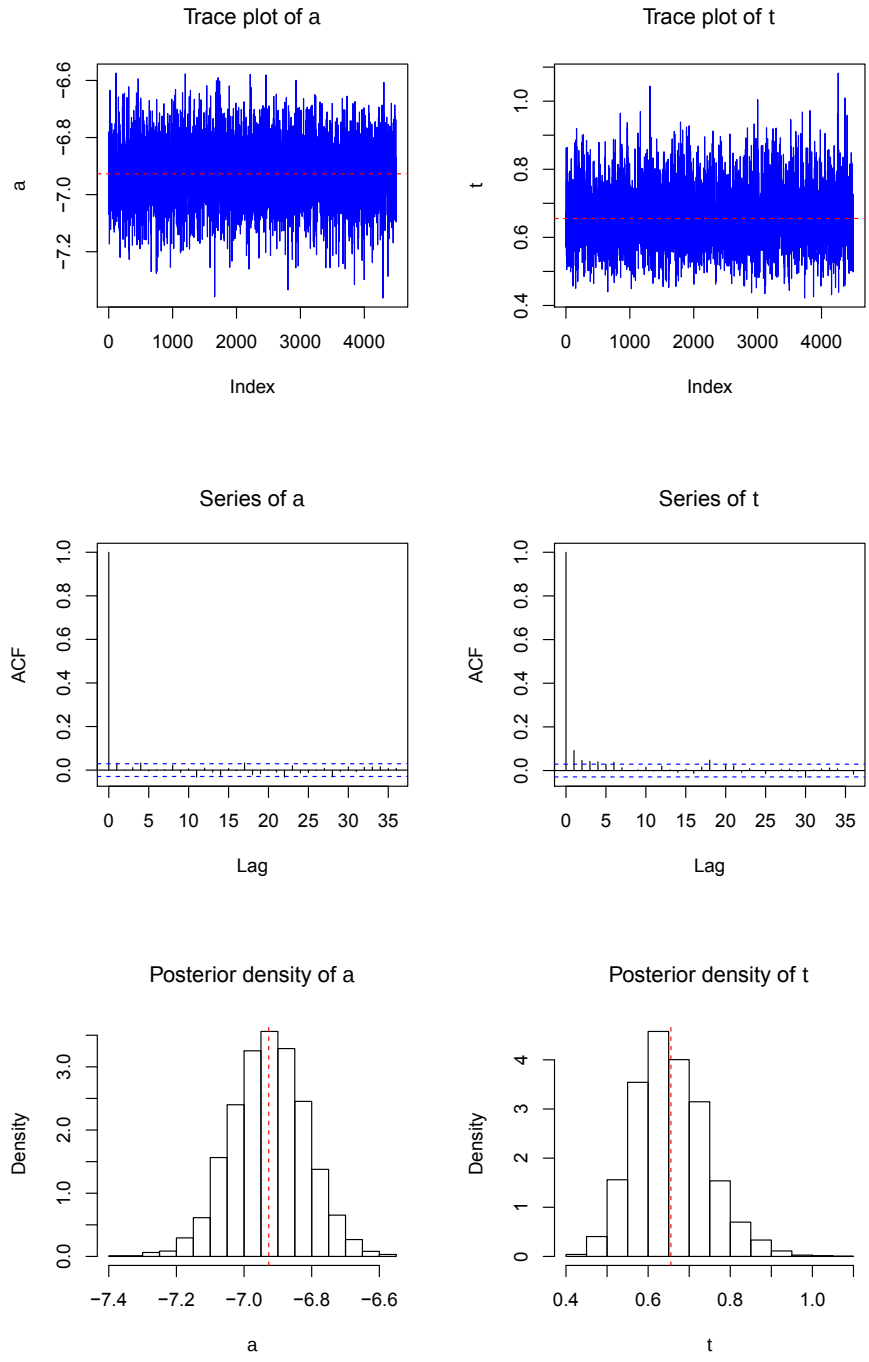


Figure B.5: The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a weakly-informative prior distributions $\tau^2 \sim$ Inv-Gamma$(0.1, 0.1)$ and $\alpha \sim$ N$(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean.

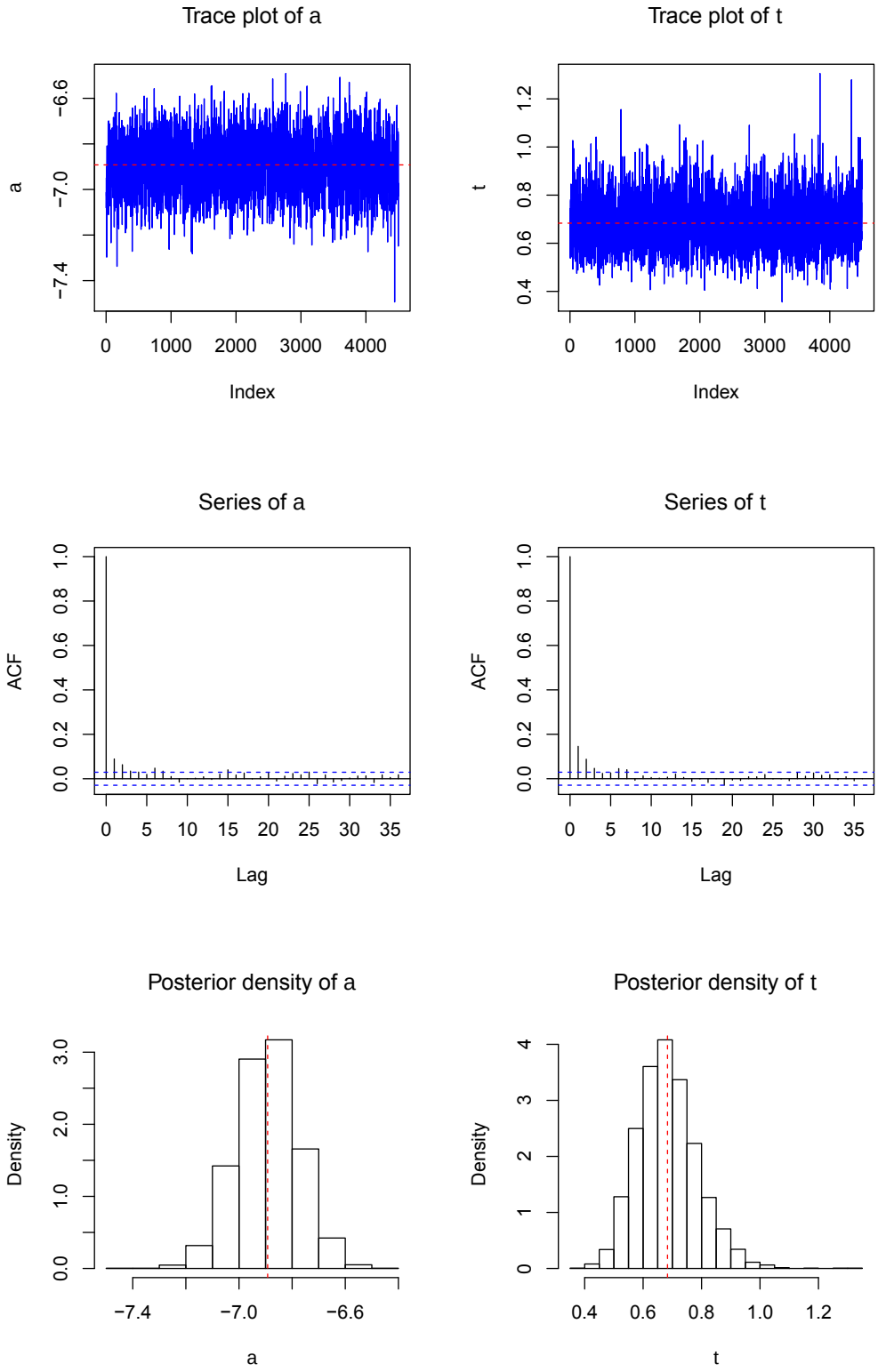


Figure B.6: The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau \sim \text{unif}(0, 100)$ and $\alpha \sim \text{N}(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean.

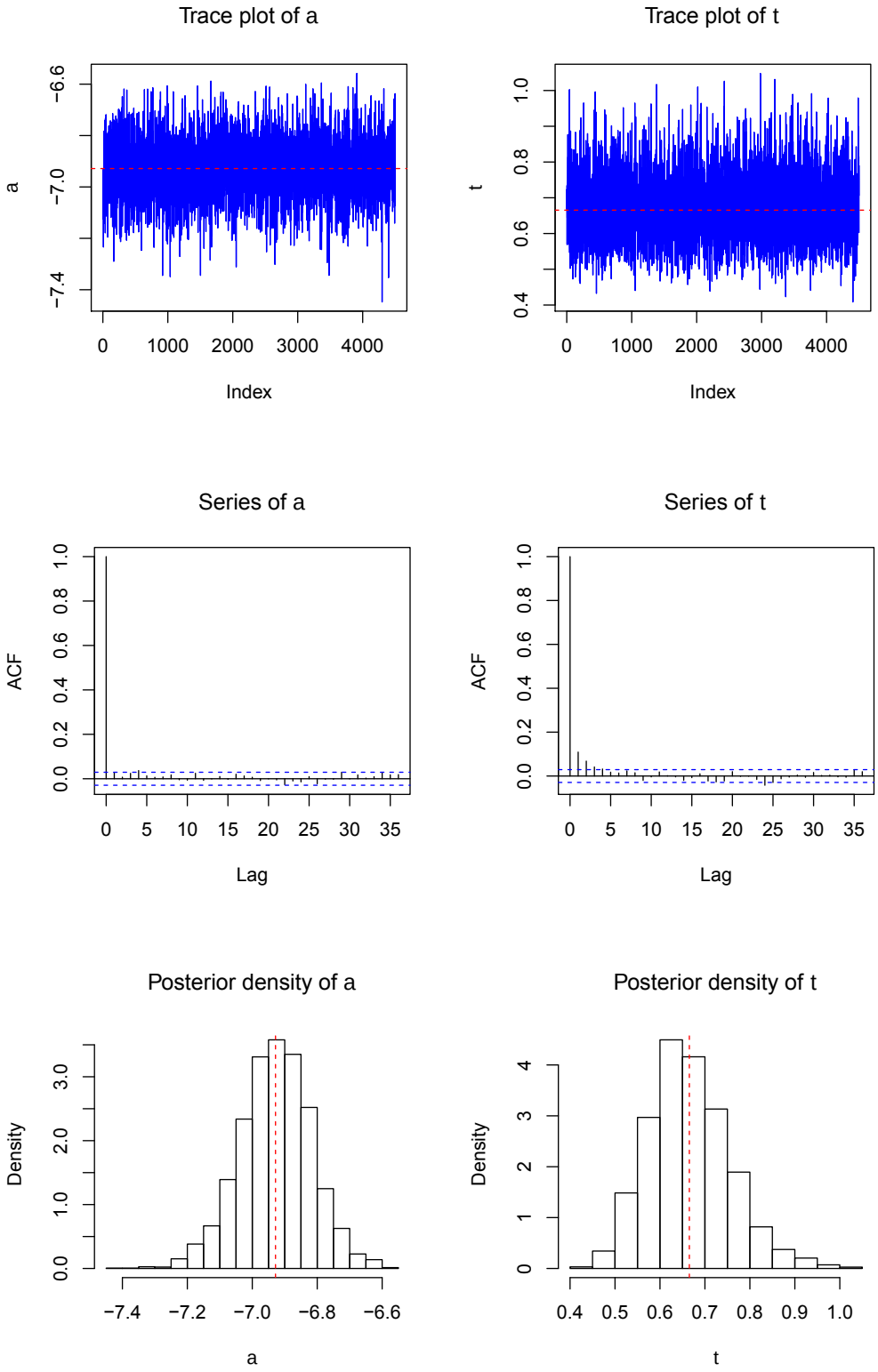


Figure B.7: The trace plots, autocorrelation function (ACF) and posterior density histgrams for three-level hierarchical model parameters α and τ under a non-informative prior distributions $\tau \sim \text{HN}(0, 0.02)$ and $\alpha \sim \text{N}(0, 100)$. 500,000 samples are generated with a burn in of 50,000 samples and a thinning 100 samples using initial values for $\alpha = 0$ and $\tau = 0.1$. The first row's graphs represent the trace plots of the parameters α and τ. The second row's graphs are the ACF of the parameters α and τ. The third row's graphs refer to the histograms of the densities of the parameters α and τ. The red dashed line is the posterior mean.

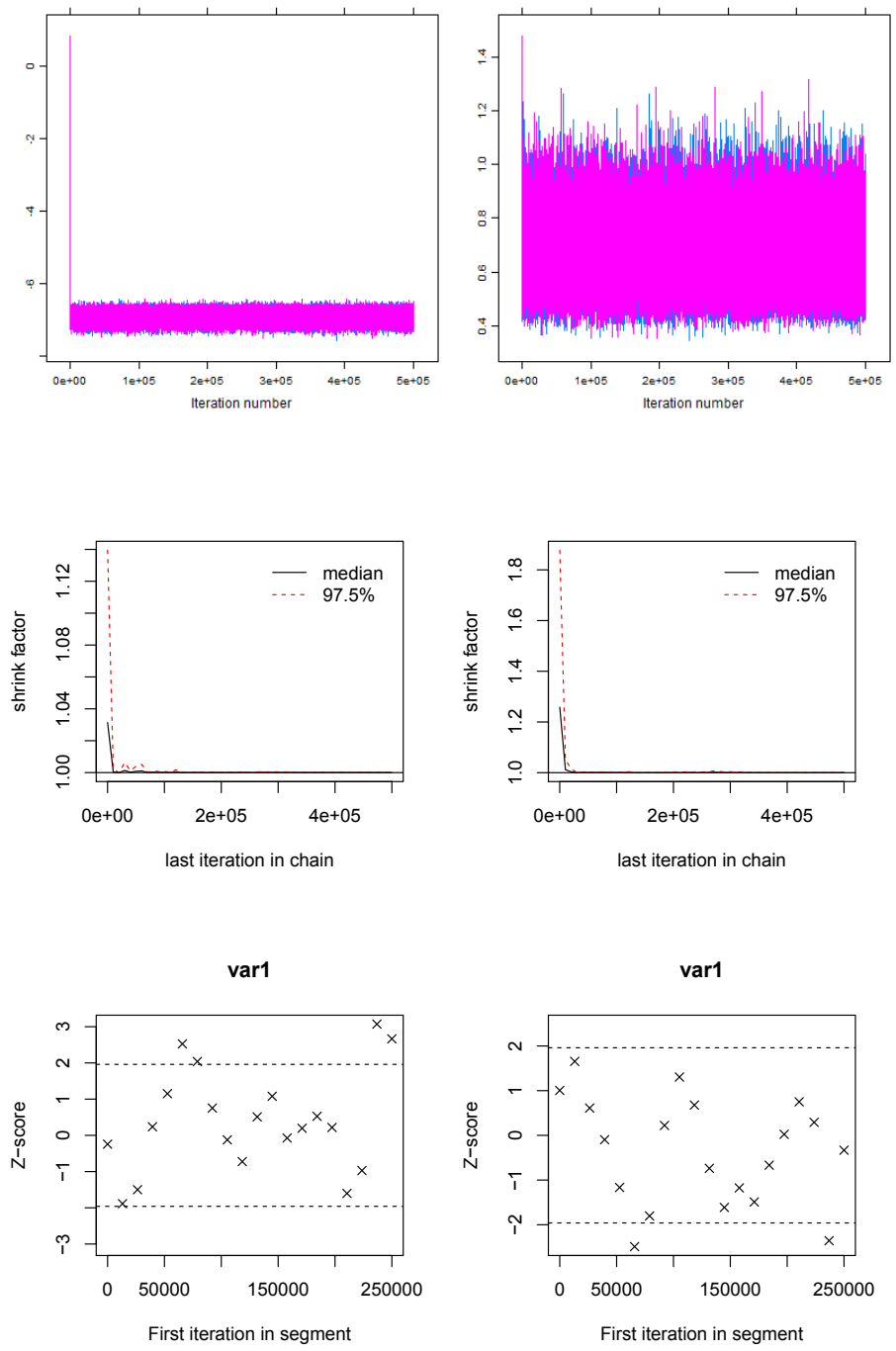


Figure B.8: The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau^2 \sim \text{Inv-Gamma}(0.1, 0.1)$. The graphs in the first row represent the trace plots of the parameters α and τ. The blue trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the red trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

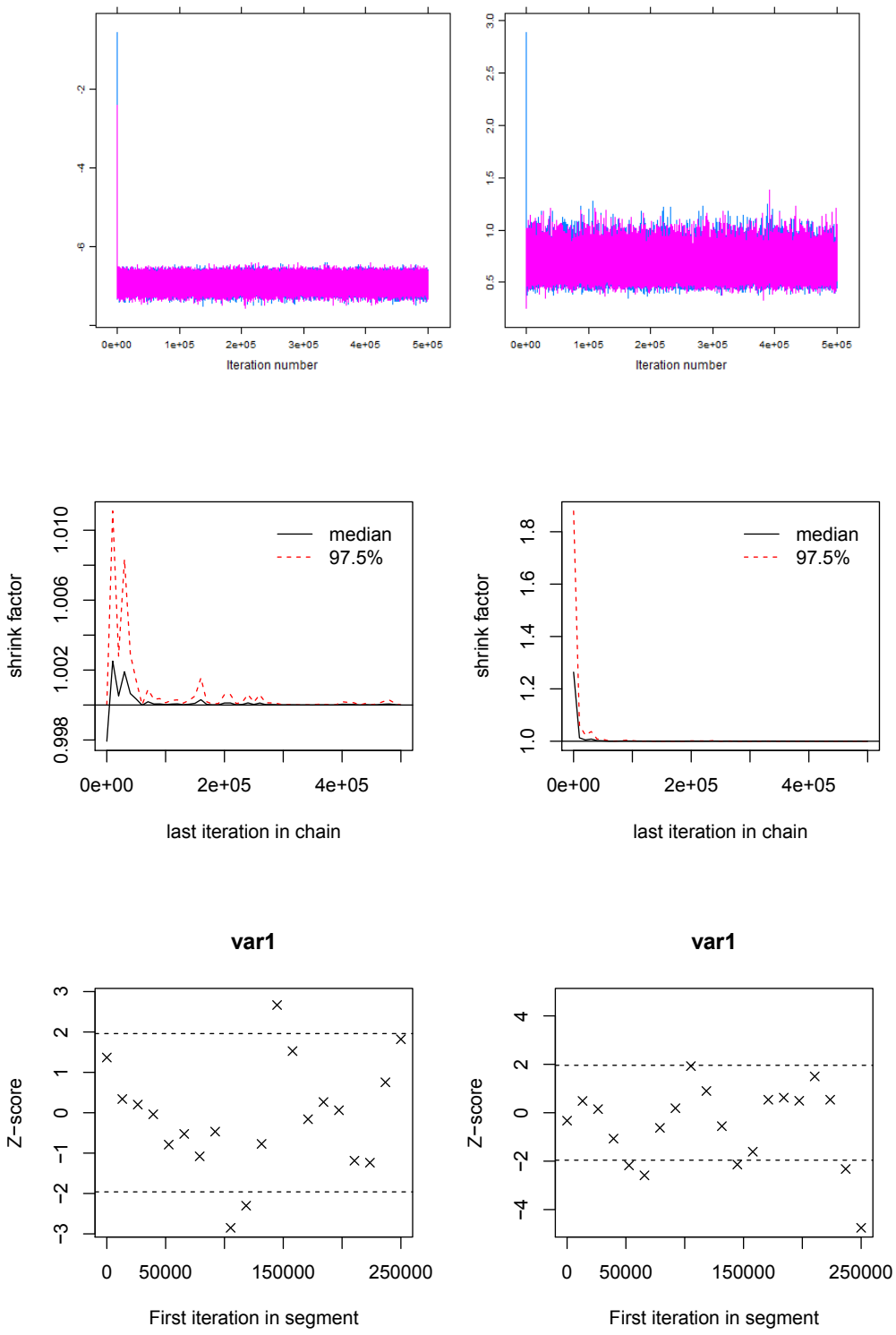


Figure B.9: The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim \text{N}(0, 100)$ and $\tau \sim \text{unif}(0, 100)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin statistic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

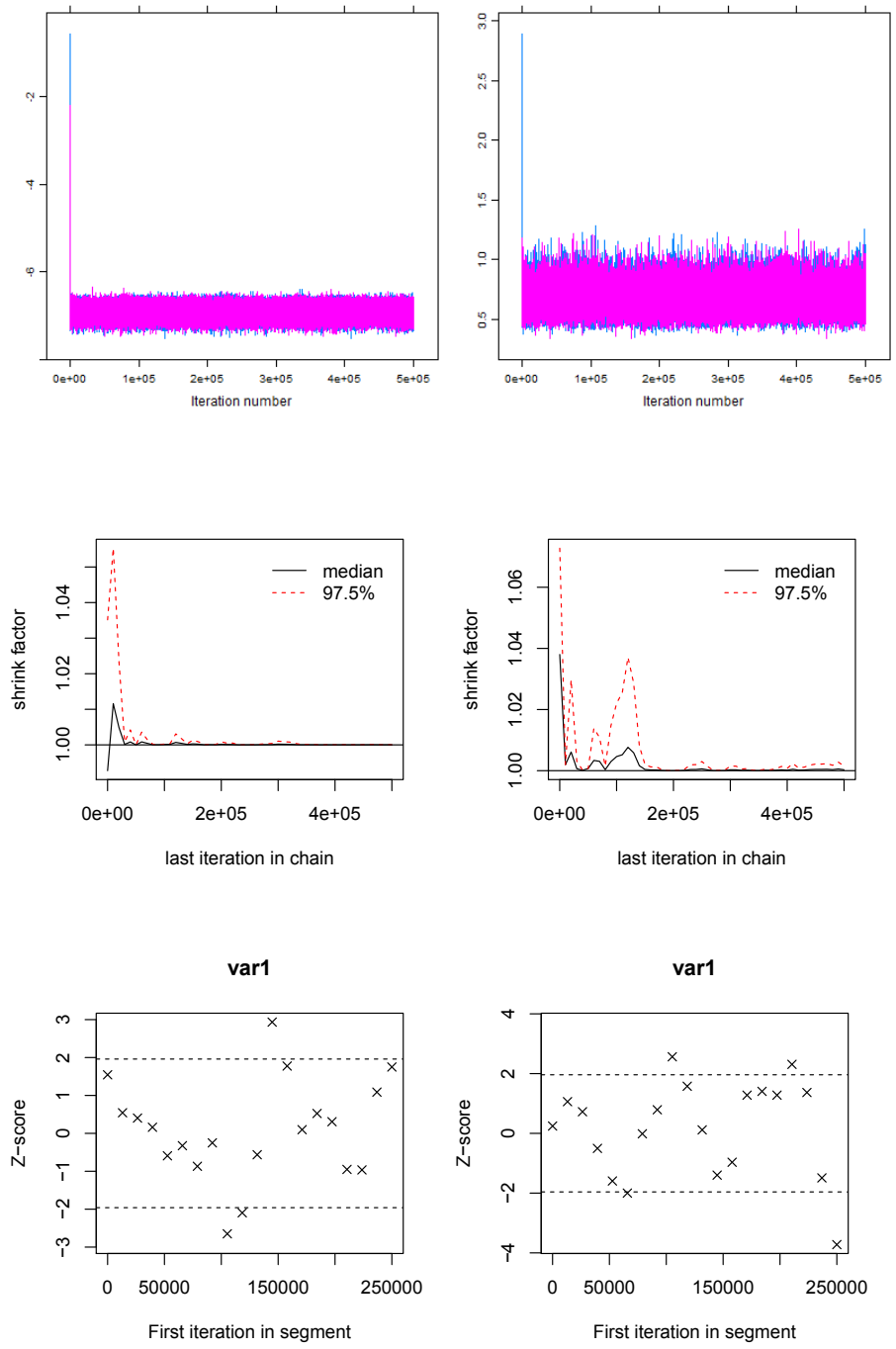


Figure B.10: The diagnostic convergence graphs of posterior parameters of the three-level hierarchical model under prior distributions $\alpha \sim \mathrm{N}(0, 100)$ and $\tau \sim \mathrm{HN}(0, 0.02)$. The graphs in the first row represent the trace plots of the parameters α and τ. The red trace plots is performed with initial values of $\alpha = -10$ and $\tau = 0.25$ and the blue trace plots is performed with starting values of $\alpha = 10$ and $\tau = 3$. The graphs in the second row represent the plots of the Gelman-Rubin statistic of the generated Markov chains of the posterior parameters of α and τ. The black sold and red dashed lines in the Gelman-Rubin diagnostic represent median and 97.5% quantile of the sampling distribution for the resulting shrink factor for stabilize around value of 1 for the last 250,000 samples of the Markov chains of α and τ. The graphs in the third row represent plots of the Geweke's diagnostic involving Z-scores. The horizontal black dashed lines in the Geweke's diagnostic plot are tails of a standard normal distribution which are ±1.96.

		$\alpha = -3$					$\alpha = -4$					$\alpha = -5.5$				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	time
0.3	α	-2.9988	0.0012	0.0028	94.6%	1039	-3.9747	0.0253	0.0033	91.0%	770	-5.3853	0.1147	0.0156	36.4%	946
	τ	0.3307	0.0307	0.0059	71.6%		0.3218	0.0218	0.0035	76.7%		0.2921	-0.0079	0.0024	88.8%	
0.7	α	-3.0159	-0.0159	0.0112	93.5%	1008	-3.9766	0.0234	0.0119	93.5%	766	-5.3630	0.1370	0.0277	70.4%	935
	τ	0.7075	0.0075	0.0078	89.5%		0.6909	-0.0091	0.0069	92.6%		0.6373	-0.0627	0.0093	92.0%	
1.1	α	-3.0269	-0.0269	0.0250	93.4%	989	-3.9731	0.0269	0.0256	92.9%		-5.3224	0.1776	0.0510	76.0%	911
	τ	1.0724	-0.0276	0.0144	94.4%		1.0539	-0.0461	0.0136	96.1%		0.9548	-0.1452	0.0317	83.4%	

Table B.1: Simulation results of frequentist method. Time is recorded in second. Note: MSE represents mean square error and CP represents the coverage probability.

		$\alpha = -6$					$\alpha = -6.5$					$\alpha = -7$				
True τ	Parameters	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time	Mean	Bias	MSE	CP	Time
0.3	α	-5.8254	0.1746	0.0328	5.6%	908	-6.2312	0.2688	0.0744	0%	865	-6.5961	0.4039	0.1650	0%	824
	τ	0.2845	-0.0155	0.0023	91.1%		0.2759	-0.0241	0.0031	92.3%		0.2650	-0.0350	0.0037	89.8%	
0.7	α	-5.7949	0.2051	0.0505	38.9%	893	-6.1832	0.3168	0.1072	3.6%	902	-6.5297	0.4703	0.2271	0%	821
	τ	0.5976	-0.1024	0.0153	81.7%		0.5537	-0.1463	0.0263	62.7%		0.4942	-0.2058	0.0472	27.1%	
1.1	α	-5.7415	0.2585	0.0860	50.8%	947	-6.1196	0.3804	0.1602	12.2%	886	-6.4496	0.5504	0.3149	0%	843
	τ	0.8915	-0.2085	0.0529	63.1%		0.8189	-0.2811	0.0886	32.8%		0.7377	-0.3623	0.1400	10.2%	

Table B.2: Simulation results of frequentist method. Time is recorded in second. Note: MSE represents mean square error and CP represents the coverage probability.

Bibliography

Ahmed, A., A. F. M. Sadullah, and A. shukri Yahya (2014). Accident analysis using count data for unsignalized intersections in malaysia. *Procedia engineering 77*, 45–52.

Ang, Q. W., A. Baddeley, and G. Nair (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics 39*(4), 591–617.

Baddeley, A., Y. Chang, Y. Song, and R. Turner (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates.

Baddeley, A., E. Rubak, and R. Turner (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Berger, J. (1985). Statistical decision theory and bayesian analysis. *Springer Series in Statistics. 24 cm. 617 p.*.

Berman, M. and T. R. Turner (1992). Approximating point process likelihoods with glim. *Applied Statistics*, 31–38.

Best, N., M. K. Cowles, and K. Vines (1995). Coda* convergence diagnosis and output analysis software for gibbs sampling output version 0.30. *MRC Biostatistics Unit, Cambridge 52*.

Black, W. R. (1991). Highway accidents: a spatial and temporal analysis. *Transportation Research Record 1318*, 75–82.

Brooks, S. P. and G. O. Roberts (1998). Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing 8*(4), 319–335.

Browne, W. J., D. Draper, et al. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis 1*(3), 473–514.

Burke, D. L., J. Ensor, and R. D. Riley (2017). Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in medicine 36*(5), 855–875.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder (2006). The design of simulation studies in medical statistics. *Statistics in medicine 25*(24), 4279–4292.

Carlin, B. P. and T. A. Louis (1997). Bayes and empirical bayes methods for data analysis. *Statistics and Computing 7*(2), 153–154.

Casella, G. and E. I. George (1992). Explaining the gibbs sampler. *The American Statistician 46*(3), 167–174.

Ceder, A. and M. Livneh (1978). Further evaluation of the relationships between road accidents and average daily traffic. *Accident Analysis & Prevention 10*(2), 95–109.

Chandler, R., A. Royle, and M. R. Chandler (2013). Package 'maxlike'.

Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *The American Statistician 49*(4), 327–335.

Chin, H. C. and M. A. Quddus (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention 35*(2), 253–259.

Collins, L. M., J. L. Schafer, and C.-M. Kam (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods 6*(4), 330.

Congdon, P. (2007). *Bayesian Statistical Modelling*, Volume 704. John Wiley & Sons.

Cowles, M. K. and B. P. Carlin (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association 91*(434), 883–904.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova 4*(5), 613–617.

Cressie, N. A. (1993). Statistics for spatial data.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics 27*(3), 567–578.

Erdogan, S., I. Yilmaz, T. Baybura, and M. Gullu (2008). Geographical information systems aided traffic accident analysis system case study: city of afyonkarahisar. *Accident Analysis & Prevention 40*(1), 174–181.

Farrell, S. and C. J. Ludwig (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic bulletin & review 15*(6), 1209–1217.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association 95*(452), 1300–1304.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). Bayesian data analysis.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., G. O. Roberts, W. R. Gilks, et al. (1996). Efficient metropolis jumping rules. *Bayesian statistics 5*(599-608), 42.

Gelman, A. and D. B. Rubin (1992a). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Gelman, A. and D. B. Rubin (1992b). A single series from the gibbs sampler provides a false sense of security. *Bayesian Statistics 4*, 625–631.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). Bayesian data analysis.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, 473–483.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). Introducing markov chain monte carlo. *Markov Chain Monte Carlo in Practice 1*, 19.

Glickman, M. E. and D. A. van Dyk (2007). Basic bayesian methods. In *Topics in Biostatistics*, pp. 319–338. Springer.

Golob, T. F., W. W. Recker, and D. W. Levine (1990). Safety of freeway median high occupancy vehicle lanes: a comparison of aggregate and disaggregate analyses. *Accident Analysis & Prevention 22*(1), 19–34.

Hamra, G., R. MacLehose, and D. Richardson (2013). Markov chain monte carlo: an introduction for epidemiologists. *International journal of epidemiology 42*(2), 627–634.

Haque, M. M., H. C. Chin, and H. Huang (2010). Applying bayesian hierarchical models to examine motorcycle crashes at signalized intersections. *Accident Analysis & Prevention 42*(1), 203–212.

Hardy, R. J. and S. G. Thompson (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine 15*(6), 619–629.

Henderson, R., P. Diggle, and A. Dobson (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics 1*(4), 465–480.

Huang, H. and M. Abdel-Aty (2010). Multilevel data and bayesian analysis in traffic safety. *Accident Analysis & Prevention 42*(6), 1556–1565.

Huang, H., H. Chin, and M. Haque (2009). Empirical evaluation of alternative approaches in identifying crash hot spots: naive ranking, empirical bayes, and full bayes methods. *Transportation Research Record: Journal of the Transportation Research Board* (2103), 32–41.

Huang, H., H. C. Chin, and M. M. Haque (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: a bayesian hierarchical analysis. *Accident Analysis & Prevention 40*(1), 45–54.

Jones, A. P. and S. H. Jørgensen (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis & Prevention 35*(1), 59–69.

Kim, D.-G., Y. Lee, S. Washington, and K. Choi (2007). Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. *Accident Analysis & Prevention 39*(1), 125–134.

Klaus, B., K. Strimmer, and M. K. Strimmer (2015). Package 'fdrtool'. *CRAN. h ttp://http1. debian. or. jp/pub/CRAN/web/packages/fdrtool/fdrtool. pdf. Accessed on October 13*, 2016.

Kontopantelis, E. and D. Reeves (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical methods in medical research 21*(4), 409–426.

Lambert, P. C., A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones (2005). How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. *Statistics in medicine 24*(15), 2401–2428.

Lenguerrand, E., J. L. Martin, and B. Laumon (2006). Modelling the hierarchical structure of road crash data—application to severity analysis. *Accident Analysis & Prevention 38*(1), 43–53.

Lesaffre, E. and A. B. Lawson (2012). *Bayesian biostatistics*. John Wiley & Sons.

Lewis, P. and G. Shedler (1976). Simulation of nonhomogeneous poisson processes with log linear rate function. *Biometrika 63*(3), 501–505.

Lewis, P. A. and G. S. Shedler (1978). Simulation of nonhomogeneous poisson processes by thinning. Technical report, DTIC Document.

Li, W., A. Carriquiry, M. Pawlovich, and T. Welch (2008). The choice of statistical models in road safety countermeasure effectiveness studies in iowa. *Accident Analysis & Prevention 40*(4), 1531–1542.

Liu, J. S. (2001). Monte carlo strategies in scientific computing.

MacNab, Y. C. (2003). A bayesian hierarchical model for accident and injury surveillance. *Accident Analysis & Prevention 35*(1), 91–102.

Marin, J.-M. and C. P. Robert (2014). *Bayesian essentials with R*. Springer.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention 26*(4), 471–482.

Mitra, S. and S. Washington (2007). On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis & Prevention 39*(3), 459–468.

Moller, J. and R. P. Waagepetersen (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC.

Ng, J. C. and E. Hauer (1989). Accidents on rural two-lane roads: Differences between seven states. *TRANSPORTATION RESEARCH RECORD 1238*(1), 1–9.

Okabe, A. and K. Sugihara (2012). *Spatial analysis along networks: statistical and computational methods*. John Wiley & Sons.

Openshaw, S. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Spatistica applications in the spatial sciences*, 127–144.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of london crash data. *Accident Analysis & Prevention 40*(4), 1486–1497.

Rathbun, S. L., S. Shiffman, and C. J. Gwaltney (2007). Modelling the effects of partially observed covariates on poisson process intensity. *Biometrika 94*(1), 153–165.

Ripley, B. D. (1991). *Statistical inference for spatial processes*. Cambridge university press.

Rizzo, M. L. (2008). *Statistical Computing with R*. CRC Press.

Robert, C. and G. Casella (1999). Monte carlo statistical methods. *Springer Texts in Statistics Show all Parts in this Series*.

Rubinstein, R. Y. and D. P. Kroese (2016). *Simulation and the Monte Carlo method*, Volume 10. John Wiley & Sons.

Sahlin, K. (2011). Estimating convergence of markov chain monte carlo simulations. *Stockholm University, Master Thesis*.

Shankar, V., R. Albin, J. Milton, and F. Mannering (1998). Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model. *Transportation Research Record: Journal of the Transportation Research Board* (1635), 44–48.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Stern, E. and Y. Zehavi (1990). Road safety and hot weather: a study in applied transport geography. *Transactions of the Institute of British Geographers*, 102–111.

Stojanovski, E., D. Nur, et al. (2011). Prior sensitivity analysis for a hierarchical model.

Symanzik, J. (2005). Statistical analysis of spatial point patterns.

Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention 28*(2), 251–264.

Thompson, S. G., T. C. Smith, and S. J. Sharp (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in medicine 16*(23), 2741–2758.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.

van de Schoot, R. and S. Depaoli (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist 16*(2), 75–84.

Viechtbauer, W. and M. W. Viechtbauer (2015). Package 'metafor'. *The Comprehensive R Archive Network. Package 'metafor'. http://cran. r-project. org/web/packages/metafor/metafor. pdf*.

Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika 95*(2), 351–363.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*(Dec), 3571–3594.

Yamada, I. and J.-C. Thill (2010). Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. *Annals of the Association of American Geographers 100*(2), 269–285.

Yoo, W. and E. H. Slate (2005). A simulation study of a bayesian hierarchical changepoint model with covariates. Technical report, Technical report, Center for Applied Mathematics and Statistics, New Jersey Institute of Technology 2005. Google Scholar.