BAYESIAN MIXTURE MODELLING WITH APPLICATION TO ROAD TRAFFIC FLOW

Submitted in partial requirement of the University of Sunderland for the degree of Doctor of Philosophy

G.J.COWBURN

May 8, 2003

Contents

1 Introduction **1**

2 Headway modelling **7**

2.1 What is a headway? 7

2.2 Who models headways? 8

2.3 Why are headways modelled? 9

2.4 Headway modelling in practice 10

2.5 The Bayesian paradigm 12

2.5.1 Thomas Bayes 1702 - 1761 12

2.5.2 Bayes' Theorem 14

2.5.3 Two Simple Examples 16

3 The three candidate models **19**

3.1 The Schuhl Model 20

3.1.1 Reason for choice 21

3.2 The Griffiths and Hunt Model 22

3.2.1 Reason for choice 23

3.3 The Gamma Exponential Model 24

3.3.1 Reason for choice 25

4 Mixture models **26**

4.1 What is a mixture model? 26

4.1.1 An example 27

4.2 Why are mixture models used? 27

4.2.1 The 'direct' use of mixture models 27

4.2.2 The 'indirect' use of mixture models 29

4.3 The use of mixtures in this thesis 31

4.4 The advantages and disadvantages of mixtures 31

4.4.1 Advantages 31
4.4.2 Disadvantages 32

5 Gibbs Sampling for Mixture Models 36

5.1 Monte Carlo integration 36
5.2 Markov chain Monte Carlo integration 40
5.3 Gibbs sampling 45
5.3.1 Full Conditional Distributions 46
5.4 Gibbs sampling for mixture models 48
5.4.1 The missing data structure 48
5.4.2 Stochastic allocation 49
5.5 Application to the models used 51
5.5.1 The Griffiths and Hunt Model 51
5.5.2 The Schuhl Model 53
5.5.3 The Gamma Exponential Distribution 54
5.6 Advantages and disadvantages 55
5.6.1 Advantages 55
5.6.2 Disadvantages 56

6 The data and the software 58

6.1 The data 58
6.1.1 Data definition 58
6.1.2 Site location 58
6.1.3 Method of collection 59
6.2 The software 61
6.2.1 Language 61
6.2.2 Gibbs sampling programs 63
6.2.3 Post-processing programs 65
6.2.4 Software testing 65

7 Implementation : Problems and Solutions **67**

7.1 Introduction 67

7.2 Preliminary considerations 67

7.3 Modelling Outcome 69

7.3.1 What would be a "successful" outcome? 69

7.4 How do we measure "success" 70

7.5 An important distinction 75

7.6 The Griffiths and Hunt model 76

7.6.1 The Base Run 76

7.6.2 $\beta_2 > \beta_1$ 81

7.6.3 Summary 85

7.7 The Schuhl Model 86

7.7.1 The Base Run : File 2 86

7.8 Further investigations 93

7.8.1 Run outcome 94

7.8.2 Summary 99

7.9 The Gamma Exponential Model 100

7.9.1 The Base Run 101

7.9.2 $E(C_1) > E(C_2)$ 105

7.9.3 R.U.G.S. 108

7.9.4 Blocking 112

7.9.5 Run outcome using File 3 115

7.10 Summary 117

8 An examination of model behaviour via Bayesian deviance **120**

8.1 Bayesian deviance 120

8.2 Run Outcomes 124

8.3 Summary 126

9 Bayesian model comparison using posterior predictive datasets and Mahalanobis distance **128**

9.1 Background . . . 129

9.2 Implementation . . . 130

9.3 Mahalanobis distance (D^2) . . . 133

9.4 Two important caveats . . . 134

9.4.1 The use of D^2 in isolation . . . 134

9.4.2 The sampling distribution of D^2 . . . 135

9.5 Summary . . . 135

10 Conclusion **136**

10.1 Research questions . . . 136

10.1.1 The Bayesian paradigm . . . 136

10.1.2 Model suitability . . . 137

10.1.3 Problems associated with the Bayesian paradigm . . . 137

10.1.4 The Bayesian paradigm and highway engineers . . . 138

10.2 Further research . . . 138

10.2.1 Model refinement . . . 138

10.2.2 A methodolgy for highway engineers . . . 140

10.2.3 Further work on mixture models . . . 141

10.2.4 A convergence / correlation analysis tool . . . 141

11 References **143**

12 Appendix **150**

12.1 1 Source code listing : headw1.c : data collection program . . . 150

12.2 Source code listing : A typical sampling routine . . . 151

12.3 Additional runs of the Gibbs sampler . . . 156

12.3.1 Making the Griffiths & Hunt model "work" . . . 156

12.3.2 Simulated data : Run 1 : Data simulated from an exponential distribution . . . 157
12.3.3 Simulated data : Run 2 : Data simulated from an gamma distribution . . . 159

List of Figures

1 A time headway . . . 7
2 Thomas Bayes . . . 13
3 The Schuhl Model . . . 20
4 The Griffiths & Hunt Model . . . 22
5 The Gamma Exponential Model . . . 24
6 Histogram of fish weight data . . . 28
7 A mixture of two normal distributions . . . 28
8 The normal distribution of X . . . 30
9 The distribution of X plus errors . . . 30
10 Two points in space . . . 41
11 Histogram of File 1 . . . 60
12 Histogram of File 2 . . . 61
13 Histogram of File 3 . . . 62
14 Block diagram of the sampling algorithm . . . 63
15 Screenshot of part of one of the Gibbs sampling programs . . . 64
16 Screenshot of part of one of the graphical output programs . . . 65
17 Autocorrelation plot 1 . . . 72
18 Autocorrelation plot 2 . . . 73
19 Autocorrelation plot 3 . . . 74
20 An example of a model/data fit diagram . . . 75
21 Raw output for base run of Griffiths & Hunt model . . . 77
22 Marginal posterior distributions . . . 79

23 Graph of β_{1i} v β_{2i} 80
24 Model fit diagram 80
25 Raw outputs for the Griffiths & Hunt model 82
26 Marginal posterior distributions for the Griffiths & Hunt model . . . 83
27 Model fit diagram 84
28 The effect of truncating a gamma density 85
29 Raw outputs for the Schuhl model 87
30 Marginal posterior distributions for the Schuhl model 87
31 Graph of β_{1i} v β_{2i} 89
32 The truncated marginal posterior distribution of k 89
33 Graph of t_{min} v iteration number 90
34 The parameters t_{min} and k 90
35 Raw outputs for the Schuhl model 95
36 Raw outputs for the Schuhl model 95
37 Model fit diagram 96
38 Raw outputs for the Schuhl model 97
39 Marginal posterior distributions for the Schuhl model 97
40 Part of the raw output for the parameter k 98
41 Model fit diagram 99
42 Graph of k_i v k_{i-1} 100
43 Raw outputs for the Gamma Exponential Model 102
44 Marginal posterior distributions for the Gamma Exponential model . 103
45 Model fit diagram 103
46 Graph β_{2i} v α_{2i} 104
47 Autocorrelation plot of the parameter p 105
48 Raw outputs for the Gamma Exponential Model 106
49 Marginal posterior distributions for the Gamma Exponential model . 106
50 Model fit diagram 107

51 Raw outputs for the Gamma Exponential Model 110
52 Marginal posterior distributions for the Gamma Exponential model . 110
53 Model fit diagram 111
54 Raw outputs for the Gamma Exponential Model 115
55 Marginal posterior distributions for the Gamma Exponential distribution 115
56 Model fit diagram 116
57 Marginal posterior distributions for the Gamma Exponential distribution 117
58 Model fit diagram 117
59 The successful algorithm 119
60 Posterior histograms for the Gamma Exponential distribution 125
61 Graph of $V_0(\alpha_2)$ v p_D 126
62 Graph of p_D v $\bar{D}$ 127
63 Posterior predictive datasets for Run 1 131
64 Posterior predictive datasets for Run 2 132
65 Posterior predictive datasets where bimodality is present 134
66 An example of a non-normal marginal posterior distribution 135
67 Direction of inferences in a genetic investigation 139
68 Direction of inferences in observation allocation 140
69 marginal posterior distributions for the Griffths & Hunt model 157
70 Model fit diagram for the Griffths & Hunt model 158
71 marginal posterior distributions for the Gamma Exponential distribution 159
72 Model fit diagram for the Gamma Exponential distribution 159
73 marginal posterior distributions for the Gamma Exponential distribution 160
74 Model fit diagram for the Gamma Exponential distribution 161

List of Tables

1 Two component h.p.d.f.'s 12

2 Headway probability density function parameter values : (DDNE = Double Displaced Negative Exponential) 20

3 Equivalance of parameters 27

4 Sampling references 38

5 Prior distributions for the Griffiths & Hunt model 51

6 Prior distributions for the Gamma Exponential model 54

7 Model parameter prior distributions 76

8 Model parameter starting values 76

9 Prior distribution parameter values 77

10 Kolmogorov - Smirnov test values 78

11 Posterior means and variances 79

12 Posterior means and variances 82

13 Kolmogorov - Smirnov test values 83

14 Kolmogorov - Smirnov test values 83

15 Posterior means and variances for the Schuhl distribution 88

16 Prior distribution parameter values 93

17 Model parameter starting values 94

18 Kolmogorov - Smirnov test values 96

19 Kolmogorov - Smirnov test values 98

20 Prior distributions for the Gamma Exponential model 101

21 Starting values for the Gamma Exponential model parameters 101

22 Prior distribution parameter values 101

23 Posterior means and variances 102

24 Posterior means and variances for the Gamma Exponential model . . 107

25 Kolmogorov - Smirnov values 108

26 Kolmogorov - Smirnov values 108

27 Posterior means and variances for the Gamma Exponential model . . 111
28 Kolmogorov - Smirnov values for raw and thinned output 111
29 Posterior means and variances for the Gamma Exponential distribution 114
30 K-S values for the Gamma Exponential Distribution 116
31 Details of run using File 3 116
32 Parameter prior distributions 123
33 Prior parameter values 123
34 Prior variance of α_2 124
35 Quantities of interest 125
36 Details of Runs 1 and 2 131
37 Further notation 133
38 The value of D^2 for Runs 1 & 2 134

Acknowledgements

I would like to thank my supervisor, Malcolm Farrow, for all his help, encouragement and patience during this project. I would also like to thank the Staff at the University of Sunderland School of Computing and Technology, The Graduate Research School and St Peters Library.

Most of all, I would like to thank my wife, Chris, whose constant support has been invaluable.

To the memory of my parents, George and Julia.

Abstract

This thesis is concerned with modelling vehicle headways on single and dual carriageway roads using two component mixture models that are estimated under the Bayesian paradigm. Vehicle headways are described as is the Bayesian paradigm with brief biographical details of Thomas Bayes. Next mixture models are defined and three particular models are subject to detailed analysis : two are taken from the highway engineering literature and in the light of difficulties encountered with these models a third is proposed by the author. This latter model is found to perform well when used with both real and simulated data.

The method of estimation is Gibbs sampling and a full description of this technique is given beginning with Markov chain Monte Carlo integration of which Gibbs sampling is an implementation. The data and the author's own software used to anaylse this data are detailed. The usual problems of slow mixing and identifiability are encountered and dealt with.

Bayesian deviance is used to explore model fit and it is found that the effective number of (model) parameters plays an important role in model behaviour. As the effective number of parameters is reduced by the use of highly informative prior distributions model fit improves initially but worsens after an optimum number of effective parameters has been reached. As an additional measure of model performance, in this case in the absence of any competing model, posterior predictive datasets are used for a qualitative assesment. For quantative purposes, Mahalonobis' distance is used in conjunction with the posterior predictive datasets.

1 Introduction

As pedestrians we have all stood by the side of the road waiting for a suitable gap in the traffic to allow us to cross. Similarly, as drivers we have waited at uncontrolled priority junctions until we could make our intended manouevre, e.g. emerge from a side road onto a main road. The phrase "a gap in the traffic" suggests a spatial distance between adjacent vehicles but this distance, together with the perceived speed of the next oncoming vehicle, is used by the motorist/pedestrian to judge if sufficient time is available for the safe execution of the intended manouevre. This "time gap" is clearly of vital importance in road safety and highway engineering and is formally referred to as the time *headway* between two vehicles. It is defined as the time between adjacent vehicles passing a fixed point and it is this quantity with which this project is concerned.

The physical distance or space headway between adjacent vehicles, although related to the time headway, will not be studied in this project and from now on the term "headway" will mean the time headway.

The fact that it is necessary to wait at the side of the road, or in a queue of traffic at a junction, for a suitable "gap" indicates that not all headways are the same. Further observation reveals that headways are a random variable that can be modelled statistically and it is this modelling that is the focus of this project. The importance of time headways in highway engineering is due to the method of modelling and simulating the behaviour of junctions at the design stage, for new roads, or assessment stage if existing roads are to be modified. Consider the following simple example. Suppose a large housing development is to be built adjacent to a main road with a single access on to the main road in the form of a simple "T" junction. The behaviour of the junction can be modelled by considering it, in mathematical terms, as a queueing model. The traffic flow on the main road is analogous to the server and the queue of vehicles on the side road analogous to the

customers. The probability distribution functions of the headways (h.p.d.f.'s) on both main road and side road are, then, of great importance. This modelling is very often carried out using two component mixtures of probability density functions from the exponential family of distributions. It is almost always carried out by highway engineers and the method of estimation is usually an *ad hoc* method. It is never Bayesian.

The primary area of interest of this project relates to the application of the Bayesian paradigm to the area of vehicle headway modelling. Until now, frequentist methodologies have always been used in this field. More specifically, it will be determined if certain two component mixture models can be used as headway probability density functions. As a result of this, the following research questions will be asked :-

- **" Can we usefully apply the Bayesian paradigm to inferences about these models?"** Since the Bayesian paradigm has not yet been applied in headway modelling, it must be determined if this is, in fact, practically possible and what advantages it offers.

- **"Are these models appropriate?"** There have been many different models used and, clearly, not every model can be examined in this project. However, three will be chosen and their suitability, or otherwise, for modelling headways will be put under scrutiny.

- **"What problems arise when the Bayesian paradigm is applied?"** It will be demonstrated that numerical difficulties can be encountered but it will also be shown that techniques exist which can circumvent them.

- **" Is the routine use of these models feasible in highway engineering"** In practise, Bayesian statistics requires its practitioners to use algebraic skills that most highway engineers have not used since their college days. The same can be said, in many cases, with respect to computer programming.

This question is concerned with finding a way of allowing highway engineers to use the Bayesian paradigm without having to do, as they see it, inordinate amounts of algebra or computer programming. To make this practical requires a methodology and software which can be reliably used in routine highway engineering work without presenting the engineer with technical statistical or computational difficulties.

These questions are dealt with in the rest of this thesis which is divided into sections as follows:-

Section 2 begins with an account of headway modelling. Some of the headway probability distribution functions (h.p.d.f.'s) used by highway engineers are described and appropriate references are given. A description of the Bayesian paradigm begins with historical details of Thomas Bayes. Next, Bayes theorem is explained and the section concludes with two examples which show how Bayes theorem is applied to modelling and the numerical complexity that can so easily arise.

Section 3 describes the three models chosen for examination in this thesis. These are

- The double displaced negative exponential distribution
- The double exponential headway distribution
- The Gamma exponential distribution

The first two, despite having similar names, are quite different and have been taken from the highway engineering literature. The third is put forward by the author on the basis of having dealt with the first two.

Section 4 deals with mixture models in general and begins with a formal definition of a mixture model. The reasons for their use is then described with the aid of examples and the specific reason for their use in this thesis is given. Finally, the advantages and disadvantages associated with this type of model are discussed with

flexibility being cited as the main advantage. In terms of disadvantages, the main focus is on the likelihood function and the issue of identifiability.

Section 5 is a key part of this thesis since it deals with the main computational technique used to estimate the models involved. The particular technique is Gibbs sampling, an implementation of Markov chain Monte Carlo integration and in this section an explanation of the underlying theory is given. Results are derived for general cases and then in particular for mixture models and the section finishes by a detailed application of this theory to the actual models used.

Section 6 is divided into two parts. The first part focusses on the data that are used and describes not only the data but also where and how they were collected. The second part, concerned with software, begins by identifying the language used and giving the reason for its choice. The programs written by the author specifically for use in connection with this thesis are then described. These fall into two groups with one group carrying out the actual Gibbs sampling and the other being concerned with what is termed "post-processing". This latter group assist in the investigation of such issues as covergence diagnosis, correlation detection and model comparison.

Section 7 is probably the most important part of the thesis since it deals with the actual implementation of the methodology and algorithms already described. It is in this section that the research questions are, in practical terms, asked. It starts with a preliminary discussion of some of the key issues involved which include

- Prior distributions, initial conditions, constraints and data
- Convergence and correlation issues
- Essential criteria for a successful run of the Gibbs sampler

Next, each model is considered in turn. It will be shown that the first two models both have their own distinct disadvantages that render them unsuitable for modelling headways. The model put forward by the author is shown to be useful provided

the problems of identifiability and convergence are dealt with. The methods used to circumvent these well known difficulties are described in depth. The section ends with a block diagram summarising the successful modelling strategy.

The next two chapters represent an attempt to add to the growing literature concerning Bayesian model fit.

Section 8 examines the use of Bayesian Deviance not only as a tool for the assessment of model fit but also to shed light on the reasons behind the notoriously poor convergence properties of mixture models. After giving a brief description of Bayesian Deviance an example is presented which not only shows that informative priors can be used to improve convergence performance but also indicates that the degree of parameterisation is crucial in terms of model fit.

Section 9 approaches the problem of Bayesian model comparison by using posterior predictive datasets and Mahalanobis' distance. A method is proposed whereby a single model can be examined in the absence of any other competing models. Again, examples are used to demonstrate the principles involved.

Section 10 concludes the main body of the thesis and serves two main functions :-

- It examines the extent to which, and with what success, the research questions have been answered.
- It suggests areas for further research which, if pursued, would be of benefit to the fields of headway modelling and mixture modelling in general.

The thesis ends with references and an appendix containing items that, although of interest, were considered better located outside the main body of the thesis.

Originality

A major part of the originality of the research is the application of the Bayesian paradigm to the field of vehicle headway modelling. As a result of this, other areas of originality arise, as listed below :-

- A single method of parameter estimation is applied to three different models, one of which is proposed by the author.
- A natural identifiability constraint is applied to the models via the sampling algorithm as opposed to using a reparameterisation.
- Bayesian deviance is used to explore the behaviour of a mixture model.
- Posterior predictive datasets are used in conjunction with Mahalanobis distance to quantatively explore the extent to which the model captures features of the data.

2 Headway modelling

The purpose of this Section is two-fold. Firstly, headways are considered and work done on modelling them is discussed. There is, here, a deliberate bias towards highway engineering. Secondly the Bayesian paradigm is described, and a short biography of Thomas Bayes is included.

2.1 What is a headway?

Strictly speaking there two type of headways : time headways and space headways. If we were to take an aerial photograph of a length of highway the distance between successive vehicles would be the space headway. Time headways, with which we are concerned, are the time intervals between successive vehicles passing a point on the highway. This is illustrated in Figure 1 below :-

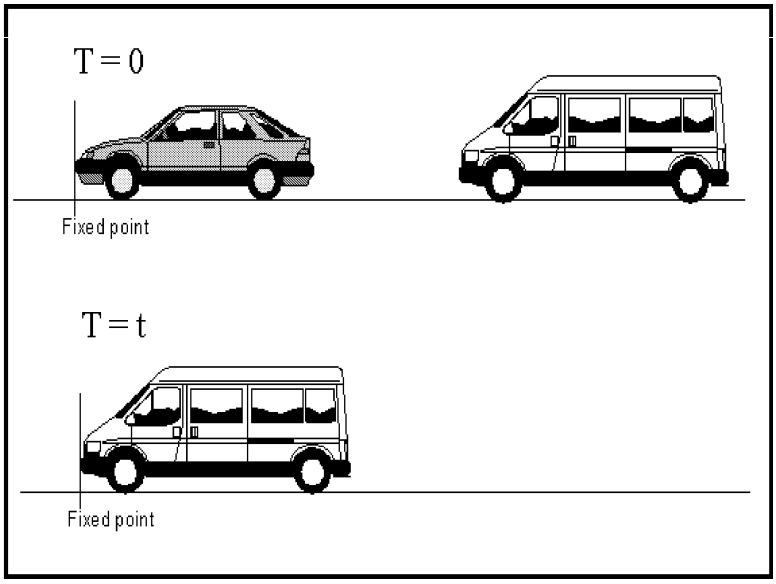


Figure 1: A time headway

The upper part of the diagram shows two vehicles travelling from right to left. The event of the saloon car passing the fixed point is deemed to have taken place at time T = 0. At a time t seconds later, the minibus now passes the same fixed point as shown in the lower part of the diagram, with the saloon car having travelled "out

of picture". This time, t seconds, is the headway between the saloon car and the minibus. All headways referred to in this thesis will be time headways.

2.2 Who models headways?

Light traffic can be modelled by a simple Poisson process which gives rise to a negative exponential distribution for the vehicle headways. The simplicity of this real-life situation makes it a good example from a teaching perspective. The Open University course "M245 : Probability and Statistics" (1984) uses free-flowing traffic as an example of a "familiar random process" in its introductory unit entitled "Chance". Although the headway modelling done in this context is, of necessity, elementary it means that this topic is familiar to many teachers and students of statistics. There are also, of course, more rigourous analyses in the statistical literature. Two good examples of this are Miller (1961) and Ashton (1971). Miller proposes what is described as a moving queue model in which he considers traffic flow as a queueing process with the slow vehicles as the service points of the classical queueing model. More relevant to this thesis is the work of Ashton who gives in-depth consideration to three models namely the single shifted exponential model, Schuhl's model (to be discussed later) and the modified semi-Poisson process. In this latter model the assumption is made that there exists behind each vehicle a "zone of emptiness" into which a following vehicle will never enter. The length of this zone, Z, is measured in seconds and is itself a random variable. A proportion of headways result from the following vehicle travelling at the edge of this zone with the remainder being exponentially distributed with a minimum headway of Z. The resulting distribution function is a two component mixture model. It is interesting to note that Ashton concludes that, in the case of low traffic flow, this model does not appear to perform better than single shifted exponential model.

Highway engineers have been involved in headway modelling since its beginning. (It seems that those who originally built the roads were required, perhaps by accident

rather than design, to take an interest in the finer points of their operation!) As a result, the majority of papers relating to this subject can be found in the highway engineering / transportation literature. Partly for this reason, and partly because the author has worked in traffic management for almost twenty years, the emphasis of this thesis will be orientated to highway engineering and subsequent discussions will reflect this.

2.3 Why are headways modelled?

There are two main reasons why headways are modelled. Firstly, we can use a headway probabilty density function (h.p.d.f.) to model a stream of traffic and then use quantities such as the mean and variance of this h.p.d.f. to describe aspects of the flow. E.g., the reciprocal of the mean headway is the average rate of vehicle flow and the relationship between mean and variance can give an indication of the amount of congestion present. Secondly, road junctions can be modelled using queuing theory and in such a situation h.p.d.f.'s become the arrival mechanism (for, say, a minor road) and the service mechanism (for the major road). Consider the traffic on the minor (side) road and, in particular, the vehicle at the front of the queue. This vehicle has to wait until a safe gap appears in the major (main) road traffic flow. This waiting time is analagous to the length of time a customer in a shop takes to be served and so is equivalent to the service time. Therefore the h.p.d.f. of the main road traffic is the service mechanism. Meanwhile, other vehicles are arriving at the side road and form a queue. This is analagous to a queue of customers waiting to be served and so the h.p.d.f. of the side road traffic is equivalent to the arrival mechanism. Examples of this can be found in Ohno & Mine (1979), Troutbeck (1986) and Troutbeck & Kako (1999).

2.4 Headway modelling in practice

Vehicle headways have been modelled since the 1930's when traffic levels were very light by today's standards and the first model formally proposed was a negative exponential distribution with p.d.f.

$$f(t) = \lambda e^{-\lambda t} \qquad (t > 0) \tag{1}$$

where $1/\lambda$ is the mean headway, measured in seconds.

An early justification of the use of this model can be found in Adams (1936) where the author begins with a simple Poisson process which itself begins with the consideration of a short time interval, δt. The arrival of a vehicle at our fixed point on the highway is, in this case, assumed to be equally likely in δt as in any other non-overlapping interval of the same length and also independent of an arrival in any other short interval of time. This is important because other models do not include these assumptions. For example, consider two short, contiguous and non-overlapping time intervals $(t, t+\delta t)$ and $(t+\delta t, t+2\delta t)$. Let A1 be the event of an arrival in $(t, t+\delta t)$ and it's probability be P(A1) and let A2 and P(A2) be similarly defined. For the Poisson process :-

$$P(A1) = P(A2) = P(A2|A1)$$

that is to say, an arrival in $(t, t+\delta t)$ does not have any influence on the probability of an arrival in $(t+\delta t, t+2\delta t)$. Other processes do not have this property i.e.,

$$P(A2) \neq P(A2|A1)$$

in which case the negative exponential model does not apply and other models must be sought.

Traffic flows, varying in volume from 70 to 1,400 vehicles per hour, were observed

and then compared to the theoretical distributions. Adams' own comments were "The agreement in these and in many other cases is sufficiently good to justify the working assumption that road traffic is normally a random series" It is, however, important to realise that the low volumes observed meant that the traffic would have been in "free flow". A stream of traffic is said to be in free flow if each vehicle is proceeding unimpeded by any other vehicle and is, therefore, free to overtake at will. As the volume of traffic increases, vehicles begin to impede one another and the flow is said to become congested. Thus congestion is defined as the extent to which vehicles impede each other. The "bumper to bumper" situation that is commonly referred to as congestion is more appropriately described as degenerate flow and is not considered in this thesis since, when this stage is reached, stochastic modelling is no longer appropriate. It is interesting to note that streets in central London (one of which was Whitehall) were used to gather observations and that these observations revealed that the traffic was free flowing. It is difficult to imagine free flowing traffic in central London today due to the huge increase in traffic volume.

As traffic volumes increased over the years, it became apparent that the negative exponential distribution was suitable in fewer and fewer cases. In other words, the assumption of the Poisson process did not apply and other types of model were needed. In 1955, Schuhl proposed a mixture for h.p.d.f.'s which is defined by :-

$$f(t) = \begin{cases} p\lambda_1 e^{-\lambda_1 t}, & (0 < t < k) \\ p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 (t-k)} & (k \leq t) \end{cases} \tag{2}$$

This distribution is described by Salter (1974) who states that the first component is the h.p.d.f. of those vehicles which are in "free- flow" and the second that of vehicles in congested flow. The weighting parameter, p, is the proportion of the total traffic volume in "free flow". Such use of a two component mixture model is now common with different authors choosing different distributions for each component. Examples of this are shown in Table 1 below :-

Author(s)	First Component	Second Component
Polus (1979)	Binomial	Binomial
Baras, Downey & Levine (1979)	Shifted Exponential	Lognormal
Tamura &Chishaki (1983)	Lognormal	Lognormal
Katti & Pathak (1986)	Shifted Exponential	Shifted Exponential
Griffiths & Hunt (1991)	Shifted Exponential	Shifted Exponential

Table 1: Two component h.p.d.f.'s

It can be seen that shifted exponential distributions figure prominently and there are two basic reasons for this :-

- the relative tractability of these models
- the ease with which samples can be drawn from these distributions for simulation purposes.

Single component shifted exponential distributions have been the subject of Bayesian analysis and examples of this can be found in Trader (1985), Calabri & Pulcini (1994) and Madi & Leonard (1996). Such methodology has not yet entered into headway modelling. Since Bayesian methods seem to have had little impact in highway engineering, we might expect to see, for example, the method of moments or maximum likelihood being used for parameter estimation. However the use of mixture models gives rise to computational difficulties for both Bayesian and frequentist alike, the chief culprit being the complicated shapes of the likelihood functions involved. This has resulted in various *"ad hoc"* methods being used, examples of which can be found in Salter (1974) and in Griffiths and Hunt (1991). In this thesis, however, Bayesian methodology will be used. Its background is described in the next section.

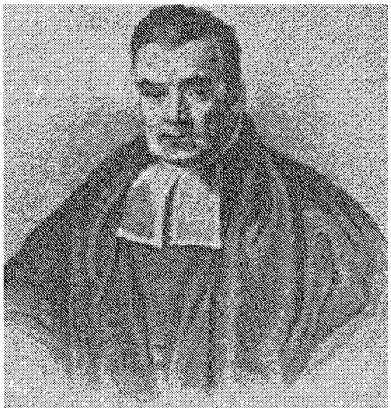
2.5 The Bayesian paradigm

2.5.1 Thomas Bayes 1702 - 1761

Born in 1702, Thomas Bayes was the eldest child of Joshua and Ann Bayes. Joshua was a Presbyterian minister and a Fellow of the Royal Society and in both respects

Thomas was to follow in his father's footsteps. Although it is not clear when Thomas was ordained, it is known that by 1731 he was working as a Presbyterian minister in Tunbridge Wells, Kent. He was elected to the Royal Society in 1742 and his reputation as a mathematician became considerable. On the 24th of August, 1746, he entertained William Whiston, a former Lucasian professor of mathematics at Cambridge University (a post now held by Stephen Hawking), who described Bayes as "a very good mathematician".

The Reverend Thomas Bayes



1702 - 1761

Figure 2: Thomas Bayes

Sometime around 1750, Bayes retired from the ministry but continued to live in Tunbridge Wells until his death, at the age of fifty nine, on the 17th of April, 1761. He was buried in the cemetary at Bunhill Fields near Moorgate, London. Also buried there are John Bunyan and Daniel Defoe. Apart from details of his will, in which he generously provided for friends and relatives, little else is known about the life of Thomas Bayes.

His contribution to statistics would have been significantly less had it not been for the actions of one of his friends, Richard Price. Whilst going through Bayes' papers after his death, Price found an essay entitled " An essay towards solving a problem in the doctrine of chance ". This paper was submitted to the Royal Society in November 1763 and read at their meeting on the 23rd of December that year.

His paper can be considered as the formal beginning of Bayesian Statistics al-

though it's author could never have of dreamed of the impact on the world of statistics it would later have. Today, Bayesian statistics is as widely applied as classical statistics. Berger (2000) gives an excellent overview of it's progress, and states that "It would be hard to find an area of human investigation in which there does not exist some level of Bayesian work". This statement is borne out by Bernado, Berger, Dawid & Smith (1998) who demonstrate the breadth of application that Bayesian statistics has in practice. Further recent examples of it's use can be found in the operations research literature e.g., Krishnan, Ramaswamy, Meyer & Damien (1999), and the medical literature e.g., Lau, Pathamanathan, Ng, Cooper, Shekan & Griffith (2000).

It must be said that the Bayesian paradigm does not yet have universal acceptance. Gullberg (1997) gives a very brief history of statistics and makes mention of 24 famous mathematicians who have made highly significant contributions to the development of the subject. Sadly, the name of Thomas Bayes does not appear.

2.5.2 Bayes' Theorem

The starting point in Bayesian statistics is the idea that probability is *"an expression of degree of belief that an event will occur"*. This is known as the subjectivist view which stands in sharp contrast to the frequentist view which defines the probability that an event will occur as *"the number of favourable outcomes divided by the number of possible outcomes"*. This definition is currently being taught to schoolchildren as a part of the National Curriculum and is due to Gerolamo Cardamo (1501-1576) (Gullberg (1997)).

There is still the need, however, to be able express Bayesian concepts in mathematical form and for this we have Bayes' Theorem which is defined as follows :-

Let A and B be two events with non-zero probability given by $Pr(A)$ and $Pr(B)$ respectively. Also, let $Pr(A|B)$ be the conditional probability of event A given event

B and let $Pr(B|A)$ be similarly defined. Bayes' Theorem states that :-

$$Pr(A|B) = \frac{Pr(B|A).Pr(A)}{Pr(B)} \tag{3}$$

This definition, concerned only with discrete events can modified in order to deal with distributions, their parameters and sets of observations.

$$f(\theta|D) = \frac{f(D|\theta).f(\theta)}{f(D)} \tag{4}$$

Where $f(.)$ is a probability density function (p.d.f.). In this case, θ can be a vector of parameters from a (p.d.f.) e.g the mean, μ, and variance, σ^2, of a normal distribution and D represents a set of observations. The remaining terms in this equation are defined as follows :-

$f(\theta|D)$ This is the p.d.f. of θ conditional upon the observed data set. In the example above, this could be written as $f(\sigma^2, \mu|D)$. In general terms, this is called the posterior distribution because it represents a p.d.f. arrived at after observations have been made and it is from this p.d.f. that inferences are made concerning θ

$f(D|\theta)$ This is the p.d.f. of the data given θ. That is to say, it is the likelihood function and is derived from the model being used.

$f(D)$ $f(D) = \int_\theta f(D|\theta).f(\theta)d\theta$

$f(\theta)$ This is called the prior distribution of θ. It represents an expression of belief about θ before any observations are taken.

We can summarise Equations 3 & 4 by saying :-

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

and so we have a method of converting a likelihood into a posterior belief. Whilst this is simple enough in theory, in practice the Bayesian statistician is very often faced with integrals of high dimensionality. The way in which this can arise is be shown in the following two examples.

2.5.3 Two Simple Examples

Suppose, in a manufacturing process, we wish to model the time interval, t, between breakdowns of a continuously running machine by using a simple negative exponential distribution, i.e.,

$$f(t) = \lambda e^{-\lambda t}$$

We wish to make inferences concerning λ. Suppose, also, that our prior belief concerning λ is such that our prior distribution is given by :-

$$f(\lambda) \propto \lambda^{\alpha-1} e^{-\lambda\beta}$$

This is a Gamma distribution with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$. If we now make observations $t_1, t_2, \ldots, t_k$ we can express the likelihood function as being :-

$$L(t) = \prod_{i=1}^{k} \lambda e^{-\lambda t_i}$$

which simplifies to :-

$$L(t) = \lambda^k e^{-\lambda \Sigma_{i-1}^{k} t_i}$$

If we now apply the formula Posterior $\propto$ Prior $\times$ Likelihood, we obtain

$$f(\lambda | t_1, t_2, \ldots, t_k) \propto \lambda^{\alpha-1} e^{-\lambda\beta} \times \lambda^k e^{-\lambda \Sigma_{i-1}^{k} t_i}$$

which simplifies to :-

$$f(\lambda|t_1, t_2, \ldots, t_k) \propto \lambda^{\alpha-1+k} e^{-\lambda(\beta+\Sigma_{i-1}^{k} t_i)}$$

The posterior distribution of λ is, thus, proportional to a Gamma distribution with parameters $\alpha + k$ and $\beta + \Sigma_{i-1}^{k} t_i$.

Whilst this example adequately demonstrates the theory it does not give any indication of the complexity that can arise but if we model t using a Gamma distribution the difficulties soon become apparent.

In such a case :-

$$f(t) = \frac{\beta^{\alpha} t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

We see, therefore, that $\Gamma(\alpha)$ is merely a normalising constant to ensure that $\int_0^{\infty} f(t) = 1$. The likelihood function, $L(t)$, is given by :-

$$L(t) = \frac{\beta^{\alpha k} (\prod_{i=1}^{k} t_i)^{\alpha-1} e^{-\beta \Sigma_{1=1}^{k} t_i}}{(\Gamma(\alpha))^n}$$

Now, if our priors for α and β are proportional to $\alpha^{a-1} e^{-b\alpha}$ and $\beta^{c-1} e^{-d\beta}$ respectively then, using the same observations as before, we arrive at the following joint posterior distribution :-

$$f(\alpha, \beta|t_1, t_2, \ldots, t_k) \propto \alpha^{a-1} e^{-b\alpha} \beta^{\alpha k} (\prod_{i=1}^{k} t_i)^{\alpha-1} \beta^{c-1} e^{-\beta(\Sigma_{1=1}^{k} t_i + d)} (\Gamma(\alpha))^{-n}$$

which simplifies to :-

$$f(\alpha, \beta|t_1, t_2, \ldots, t_k) \propto \alpha^{a-1} e^{-b\alpha} (\prod_{i=1}^{k} t_i)^{\alpha-1} \beta^{\alpha k+c-1} e^{-\beta(\Sigma_{1=1}^{k} t_i + d)} (\Gamma(\alpha))^{-n}$$

To evaluate, say, the marginal posterior distribution of α, we require :-

$$f(\alpha|t_1, t_2, \ldots, t_k) \propto \frac{\alpha^{a-1}e^{-b\alpha}(\prod_{i=1}^{k} t_i)^{\alpha-1}}{(\Gamma(\alpha))^{-n}} \int_0^\infty \beta^{\alpha k+c-1}e^{-\beta(\Sigma_{1=1}^{k} t_i+d)}d\beta$$

which gives :-

$$f(\alpha|t_1, t_2, \ldots, t_k) \propto \frac{\alpha^{a-1}e^{-b\alpha}(\prod_{i=1}^{k} t_i)^{\alpha-1}\Gamma(\alpha k+c)}{(\Gamma(\alpha))^{-n}(\Sigma_{1=1}^{k} t_i+d)^{\alpha k+c-1}}$$

Clearly, this function cannot easily be dealt with and yet it arises out of the use of a simple Gamma model and the use of simple prior distributions for the model parameters. There is no reason (except tractability) why conjugate priors should be used and it may well be that a non-conjugate distribution better expresses our prior beliefs. (A prior is conjugate if the the posterior distribution belongs to the same family)

If a model has n parameters, then the dimensionality of the integration is also n. Whilst this high dimensionality does constitute a problem it is more often the intractable nature of the integrals that causes difficulties. Techniques do exist, however, which can overcome these problems although the choice of an appropriate method may not be straightforward even when the dimensionality of the integral is relatively low. The method used in the case of high dimensionality is detailed in Section 5.

3 The three candidate models

This section briefly describes the three headway models that will be examined. Two of them are from the highway literature with the third being proposed by the author. The number of headway models to be found in the literature is, of course, far greater than three but a balance must be struck between the number of models chosen and the depth in which they can be analysed.

Also, it is felt that the two chosen models are representative of those used by highway engineers in that simple shifted exponential distributions figure prominently. The model proposed by the author does differ significantly from the other two but it will be shown that this model is, computationally at least, superior. All three are two component mixture models with the distributions coming from the exponential family.

A distribution is said to come from the exponential family if it can be expressed in the form :-

$$f(y;\theta) = a(y)b(\theta)e^{c(y)d(\theta)} \tag{5}$$

In fact, the distributions are gamma distributions but the parameterisations used are such that the first component in each mixture is an exponential distribution and in two of the models both components are exponential. The three models to be considered can be summarised by the following equation :-

$$f(t) = \begin{cases} 0 & t < k_1, \\ p.f(t-k_1,\alpha_1,\beta_1) & k_1 \leq t < k_2, \\ p.f(t-k_1,\alpha_1,\beta_1) + (1-p).f(t-k_2,\alpha_2,\beta_2) & t \geq k_2 \end{cases} \tag{6}$$

where

$$f(t,\alpha,\beta) = \frac{\beta^{\alpha}t^{\alpha-1}e^{-\beta t}}{\Gamma(\alpha)}.$$

and $k_2 \geq k_1 \geq 0$. By altering the values of the parameters α_1, α_2, k_1 and k_2 each

separate distribution is obtained. This is demonstrated in Table 2 below :-

Model	α_1	α_2	k_1	k_2
DDNE	1	1	k	k
Schuhl	1	1	0	k
Gamma/Exponential	1	α_2	0	0

Table 2: Headway probability density function parameter values : (DDNE = Double Displaced Negative Exponential)

3.1 The Schuhl Model

This model, also referred to as the Double Exponential Headway Model, is defined as follows :-

$$f(t) = \begin{cases} p\beta_1 e^{-\beta_1 t}, & (0 < t < k) \\ p\beta_1 e^{-\beta_1 t} + (1-p)\beta_2 e^{-\beta_2(t-k)} & (k \leq t) \end{cases} \tag{7}$$

Its use for modelling headways on dual carriageway roads will be examined in this thesis. Figure 3 below shows a plot of the model in which the parameters are $p = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.5$ and $k = 1.5$.

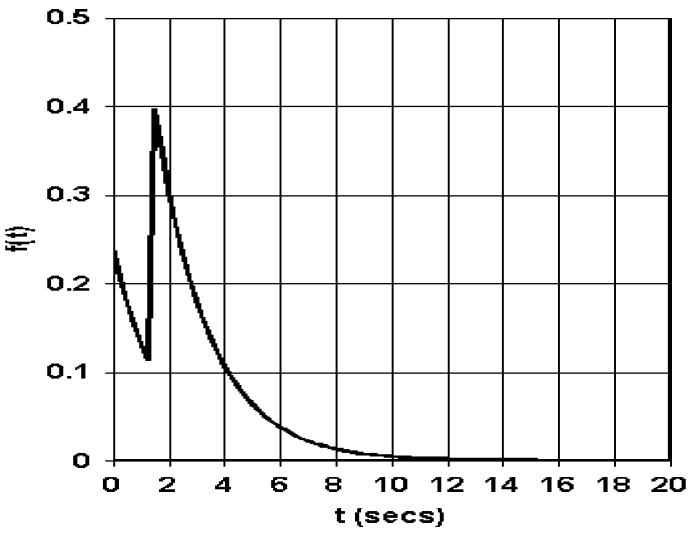


Figure 3: The Schuhl Model

Having been proposed by Schuhl in 1955, this model appears in Salter (1974) and

was used, for teaching purposes, in the now defunct Department of Civil Engineering at The University of Sunderland. It is also discussed by Ashton (1971) and whilst neither uses Bayesian analysis, a comparison of the two analyses is interesting in that the two approaches, despite having one similarity, are quite different. Each author uses an *"ad hoc"* method but only Ashton gives a full explanation of the difficulties that arise from the use of the Method of Moments and the method of Maximum Likelihood. In the case of the former, negative parameter estimates can result for positive value parameters and the latter can give rise to intractable equations. One method mentioned by Ashton, though not used, depends upon the assumption that large headways must belong to the free flowing component of the model used. This assumption lies behind the method employed by Salter and is also included in the work done by Wasielewski (1979). The presentation of the algebra is more thorough in Ashton whose approach concentrates on the methodology. Salter, for example, uses the letter e to denote the shift of the second component (as opposed to k above) yet gives a detailed example of the graphical method employed to estimate the model parameters. It is interesting to note that, in Salter's example, the model predicts more headways in the range 0 to 1 second than are actually observed. It will be seen later that this is often the case in this thesis. Also, Salter goes as far as attaching realistic interpretations to parameter estimates whereas Ashton does not. This will be discussed in greater depth later and it will be shown to be somewhat unreliable. This comparison highlights the difference in approaches of statisticians and highway engineers : the emphasis of the former is nearly always more methodological whereas it is computational in the case of the latter. But, perhaps more importantly, it can be seen that the same difficulties are encountered in both disciplines.

3.1.1 Reason for choice

There are two reasons why the Schuhl model has been included in this study :-

1. It is typical of the type of headway model used by highway engineers.

2. It has been studied by two other authors (Schuhl and Ashton), each using a different method of estimation. The method of estimation used in this project will constitute a third, believed by the author to be superior to the other two.

3.2 The Griffiths and Hunt Model

This model is proposed for modelling vehicle headways on single carriageway roads only. It appears only once in the literature (Griffiths & Hunt,1991), where it is referred to as the Double Displaced Negative Exponential model (DDNE) and is defined by :-

$$f(t) = \begin{cases} 0, & (t < k) \\ p\beta_1 e^{-\beta_1(t-k)} + (1-p)\beta_2 e^{-\beta_2(t-k)} & (k \leq t) \end{cases} \tag{8}$$

Figure 4 below shows a plot of the model in which the parameters are $p = 0.5$, $\beta_1 = 0.5$, $\beta_2 = 0.25$ and $k = 1.5$. This model is fairly typical of those used by highway

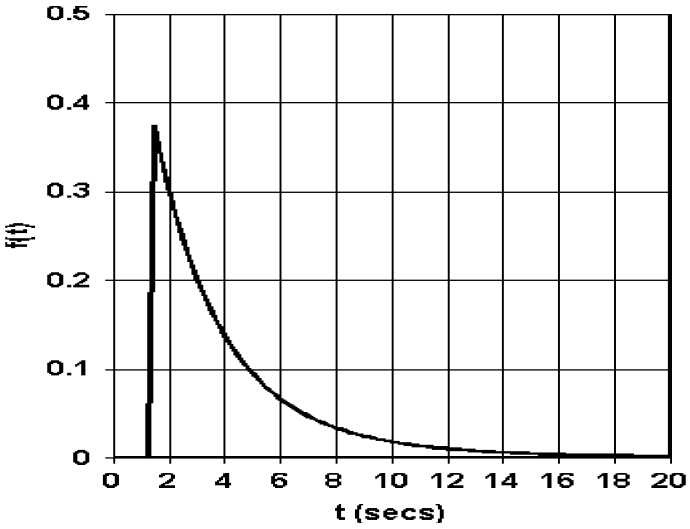


Figure 4: The Griffiths & Hunt Model

engineers in that it is a two component mixture model in which shifted exponentials feature prominently. There are, however, other factors of interest which justify its inclusion in this work.

At the time of writing their paper one of the authors of the paper in which it appears, J.D. Griffiths, was a professor of mathematics and the other, Dr J.G.Hunt, was an engineer. This collaboration across the two disciplines, although not unique, is by no means common and in the results, as one might expect, features of both can be seen.

In their methodology are several points of note, the first being a constraint placed on the weighting parameter, p, which is only allowed to take values in the range 0 to 0.5. No reasons are given for this but the use of constraints is commonplace when Bayesian methodology is applied to mixture models. Usually, however, it is one of the model parameters that is constrained rather than the weighting parameter. An important side effect of this is that no realistic interpretation can be inferred from the estimated values of the model parameter and the use of the model becomes non-parametric in nature. Another constraint used was that the value of k was never allowed to be smaller than the smallest observed headway.

Maximum Likelihood techniques and the Method of Moments were tried by Griffiths & Hunt but did not result in good enough fits between theoretical frequency and observed frequency of headways and, as a result, a hybrid approach was adopted. This hybrid method nearly always resulted in values for k which met the above constraint. When this was not the case, poor fits resulted and this was blamed on observers being, in the words of the authors "*trigger happy*" and recording very short headways. Their claim that the model is appropriate for modelling vehicle headways on single carriageway roads in urban areas will be placed under scrutiny later in this thesis.

3.2.1 Reason for choice

As with the Schuhl model, there are two reasons why the Griffiths & Hunt model is included here :-

1. It is, again, typical of the type of headway model used by highway engineers.

2. Griffiths & Hunt make significant claims for this model. However, it will be shown that this model is not suitable for use with vehicle headways.

3.3 The Gamma Exponential Model

This distribution, proposed by the author for modelling headways on dual carriageway roads, is defined by :-

$$f(t) = p\beta_1 e^{-\beta_1 t} + (1-p)\frac{\beta_2^{\alpha_2} t^{\alpha_2 - 1} e^{-\beta_2 t}}{\Gamma(\alpha_2)} \qquad (t > 0) \tag{9}$$

Figure 5 below shows a plot of the model in which the parameters are $p = 0.3$, $\beta_1 = 0.2$, $\beta_2 = 0.9$ and $\alpha_2 = 3$.

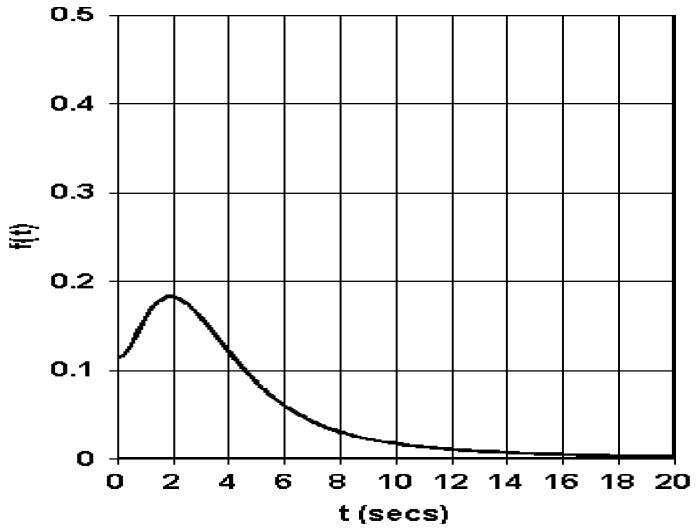


Figure 5: The Gamma Exponential Model

There are two basic reasons why this distribution has been proposed as an h.p.d.f. Firstly, there is a realistic interpretation for each component. The negative exponential first component models that part of the flow which is in free flow and the second, gamma, component models the congested part. The weighting parameter, p, can then be seen as the proportion of the traffic stream which is in free flow. Each of these distributions has, in the past, been used separately to model a traffic stream that is either free-flowing or congested (negative exponential for free flowing

and gamma for congested) and so a mixture of the two seems, at the outset, quite a reasonable proposition. However, it has been shown that, at least in certain cases, it is wrong to attach such interpretations to model parameters (Robert & Mengersen, 1997) and the values obtained are merely a means to achieving good model/data fit.

Secondly, there is no discontinuity in the p.d.f. as there is in the case of the shifted exponential distributions. Later, it will be shown that the shift parameter, k, proves extremely troublesome in the case of the Schuhl model and so its absence is, from a computational view point at least, an advantage. It could also be argued that such a sharp discontinuity is not entirely realistic with the "zone of emptiness" (Ashton, 1971) being more reasonable.

3.3.1 Reason for choice

Having dealt with the previous two models, the Gamma-Exponential distribution represents the authors own answer to the difficulties previously encountered. The model has two components, a simple exponential distribution to represent the free-flowing part of the flow and a gamma distribution for the congested. Neither component contains a shift parameter and this, it will be shown, leads to significant computational advantages.

4 Mixture models

This section begins by giving a general definition of a mixture model and moves on to illustrate their main uses, with the help of examples. The advantages and disadvantages associated with their use is discussed as is the reason for their use in this thesis. Mixture models have a wide field of application with their use being found in areas including medical statistics (Vounatsou & Smith, 1998), (Thompson et al., 1998) and recidivism studies (Copas & Heydari, 1997).

4.1 What is a mixture model?

Let X be a random variable. The probability density function (p.d.f.) of X, $\pi(x)$, is said to be a *finite mixture model* if

$$\pi(x) = \sum_{i=1}^{i=n} w_i.f_i(x|\theta_i) \tag{10}$$

where

$$\sum_{i=1}^{i=n} w_i = 1$$

and each $f_i(x|\theta_i)$ is itself a p.d.f. and is referred to as the i^{th} component of the mixture. Also, θ_i is the parameter, or vector of parameters, associated with this component and w_i is referred to as the weighting parameter.

In practice, the word *"finite"* is often omitted and this will be case in this thesis.

Robert (1996) expresses the view that it is, strictly speaking, more accurate to say that we are actually *approximating* the true p.d.f. of X to a mixture i.e.

$$\pi(x) \approx \hat{\pi}(x) = \sum_{i=1}^{i=n} w_i.f_i(x|\theta_i) \tag{11}$$

but in this thesis, for the sake of simplicity and clarity, the form of Eqn(10) will be used.

w_1	p
w_2	$(1-p)$
$f_1(x\|\theta_1)$	$\phi(x\|\mu_1, \sigma_1^2)$
$f_2(x\|\theta_1)$	$\phi(x\|\mu_2, \sigma_2^2)$
θ_1	(μ_1, σ_1^2)
θ_2	(μ_2, σ_2^2)

Table 3: Equivalance of parameters

4.1.1 An example

A two component mixture of normal distributions can be written as :-

$$\pi(x) = p\phi(x|\mu_1, \sigma_1^2) + (1-p)\phi(x|\mu_2, \sigma_2^2)$$

where

$$\phi(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Table 3 shows how the terms in Eqn(10) relate to those in this equation given that, in this case, $i = 2$.

4.2 Why are mixture models used?

4.2.1 The 'direct' use of mixture models

Principally, mixtures are used when classical distributions, e.g. normal, poisson, gamma, binomial etc, cannot adequately model the data observed. A common instance of this is when the data are multi- modal which may be the case when the data sampled are composed of two or more sub-populations, each having a different location.

This phenomenom can be found in zoological surveys as the following example shows.

Suppose observations are made of the weights (in kilogrammes) of a certain species of fish. The data, displayed as a histogram, are shown in Figure 6.

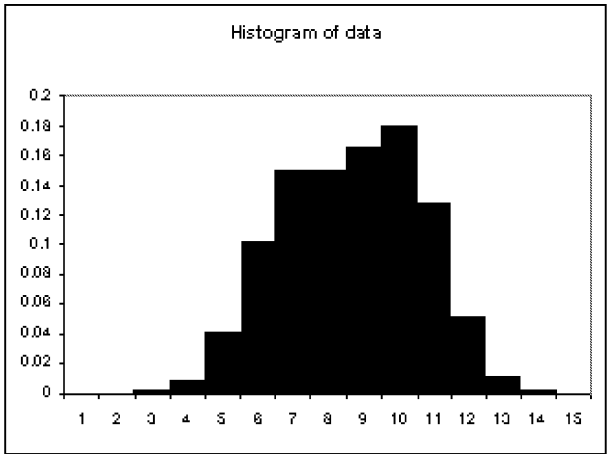


Figure 6: Histogram of fish weight data

We can see that a classical distribution is not implied by the shape of the histogram which has two modes, one at about 7 or 8 kg, the other at 10kg. The presence of two modes is often indicative that a two component mixture model would be appropriate. Also, it is often the case that zoological data of this type can be modelled using normal distributions (Titterington et al., 1983) and so we choose a mixture of two normal distributions to model the data. Such a mixture is shown in Figure 7.

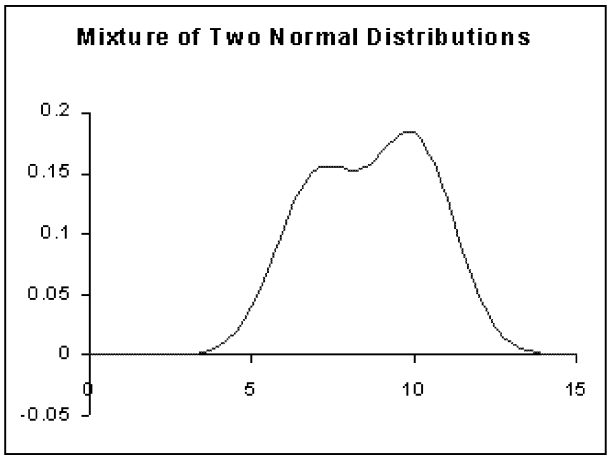


Figure 7: A mixture of two normal distributions

The parameter values used here were $\mu_1 = 10$, $\mu_2 = 7$, $\sigma_1^2 = \sigma_2^2 = 1.25$ and $p = 0.55$. In a case like this there may be biological reasons for interpreting parameter values in a realistic manner. For example, it may be valid to interpret the value of p as being the proportion of males in the population. This type of use of mixture

models is referred to as *direct* use (Titterington et al., 1983).

The data, however may not always be multimodal. One of the very earliest uses of mixture models analyses a dataset that has just one mode. The data were the ratios of forehead breadth to body length for 1000 crabs sampled at Naples by Professor W.F.R.Weldon and the analysis was carried out by Pearson (1894) who used the method of moments. This calculation would have been formidable given that no computing machinery were available at that time. Pearson's analysis was not Bayesian but there is one important point about his work : he viewed the presence of two components as evidence for the existence of two species of crab and so he interpreted his results in a physical manner and did not view the model parameters purely in model-fitting terms.

4.2.2 The 'indirect' use of mixture models

When mixture models are used in a manner that does not seek to assign such physical interpretations to parameter values, their use is said to be *indirect* (Titterington et al., 1983). One example of this is in the treatment of outliers. The contaminated normal model is defined by :-

$$\pi(x) = p\phi(x|\mu, \sigma^2) + (1-p)\phi(x|\mu, k\sigma^2) \tag{12}$$

where

$$\phi(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and $k > 1$. Here, the weighting parameter, p, is close to 1 and the two densities have the same mean. The outliers, or contaminants, have a much larger variance and can, therefore, have a very significant effect on the inferences drawn from such a sample of observations. By "separating" the outliers from the correct observations (in component 1) more accurate inferences are able to be drawn not only concerning

the data observed but also, if required, the errors made in observation.

Consider the following example. Suppose X is normally distributed with a mean of 15 and a variance of 2, as shown in Figure 8.

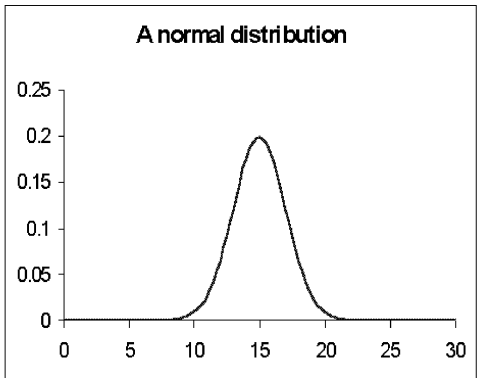


Figure 8: The normal distribution of X

Suppose now that observations of X are made. These observations are composed of correct observations, which are the majority, and erroneous ones, or outliers, whose values lie at the extremities of the true distribution of X. When the observed distribution is plotted the result is a distribution which appears normal but has heavier tails due to the presence of these outliers. The resulting distribution, which could be called "true X plus outliers", is shown in Figure 9.

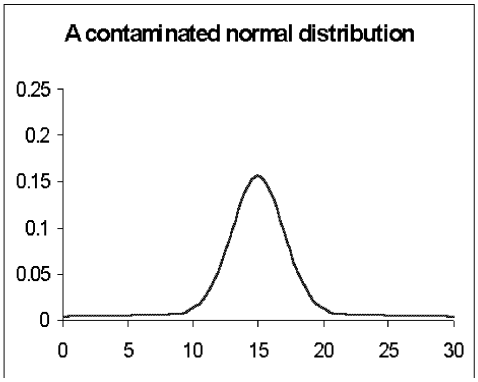


Figure 9: The distribution of X plus errors

The distribution plotted in Figure 9 can be represented by a two component mixture of normal distributions as in Equation 12 where $p = 0.85$, $\mu = 15$, $\sigma^2 = 2$ and $k = 8$. In this way it can be seen that the presence of outliers can be accommodated by the use of mixture models.

It is, perhaps, worth mentioning another use of mixture models that only occurs in Bayesian statistics. We have already summarised Bayes' Theorem by using the expression

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

and, so far, it has always been the case that the mixture is used in the likelihood. There are times, however, when a mixture model can be used to express our prior belief concerning a parameter.

Titterington et al. (1983, pp16 - 20) gives an indication of the many and varied uses to which mixture models have been put.

4.3 The use of mixtures in this thesis

In most cases where mixtures are used, the data are clearly multimodal which initially suggests that a mixture model may well be appropriate. In this thesis, however, mixture models are proposed because it is the prior belief of the experimenter(s) that the data are drawn from two sub- populations, as described in Section 2. Furthermore, we cannot state that the locations of the components are always well-separated and thus we have a case which falls outside the usual notion of mixture model estimation (Richardson & Green, 1997). We are, in this case, starting with a belief about the data and using this belief to design the model. Subsequent investigation and analysis will determine the validity, or otherwise, of this belief.

4.4 The advantages and disadvantages of mixtures

4.4.1 Advantages

Mixture models provide sufficient flexibility to model non-homogenous populations such as the zoological data previously discussed. We have seen that it is often the case that a mixture does not simply model the data *better* than a classical

distribution, but that the latter cannot model the data at all. But this flexibilty is not solely possessed by mixtures.

When the underlying probabilistic structure of a data set is too complex to be modelled by a classical distribution, mixture modelling is not the only answer. Other methods do exist but are often much more complicated to implement. An example of this is the non-parametric method of kernel density estimation. The case illustrated below shows a non-Bayesian estimate, $\hat{\pi}(x)$, of a density.

$$\hat{\pi}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where n is the number of observations, h is a constant known as the bandwidth, x_i is the ith observation and $K(.)$ is a symmetrical probability density function, having it's mode at zero. We can see that this is itself a mixture model having equally weighted components whose number is equal to the number of observations. It is clear that this approach can, in all cases except the trivial, become very cumbersome indeed.

We see, therefore, that a combination of flexibilty and computational convenience (the word *"ease"* is deliberately avoided) are the chief advantages of mixture models.

4.4.2 Disadvantages

It is, perhaps, worth mentioning that the disadvantages described here are those which apply irrespective of the method of estimation or the family of distributions from which the individual components of the mixture are drawn. Issues such as identifiability, the problem of uncertain component membership and related computational difficulties, so often associated with mixtures, will discussed later.

The Likelihood Function

The likelihood function is important to both Bayesian and frequentist alike. The latter needs to find the parameter values which maximise it. The former, however,

must multiply it by the prior distribution and then integrate it to find the posterior distributions of the parameters involved. When mixture models are involved each of these processes assumes a complexity not encountered when dealing with the classical distributions. This can be demonstrated by considering a general two component mixture model :-

Let the mixture model be denoted by

$$f(x) = pf_1(x) + qf_2(x)$$

where $p + q = 1$. Suppose, now, that we have observations $x_1, x_2, \ldots x_n$. The likelihood function is, then, given by

$$L_n = \prod_{j=1}^{n} (pf_{1j} + (1-p)f_{2j}) \tag{13}$$

It can be seen that as the number of observations rises, so the complexity of Equation 13 also increases. The data files used in this thesis contain about 200 observations and we can write the likelihood function in this case as being :-

$$L_{200} = (pf_{1,1} + qf_{2,1})(pf_{1,2} + qf_{2,2})(pf_{1,3} + qf_{2,3}) \ldots (pf_{1,200} + qf_{2,200})$$

It must be remembered that we are, in this case, considering only a two component mixture : the complexity of the likelihood function would again increase as more components are added to the model.

The missing data problem

Mixture models are an example of a "missing data" problem. The word *"missing"* is used in the sense that one item of data is unobserved. The item in question is the component membership of each observation. This is fundamental to the estimation of mixtures in as much as if the component membership were known then this process would be radically different since, given this knowledge, estaimation of the

model parameters would usually be straightforward. In Section 5 it will be shown how the Gibbs sampler can be applied to such problems.

Identifiability

Let $X \in \{0, 1, 2\}$ be a random variable that we wish to model using a mixture of two binomial distributions, where the parameters of the distributions are θ_1 and θ_2 and the weighting parameter is π. We can write :-

$$p(X = 0) = \pi(1 - \theta_1)^2 + (1 - \pi)(1 - \theta_2)^2$$

and

$$p(X = 1) = 2\pi\theta_1(1 - \theta_1) + 2(1 - \pi)\theta_2(1 - \theta_2)$$

There is a third equation, for $p(X = 2)$, i.e.,

$$p(X = 2) = 1 - p(X = 1) - p(X = 0)$$

but, since it is dependent on the previous two, it is obviously not independent. There are, however, three unknowns, i.e. π, θ_1 and θ_2. We are left with the situation where we have three unknowns but only two independent equations. There is, therefore, insufficient information for the solution of the equations and in such a case the model is referred to as being *non- identifiable*

The way in which a two component mixture is affected by this problem can be illustrated by considering a mixture of two normal distributions as in Equation 14 below :-

$$\pi(x) = p\phi(x|\mu_1, \sigma_1^2) + (1 - p)\phi(x|\mu_2, \sigma_2^2) \tag{14}$$

Now, let $p' = 1 - p$, $\mu_1' = \mu_2$, $\sigma_1^{2'} = \sigma_2^2$, $\mu_2' = \mu_1$ and $\sigma_2^{2'} = \sigma_1^2$. If these values are

substituted into Equation 14, the following is obtained :-

$$\pi(x) = p'\phi(x|\mu_1', \sigma_1^{2'}) + (1 - p')\phi(x|\mu_2', \sigma_2^{2'}) \tag{15}$$

which gives exactly the same p.d.f. as Equation 14. This effect is common in mixture models and Diebolt & Robert (1994) state quite bluntly that "Mixture models are not identifiable".

Interpretation of parameter values

There is a certain amount of evidence that, in come cases, it is invalid to assign a realistic interpretation to parameter values (Robert & Mengerson, 1997). However, the ability to interpret parameters is of sufficient value that attempts to do so will be made in this thesis.

Given the above, one could easily become pessimistic and conclude that the disavantages of mixture models outweigh their advantages and reject them as a worthwhile tool. However, there are techniques that can be used to overcome each drawback and it will be shown that mixtures are well worth using despite these difficulties.

5 Gibbs Sampling for Mixture Models

In Section 2 it was shown how intractable integrals, sometimes of high dimensionality, can arise in Bayesian statistics. It was also stated that techniques exist whereby the difficulties presented by these integrals can be overcome. A good review of some of these techniques can be found in Swartz & Evans (1995) although the continual increase in computational power must be borne in mind when comparing different methods.

The chosen computational technique for use in this thesis is Gibbs sampling. There are numerous reasons for its choice and these will be stated later but before that we will describe those techniques familiarity with which is necessary in order to understand Gibbs sampling itself.

5.1 Monte Carlo integration

Monte Carlo integration is not new : by the time Hammersley & Handscomb published their wide-ranging monograph on the subject in 1964 it was already well established. It was not, however, until 1971 when it was first used in a Bayesian context by Stewart & Johnson (1971), although Kloek & van Duk (1978) brought it to the attention to the wider Bayesian community.

It can be demonstrated with a simple example.

Suppose we wish to evaluate the following integral :-

$$k = \int_{-\infty}^{\infty} f(x)dx$$

We can rewrite this equation as follows :-

$$k = \int_{-\infty}^{\infty} g(x)\frac{f(x)}{g(x)}dx$$

where g(.) is a probability density function. Now,

$$k = E[\frac{f(x)}{g(x)}]$$

provided X has the distribution $g(x)$. We now generate a random sample from $g(x)$, $x_1, x_2, x_3, \ldots, x_n$, and for each x_i evaluate $z_i = f(x_i)/g(x_i)$. Let the mean of this sample be denoted by $\bar{r}$. We know that

$$\bar{r} = \frac{\sum_{i=1}^{n} z_i}{n} \rightarrow k \quad \text{almost surely, as } n \text{ tends to } \infty.$$

By the term *almost surely* we mean that the convergence is stochastic rather than deterministic.

Although some authors have claimed that Monte Carlo can only ever be a method of choice for rough estimates of numerical quantities (Kalos & Whitlock, 1986) and despite the fact that O'Hagan (1987) raises fundamental objections to its use, Monte Carlo integration does form the basis for techniques very widely used in Bayesian statistics. However, also fundamental to these techniques is the question of sampling from a probability density function.

Sampling random variables.

Suppose that X is a random variable whose p.d.f. is $f(x)$ and that we wish to draw a random sample $x_1, x_2, x_3 \ldots x_n$ from $f(x)$. There are numerous methods by which this can be achieved and for any given problem the choice must be made carefully. It would not be appropriate to detail all current methods but Table 4 below provides some helpful references.

One method, however, that can be detailed here is that of rejection sampling since, at this point, we can introduce two important points. If we wish to sample values from a p.d.f. $f(x)$ or from a distribution that is proportional to $f(x)$ rejection sampling requires an envelope function $g(x)$ such that, for all x, $g(x) > f(x)$. It

Author(s) & Date	Subject
Hammersley & Handscome, 1964	Theoretical background to sampling
Kalos & Whitlock, 1986	As above
Gilks, W, 1996 (p79)	An introduction to Rejection Sampling
Gilks, W. & Wild, P, (1992)	Adaptive Rejection Sampling

Table 4: Sampling references

must also be the case that we can sample from $g(x)$, where by this we mean that we can sample from the distribution with pdf proportional to $g(x)$. The algorithm for rejection sampling can be written as follows :-

```
Repeat {
   Sample a value Q from g(.)
   Sample a value U from a uniform U(0, 1)
   if U <= f(Q)/g(Q) then accept Q
   }
Until one Q is accepted
```

The two points arising from this are

1. We still have to sample from g(x) and, in a sense, we are no further forward since we have merely substituted $g(x)$ for $f(x)$.

2. Even when this hurdle is overcome, we still have to sample from $U(0, 1)$.

To overcome the first difficulty we have to use a function $g(x)$ such that a method such as the following can be used. We have already stated that $g(x)$ is *proportional* to a p.d.f., i.e.

$$\int_{\Omega_X} g(x)dx = k \quad \text{where} \quad k \geq 1 \qquad \text{and } g(x) \geq 0.$$

Now, let

$$\int_0^x g(u)du = G(x)$$

and let u be a value sampled from $U(0,1)$. We can sample a value from $g(x)$ by solving the equation

$$G(x) = uk$$

Choosing an appropriate $g(x)$ for a particular $f(x)$ is clearly crucial: not only must $g(x) > f(x)$, for all x, or else the method will not work but also, for all x, the ratio $f(x)/g(x)$ must not be too small or only a correspondingly small number of candidate points will be accepted and the sampling will take longer. Gilks and Wild (1992) detail a very powerful general method that is valid for a large class of $f(x)$, and their method was used in this project.

Secondly, we have to sample from $U(0,1)$. In practice, this can be very easily done by using what is a standard library function of most high level computer languages. It must be remembered, however, that the values generated here are not truly random and are described as *pseudo*-random because they are produced by deterministic algorithms. A general form of a popular algorithm can be expressed as follows :-

$$x_i \equiv ax_{i-1} + c \pmod{m}$$

where m is a large integer and a, c, and x_i are integers between 0 and $m-1$.

(The notation signifies that x_i is the remainder when $ax_{i-1} + c$ is divided by m.)

The choice of the constants a, c and m is crucial to the success of the algorithm and this is demonstrated by Kalos & Whitlock (1986). These authors conclude, however, that sufficient work has been done in this field that the values of these parameters can be chosen with confidence. A feature of this type of pseudorandom generator (prn) is that after m steps (at most) the sequence will be repeated and is, therefore, described as periodic. A full description of prn's can be found in Hammersley & Handscome (1964) where the above points are discussed in detail.

5.2 Markov chain Monte Carlo integration

Perhaps the biggest single development in Monte Carlo integration came about with the publication of a paper in a journal that is possibly not well known to the majority of statisticians. In June, 1953 the Journal of Chemical Physics published a paper entitled "Equation of State Calculation by Fast Computing Machine" by five physicists one of whom was Edward Teller. (Metropolis, et al., 1953)

In this paper a sample from the required distribution (known as the *target* distribution) is obtained by simulating values from a Markov chain whose stationary distribution is the target distribution. At first sight this ingenious method seems as if it ought to be highly complex but the reverse is true. The following explanation of its operation is based on that contained in Kalos & Whitlock (1986). This explanation, it should be noted, goes somewhat beyond the original Metropolis algorithm and is more generalised in its form and is, therefore, more akin to that of Hastings (1970).

Consider a space Ω in which a particle can perform a random walk. Let the positions of the particle as it performs the random walk be denoted by $X_1, X_2, X_3, \ldots X_n$. Now, each X_i, where $1 \leq i \leq n$, has it's own p.d.f. which we will denote by $\phi(X_i)$. Each X_i is, in stochastic terms, a function of X_{i-1} only and so the random walk is a regular Markov chain. Therefore, since this is the case, we can state that :-

$$\phi(X_n) \rightarrow f(X) \text{ as } n \rightarrow \infty$$

where $f(X)$ is the p.d.f. of the particle being at X when the random walk has reached equilibrium. Now, consider any two arbitrary points in Ω say, X and Y. Figure 10 may be helpful :-

The random walk is such that the probability of a move from X to Y is exactly

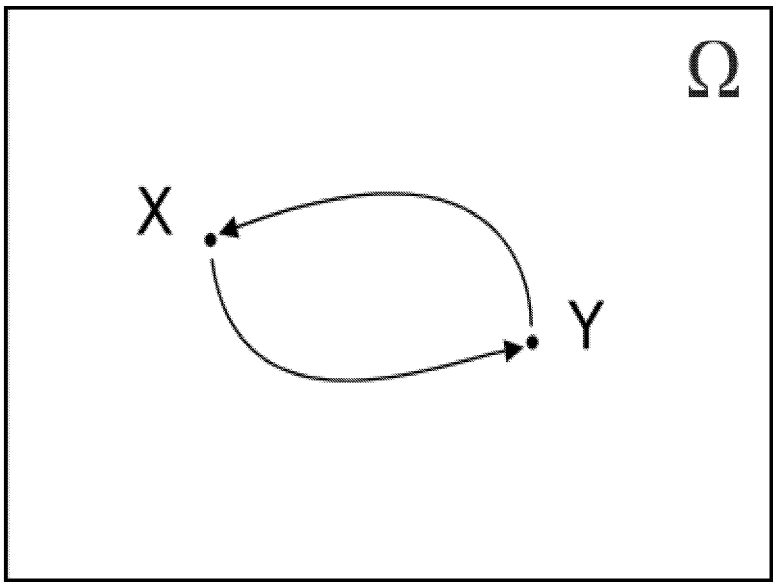


Figure 10: Two points in space

the same as the reverse move. This can be expressed by the following equation.

$$K(X|Y)f(Y) = K(Y|X)f(X)$$

This relationship is called *detailed balance* and must be satisfied for this algorithm to work. The terms are explained below :-

$K(X\|Y)f(Y)$	This is probability of the particle moving from Y to X.
$f(Y)$	This is the *a priori* probability of finding the particle at Y.
$K(X\|Y)$	This is the *conditional* probaility that the particle will move to X **given** that it is currently at Y.

Now, let

$$K(X|Y) = A(X|Y)T(X|Y)$$

where $T(X|Y)$ is, theoretically, *any* p.d.f. If we now sample a value from $T(X|Y)$ and accept it with probability $A(X|Y)$, we can progress the random walk on this

basis.

Let us now consider the limiting, or stationary, distribution of the random walk by considering a the particle being at the point X after $n+1$ moves. There are two ways this event can happen :-

1. The particle can be at Y, a move can be proposed from $T(X|Y)$ and accepted with probability $A(X|Y)$. Since we are considering all Y's, the expression for this is $\int A(X|Y)T(X|Y)f(Y)dy$, i.e., we sum over all Y.

2. Conversely, the particle can already be at X and the proposed move is rejected. The expression for this is $\int(1 - A(Y|X))T(Y|X)f(X)dy$

We can now write

$$\phi(X_{n+1}) = \int A(X|Y)T(X|Y)f(Y)dy + \int(1 - A(Y|X))T(Y|X)f(X)dy$$

Since detailed balance holds, expansion and simplification yield the result

$$\phi(X_{n+1}) = \int T(Y|X)f(X)dy$$

and, because $T(Y|X)$ is a p.d.f., we are left with the important result

$$\phi(X_{n+1}) = f(X)$$

We have, thus, shown that the stationary distribution of the randon walk, which is a Markov chain, is $f(X)$ All that is required now is to find a suitable $A(X|Y)$ so that detailed balance is satisfied. Recall that by this we mean that

$$A(X|Y)T(X|Y)f(Y) = A(Y|X)T(Y|X)f(X)$$

since

$$K(X|Y) = A(X|Y)T(X|Y)$$

A very common choice is

$$A(X|Y) = \min(1, q(X|Y))$$

where

$$q(X|Y) = \frac{T(Y|X)f(X)}{T(X|Y)f(Y)} \geq 0$$

We can now state the algorithm for sampling from $f(X)$ completely :-

1. Choose a suitable proposal distribution, $T(.|.)$ Note that the requirement $T(.|.) \geq f(.)$ does not apply here, unlike in simple rejection sampling.

2. Choose a starting value, X_0, for the process.

3. Sample a possible next value for X, in this case X_1' from $T(X_1'|X_0)$

4. Compute $q(X_1'|X_0)$ and, hence, $A(X_1'|X_0)$

5. With probability $A(X_1'|X_0)$, set $X_1 = X_1'$, otherwise set $X_1 = X_0$

6. Repeat steps 3, 4 & 5 until the Markov chain reaches equilibrium, but do not save any simulations.

7. Repeat steps 3, 4 & 5 saving as many simulations as required.

The brevity of the above algorithm conceals the care that must be taken at each step, the majority of which have their own special difficulties.

- A proposal distribution must be such that it can be sampled from with a minimum of difficulty but the closer it is to the target distribution then the quicker the Markov chain will reach it's state of equilibrium. These two factors need to be balanced for optimum performance of the algorithm.

- The choice of starting value can be of crucial significance to the success, or otherwise, of the algorithm. A good strategy is to use several dispersed starting values and then compare results.

- At the present time, assessment of when a chain has reached equilibrium is not a straightforward process. The portion of the chain before equilibrium has been reached is known as the "burn-in" and it's length is difficult to determine. Graphical plots of the output of the chain can be useful and will be used in this thesis.

- Likewise, the number of values to be saved from a simulated chain is not easy to determine and, in some cases, it can be helpful to save not every value but every kth value to minimise the effect of correlation.

The above explanation is necessarily brief but there is a large, and still growing, literature on the subject and very good detailed discussions on these points can be found in Gelman (1996), Raftery & Lewis (1996) and Gilks & Roberts (1996).

So far, it has not mattered whether or not the X we have considered is multidimensional : the algorithm holds good for any dimensionality. In practice, however, we specifically need to consider the case of a multidimensional X and it is also helpful to modify our notation slightly so that it is in line with the more recent literature. For the target distribution we will now use $\pi(.|.)$ instead of $f(.|.)$, for the proposal distribution $q(.|.)$ instead of $T(.|.)$ and for the acceptance probability $\alpha(.|.)$ instead of $A(.|.)$.

Suppose that the dimensionality of X is h, i.e.

$$X = \{X_1, X_2, X_3, \ldots X_h\}$$

and let

$$X_{-i} = \{X_1, X_2, \ldots X_{i-1}, X_{i+1} \ldots X_h\}$$

i.e. X_{-i} is simply X without X_i. Every iteration of the Markov chain consists of h separate steps each of which update a component, X_i of X. Thus, we are treating each component separately which is clearly easier than considering X as a whole.

After t such iterations, we can define the state of the Markov chain by

$$X_t = \{X_{t,1}, X_{t,2}, X_{t,3}, \ldots X_{t,h}\}$$

Suppose, now, that we have completed t complete iterations and have updated $i-1$ components during iteration $t+1$, i.e., the next component to be updated is $X_{t,i}$. At this point the chain will be in the following state :-

$$X_{t,i-1} = \{X_{t+1,1}, X_{t+1,2} \ldots X_{t+1,i-1}, X_{t,i}, X_{t,i+1} \ldots X_{t,h}\}$$

Next a candidate, Y_i is sampled from the proposal distribution

$$Y_i \sim q(Y_i | X_{t,i}, X_{t,-i})$$

Notice that this distribution only generates candidates for the ith component and that it is conditional on the state of X at the "time" of sampling. The candidate is accepted with probability $\alpha(Xi, X_{-i}, Y_i)$ where

$$\alpha(X_i, X_{-i}, Y_i) = \min\left(1, \frac{\pi(Y_i|X_{-i})q(X_i|Y_i, X_{-i})}{\pi(X_i|X_{-i})q(Y_i|X_i, X_{-i})}\right)$$

If the candidate is accepted then $X_{t+1,i}$ is set to Y_i. If the candidate is not accepted then $X_{t+1,i}$ is set to $X_{t,i}$.

5.3 Gibbs sampling

Once again, an important technique has entered statistics from statistical physics. The *heat bath algorithm* was used by Geman & Geman (1984) in their work on image analysis and renamed the Gibbs sampler. Several other papers introduced the technique into mainstream statistics and Smith & Roberts (1993) provide a lucid explanation of the algorithm.

In Gibbs sampling **all** candidate points are accepted because of the choice of

proposal distribution, i.e.,

$$q(Y_{.i}|X_{.i}, X_{-i}) = \pi(Y_{.i}|X_{-i})$$

The right hand side of this equation is referred to as a full conditional distribution because it is the distribution of the candidate point *conditional* upon the values of *all* the other components of X at the "time" of sampling. This important concept can, perhaps, best be demonstrated by a simple example.

5.3.1 Full Conditional Distributions

In Section 2.5.3 two examples were used to show how complex integrals routinely arise in Bayesian statistics. Recall that the second of these examples used a gamma distribution with scale and index parameters β and α respectively. Given a set of observations $(t_1, t_2 \ldots, t_k)$, we arrived at the following joint posterior for α and β :-

$$f(\alpha, \beta|t_1, t_2, \ldots, t_k) \propto \alpha^{a-1} e^{-b\alpha} \beta^{\alpha k} (\prod_{i=1}^{k} t_i)^{\alpha-1} \beta^{c-1} e^{-\beta(\Sigma_{1=1}^{k} t_i + d)} (\Gamma(\alpha))^{-k}$$

where a and b are the prior parameters for α and c and d are those for β. (Note that the unsimplified form is used here ; the reason for this will soon become apparent.) In terms of what has gone before, we can say that $X = (\beta, \alpha)$. The sampling proceeds as follows :-

1. Choose arbitrary starting values for β and α and let these be denoted by β_0 and α_0.

2. We now have to sample a value for β_1, i.e., we need to sample from $\pi(\beta_1|X_{-\beta_1})$ which is equal to $\pi(\beta_1|\alpha_0)$. This distribution, the full conditional distribution of β_1, can be found by substituting α_0 into the joint posterior for α and β above, which gives :-

$$f(\beta_1|\alpha_0, t_1, t_2, \ldots, t_k) \propto \alpha_0^{a-1} e^{-b\alpha_0} \beta_1^{\alpha_0 k} (\prod_{i=1}^{k} t_i)^{\alpha_0-1} \beta_1^{c-1} e^{-\beta_1(\Sigma_{1=1}^{k} t_i + d)} (\Gamma(\alpha_0))^{-k}$$

At this point a "term-by-term" explanation will be helpful :-

α_0^{a-1}. The value of α_0 has already been chosen and so this term is a constant. Now because we can sample from a function which is *proportional to* a p.d.f, this term can be ignored.

$e^{-b\alpha_0}$ Because b is also a fixed prior parameter, this term is also fixed and, for our purposes here, can be ignored.

$(\prod_{i=1}^{k} t_i)^{\alpha_0 - 1}$. Again, this is a constant and can be ignored.

$(\Gamma(\alpha_0))^{-k}$. Again, a constant.

3. If we now combine all the remaining terms, we can write

$$f(\beta_1|\alpha_0, t_1, t_2, \ldots, t_k) \propto \beta_1^{\alpha_0 k + c - 1} e^{-\beta_1(\Sigma_{1=1}^{k} t_i + d)}$$

which is a simple gamma distribution whose parameters are $\alpha_0 k + c$ and $\Sigma_{1=1}^{k} t_i + d$.

4. We now sample $f(\beta_1|\alpha_0, t_1, t_2, \ldots, t_k)$ from a distribution that is *proportional* to Gamma$(\alpha_0 k + c, \Sigma_{1=1}^{k} t_i + d)$ which, given the method proposed by Gilks & Wild (1992), is quite straightforward. Remember that we can sample from a function that is *proportional* to a p.d.f. using this method and it is not necessary to know the value of the normalising constant.

5. Lastly, as far as this iteration is concerned, we have to sample a value for α_1, i.e., we need to sample from $\pi(\alpha_1|X_{-\alpha_1})$ which is equal to $\pi(\alpha_1|\beta_1)$. In this case we substitute α_1 and β_1 into the joint posterior but this time we ignore terms which do not involve α_1. We can then write :-

$$f(\alpha_1|\beta_1, t_1, t_2, \ldots, t_k) \propto \alpha_1^{a-1} e^{-b\alpha_1} \beta_1^{\alpha_1 k} (\prod_{i=1}^{k} t_i)^{\alpha_1 - 1} (\Gamma(\alpha_1))^{-k}$$

Whilst this function is not standard, it can be sampled from without too much

difficulty.

6. To continue, we start the cycle at 2. and repeat the process, starting by sampling from $\pi(\beta_2|\alpha_1)$.

It is worth pointing out that all we do to find the full conditional distribution of a parameter is to eliminate from the joint posterior distribution those terms which do not contain the parameter in question. Note also that the full conditional distribution (f.c.d.) for each parameter, whilst being different at each iteration, is of the same form at each iteration. For example, in this case the f.c.d. for β is always proportional to a Gamma distribution but with a different parameters at each iteration.

5.4 Gibbs sampling for mixture models

5.4.1 The missing data structure

We begin by assuming that each observation "belongs to" or is "generated by" a particular component whether we are using the mixture model parametrically or otherwise. Suppose, now, that for each observation x_i there exists a corresponding z_i whose value indicates the component membership of the observation, as shown below. If there are k components in the mixture, then z is an integer where $1 \leq z \leq k$.

$$\begin{array}{ccccccc} x_1, & x_2, & x_3, & \ldots & x_n \\ \downarrow & \downarrow & \downarrow & \ldots & \downarrow \\ z_1, & z_2, & z_3, & \ldots & z_n \end{array}$$

Now, let $Z = \{z_1, z_2, z_3, \ldots, z_n\}$, and suppose that a two component mixture model is being used, i.e. $k = 2$. Each z can take the value of either 1 or 2. For example, if we had 10 observations Z could be equal to $\{1, 1, 2, 2, 1, 2, 1, 2, 2, 1\}$ and this would mean that observation 1 would belong to component 1, observation 2 would belong to component 1, observation 3 would belong to component 2 and so on.

Unfortunately, the vector $Z = \{z_1, z_2, z_3, \ldots, z_n\}$ is not observed and is referred to as the missing data.

5.4.2 Stochastic allocation

At each sweep of the Gibbs sampler observations do need to be temporarily allocated to a particular component and, since component membership is not observed, it is done stochastically.

Let I be an indicator variable and suppose, again, that we are using a two component mixture model defined by :-

$$\pi(x) = pf_1(x) + (1-p)f_2(x)$$

If we let $I \in \{1, 2\}$ we can say, for an observation, x_i, that if $I_i = 1$ then the observation belongs in the first component and if $I_i = 2$ then it belongs in the second. To determine the probability of an observation being in component 1 we proceed as follows :-

We require $Pr(I_i = 1|x_i)$. From Bayes Theorem, we have

$$Pr(I_i = 1|x_i) = \frac{Pr(x_i|I_i = 1).Pr(I_i = 1)}{Pr(x_i)}$$

which gives

$$Pr(I_i = 1|x_i) = \frac{pf_1(x_i)}{pf_1(x_i) + (1-p)f_2(x_i)}$$

It is, perhaps, worth stating that the use of Bayes Theorem is valid here because of the conditional independence of x_i with respect to all the other unknowns.

If we let the RHS of the above equation equal a then we say that we allocate each x_i to the first component with probability a. Before we describe each sweep (or iteration) of the Gibbs sampler we need, firstly, to be rather more specific in our

definition of the mixture model involved. Let

$$\pi(x) = pf_1(x|\theta_1) + (1-p)f_2(x|\theta_2)$$

Also, let the prior distribution for θ_1 be $\pi_0(\theta_1)$ and let the prior distribution for θ_2 be $\pi_0(\theta_2)$ and for p, $\pi_0(p)$

Step 1 Choose starting values for all parameters.

Step 2 Using current parameter values, calculate the allocation probability, a, for each observation and allocate to either component 1 or 2 accordingly. This will mean that n_1 observations will be allocated to component 1 and n_2 to component 2.

Step 3 Simulate a value of p from it's full conditional distribution. This will be denoted by $fcd(p)$ and again we use the relationship *Posterior* $\propto$ *Prior* $\times$ *Likelihood*.

$$fcd(p) \propto \pi_0(p) \times p^{n_1}.(1-p)^{n_2}.$$

Step 5 Simulate a value of θ_1 from it's fcd.

$$fcd(\theta_1) \propto \pi_0(\theta_1) \times \prod_{x \in C_1} f_1(x|\theta_1)$$

Here, the notation $x \in C_1$ means " each x that has been allocated to the first component". Also, the assumption is that θ_1 is a single parameter and not a vector. If it were a vector, the fcd of each constituent parameter would be found by the method previously described.

Step 6 Simulate a value of θ_2 from it's fcd.

$$fcd(\theta_2) \propto \pi_0(\theta_2) \times \prod_{x \in C_2} f_1(x|\theta_2)$$

Repeat steps 2 - 6 as required.

5.5 Application to the models used

So far, the variable modelled has been denoted by x. Now, because the discussion is focused on the actual models used, the variable will be denoted by t since the quantity modelled is a unit of time. Also, we will use the notation $B(a, b)$ to represent a beta distribution with parameters a, b and $G(c, d)$ to represent a gamma distribution with parameters c, d.

5.5.1 The Griffiths and Hunt Model

This model was defined in Section 3 by :-

$$f(t) = \begin{cases} 0, & (t < k) \\ p\beta_1 e^{-\beta_1(t-k)} + (1-p)\beta_2 e^{-\beta_2(t-k)} & (k \leq t) \end{cases} \tag{16}$$

Prior Distributions

Table 5 below sets out the prior distributions chosen for the various parameters :-

Parameter	Prior Distribution
p	$B(\phi, \psi)$
β_1	$\propto \beta_1^{\gamma-1} e^{-\delta\beta_1}$, i.e. $G(\gamma, \delta)$
β_2	$\propto \beta_1^{\gamma-1} e^{-\delta\beta_1}$, i.e. $G(\gamma, \delta)$
k	$\propto k^{\theta-1} e^{-\nu k}$, i.e. $G(\theta, \nu)$

Table 5: Prior distributions for the Griffiths & Hunt model

In Section 7 experimental runs of the Gibbs sampler will be described using particular values for prior distribution parameter values.

Sampling

After each prior parameter has been given a value, and each model parameter an initial value, sampling can begin. Each "sweep" of the sampler comprises two main

steps :-

1. Partition the data into two sub-samples, S_1 and S_2, each comprising n_1 and n_2 observations respectively. This is achieved by assigning each observation, t_i, to the first sub-sample with probability a where

$$a = \frac{p\beta_1 e^{-\beta_1(t_i-k)}}{p\beta_1 e^{-\beta_1(t_i-k)} + (1-p)\beta_2 e^{-\beta_2(t_i-k)}}$$

   If an observation is not assigned to the first component, it is assigned to the second.

2. - Sample p from the $B(n_1+\phi, n_2+\psi)$
   - Sample β_1 from $G(n_1+\gamma, \Sigma_{i\in S_1}t_i - n_1k + \delta)$
   - Sample β_2 from $G(n_2+\gamma, \Sigma_{i\in S_2}t_i - n_2k + \delta)$
   - Sample k from the distribution with density proportional to $k^{\theta-1}\exp[-k(\nu - n_1\beta_1 - n_2\beta_2)] \quad (k \leq t_{\min})$ where $t_{\min}$ is the smallest observation.

The sampling distribution for k has not been described as a gamma distribution because the coefficient of $-k$ can, under certain circumstances, be less than zero. This, however, does not require any fundamental change to the sampling procedure, since samples are obtained using a method similar to that described in Gilks & Wild (1992).

When examining the fcd's, it can be seen that when the data do not give any information about a particular parameter, it is the prior distribution of that parameter from which a value is sampled. This entirely consistent with the principles of Bayesianism and also serves to verify the algebra used to obtain an fcd.

5.5.2 The Schuhl Model

Again, this model was defined in Section 3 by :-

$$f(t) = \begin{cases} p\beta_1 e^{-\beta_1 t}, & (0 < t < k) \\ p\beta_1 e^{-\beta_1 t} + (1-p)\beta_2 e^{-\beta_2(t-k)} & (k \leq t) \end{cases} \tag{17}$$

We see here that this model has at least the possiblility of modal separation.

Prior Distributions

The similarity between the models means that we can give the same prior distributions to both sets of parameters.

Sampling

Again this is very similar to the Griffith & Hunt case but there are important differences. In Step 1, the allocation probability, a, is determined as follows :-

Let t_{min2} be the smallest observation in the second component. If $t_i < t_{min2}$ then $a = 1$. If $t_i \geq t_{min2}$ then the following equation is used :-

$$a = \frac{p\beta_1 e^{-\beta_1 t_i}}{p\beta_1 e^{-\beta_1 t_i} + (1-p)\beta_2 e^{-\beta_2(t_i - k)}}$$

and Step 2 then proceeds as follows :-

- Sample p from $B(n_1 + \phi, n_2 + \psi)$
- Sample β_1 from $G(n_1 + \gamma,\ \Sigma_{i \in S_1} t_i + \delta)$
- Sample β_2 from $G(n_2 + \gamma,\ \Sigma_{i \in S_2} t_i + \delta)$
- Sample k from the distribution with density proportional to $k^{\theta-1} \exp[-k(\nu - n_2\beta_2)] \quad (k \leq t_{min2})$,

where $t_{\min 2}$ is the value of the smallest observation in S_2.

5.5.3 The Gamma Exponential Distribution

This distribution, proposed by the author, was defined in Section 3 by :-

$$f(t) = p\beta_1 e^{-\beta_1 t} + (1-p)\frac{\beta_2^{\alpha_2} t^{\alpha_2 - 1} e^{-\beta_2 t}}{\Gamma(\alpha_2)} \qquad (t > 0)$$

The constraint that $\alpha_2 > 1$ is imposed on the model.

Prior Distributions

The prior distributions used in Table 6 below :-

Parameter	Prior Distribution
p	$B(\phi, \psi)$
β_1	$G(\gamma, \delta)$
β_2	$G(\theta, \nu)$
α_2	$G(\omega, \kappa)$.

Table 6: Prior distributions for the Gamma Exponential model

Sampling

Again at each sweep the sample is partitioned into two sub-samples, S_1 and S_2 by allocating each observation (indexed by i) to S_1 with probability a where

$$a = \frac{p\beta_1 e^{-\beta_1 t_i}}{p\beta_1 e^{-\beta_1 t_i} + (1-p)\frac{\beta_2^{\alpha_2} t_i^{\alpha_2 - 1} e^{-\beta_2 t_i}}{\Gamma(\alpha_2)}}$$

As before S_1 will contain n_1 observations and S_2 will contain n_2, and their sums will be $\sum_{i \in S_1} t_i$ and $\sum_{i \in S_1} t_i$ respectively. The sampling step can now take place and is as follows :-

- Sample p from $B(\phi + n_1, \psi + n_2)$
- Sample β_1 from $G(\gamma + n_1, \delta + \sum_{i \in S_1} t_i)$
- Sample β_2 from $G(\theta + \alpha_2 n_2, \gamma + \sum_{i \in S_2} t_i)$

- Sample α_2 from the distribution with density proportional to

$$\frac{\alpha_2^{\omega-1}.\exp[-\alpha_2\kappa].\beta_2^{n_2\alpha_2}.(\prod_{i\in S_2} t_i)^{\alpha_2-1}}{[\Gamma(\alpha_2)]^{n_2}}$$

  where $\alpha_2 > 1$

It is immediately clear that the fcd of α_2 is not a standard distribution. However, its log-concavity is verified by Dey, Kuo & Sahu, (1995) and it is sampled from using the method of Gilks & Wild (1992).

It should be noted that the above algorithm represents what might be termed the *default* algorithm, as far as this model is concerned, in that others will also be considered. These variants will be described in full in subsequent section since the point here is solely to illustrate the theory of Gibbs sampling for mixture models.

5.6 Advantages and disadvantages

Mixture modelling is only one use of Gibbs sampling and the following sub-sections apply to all uses of this technique. It has to be pointed out, however, that the advantages and disadvantages associated with Gibbs sampling are exaggerated when mixture models are involved. For instance when the discussion centres on the necessity of checking for convergence then, in the case of mixture models, even greater care must be taken with this part of the analysis.

5.6.1 Advantages

The chief advantage of Gibbs sampling is its simplicity. This is largely due to the fact that the proposal distribution is simply the fcd of the parameter concerned. This means that the practitioner is not faced with the usual dilemma of McMC of choosing an appropriate proposal distribution. Also, because all candidate points are accepted, there is a gain in computational efficiency.

Also, a large number of fcd's that arise from the use of common models can quite

easily be sampled from using the method of Gilks & Wild (1992). All the fcd's in this project were sampled from using this method.

Another advantage of Gibbs sampling is that it is straightforward to obtain marginal posterior distributions of functions of parameters. Suppose that the model concerned has two parameters, θ_1 and θ_2, whose fcd's are known and whose marginal posterior distributions are required. Suppose, also, that we require the marginal posterior distribution of $(\theta_1 + \theta_2)^2$. It can be found by adding just one step to every sweep of the Gibbs sampler, as shown below :-

- Sample a value for θ_1 from its fcd and write to file
- Sample a value for θ_2 from its fcd and write to file
- Compute $(\theta_1 + \theta_2)^2$ and write to file

This facet of Gibbs sampling will be used in Section 7 where the means of the two model components will be calculated and compared at every sweep as sampling progresses.

Also, although the algebra can sometimes appear complex, there is always one check that can be applied not only here but in any Bayesian analysis. This is that if the data do not give any information about a parameter, then what remains is the prior distribution for that parameter.

5.6.2 Disadvantages

We know that a regular Markov chain converges to its target distribution but it can be, in practice, difficult to determine when this has actually happened. Also, it is known that this convergence can be slow to occur when mixture models are used and that the problem of identifiability gives rise to the phenomenom of "label switching", a good description of which can be found in Stephens (2000). Label switching arises from the non-identifiability of the mixture model. Equations 14 & 15 in Section 4.4.2 were used to show that non-identifiability can give rise to

the interchangability of component parameters. If, as is the case in the example given, all the parameters change, then what has effectively happened is that the components themselves have changed labels. In terms of Gibbs sampling, we can see the effect of this phenomenom in Figure 20 in Section 7.6.1. where the parameters β_1 and β_2 often make co-ordinated "jumps" between different zones of their respective parameter spaces.

At the present time, there is no universal test that can be applied to all chains and the process known as convergence diagnostics is still something of an art (Stephens, 2001). A side effect of this is that many computer runs, each generating quite a lot of data, may need to be made. Although this no longer means that storage presents a problem, care must be taken in choosing a suitable nomenclature for the files generated and documentation must be precise.

6 The data and the software

Having described the theoretical basis for the computational algorithm to be used, the data that will be subjected to that algorithm and the software that will perform the computations are now described.

6.1 The data

6.1.1 Data definition

The data consist of vehicle time headways measured in seconds and stored sequentially as an ASCII file. The first ten observations from File 2 are shown below :-

```
6.49
2.43
3.73
6.16
6.38
0.70
1.19
5.78
3.68
5.73
```

Three data files have been used in this project, referred to as Files 1, 2 & 3. The number of observations in each file are 150, 200 and 205 respectively. The data were collected at two sites in Sunderland, Tyne and Wear.

6.1.2 Site location

File 1 was collected on a single carriageway local distributor road in a suburb in the north of the city towards the end of the evening rush hour in July 2000. Files

2 and 3 were both collected at the same site, but on different days, on a high speed dual carriageway to the west of the city which connects the city centre to the motorway and trunk road network as well as acting as a primary distributor. File 2 was collected immediately before the evening rush hour while File 3 was collected during it.

6.1.3 Method of collection

The data were collected by a single observer seated in a stationary vehicle located close to the site in question. In the case of Files 2 and 3 the vehicle was located in a layby at the side of the road. A computer program was used to capture the time between vehicles observed passing a fixed point. (The program was written in C and ran on a laptop computer : a listing of the source code can be found in the Appendix.)

When the first vehicle passes the fixed point the space bar is pressed and timing begins. (The DOS clock is utilised) When the next vehicle passes, the space bar is pressed again and the time between these two events is calculated and written to file. The process is repeated for as long as required and terminates when the observer presses Ctrl-Q. Between events what are actually counted are "ticks" of the DOS clock. These are then converted to seconds by dividing by 18.5.

It is acknowledged that this method is far from sophisticated and that there are two main sources of error :-

1. The measurements depend on the reaction time of the observer. This may not be as significant as first thought since it is the time between successive events that is being measured and not any measure of "absolute" time at which an event occurs. In this way errors can cancel out.

2. Since each "tick" of the DOS clock occurs every 1/18.5 of a second, time can only be measured to an accuracy of 0.05 (2 d.p.) of a second.

This method does, however, have distinct advantages ;-

1. No specialist equipment is required. Laptop computers are now widely available to all who would wish to carry out this kind of work.

2. No special skills are needed and a person can be instructed in data gathering in just a few minutes.

3. Data can be collected very quickly and do not need any "post- processing" such as may be the case if film or video methods are used.

It is believed that, for the purposes of this project, the advantages outweigh the disadvantages.

One final point relating to data collection must be made in that issues relating to health and safety are, of course, important and must not be neglected.

Histograms

Histograms of the three data files are shown below.

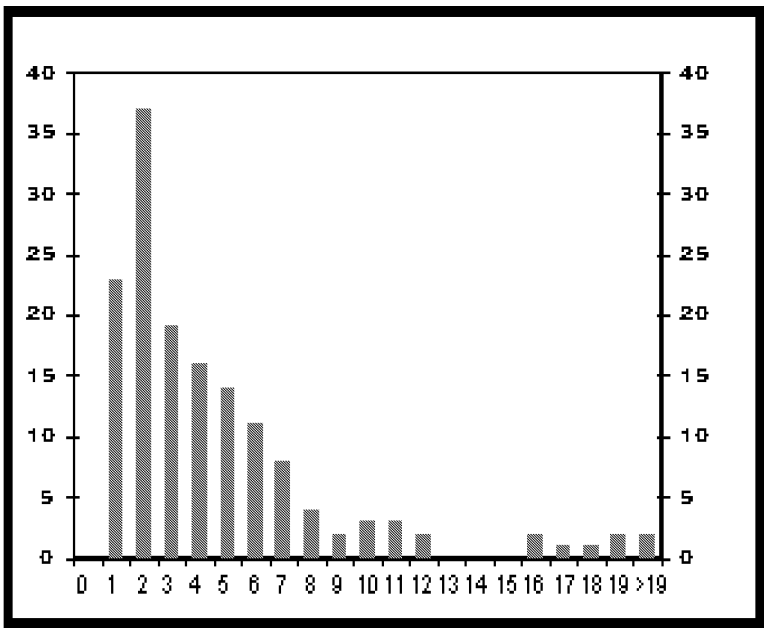


Figure 11: Histogram of File 1

The intervals on the horizontal axis are to be interpreted such that, for example, the interval labelled "9" contains observations ranging from 9.000 to 9.999. The

interval labelled ">19" contains all observations greater than or equal to 20.000. The vertical axis indicates the number of observations in each interval. Thus, in File 2 below, there are 35 observations recorded in the range 1.0 to 1.999.

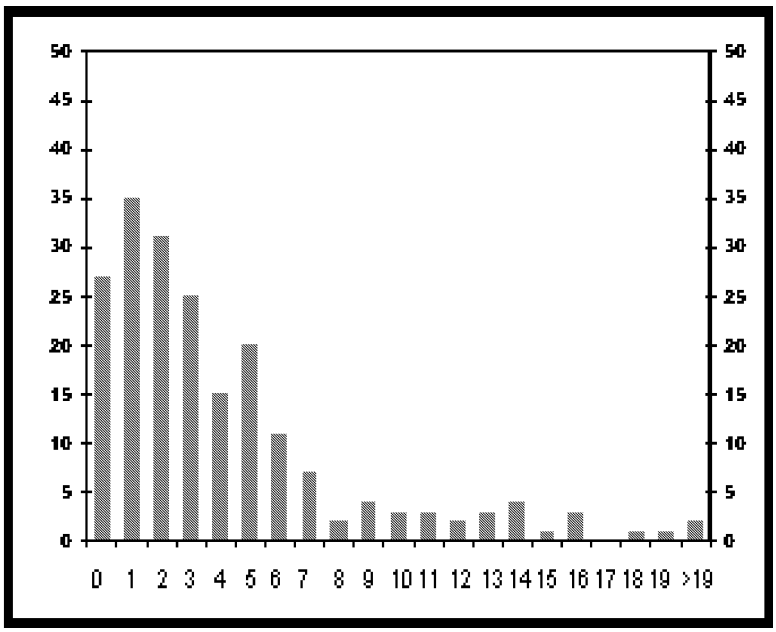


Figure 12: Histogram of File 2

Some may comment that the data of File 2 may be better modelled using a mixture model with more than two components. From the outset, the number of components in the models used in this project has been fixed at two. However a future project could be to investigate the number of components that best fits the data, possibly using reversible jump McMC methodology.

6.2 The software

6.2.1 Language

With the exception of the data collection program already mentioned all the software used for the purpose of calculation etc in this project was written by the author using Delphi for Windows, v1.0. ("Delphi" is a registered trade mark of the Borland corporation.)

Delphi is a visual programming environment system for Windows. At its very

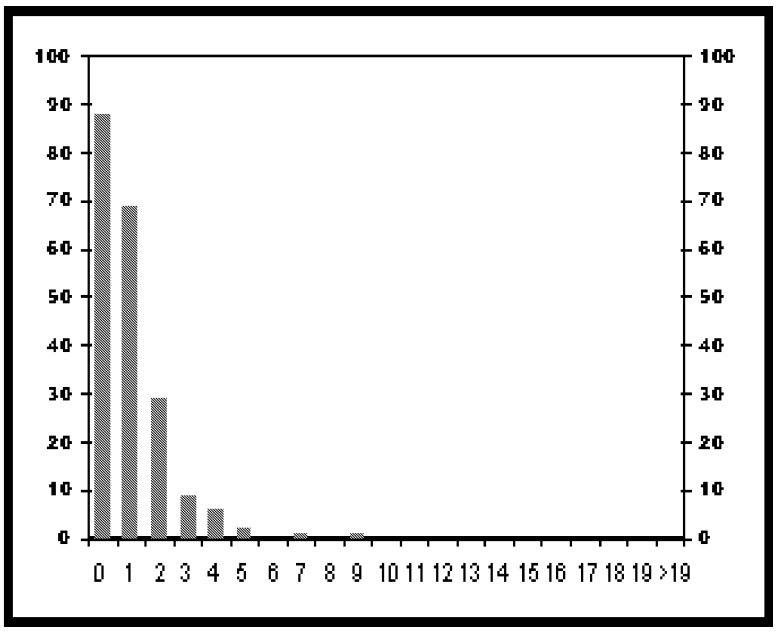


Figure 13: Histogram of File 3

simplest this means that the various components of the graphical user interface i.e., buttons, edit boxes etc. are placed onto a form (i.e.window) using the "drag and drop" method. This means that the programmer does not have to write any program code to position these components and so is free to concentrate on matters such as sampling algorithms etc. The high level language used by the system is Pascal in its object oriented form although, in this project, it was only ever necessary to use Pascal in its basic form.

Although C and its derivatives are currently in vogue, it is very interesting to note that Peter Norton has stated *" By itself I consider Pascal to be the better language, cleaner and less error prone;"* (Norton, 1987)

Delphi v1.0, like most of its successors was available free of charge on the cover of a computer magazine. Such are the vagaries of the software industry that a package sold for £349 in 1995 was given away for free in 1997, with complete documentation. Under the name Kylix, the package is available for Linux systems. ("Kylix" is a registered trademark of the Borland corporation)

The software written for this project, with the exception of the data collection program, divides itself quite naturally into two parts :-

1. Gibbs sampling programs

2. Post-processing programs

6.2.2 Gibbs sampling programs

Even with the benefit of a visual programming environment, the Gibbs sampling programs used in this project consisted of well in excess of 2000 lines of code. Although some of this was taken up by the graphical outputs used, it still remains true that this type of program can be quite complex. As an aid to understanding, the program can be broken down into three stages :-

1. Input

2. Processing

3. Output

and also expressed in terms of a block diagram as shown below :- In total, three

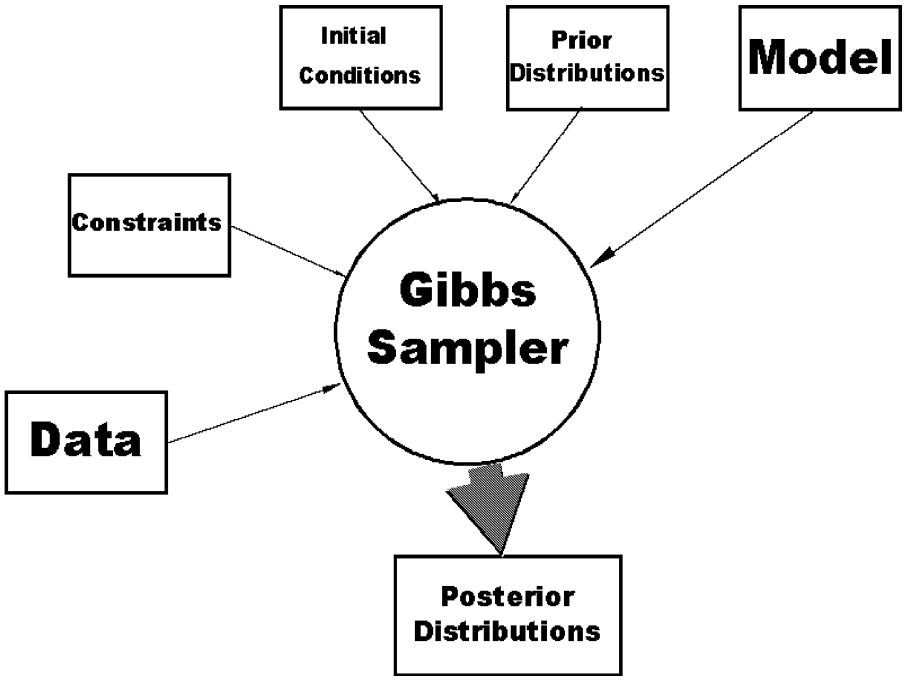


Figure 14: Block diagram of the sampling algorithm

Gibbs sampling programs were written :-

- A single program for both shifted exponential models.
- A program for the Gamma/Exponential Distribution.
- A program for the Gamma/Exponential Distribution that uses "blocking" in the sampling algorithm.

Figure 15 below shows a screen shot from the last of the above mentioned programs.

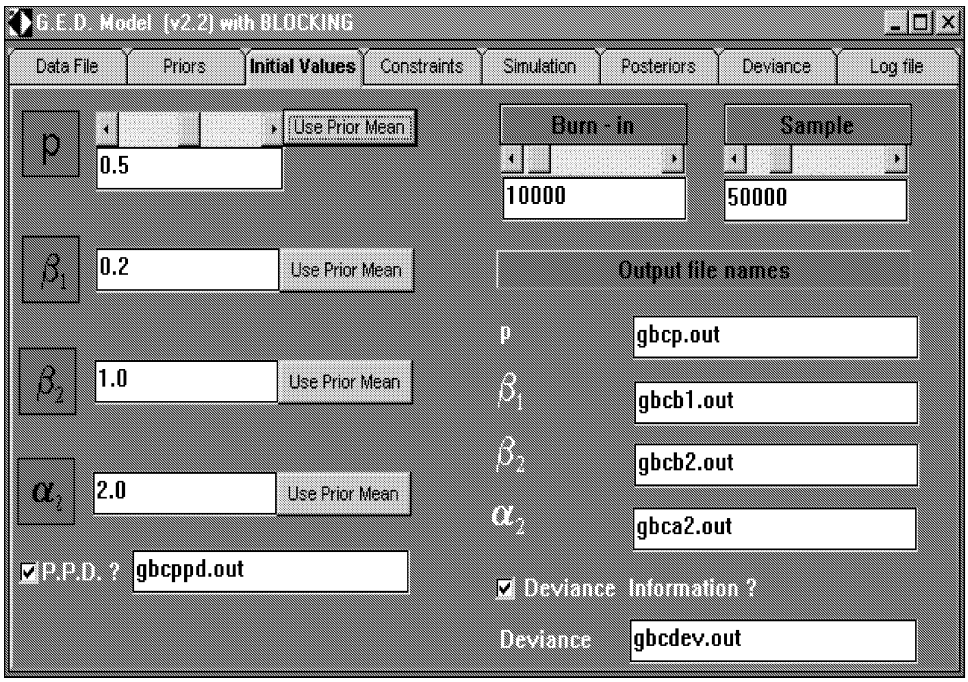


Figure 15: Screenshot of part of one of the Gibbs sampling programs

Further screenshots and extracts of source code from the above programs can be found in the Appendix.

Much work with Gibbs samplers is done using standard software such as "BUGS" (Spiegelhalter et al, 1994) However, it was not found possible to accommodate a *shifted* exponential model within the B.U.G.S. package. Having written a program that could deal with the model it seemed a logical step to carry out two extensions. Firstly graphical ouput was included and secondly another very similar program to deal with the Gamma/Exponential model was written.

6.2.3 Post-processing programs

Again these programs can be divided into three groups :-

1. Further graphical examination of output.

2. Analysis of convergence.

3. Model/data comparison i.e., goodness of fit.

Figure 16 below shows a screen shot from one of the programs from 1) above.

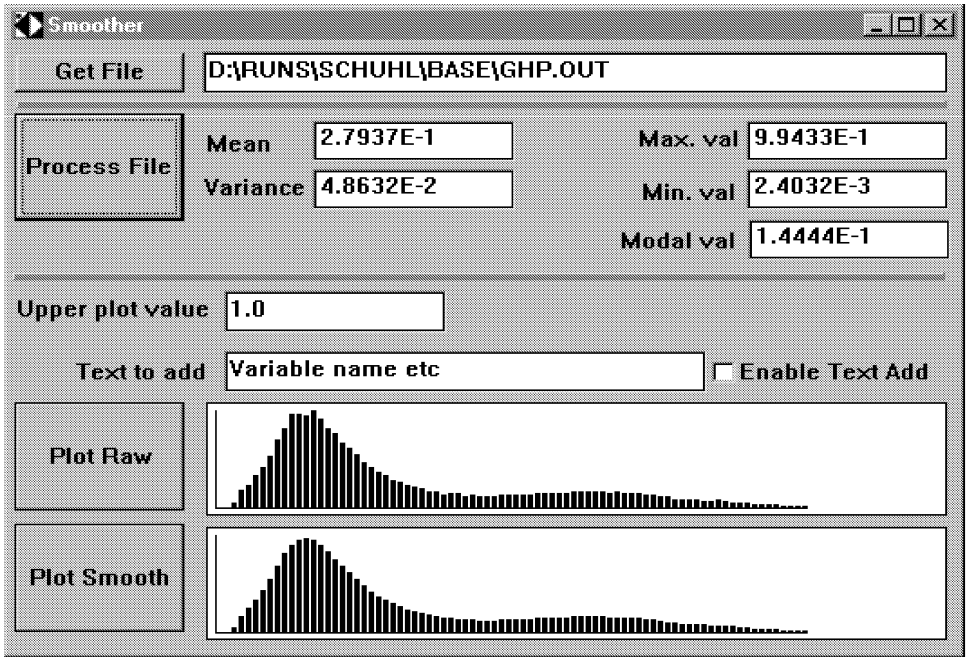


Figure 16: Screenshot of part of one of the graphical output programs

6.2.4 Software testing

Several of the programs used in this project involve complex procedures such as Gibbs sampling or matrix inversion and, therefore, testing these programs prior to use is obviously essential. The design of the programs was modular, that is to say, each program was composed of a number of functions or procedures, designed to carry out a specific task, and program execution consisted of each of the functions or procedures being executed in the correct order. It was, therefore, natural to test

each module separately to ensure its correct operation before testing the program as a whole. On occasion this necessitated writing a separate testing program the sole function of which was to test a single function or procedure. By way of example, consider the function that returns a single value sampled from a Gamma(a,b) distribution. A program was written to sample any number of values from any gamma distribution and write them to file. This program could be run so that, say, 50,000 values were sampled from a Gamma(10,5) distribution. The resulting file could then be examined to see if the mean and variance of the sampled values were approximately those of a Gamma(10,5) distribution i.e., 2 and 0.4 respectively. A histogram of sampled could be plotted and its shape assessed visually. This test procedure was carried out many times, for Gamma distributions having different parameters and, after the removal of initial errors, the routine was found to work well.

The program which carried out the estimation of the Gamma/Exponential model was tested in its complete form by comparing the results it gave with those obtained when the same data was analysed using the previously mentioned B.U.G.S. package. The resulting comparison was found to be very good.

Program testing was often time consuming and laborious but its necessity cannot be overstated

The data and software have now been described. Further screen shots and an example of source code can be found in the Appendix to this thesis. In the next Section experimental runs of the Gibbs sampler will carried out and the results will be discussed.

7 Implementation : Problems and Solutions

7.1 Introduction

So far, the basic research questions have been asked (Introduction) and the background to both headway modelling and Bayesian statistics has been outlined (Section 2). The models to be used have been defined (Section 3) and a general treatment of mixture models has been given (Section 4). The theoretical basis for Gibbs sampling has been described (Section 5) as has the data and software in Section 6. Now the actual implementation of the algorithms previously discussed is presented and described.

For each model a number of runs of the Gibbs sampler will be carried out in order to test the behaviour of that model and the sampler. The problems encountered will be focussed upon and possible solutions will be proposed. In some cases, however, it will be shown that none of the proposed solutions bring about a sufficient improvement in model performance so that use of the model can be recommended. That is to say, no way was found to fully overcome the difficulties encountered and the conclusion must be drawn that the model is unsuitable.

7.2 Preliminary considerations

It was pointed out, in the previous section, that the Gibbs sampler has, apart from the model itself, four "inputs".

- Prior distributions
- Initial conditions
- Constraints
- Data

Each of these could, theoretically change at each run of the Gibbs sampler but this is neither necesssary nor appropriate in practice. Each of these remaining inputs is

discussed below :-

1. **Prior distributions** For each model parameter the form of the prior distribution will remain unchanged at each run of the sampler. The values of the prior parameter values will, in most cases, be such that the prior distributions will be "mildly informative" except in the case of the parameter p. Here, the example of Robert & Mengersen (1999) will be followed and a $U(0,1)$ distribution will be used. However, in order to build into the sampling program a measure of flexibility that may be used in the future, the prior distribution for p will be expressed as as $\beta(\phi,\psi)$ where $\phi = \psi = 1$.

   Prior distributions will be tabulated for each model whereas prior distribution parameters will be tabulated before each run.

2. **Initial Conditions** In order to test the robustness of Gibbs sampling to initial values of parameters, numerous exploratory runs were carried out using widely varying values of the model parameters. It was found that after a "burn-in" of 10,000 iterations the effect of the initial values on sampling was not present and that the sampler had started to move throughout the support of the posterior distribution. As well as the model parameters themselves, one other parameter was tested in this manner.

   Recall that, in the algorithm used, the data are partitioned at every iteration with each observation being allocated to one of the two components. In the case of the Schuhl model, the lowest observation allocated to the second component, denoted by t_{min2} has two vital functions :-

   (a) At the allocation step, if an observation is less than the current value of t_{min2} then that observation is allocated to component 1.

   (b) At the sampling step the current value of t_{min2} is the upper limit of the full conditional distribution of the parameter k.

A value of t_{min2} must, therefore, be chosen so that the allocation step of the very first iteration can be carried out. Again, this was tested using widely varying values and the result was that the choice of initial value did not affect the sampling outcome, given a "burn- in" of 10,000 iterations. Initial Conditions will be tabulated for each model.

3. **Constraints** A constraint is a condition that we apply on the Gibbs sampler and is really a part of the prior beliefs. When we apply such a constraint we are stipulating that certain outcomes, which may be parameter posterior values, have either very low or zero probabilities and thus we are reflecting our prior beliefs. It will be shown that constraints can have a great influence on the modelling outcome. They will, therefore, be tabulated and discussed before each run.

4. **Data** The data file used could, in theory, affect the choice of prior distribution parameter values or constraint and will be identified before each run. This is due to the fact that, in this case, prior knowledge exists concerning the data in terms of where and when it was collected.

7.3 Modelling Outcome

7.3.1 What would be a "successful" outcome?

Having completed a particular run, analysis of the output is required and the success, or otherwise, of the run is evaluated. At the outset, clarity is needed as to what constitutes a successful run and consistency is required in terms of measuring it. Therefore, the basic question needs to be asked *"What features do we wish to see in the output that will cause us to view the run as a success?* For the purposes of this project, if the following features are all present then the run will be deemed a success.

- **Unimodal posterior distributions** Although the posterior distribution for a given parameter can sometimes be bimodal, for a given set of prior distributions, and likelihood, this type of distribution does give rise to difficulties in practical terms. This is because the mean of such a distribution usually occurs between two modes at a place where the probability density function is quite low and, sometimes, at a minimum. This makes such distributions difficult to summarise as is also the case with distributions which, although unimodal, have a "ridge". This particular effect can sometimes be remedied by a reparameterisation. However, it must be stated that a unimodal posterior distribution of a particular parameter is desirable rather than absolutely necessary. Also, it will be shown that unimodal posterior distributions are achievable where multimodality arises from the non-identifiability of the mixture model in question. The requirement here could be stated as being for distributions with a clear single mode.

- **Convergent chains** We must be sure that the apparent posterior distribution for each parameter is, in fact, the stationary distribution for the Markov chain concerned. This important aspect of Gibbs sampling has considerable literature devoted to it with Mengersen, Robert & Guihenneuc-Jouyaux (1998) providing a valuable study.

- **Good model fit** We obviously require the model to be a good reflection of the data.

7.4 How do we measure "success"

In order to adopt a consistent approach to evaluating the modelling outcome, we will use the following methods :-

- **Evaluating unimodality** The modality of a posterior distribution can be seen from its graph and so this criterion is easily assessed.

- **Diagnosing convergence** Unfortunately there is, to date, no "black box" method for assessing convergence and diagnosis remains something of an art as well as a science (Stephens, 2001). That is, there is no algorithm which takes the sampled chain as its input and generates a "true/false" statement regarding convergence as its output. The problems here are already well documented with several authors giving good accounts of the difficulties involved. Gelman (1996) and Raftery & Lewis (1996) give good descriptions. The particular difficulties relating to the use of mixture models are described by Brooks (1998) and it is gratifying to observe similarities between the example used by Brooks and the cases considered in this project. At this point the chosen convergence diagnostic will be described.

  If the post burn-in chain for a given parameter has reached its stationary distribution, then all values sampled from it will be from the same distribution. One method we can use to determine whether or not this has happened is as follows :-

  Suppose there are n values in the sampled chain after the burn-in has been discarded. We divide this into two sub-samples. The first containing the first $n/2$ values and the second containing the last $n/2$. A simple Kolmogorov-Smirnov two sample test is applied to determine whether or not the two sub-samples are drawn from the same distribution. If the test indicates that the two sub-samples are drawn from the same distribution then the chain is deemed to have converged. The level of significance was chosen at 1 per cent. There are two main reasons for using this particular diagnostic :-

  1. The test does not depend on the stationary distribution being normal. This is useful since normality cannot always be guaranteed and this is true in the cases considered here.
  2. There are no significant computational difficulties to be overcome when using this method.

Having decided upon a strategy for convergence diagnosis, it was interesting to discover that this test is almost identical to that proposed by Robert, Ryden & Titterington (1999). The difference between the two is the way in which independence, or lack of independence, is dealt with. It is known that samples generated in McMC are often highly correlated and this is especially true in the case of mixture models. In this project this problem is dealt with by means of a technique known as *"thinning"* which is described by Gelman (1996).

Suppose the sampler is run and a sample of size 50,000 is used. Consider the chain which is the output for the parameter α_2. Figure 17 below shows a plot of $\alpha_{2.i}$ v $\alpha_{2.i-1}$, where i is the iteration number, i.e. $i \in \{0..50,000\}$. (For terminology, we shall say that α_2 is plotted with a *"lag"* of 1.) The correlation

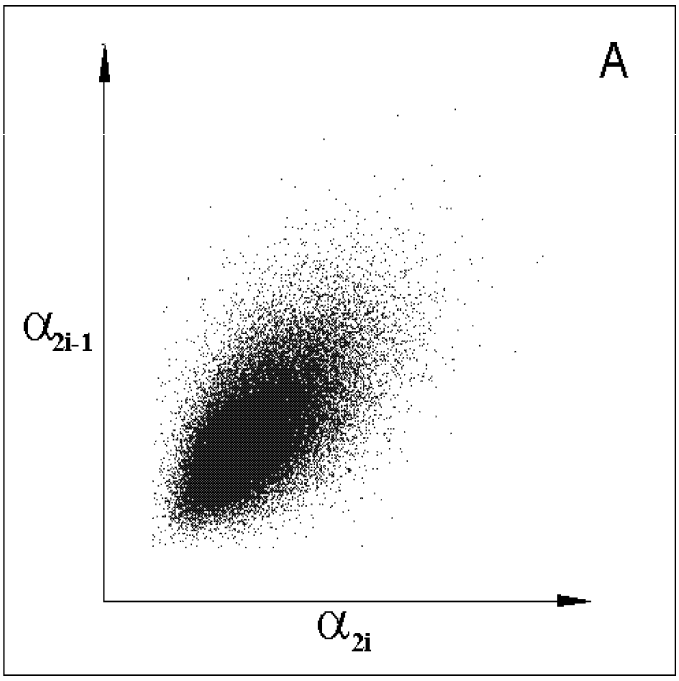


Figure 17: Autocorrelation plot 1

is evident and the estimated value of the coefficient of correlation, r, is 0.62663. The test indicates that this chain has not converged. If α_2 is now plotted with a lag of 5 (i.e. α_{2i} v α_{2i-5} is now plotted) it can be seen, as in Figure 18 below, that there is visibly less correlation. The value of r was 0.23311 in this case. If the chained is thinned by a factor of 5, i.e. only every fifth value is accepted

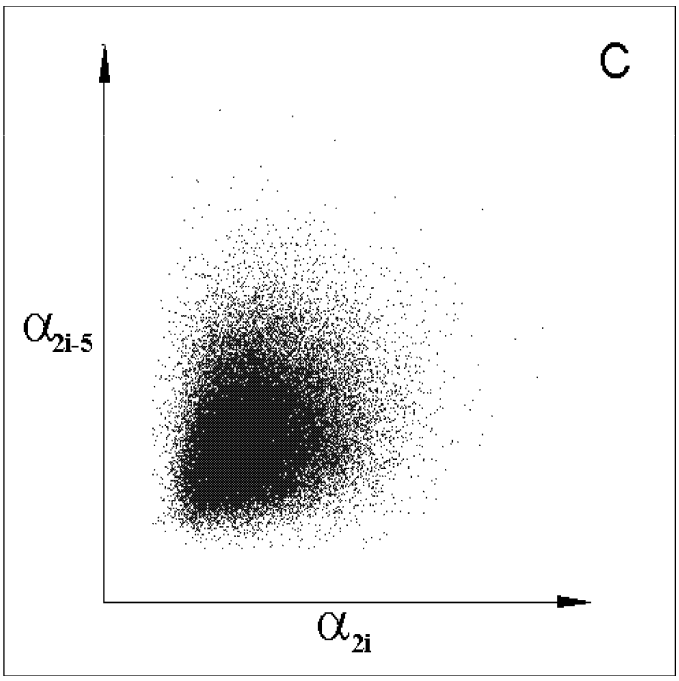


Figure 18: Autocorrelation plot 2

and all others are discarded, it is found that this thinned chain passes the test for convergence. The reason for this is that the test for convergence requires independent observations without any correlation. If correlation is present then the effective number of independent observations is reduced and, since the actual number of observations is used, the test is rendered inaccurate. The effect of thinning is to reduce to zero the discrepancy between the *actual* and *effective* number of observations and, by doing so, restore accuracy to the test.

If α_2 is now plotted with a lag of 1 for the thinned chain, as in Figure 19, it is apparent that thinning has reduced the correlation. Here, r was found to be 0.23844.

Thus, the relationship between correlation (more accurately referred to as autocorrelation since a parameter co-relates with itself), convergence and thinning is demonstrated. Autocorrelation arises because the sampler only moves slowly throughout the support of the posterior distribution. For a further explanation of this phenomena, see Gilks & Robert (1996).

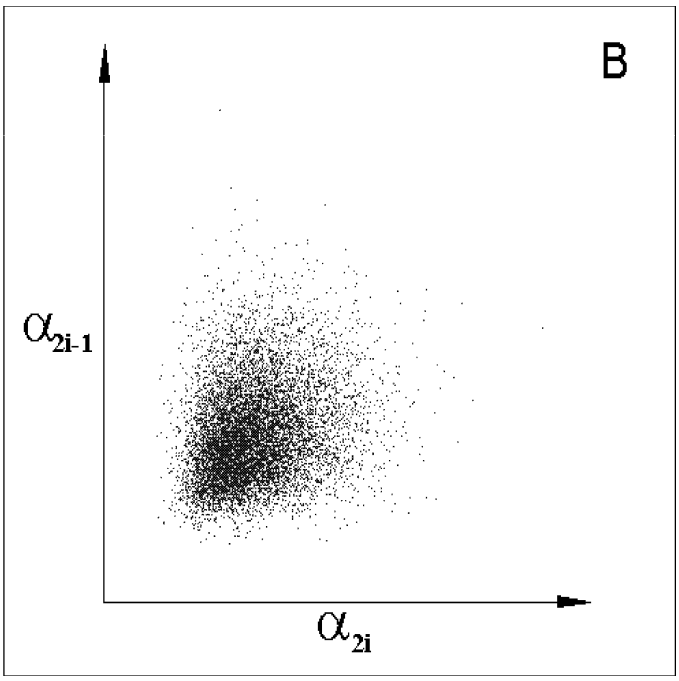


Figure 19: Autocorrelation plot 3

- **Testing model fit** For the purposes of this section a straightforward test of model fit will be used. A histogram of the data will be compared to a histogram of model simulations. The parameter values used will be the posterior density means, otherwise known as "plug-in" values, and 50,000 simulations will be used. Others, e.g. Kass & Raftery (1995) have advocated the use of the posterior predictive distribution (p.p.d.) to assess model fit and such a method will be discussed in Section 9 of this project. The reason for using "plug-in" values here is as follows. At each sweep of the Gibbs sampler values for the model parameters are sampled which can be expressed as Θ_i where i is the iteration number and Θ_i is the vector of parameters, i.e. $\Theta_i = \{p, \beta_1, \beta_2, \alpha_2\}$. The p.p.d. is created by sampling, at each iteration, a value of t from $\pi(t|\Theta_i)$, where $\pi(t|.)$ is the model under consideration. Clearly, at some iterations, Θ_i will contain parameter values from the tails of their respective posterior distributions which will give rise to values of t that are also unlikely. Thus, if the posterior densities are symmetrical, the "most likely model" will be compared with the data sample. Figure 20 shown below is a typical diagram

of the type described above.

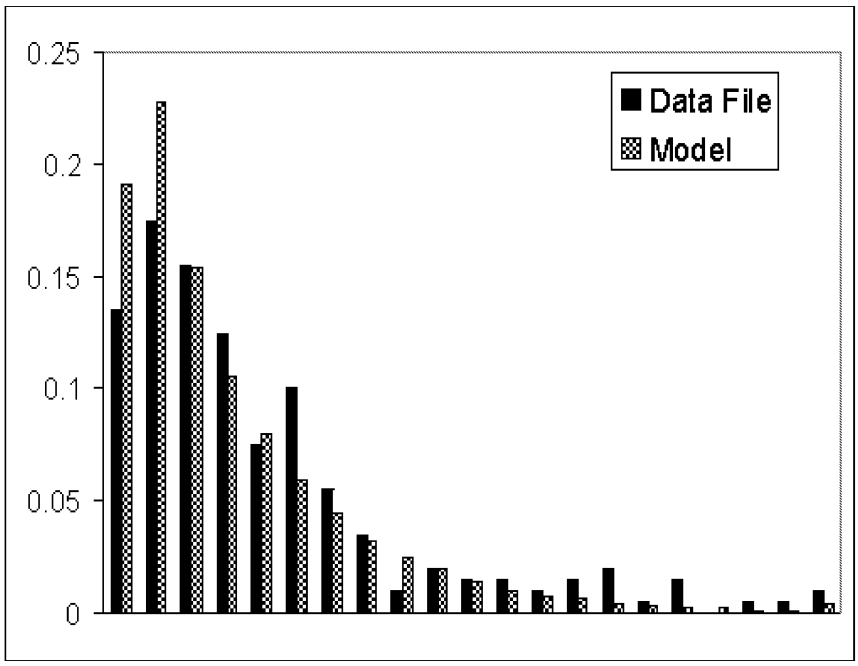


Figure 20: An example of a model/data fit diagram

It can be seen from this diagram that the model does not reflect the data very well, especially in the first two intervals.

7.5 An important distinction

If, after a particular run of the sampler, the posterior distributions are unimodal, the parameter chains converge and the model adequately reflects the data then it can be assumed that the choice of model and prior were correct and that the sampling algorithm has functioned properly. However, if these features are not all present then careful consideration must be given to the cause and an important distinction must be made. That is to say, it must be determined if the choice of model and/or prior has been incorrect or if the sampling algorithm is, in some way, defective. Remedial action should not be taken on a trial and error basis.

Parameter	Prior Distribution
p	Uniform(0,1)
β_1	$\propto \beta_1^{\gamma-1} e^{-\delta\beta_1}$
β_2	$\propto \beta_2^{\gamma-1} e^{-\delta\beta_2}$
k	$\propto k^{\theta-1} e^{-\nu k}$

Table 7: Model parameter prior distributions

7.6 The Griffiths and Hunt model

Recall that this model is defined by :-

$$f(t) = p\beta_1 e^{-\beta_1(t-k)} + (1-p)\beta_2 e^{-\beta_2(t-k)}$$

The prior distributions, which are independent, used for its parameters are shown in Table 7 . The starting values of the model parameters are also set out in Table 8

Parameter	Starting value
p	0.5
β_1	0.2
β_2	0.2
k	0.2

Table 8: Model parameter starting values

7.6.1 The Base Run

In order to demonstrate the type of problems encountered when dealing with mixture models a run will be carried out, for each model, where no constraints are used. Also, in the base run, mildly informative priors will be used and their values are set out in Table 9 Since this model is intended for use on single carriageway roads, File 1 will be used as the data file.

Run Outcome

Figure 21 shows the raw output plots for the model parameters. The iterations 0 to 10,000 form the "burn-in" and 10,001 to 60,000 form the sample. There are two

Parameter	Value
γ	5.0
δ	5.0
θ	5.0
ν	5.0

Table 9: Prior distribution parameter values

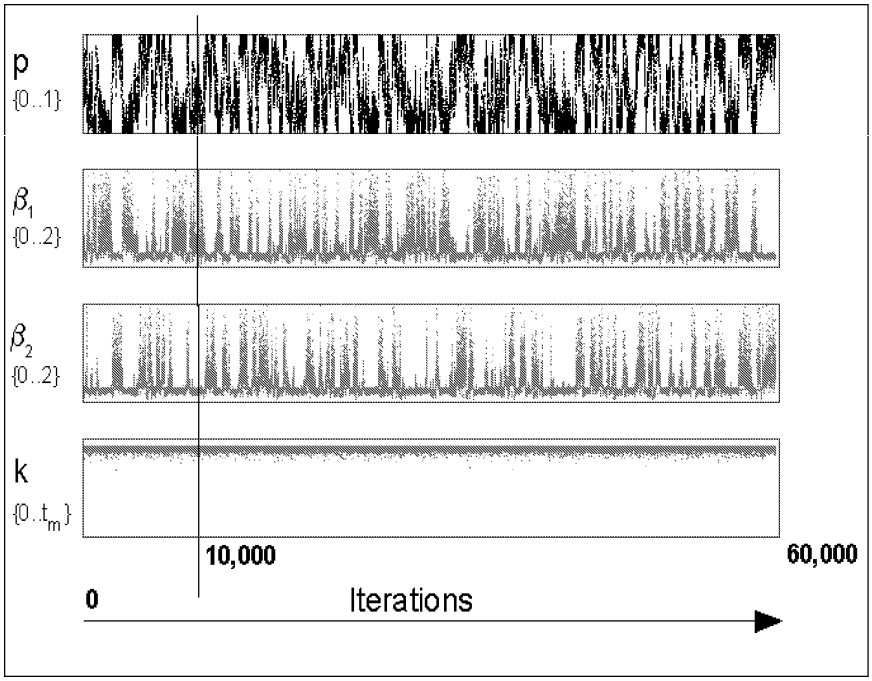


Figure 21: Raw output for base run of Griffiths & Hunt model

obviously noteworthy points :-

- The parameters p, β_1, and β_2 have clearly not converged to a uni-modal stationary distribution. Instead, all three parameters alternate between two states which appear to be related. When p is in its "higher" region, β_1 is in its "lower" state and has a relatively low variance but β_2 has a much higher variance. This can be explained by the behaviour of the allocation step described in Section 5. When p is high, a higher proportion of observations are allocated to component one : this will mean that there is sufficient information in the sample for the posterior distribution of β_1 to have a low variance. However, component two will contain relatively few observations and so β_2 will be sampled from a distribution that is much closer to its prior distribution. Recall that the fcd is

given by :-

$$fcd(\beta_2) \propto \Gamma(\gamma + n_2, \Sigma_2 t_i + \delta - n_2 k)$$

When few components are allocated to component 2 the terms n_2, $\Sigma_2 t_i$ and k will be small enough for the prior parameters γ and δ to dominate. When p is in its "lower" state we see that the reverse in true. In Figure 22 we can see that each of the posterior distributions of β_1 and β_2 can be viewed as a superposition of two distributions. In each case, the two distributions are caused by p being in one of its two states.

- The output for the parameter k, however, behaves quite diffently and k is the only parameter that, according to the Kolmogorov-Smirnov Test, converges. This is because k takes a part of its likelihood from both components and so is not subject to the same alternating that we see in the behaviour of the other parameters. The critcal KS value is 1.3238E-2 and Table 10 shows the actual KS values for the model parameters.

Parameter	K - S value
p	2.6388E-2
β_1	1.4612E-2
β_2	1.9100E-2
k	4.1403E-3

Table 10: Kolmogorov - Smirnov test values

The bimodality of the parameters p, β_1 and β_2 can be clearly seen in Figure 22. It is also apparent which areas of each output give rise to corresponding regions in the histogram. A table of posterior means and variances is shown in Table 11. It

Parameter	Posterior Mean	Posterior Variance
p	0.47710	0.11243
β_1	0.43608	0.094821
β_2	0.40956	0.084553
k	1.21520	6.3105E-2

Table 11: Posterior means and variances

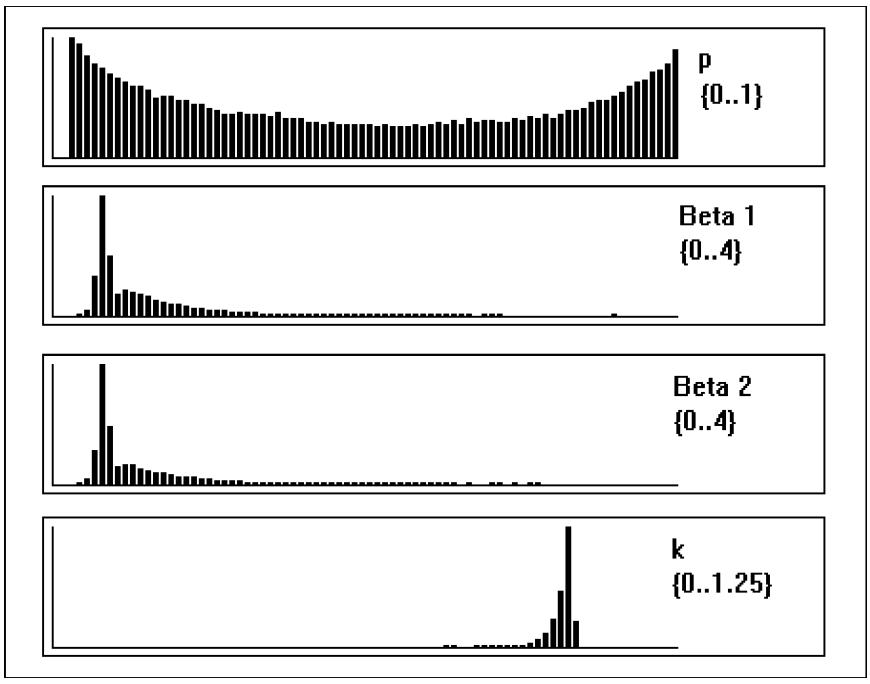


Figure 22: Marginal posterior distributions

is clear from Table 11 that, particularly in the case of the parameter p, difficulties arise when attempting to summarise a bimodal posterior distribution.

Also if we plot a graph of β_{1i} against β_{2i}, where i is the iteration number we see the relationship between these two parameters during sampling. This type of graph can be very informative (particularly for demonstrating correlation between parameters) and will be used again.

There is already sufficient evidence to deem this run a failure but, for the sake of completeness, a graph of model versus data file is shown in Figure 24.

As expected, the fit is very poor and provides further evidence to support the assertion that the run has been a failure. The reasons for such a clear failure are also important and for these it is necessary to look at the model with regard to its identifiability.

In the base run, there are no constraints applied to the model and so there is no way the sampler can distinguish between the components. In other words, the labelling of the two components can be switched quite arbitrarily without any loss of accuracy. To illustrate the point further, let $p = 1 - p'$, $\beta'_1 = \beta_2$ and $\beta'_2 = \beta_1$.

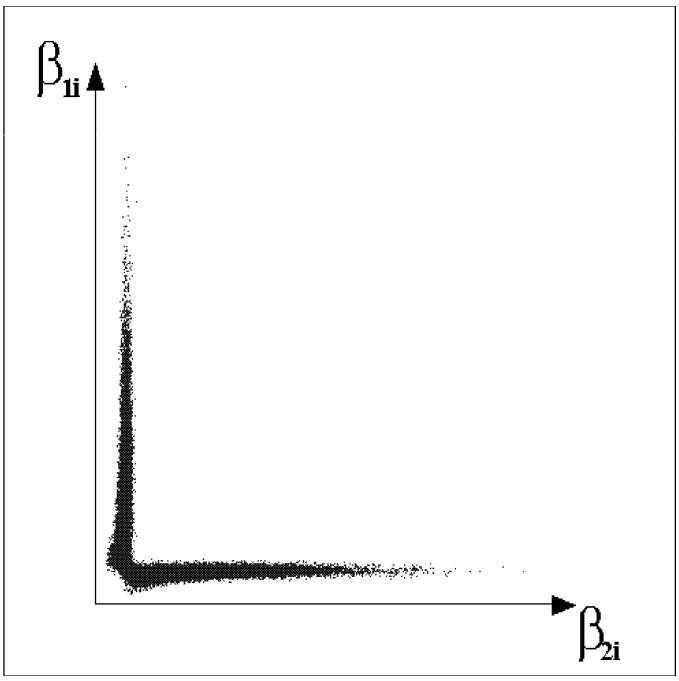


Figure 23: Graph of β_{1i} v β_{2i}

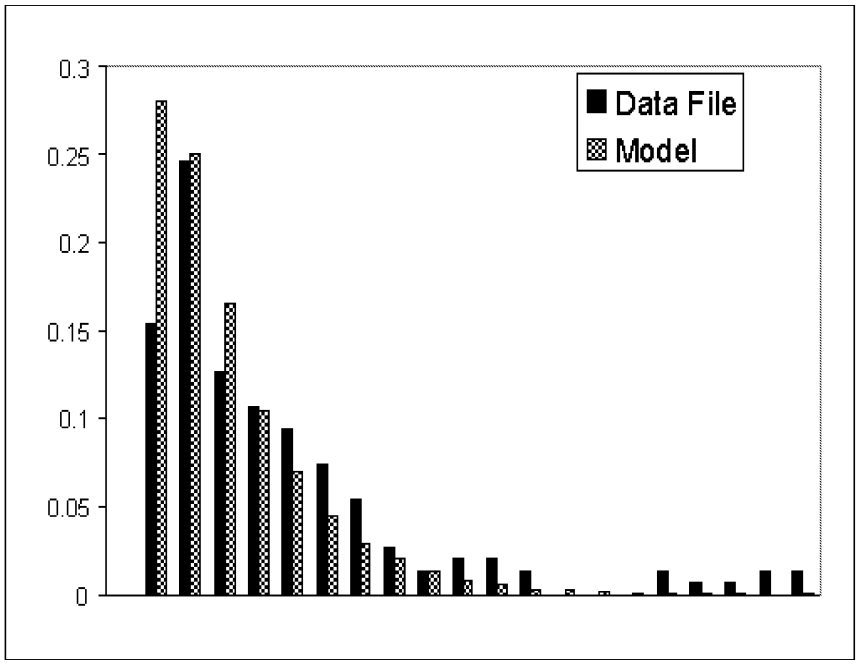


Figure 24: Model fit diagram

Suppose we now restate the model as :-

$$f(t) = (1 - p')\beta_1' e^{-\beta_1'(t-k)} + p'\beta_2' e^{-\beta_2'(t-k)}$$

It can be seen that we have moved the parameters from one component to the other but have not altered the model. Thus, there is more than one set of model parameters that will result in the same likelihood.

If we now look once more at the histograms of the parameter outputs with the above in mind we can see that they appear to indicate the existence of a single component but, due to label switching, it is impossible to determine which one. It is also clear that it would have been completely wrong to have accepted the posterior mean of p (or any other parameter) in this case, hence the requirement for unimodal posteriors. We conclude that the run has been a failure.

7.6.2 $\beta_2 > \beta_1$

Given the above, it would appear advantageous if we could apply a constraint that would create a distinction between the two components. The obvious one would be $\beta_2 > \beta_1$ but it must be decided how to apply this constraint. The constraint is applied at the sampling stage by testing, when all the parameters have been sampled during a given iteration, if the mean of the first component is greater than the mean of the second. The condition tested for, namely

$$\frac{1}{\beta_1} + k > \frac{1}{\beta_2} + k$$

is equivalent to $\beta_2 > \beta_1$. If this condition is not met, then all the parameters are sampled again from the same full conditional distributions. In this way we are holding the Markov chain at one point before moving it on when the condition is met. Thus the chain is not violated.

A brief examination of the raw outputs in Figure 25 reveals that the constraint has had an effect. The parameter p does not move between the two extremes of its space as before but clearly favours the upper part. Also the parameters β_1 and β_2 show quite different behaviour as sampling progresses. There is no jumping between two states and the resulting histograms are unimodal as shown in Figure 26.

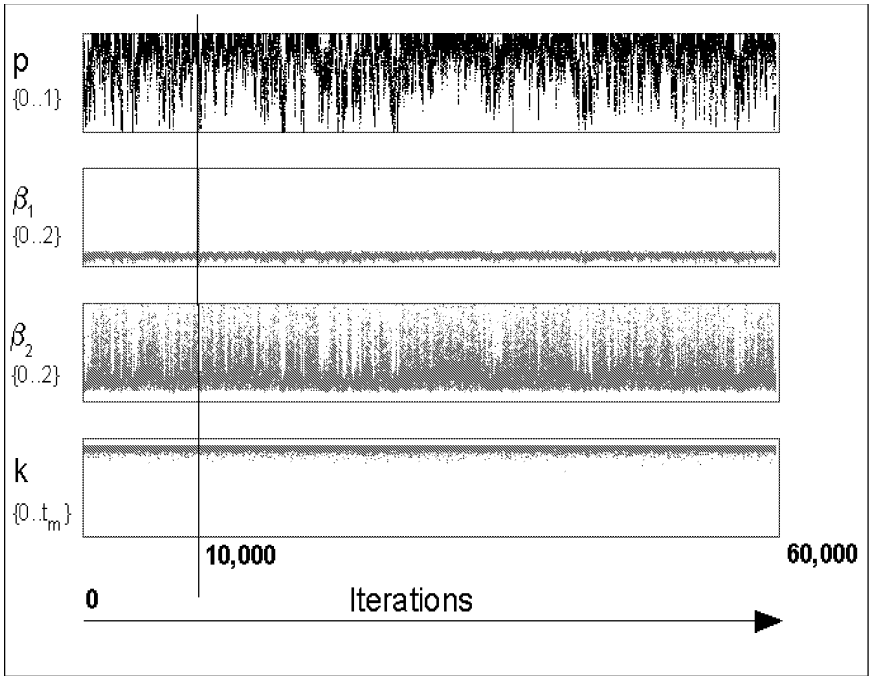


Figure 25: Raw outputs for the Griffiths & Hunt model

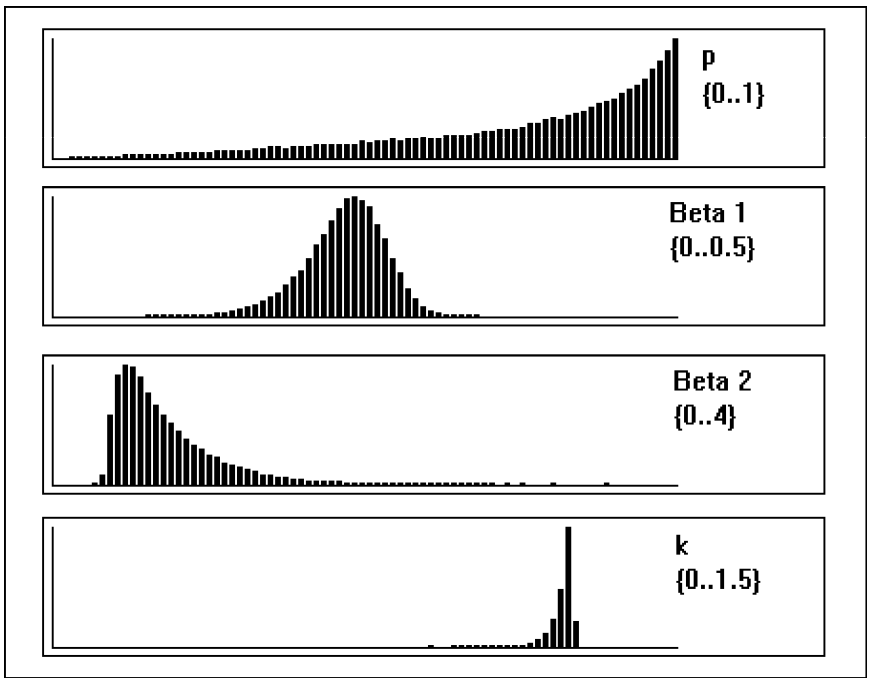


Figure 26: Marginal posterior distributions for the Griffiths & Hunt model

A table of posterior means and variances is shown in Table 12

From a point of view of convergence, the results are also interesting and the K-S values for all parameters are tabulated in Table 13

The values in Table 13 represent an improvement, in terms of covergence, over

Parameter	Posterior Mean	Posterior Variance
p	0.76246	4.7100E-2
β_1	0.22619	1.0489E-3
β_2	0.63001	1.0403E-1
k	1.21510	6.2957-4

Table 12: Posterior means and variances

Parameter	K - S value
p	2.6388E-2
β_1	1.4612E-2
β_2	1.9100E-2
k	4.1403E-3

Table 13: Kolmogorov - Smirnov test values

the base run but only the parameter k passes the KS test since the critical value here is 1.3238E-2. If the parameter output chains are thinned by a factor of 5, then each chain passes the KS test for convergence as demonstrated in Table 14. The critical value, in this case 2.9603E-2, is not exceeded by any of the parameters and so the run has satisfied the criterion regarding convergence.

Parameter	K - S value
p	2.6906E-2
β_1	2.0840E-2
β_2	1.9265E-2
k	7.5854E-3

Table 14: Kolmogorov - Smirnov test values

Two of the required criteria are now met but has model fit improved? Figure 27 shows the graph of model against data for this run.

We see some improvement here but the fit is still poor even though the marginal posterior distributions are unimodal and convergence has been achieved. This poor model fit must, therefore, be attributed to a poor choice of model.

An obvious question must now be answered :- *"If this is the case, why are such claims made for the model by Griffiths & Hunt?"*

The answer can be found within their paper where they report that, initially,

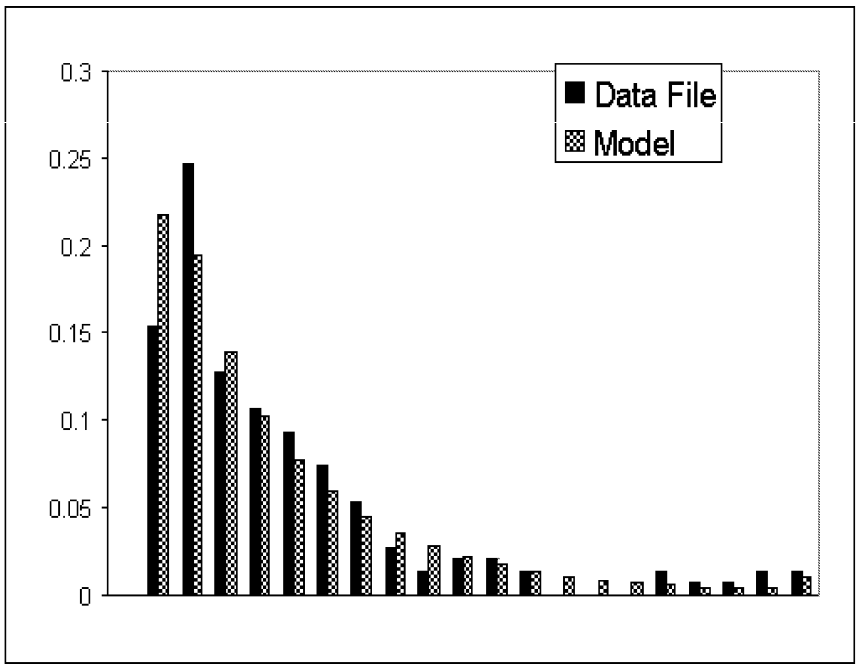


Figure 27: Model fit diagram

their model also exhibited poor model / data fit when estimated using their *ad hoc* method. They attribute this to estimated values of k (in their paper, d) being too small which is in turn attributed to *" rogue values being recorded in observing very short headways, with observers tending to be a little 'trigger happy' when using the event recorders"*. The authors circumvent this difficulty by simply removing headways of less than 0.5 seconds from their data set and re-estimating the model. In this case the model / data fit is very good. There are, however, no such doubts concerning the data gathering in this project and we conclude that this model is not suitable for this particular use. It is possible, however, to offer an explanation of how increasing the accepted value of k can give rise to better model / data fit.

Figure 28 shows an approximately gamma density representing the data. It can be seen that by rejecting low headways we truncate the data and remove the left-hand tail of the gamma density. If we now fit a shifted exponential distribution to this adjusted data the model / data fit will be much better. It seems that Griffiths & Hunt were attempting to fit a shifted exponential distribution to data was probably better modelled by a gamma distribution.

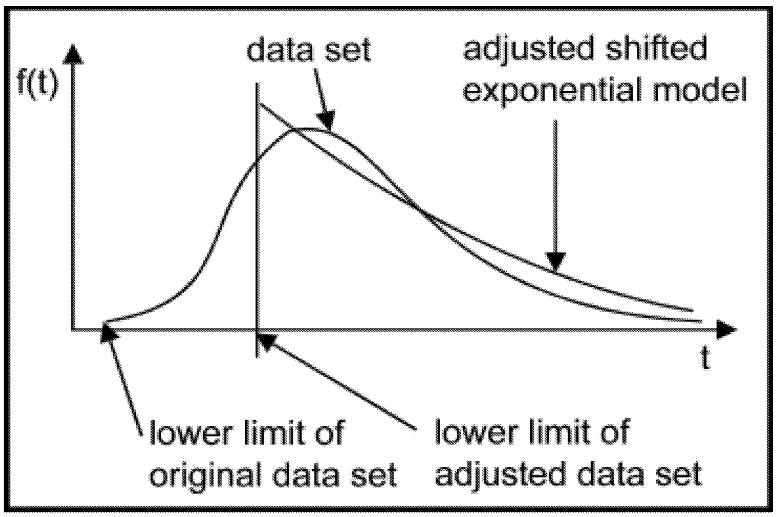


Figure 28: The effect of truncating a gamma density

7.6.3 Summary

It has been shown that this model, like all mixtures, carries with it the problem of identifiability. By using a suitable constraint, this can be overcome as can the difficulty of slow mixing. Unfortunately poor model fit is still evident and the conclusion is drawn that this particular model was a poor choice for modelling headways.

7.7 The Schuhl Model

Brief inspection of the Schuhl model, defined by

$$f(t) = \begin{cases} p\beta_1 e^{-\beta_1 t}, & (0 < t < k) \\ p\beta_1 e^{-\beta_1 t} + (1-p)\beta_2 e^{-\beta_2 (t-k)} & (k \leq t) \end{cases} \tag{18}$$

gives some optimism since we see immediately that the two components do not share the same mode and this may enable the sampler to differentiate between the two components i.e., the identifiability problem found in the Griffiths & Hunt model may not be present to the same extent with this model. Since the Schuhl model is proposed for use on roads of more than one carriageway Files 2 & 3 will be used as

data.

7.7.1 The Base Run : File 2

The prior distributions and their parameters are the same here as in the case of the Griffiths & Hunt model and, again, no constraints were used.

Run Outcome

As expected, this model does perform better but, as Figure 29 reveals, the sampler still jumps between two states. These "jumps" do not occur with the same

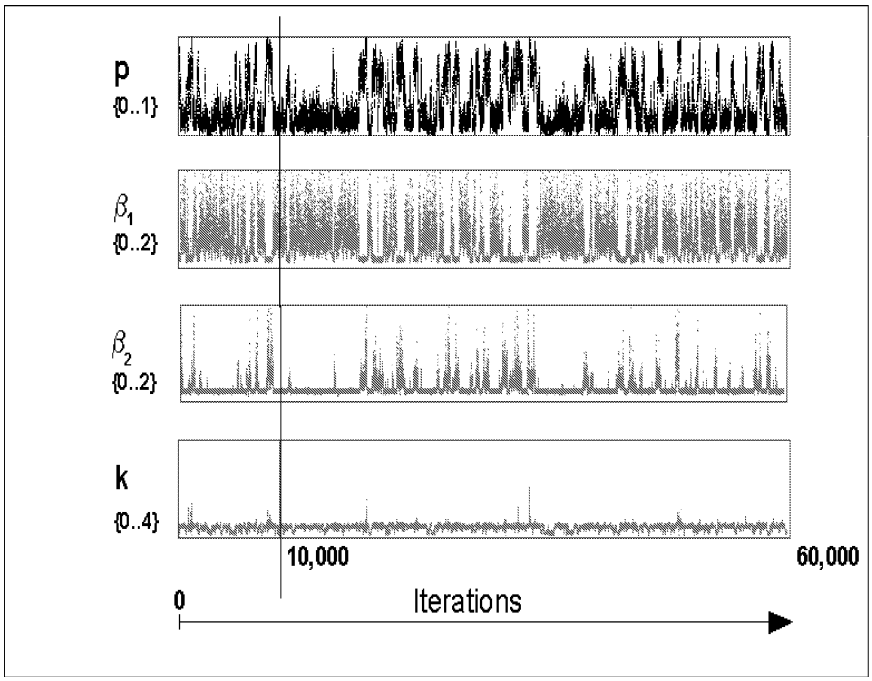


Figure 29: Raw outputs for the Schuhl model

regularity as in the case of the Griffiths & Hunt model but they still suggest the same indentifiability problems previously encountered. In this case, however, the sampler shows a bias towards the second component as can be seen from the output of the weighting parameter p. Here, the sampler stays longer at the lower end of the range of p. This is clearly evident when the marginal posterior distributions of the parameters are examined. These are shown in Figure 30.

A table of posterior means and variances is shown in Table 15

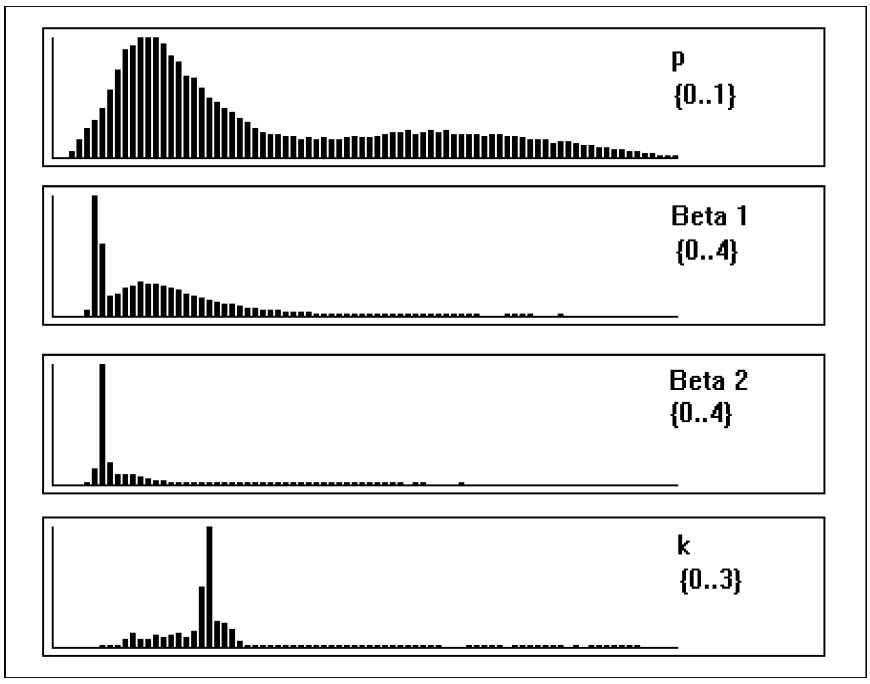


Figure 30: Marginal posterior distributions for the Schuhl model

Parameter	Posterior Mean	Posterior Variance
p	0.31067	5.6154E-2
β_1	0.55787	1.3866E-1
β_2	0.29861	2.6381E-2
k	0.64257	2.1666E-2

Table 15: Posterior means and variances for the Schuhl distribution

Examination of the histograms shows that each of the parameters has a pronounced right hand tail. In the case of p, β_1 and β_1, each posterior distribution can be regarded as the superposition of one posterior distribution onto another, with each one corresponding to a "state" that the sampler spends time in. Consider, for example, β_1, and recall that its full conditional distribution is given by

$$fcd(\beta_1) \propto \Gamma(\gamma + n_1, \Sigma_1 t_i + \delta)$$

where :-

- γ and δ are the parameters of the prior distribution for β_1,
- n_1 is the number of observations allocated to component 1 at any given sweep

of the Gibbs sampler and

- $\Sigma_1 t_i$ is the sum of the observations allocated to component 1 at any given sweep of the Gibbs sampler.

The posterior distribution of β_1 can now be explained as follows :- When, at a given sweep of the Gibbs sampler, most of the observations are allocated to component 2, n_1 and $\Sigma_1 t_i$ are both small and so the full conditional distribution of β_1 will be close to its prior distribution. This is seen as the right hand part of the bimodal posterior. On the occasion that component 1 is allocated most of the observations, n_1 and $\Sigma_1 t_i$ will dominate and, because their individual values are much higher than those of γ and δ, the resulting full conditional distribution will have a much lower variance than the prior distribution. This effect is seen as the left hand part of the histogram of β_1. In the majority of iterations, however, most observations are allocated to component 2 and this is why the left hand part of the histogram dominates in the cases of the parameters p and β_2.

Figure 31, β_{1i} v β_{2i} for all post burn-in iterations, again shows the way in which the sampler "jumps" between states.

As expected, the parameters p, β_{1i} and β_{2i} fail to converege and in this we see another similarity to the Griffiths & Hunt model. The shift parameter, k, also fails the test for convergence but detailed examination of its behaviour reveals a phenomenom not previously encountered.

Figure 32 shows the marginal posterior distribution of the parameter k truncated such that the range plotted is $\{0..0.8\}$ and it can be seen that this distribution is multimodal and although one mode clearly dominates, a further eight modes can be observed.

There are two factors which contribute to this behaviour :-

1. In this model, k only features in the second component and so it takes its likelihood only from that component. This means that only those observations

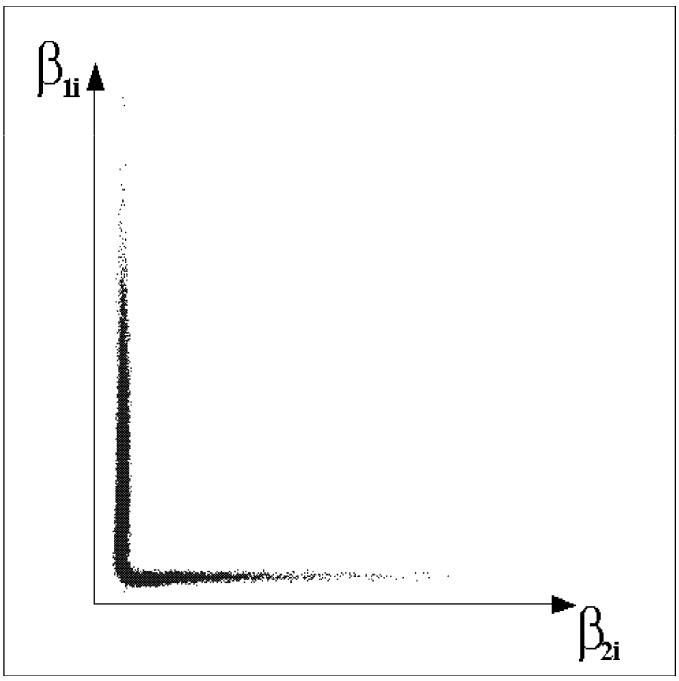


Figure 31: Graph of β_{1i} v β_{2i}

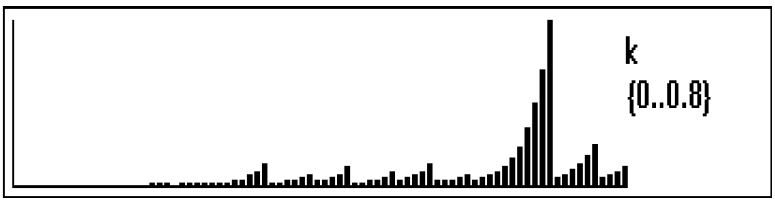


Figure 32: The truncated marginal posterior distribution of k

allocated to that component contribute to the full conditional distribution of k.

2. The lowest observation in the second component becomes the upper limit of the posterior distribution of k and this value can change with each iteration of the sampler. In practise, however, this value remains constant for a number of iterations before changing. This can be seen from the Figure 33 which plots the values of the lowest observation in component two for iterations 1 to 500 of the sampler (after burn-in).

The relationship between these two quantities is clearly seen through Figure 34. We see that t_{min} is not a continuous variable but that it takes set values on the real line. This causes, in effect, k to have numerous posterior distributions each

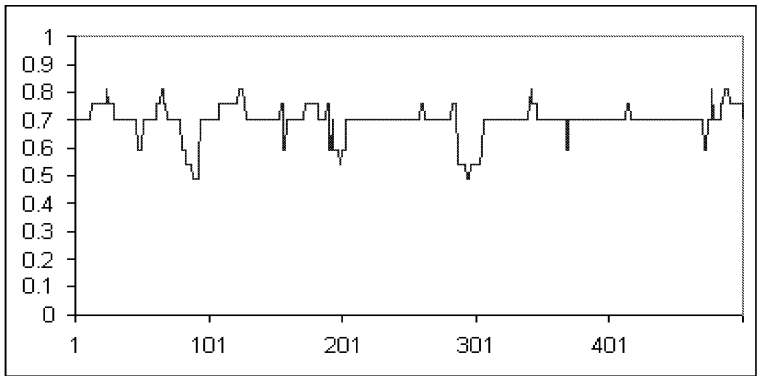


Figure 33: Graph of t_{min} v iteration number

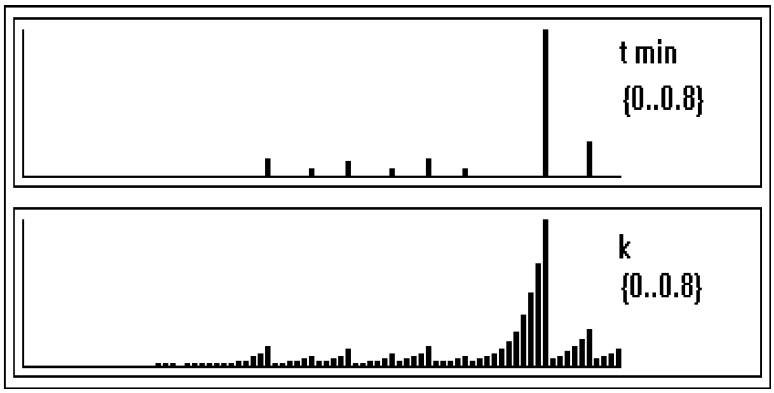


Figure 34: The parameters t_{min} and k

depending on the current value of t_{min}. This is disturbing since it means that a uni-modal posterior distribution for k can never exist. Thus one of the criteria for a succesful run can never be met.

Again, however, the question must be asked *" Was it the algorithm, model, prior distribution, data, or a combination of some or all of these that is responsible for this effect?"* It is, perhaps, worthwhile to comment on the above factors individually :-

- **The algorithm** Gibbs sampling is the algorithm of choice for this project and no other will be considered at this time.

  We could, as in the case of the previous model, introduce the constraint that the mean of the first component is always greater than that of the first. Strictly speaking a constraint such as this forms a part of the prior beliefs but it is

implemented by amending the sampling algorithm. This would mean sampling all the parameters, testing for the condition and re-sampling all the parameters if the condition is not met. It is, perhaps, tempting to shorten this procedure by only re-sampling the last of the parameters, instead of all of them, until the condition is met. In this case, the parameter in question is k since the sampling order is p, β_1, β_2 and k but careful consideration of the detail of such a strategy reveals that a major problem can arise.

The condition that we require for this constraint to apply is that

$$\frac{1}{\beta_1} > \frac{1}{\beta_2} + k$$

This inequality can be rearranged in several ways but consider the following

$$k < \frac{\beta_2 - \beta_1}{\beta_2 \beta_1}$$

and also the order in which the parameters are sampled :-

1. Sample p
2. Sample β_1
3. Sample β_2
4. Sample k

We see that when k is sampled β_1 and β_2 have already been sampled and if β_1 is greater than β_2 then the condition cannot be met no matter how many times we sample k. By adopting the method of re-sampling all parameters, we avoid any difficulties that parameter sampling order can potentially cause. This is, perhaps, a good example of the care that must be taken when constructing Gibbs sampling algorithms.

- **The model** It may be that the discontinuity in the model, caused by the parameter k in the second component, renders the model unusable in that

no method can be found to circumvent the multimodality in the marginal posterior distribution of k.

- **The prior distributions** We could use more informative prior distributions, particularly for the parameter k and, if the variance of this particular prior is small enough, the problem may be circumvented. The difficulty here is that in making the variance very small, very definite prior knowledge is implied which may not actually exist.

- **The data** The data file already used represents fairly light traffic : File 3 is composed of observations recorded during the evening rush hour and so we might expect the model to behave differently when this data is used.

7.8 Further investigations

Based on the evidence provided by the Base Run a number of exploratory runs of the Gibbs sampler were carried out. From these, the following conclusions were drawn :-

- In order to ensure that the parameters p, β_1 and β_2 have unimodal marginal posterior distributions, it is necessary to impose the condition that $E(C_1) > E(C_1)$, i.e. the mean of the first component is greater than that of the second.

- If informative priors are used for the parameters β_1 and β_1, then the convergence properties of that particular run will be improved. A satisfactory prior distribution for these parameters was found to be $\Gamma(5, 15)$.

- As the variance of the prior distribution for the parameter k is progressively reduced, the sampler becomes increasingly likely to become "trapped" at a random iteration. This causes computer run times to become unacceptably long. A prior distribution of $\Gamma(6, 12)$ was found to have a sufficiently low variance (0.04167 to 5 d.p.) without causing the sampler to become "trapped".

Given these conclusions, a run of the Gibbs sampler was carried out as follows :-

Prior distributions

The prior distributions are the same as those use used in previous runs but the values of the prior parameters have been changed. These are shown in Table 16.

Parameter	Value
γ	5.0
δ	15.0
θ	6.0
ν	12.0

Table 16: Prior distribution parameter values

Starting values

The starting values for the run are shown in Table 17.

Parameter	Starting value
p	0.5
β_1	0.2
β_2	0.2
k	0.2

Table 17: Model parameter starting values

Constraints

The constraint used was the same as in previous runs, i.e. $E(C_1) > E(C_2)$.

Burn-in and sample sizes

The burn -in consisted of 10,000 iterations and the sample 50,000 as before.

7.8.1 Run outcome

The raw outputs from the run are shown in Figure 35.

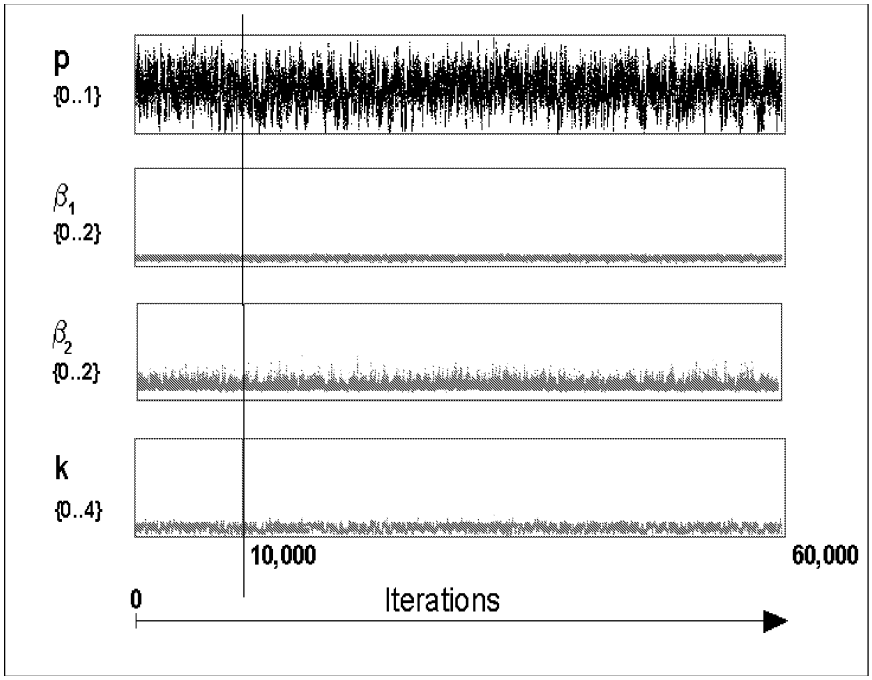


Figure 35: Raw outputs for the Schuhl model

It can be seen from Figure 35 that the sampler does not jump from one region of the parameter space to another and in Figure 36 it can also be observed that the parameters p, β_1 and β_2 have unimodal marginal posterior distributions.

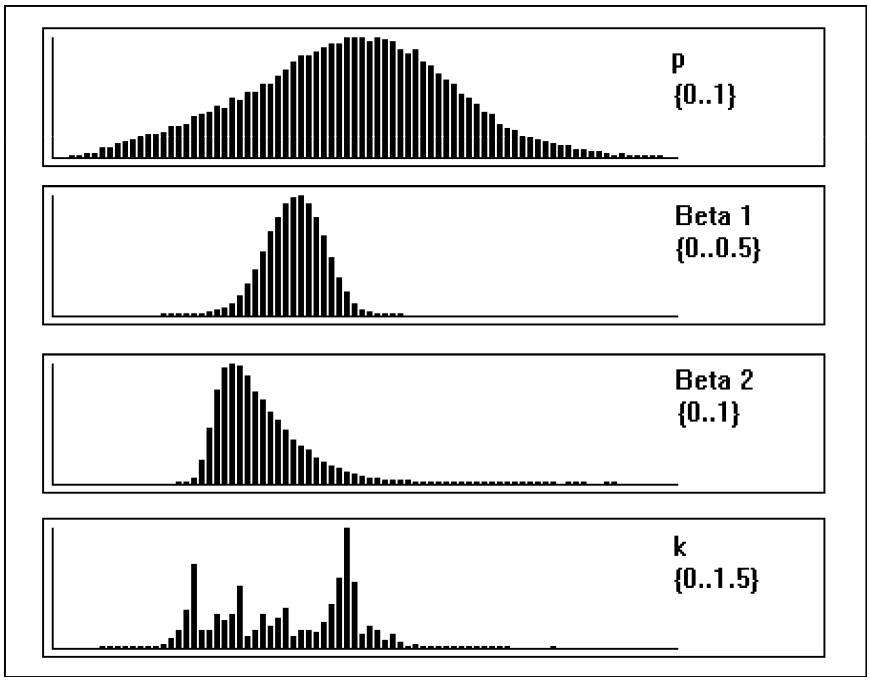


Figure 36: Raw outputs for the Schuhl model

Unfortunately, however, the parameter k clearly still has a multimodal marginal posterior distribution and so one of the criteria for a successful run has not been met. However, examination of Figure 37 shows that, in spite of this failing, model fit appears to be quite acceptable.

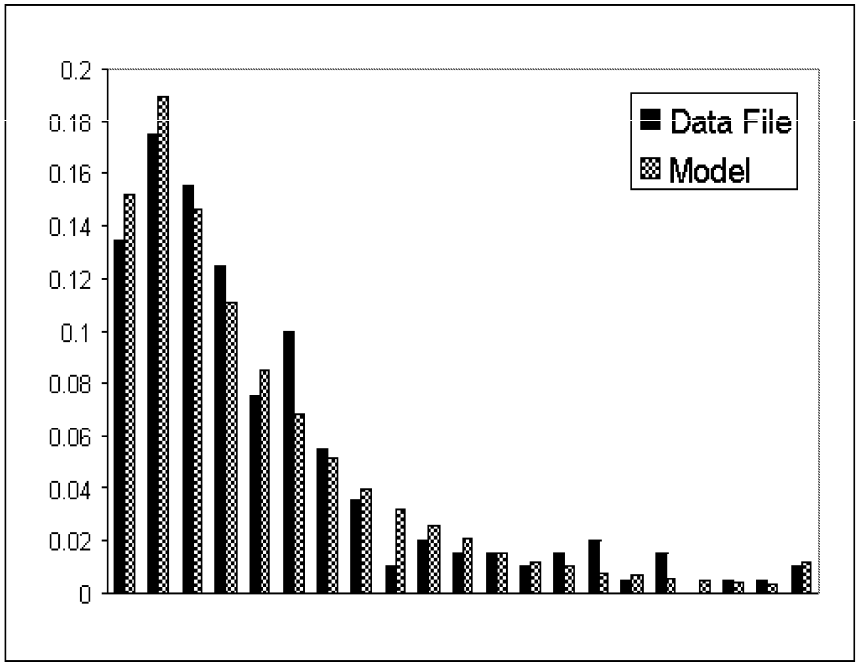


Figure 37: Model fit diagram

In addition to satisfactory model fit, the Kolmogorov-Smirnov test reveals that the sampler has converged. Test values for each parameter are shown in Table 18, together with the critical value.

Parameter	K - S test value
p	6.7840E-3
β_1	9.2170E-3
β_2	4.7099E-3
k	7.4953E-3
Critical value	1.3238E-2

Table 18: Kolmogorov - Smirnov test values

Consideration of these results begs the question *" In the light of the other criteria being met, does presence of a single multimodal marginal posterior distribution necessarily mean that the run must be considered a failure?"* In the particular case

of this specific run the answer would appear to be "No" the answer must be sought in general terms. A step in this direction would be to repeat this run but with a different data file such as File 3 which corresponds to a fairly congested stream of traffic.

Run outcome

The raw outputs for this run are shown in Figure 38 and it can be immediately seen that there are difficulties present.

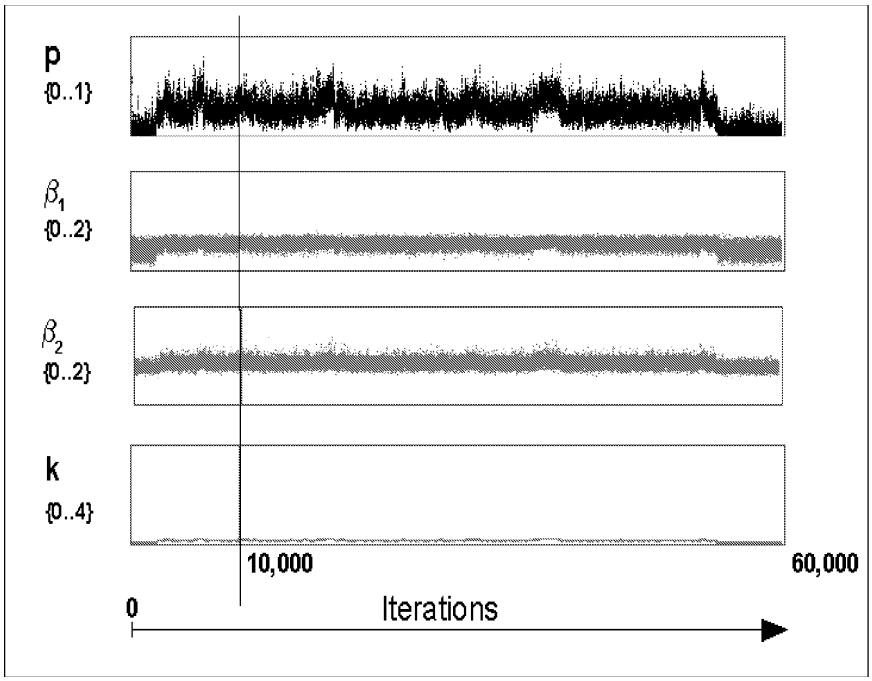


Figure 38: Raw outputs for the Schuhl model

The parameter p moves between three regions of its parameter space. These movements are not as distinct as in the case of some other runs and so are not referred to as "jumps" but their presence still gives rise to an unacceptable marginal posterior distribution. All the marginal posterior distributions for this run are shown in Figure 39 and it can be seen that p has a bimodal marginal posterior distribution.

The likely reason for its histogram having two modes as opposed to three is that the movements between the upper two states are relatively smooth transitions compared to the movement between the first and middle states. It may be the case

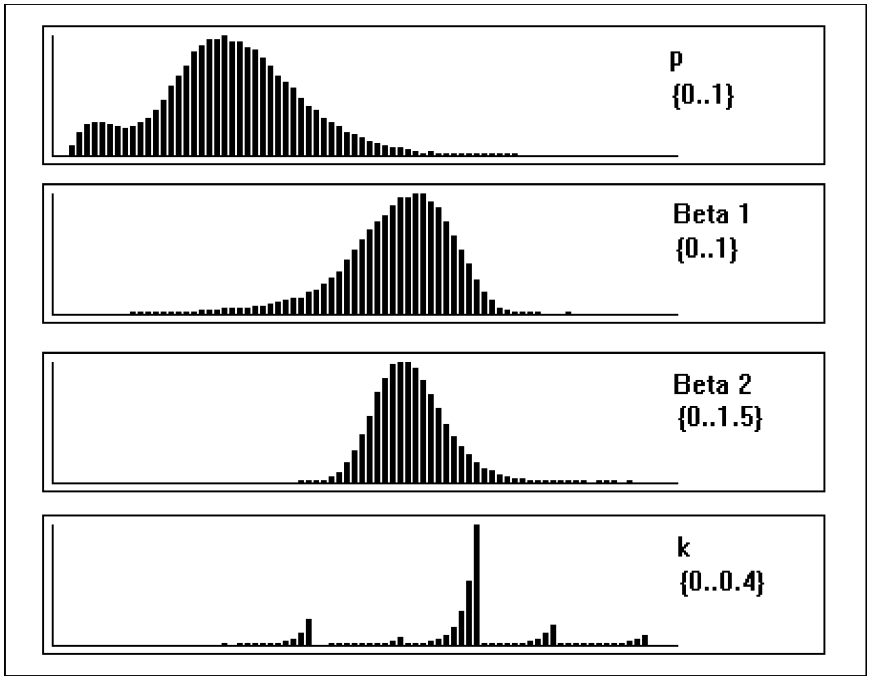


Figure 39: Marginal posterior distributions for the Schuhl model

that the upper state corresponds only to the right hand tail of the distribution and so is not an altogether separate region of the parameter space.

A similar problem is encountered in the case of the parameters β_1 and β_2 where it appears that both have tails of their marginal posterior distributions corresponding to a region of thier parameter spaces which is not frequently visited. The overall effect is similar to that observed in the Base Runs carried out so far, although not as dramatic. It is, however, the parameter k that displays the more curious behaviour. The multimodal nature of its marginal posterior distribution is clearly visible in Figure 39 and is also apparent if we plot a part of the raw ouput for k on a scale of $\{0..0.4\}$. Figure 40 shows the first 25,000 post burn-in iterations plotted in this way and distinct regions of the parameter space that the sampler moves to are clear.

Examination of Figure 40, however, appears to reveal only four distinct regions of the parameter space while Figure 39 shows five modes. This is because two regions are so close that, plotted on the scale $\{0..0.4\}$, they appear to merge into one.

It is, perhaps, not surprising that the convergence properties of this run are poor. The Kolmogorov - Smirnov values are shown in Table 19 and it is clear that the

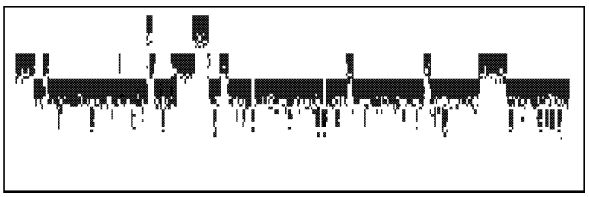


Figure 40: Part of the raw output for the parameter k

sampler is far from convergence.

Parameter	K - S test value
p	2.0976E-1
β_1	1.2979E-1
β_2	1.0225E-1
k	7.4953E-1
Critical value	1.3238E-2

Table 19: Kolmogorov - Smirnov test values

Even if the output files for each parameter are thinned by a factor of 10, there is little improvement and this in itself is sufficient reason for considering this particular run a failure even though the model fit, as shown in Figure 41 is quite good.

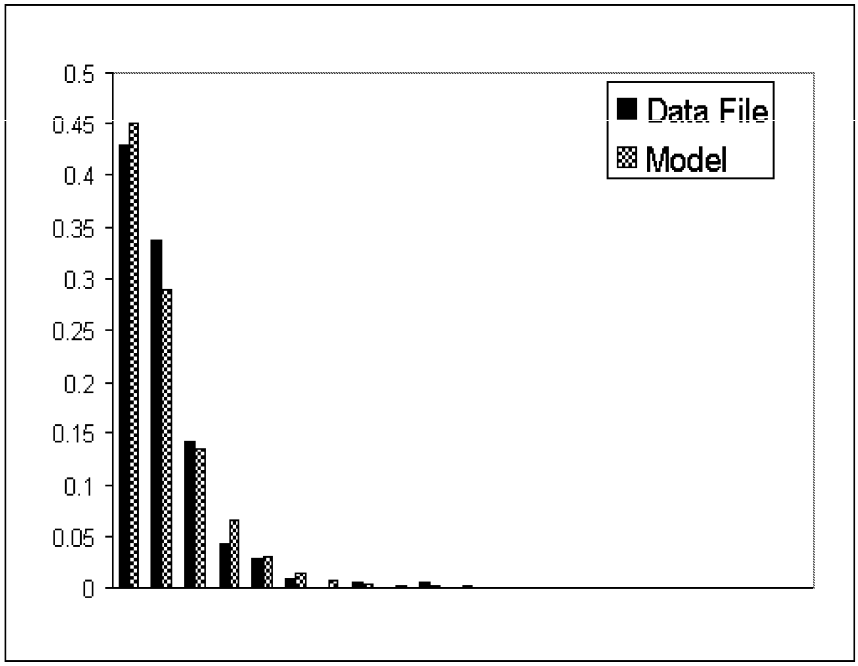


Figure 41: Model fit diagram

Finally, as another demostration of the strange behaviour of the parameter k, Figure 42 shows a graph of k_i v k_{i-1} for the above run of the Gibbs sampler. The briefest inspection of this graph reveals that k behaves quite differently to any parameter previously encountered.

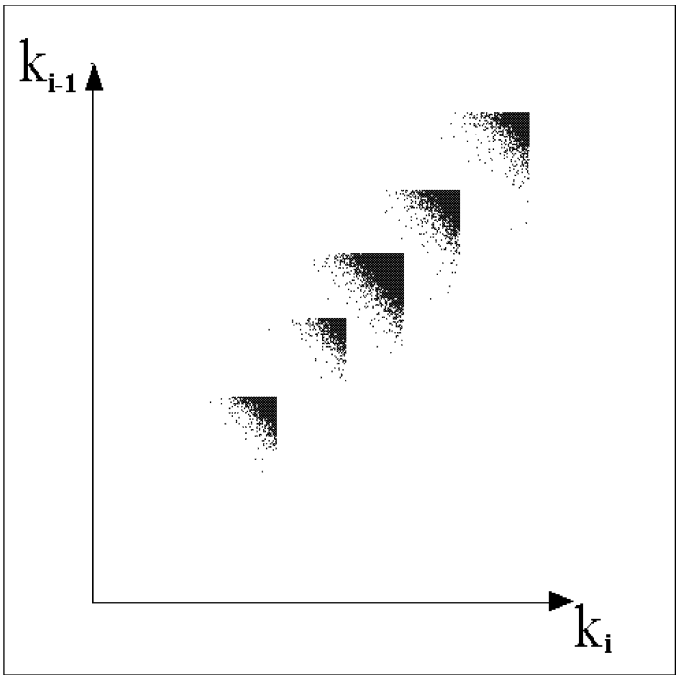


Figure 42: Graph of k_i v k_{i-1}

7.8.2 Summary

The Schuhl model performs better than the Griffiths & Hunt model in terms of reflecting the data used but the parameter k presents a problem that cannot, at present, be overcome. Not only is there the ever present multimodal marginal posterior distribution for k which is difficult to summarise but there also exist severe convergence problems when more congested data is analysed. Whilst some practitioners may choose to use the model in the context of light traffic, and ignore the multimodality of k, its use with congested streams of traffic is not recommended.

Parameter	Prior Distribution
p	$\beta(0,1)$
β_1	$\propto \beta_1^{\gamma-1}e^{-\delta\beta_1}$
β_2	$\propto \beta_2^{\theta-1}e^{-\nu\beta_2}$
α_2	$\propto \alpha_2^{\omega-1}e^{-\kappa\beta_2}$

Table 20: Prior distributions for the Gamma Exponential model

7.9 The Gamma Exponential Model

Any model now proposed must possess certain properties in order that the previously encountered problems can be overcome. These are as follows :-

- The model must be of a form such that the identifiability problem encountered in all mixture models can be dealt with and an advantageous feature of any proposed model would be that, for at least some parameter values, the modes of the components are different.

- The fundamental failing of the Schuhl model was the presence of the shift parameter k in one component. No such parameter should be present in any proposed model.

- In preparation for a future attempt to assign realistic interpretations to parameter values, the model proposed should have two components

The model proposed by the author, to satisfy all the above requirements, is as follows :-

$$f(t) = p\beta_1 e^{-\beta_1 t} + (1-p)\frac{\beta_2^{\alpha_2} t^{\alpha_2-1} e^{-\beta_2 t}}{\Gamma(\alpha_2)}$$

where $\alpha_2 > 1$. The prior distributions used for its parameters are shown in Table 20. The starting values of the model parameters are also set out in Table 21.

7.9.1 The Base Run

We begin the modelling process by running the Gibbs sampler on an unconstrained model using File 2 as data. Also, mildly informative priors will be used and their

Parameter	Starting value
p	0.5
β_1	0.2
β_2	1.0
α_2	2.0

Table 21: Starting values for the Gamma Exponential model parameters

parameter values are set out in Table 22.

Parameter	Value
γ	5.0
δ	5.0
θ	5.0
ν	5.0
ω	3.0
κ	1.0

Table 22: Prior distribution parameter values

Run Outcome

The raw output from the base run is shown in Figure 43. Once again we immediately

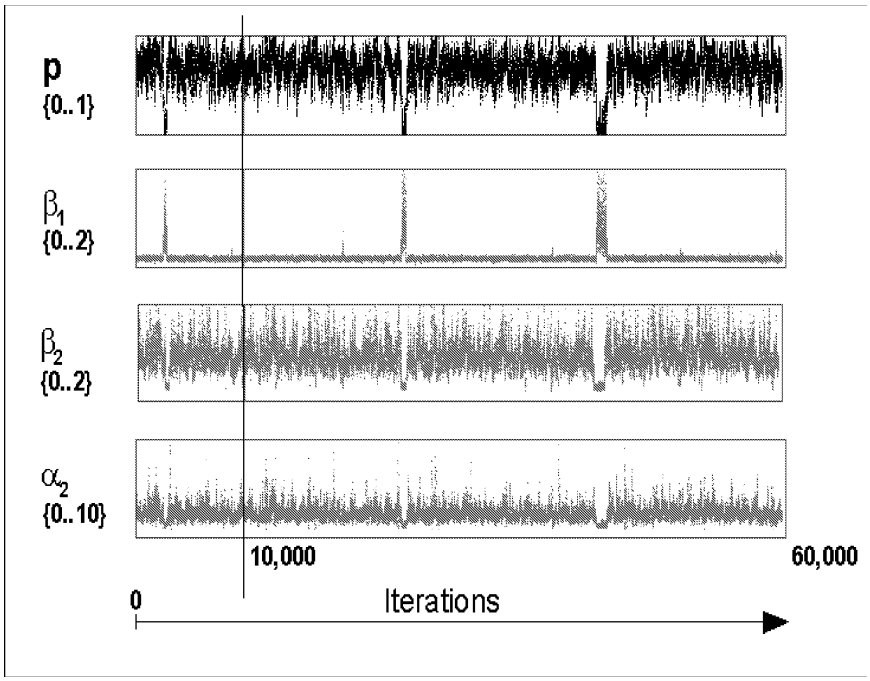


Figure 43: Raw outputs for the Gamma Exponential Model

see that use of the unconstrained model gives rise to a difficulty already encountered. Although not as distinct as in the case of the Griffiths & Hunt model the now familiar "jumps" between states can be observed in the outputs for all parameters and most clearly for p and β_1. The resultant bimodal posterior distributions are clearly visible in the histograms shown in Figure 44 although in the case of β_1 the effect is observed by way of a very long right tail of the posterior.

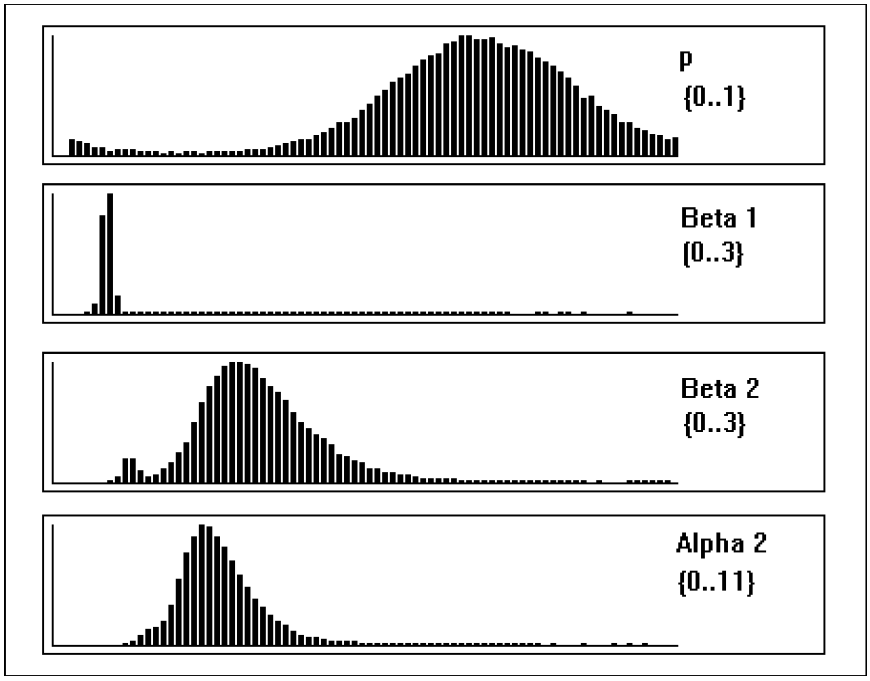


Figure 44: Marginal posterior distributions for the Gamma Exponential model

A table of posterior means and variances is shown in Table 23.

Parameter	Posterior Mean	Posterior Variance
p	0.66036	2.8538E-2
β_1	0.21062	1.7203E-2
β_2	0.92767	8.3508E-2
α_2	2.7319	5.2348E-1

Table 23: Posterior means and variances

The model/data graph is shown in Figure 45 and it can be seen that the model appears to reflect the data reasonably in spite of the identifiability problem observed. The reason for this is that one state dominates, i.e. the sampler spends most of the

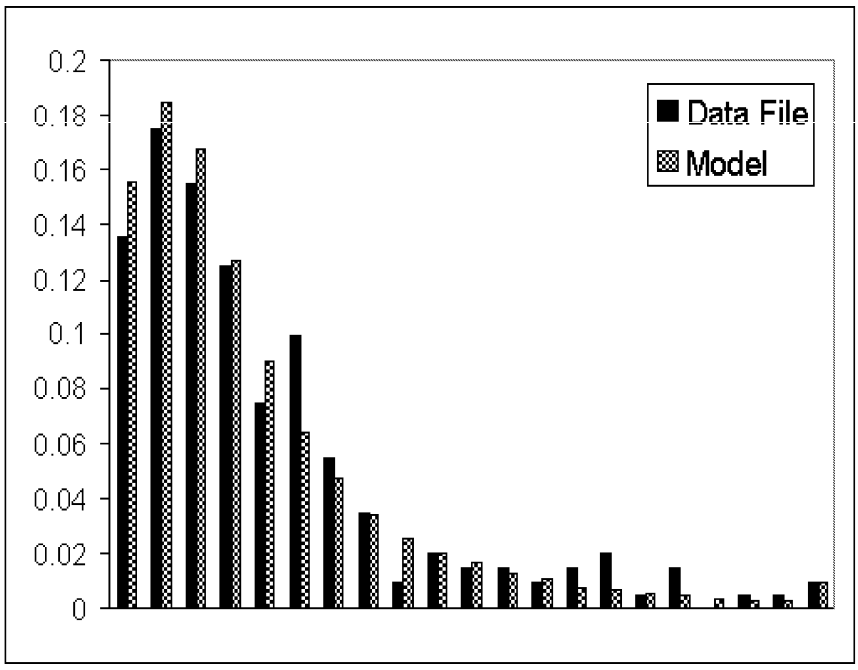


Figure 45: Model fit diagram

time in one zone of the parameter space, and the jumps into a different zone are infrequent and short lived. Thus, the parameter posterior mean values still provide a reasonable summary of the posterior distribution.

In addition to the bimodality it can also be seen that in the cases of β_2 and α_2, there is another effect that can be observed that has not been encountered before. Examination of the raw output plots for the two parameters shows that there appears to be correlation between them. This becomes clear when we plot β_{2i} against α_{2i} as shown in Figure 46, with both axes representing the range $\{0..10\}$. In addition to this correlation between two parameters, strong autocorrelation is also observed when parameters are plotted against themselves as described previously. Figure 47 shows the parameter p plotted agianist itself with a lag of 1 and the autocorrelation is striking, with the value of r being 0.95693. Both axes of Figure 47 represent the range. $\{0..1\}$.

With such a high value for r, combined with the identifiability problem observed, it is not surpising that the sampled chains did not pass the K-S test for converegence.

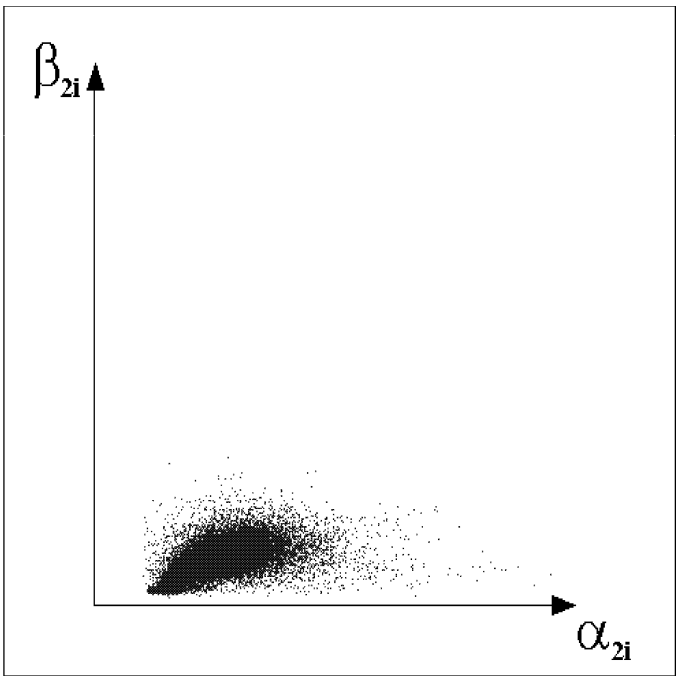


Figure 46: Graph β_{2i} v α_{2i}

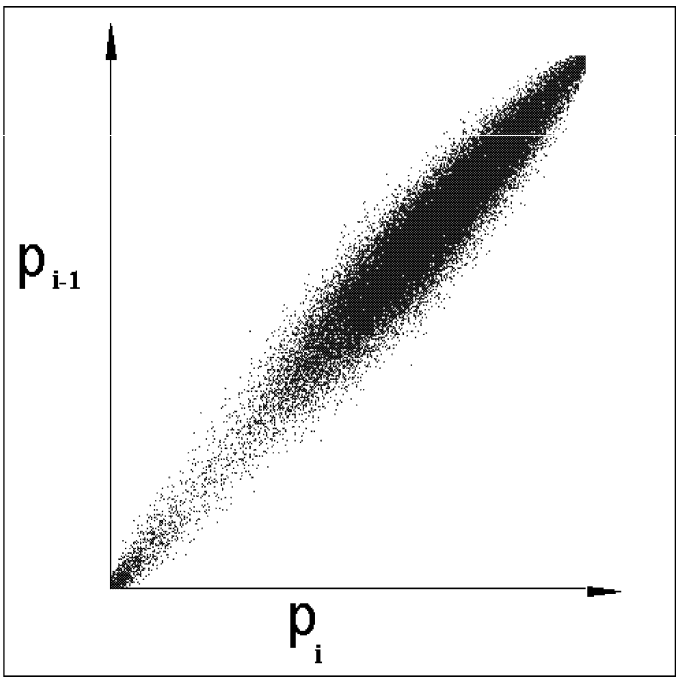


Figure 47: Autocorrelation plot of the parameter p

7.9.2 $E(C_1) > E(C_2)$

In the case of this model this particular constraint is an expression of prior belief. The first component, parameterised as a simple exponential distribution, represents

free-flowing traffic which will inevitably contain some large headways. Salter (1974) goes as far saying that any headway greater than seven seconds will be from the free-flowing component. In this project, however, the prior belief will not be so informative and the condition $E(C_1) > E(C_2)$ will be used where $E(C_1)$ denotes the mean of the first component of the model, in this the exponential component, and $E(C_2)$ is similarly defined.

The prior distributions, starting values and prior distribution parameters are the same for the base run. The constraint is applied using the method already described.

Run Outcome

The raw ouput from this run is shown in Figure 48

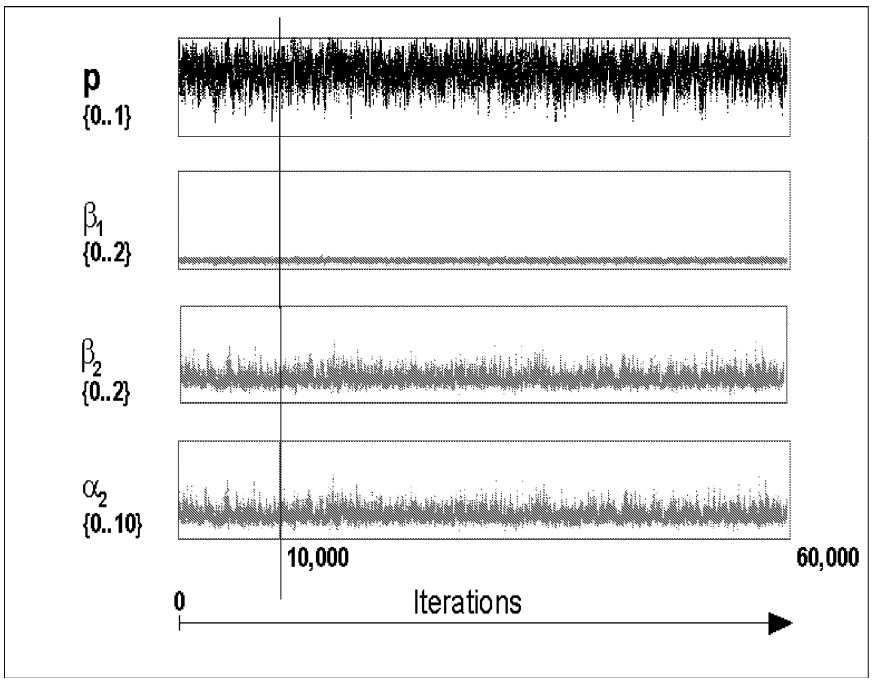


Figure 48: Raw outputs for the Gamma Exponential Model

It is immediately apparent that there is a significant improvement gained by the use of this constraint and this is reinforced by inspection of the posterior histograms which are shown in Figure 49. A table of posterior means and variances is shown in Table 24

In Figure 49 it can be observed that one of the criteria by which the modelling

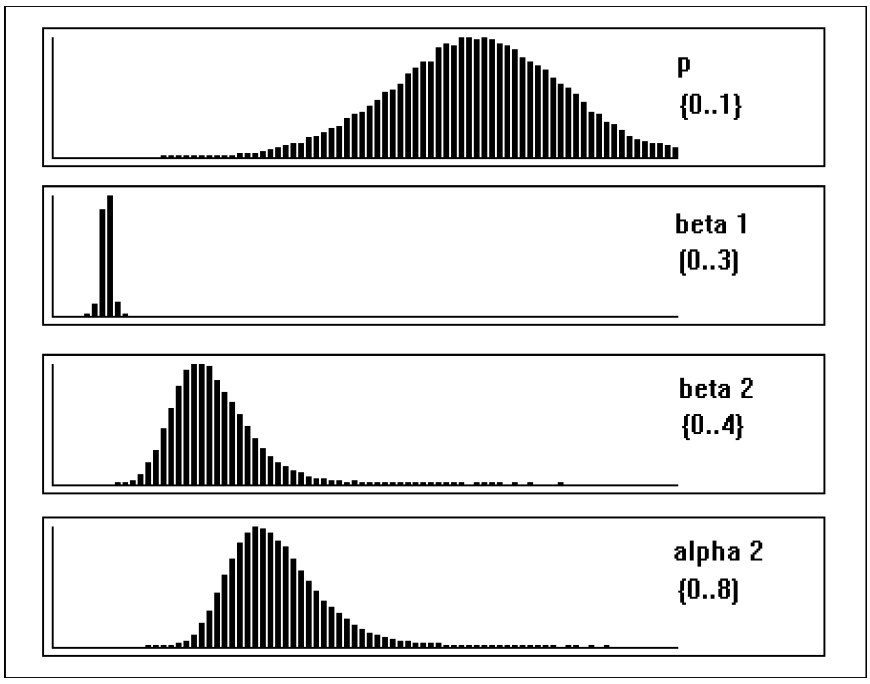


Figure 49: Marginal posterior distributions for the Gamma Exponential model

Parameter	Posterior Mean	Posterior Variance
p	0.66569	1.8468E-2
β_1	0.18909	5.0767E-2
β_2	0.95174	6.6663E-2
α_2	2.72840	3.6125E-1

Table 24: Posterior means and variances for the Gamma Exponential model

process is judged, i.e. the presence of unimodal posterior distributions, has been satisfied.

Another criterion is that of model fit and Figure 50 shows how the model reflects the data. Although the model does not reflect the data perfectly, the fit can still be described as satisfactory and this represents another criterion for successful modelling met.

Unfortunately, however, convergence remains a problem and it can be seen from Table 25 that the sampler has not converged.

The maximum allowable value of the K-S statistic, for samples of size 50,000, is 1.3238E-2 and so it is clearly visible that convergence has not been achieved. If the output chains are "thinned", i.e. only every fifth sampled value is saved,

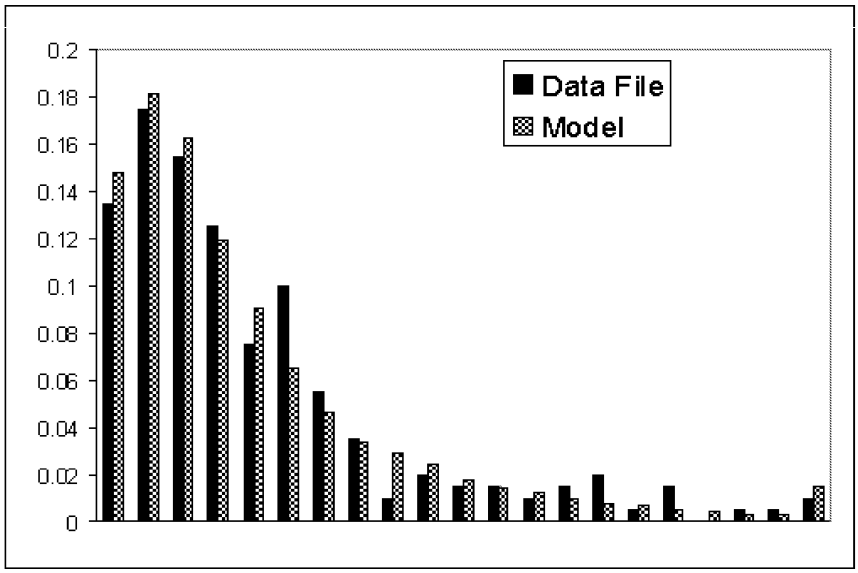


Figure 50: Model fit diagram

Parameter	K - S value
p	6.1128E-2
β_1	4.1281E-2
β_2	5.4115E-2
k	5.6185E-2

Table 25: Kolmogorov - Smirnov values

then the sample size is reduced to 10,000 and the maximum allowable value of the K-S statistic is 2.9603E-2. Table 26 shows that even under this condition the sampler does not converge. So far, two of the three criteria required for a successful

Parameter	K - S value
p	5.7792E-2
β_1	3.4175E-2
β_2	6.0568E-2
k	6.3537E-2

Table 26: Kolmogorov - Smirnov values

modelling outcome have been met but the problem of non-convergence remains. This particular difficulty is very common when modelling with mixtures and there are numerous techniques that can be used to overcome it. The two that will be considered here are the use of a *" Randomly Updated Gibbs Sampler*, or R.U.G.S.,

and a technique known as *"Blocking"*. A full general treatment of these two updating schemes is given by Roberts & Sahu (1997). Here an explanation will be given of how R.U.G.S. and bocking apply in the present modelling context.

7.9.3 R.U.G.S.

The acronym R.U.G.S. stands for Randomly Updated Gibbs Sampler which is also known as the Random Scan Gibbs Sampler. So far, all sampling has been carried out in such a manner that at each sweep of the Gibbs sampler the parameters have been sampled in strict order, ie., p, β_1, β_2 and α_2. This technique is known as a Deterministically Updated Gibbs Sampler or D.U.G.S. and is used in the programs BUGS and WinBUGS. However, to understand the purpose of R.U.G.S., it is necessary to recall the conditional nature of the sampling process.

Suppose the sampler is about to make its i^{th} sweep. The sampling proceeds as follows :-

1. Sample p_i from $fcd(p|\beta_{1.i-1}, \beta_{2.i-1}, \alpha_{2.i-1})$
2. Sample $\beta_{1.i}$ from $fcd(\beta_1|p_i, \beta_{2.i-1}, \alpha_{2.i-1})$
3. Sample $\beta_{2.i}$ from $fcd(\beta_2|p_i, \beta_{1.i}, \alpha_{2.i-1})$
4. Sample $\alpha_{2.i}$ from $fcd(\alpha_2|p_i, \beta_{1.i}, \beta_{2.i})$

In R.U.G.S., however, the order of updating the parameters is chosen at random at every sweep and this technique has been found to have a bearing on convergence (Roberts & Sahu, 1997). The algorithm for R.U.G.S. can be expressed as follows :-

```
REPEAT {
        pick a parameter at random
        IF the parameter has already been sampled  THEN
            do nothing
        ELSE
```

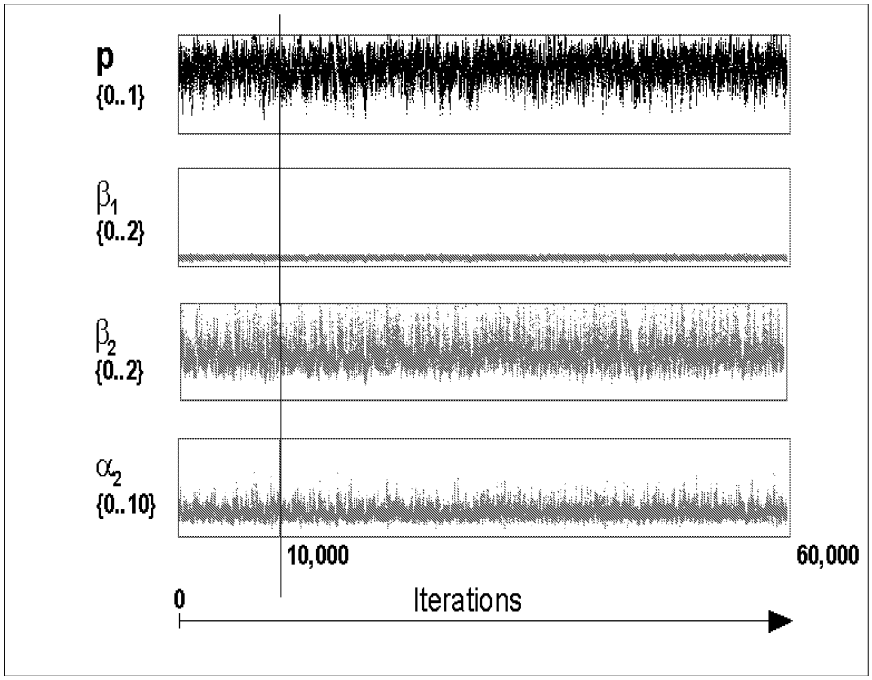


Figure 51: Raw outputs for the Gamma Exponential Model

```
        BEGIN
            sample from the fcd of the parameter
            mark the parameter as sampled
        END
    }   UNTIL all parameters have been sampled
```

Run Outcome

The raw output for the Gibbs sampler using R.U.G.S, is shown in Figure 51

There is little difference to be observed between Figures 48 and 51 and this is also the case when marginal posterior distributions are compared (Figures 49 and 52)

We see, as expected, that the posterior distributions are unimodal and, from Figure 53, that the model fit is satisfactory. However, the main purpose of this particular run was to determine if use of the R.U.G.S. algorithm was a useful aid to convergence. To do this the K-S values need to be examined and they are shown

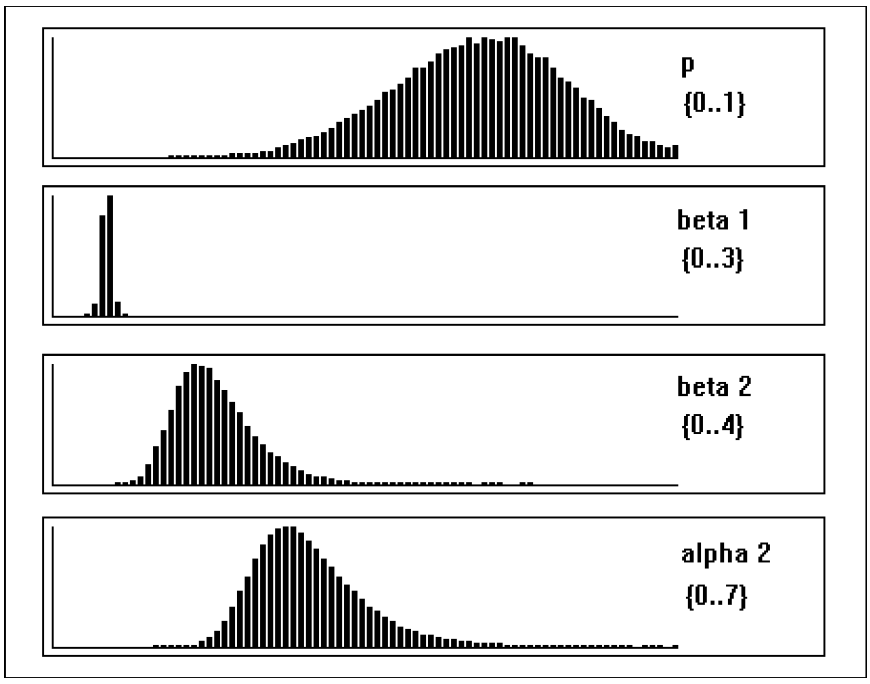


Figure 52: Marginal posterior distributions for the Gamma Exponential model

Parameter	Posterior Mean	Posterior Variance
p	0.67088	1.8636E-2
β_1	0.18949	5.0876E-2
β_2	0.96490	7.0801E-2
α_2	2.75410	3.8526E-1

Table 27: Posterior means and variances for the Gamma Exponential model

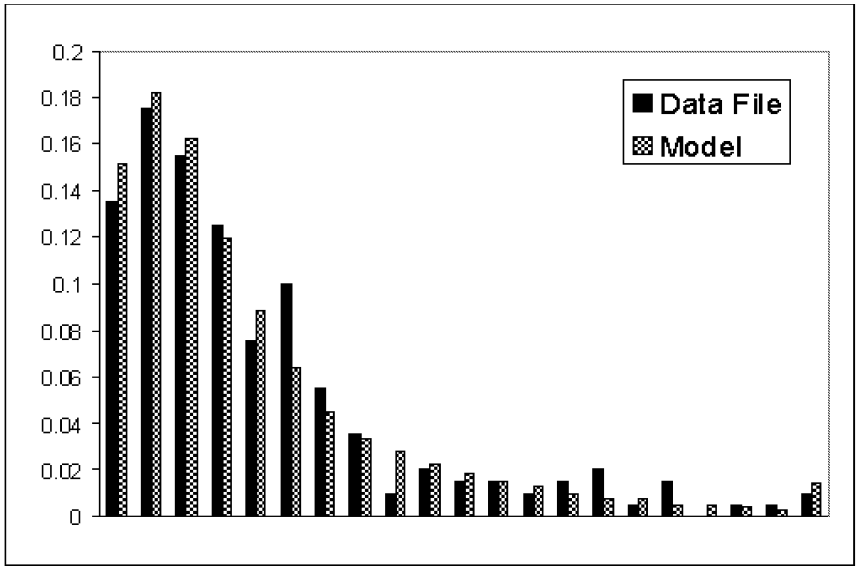


Figure 53: Model fit diagram

for both the raw and thinned output in Table 28

Parameter	Raw Output	Thinned output
p	2.0462E-2	2.1941E-2
β_1	1.0207E-2	0.6112-2
β_2	3.1322E-2	3.5429E-2
α_2	3.4449E-2	3.1073E-2
Critical value	1.3238E-2	2.9603E-2

Table 28: Kolmogorov - Smirnov values for raw and thinned output

Although there has been a clear improvement the sampler does not converge, even when the thinned chains are used. Furthermore, it can be seen that it is the parameters of the second component that are, in both the raw and thinned cases, the furthest from convergence in that their K-S values exceed the critical value by the greatest amount. This would suggest that any method of improving the performance of the model in terms of convergence should focus on the second component and its parameters β_2 and α_2.

7.9.4 Blocking

The purpose of the technique known as "blocking" is to sample the parameters of a particular component from their joint probaility density function. In the second component of the model under consideration there are two parameters, β_2 and α_2, and so it required to sample from $f(\alpha_2, \beta_2|t_1, \ldots, t_k)$. This operation can be broken down into two stages as follows :-

Firstly recall that, for events A, B, we have $\Pr(A \wedge B) = \Pr(A)\Pr(B|A)$. Similarly the joint pdf of two continuous variables x, y can be written $f(x, y) = f(x)f(y|x)$.

In the same way

$$f(\alpha_2, \beta_2|t_1, \ldots, t_k) = f(\alpha_2|t_1, \ldots, t_k) \times f(\beta_2|\alpha_2, t_1, \ldots, t_k).$$

Now, if our priors for α_2 and β_2 are proportional to $\alpha_2^{\omega-1}e^{-\kappa\alpha_2}$ and $\beta_2^{\theta-1}e^{-\nu\beta_2}$ respectively then, using the same observations as before, we arrive at the following

joint posterior distribution :-

$$f(\alpha_2, \beta_2|t_1, t_2, \ldots, t_{n_2}) \propto \alpha_2^{\omega-1} e^{-\kappa\alpha_2} \beta_2^{\alpha_2 n_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2-1} \beta_2^{\theta-1} e^{-\beta_2(\Sigma_{1=1}^{n_2} t_i+\nu)} (\Gamma(\alpha_2))^{-n_2}$$

which simplifies to :-

$$f(\alpha_2, \beta_2|t_1, t_2, \ldots, t_{n_2}) \propto \alpha_2^{\omega-1} e^{-\kappa\alpha_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2-1} \beta_2^{\alpha_2 n_2+\theta-1} e^{-\beta_2(\Sigma_{1=1}^{n_2} t_i+\nu)} (\Gamma(\alpha_2))^{-n_2}$$

To evaluate the marginal posterior distribution of α_2, we require :-

$$f(\alpha_2|t_1, t_2, \ldots, t_{n_2}) \propto \frac{\alpha_2^{\omega-1} e^{-\kappa\alpha_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2-1}}{(\Gamma(\alpha_2))^{n_2}} \int_0^\infty \beta_2^{\alpha_2 n_2+\theta-1} e^{-\beta_2(\Sigma_{1=1}^{n_2} t_i+\nu)} d\beta_2$$

which gives :-

$$f(\alpha_2|t_1, t_2, \ldots, t_{n_2}) \propto \frac{\alpha_2^{\omega-1} e^{-\kappa\alpha_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2-1} \Gamma(\alpha_2 n_2+\theta)}{(\Gamma(\alpha_2))^{n_2} (\Sigma_{i=1}^{n_2} t_i+\nu)^{\alpha_2 n_2+\theta}}$$

Now

$$f(\alpha_2, \beta_2|t_1, t_2, \ldots, t_{n_2}) = f(\alpha_2, |t_1, t_2, \ldots, t_{n_2}).f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}).$$

Therefore

$$f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}).\frac{f(\alpha_2, \beta_2|t_1, t_2, \ldots, t_{n_2})}{f(\alpha_2, |t_1, t_2, \ldots, t_{n_2}).}$$

Which, after some algebra gives us :-

$$f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}) = \frac{e^{-\beta_2(\Sigma_{i=1}^{n_2} t_i+\nu)} \beta_2^{\alpha_2 n_2+\theta-1} (\Sigma_{i=1}^{n_2} t_i+\nu)^{\alpha_2 n_2+\theta}}{\Gamma(\alpha_2 n_2+\theta)}$$

The above equation is a Gamma distribution whose parameters are $\alpha_2 n_2 + \theta$ and $\Sigma_{1=1}^{n_2} t_i + \nu$ so we can write :-

$$f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}) = \Gamma(\alpha_2 n_2+\theta, \Sigma_{i=1}^{n_2} t_i+\nu)$$

The difficulty of sampling from a bivariate distribution, i.e. $f(\alpha_2, \beta_2|t_1, \ldots, t_k)$ has been circumvented by dividing the operation into two more manageable steps i.e.,

1. Sample α_2 from

$$f(\alpha_2|t_1, t_2, \ldots, t_{n_2}) \propto \frac{\alpha_2^{\omega-1} e^{-\kappa\alpha_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2 - 1} \Gamma(\alpha_2 n_2 + \theta)}{(\Gamma(\alpha_2))^{n_2} (\Sigma_{i=1}^{n_2} t_i + \nu)^{\alpha_2 n_2 + \theta}} \tag{19}$$

2. Sample β_2 from

$$f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}) = \Gamma(\alpha_2 n_2 + \theta, \Sigma_{i=1}^{n_2} t_i + \nu)$$

Fortunately, the distribution in Equation 19 above is log-concave and so can be sampled from the method already in use.

At each sweep of the Gibbs sampler, after the sample has been partitioned, sampling proceeds as follows :-

- Sample p from $B(\phi + n_1, \psi + n_2)$
- Sample β_1 from $G(\gamma + n_1, \delta + \sum_{i=1}^{i=n_1} t_i)$
- Sample α_2 from

$$f(\alpha_2|t_1, t_2, \ldots, t_{n_2}) \propto \frac{\alpha_2^{\omega-1} e^{-\kappa\alpha_2} (\prod_{i=1}^{n_2} t_i)^{\alpha_2 - 1} \Gamma(\alpha_2 n_2 + \theta)}{(\Gamma(\alpha_2))^{n_2} (\Sigma_{i=1}^{n_2} t_i + \nu)^{\alpha_2 n_2 + \theta}}$$

- Sample β_2 from

$$f(\beta_2|\alpha_2, t_1, t_2, \ldots, t_{n_2}) = \Gamma(\alpha_2 n_2 + \theta, \Sigma_{i=1}^{n_2} t_i + \nu)$$

Run Outcome

This particular run gave very encouraging results in that, for the first time in this project, all criteria for a successful run were met. Data from this run are shown in Figures 54, 55 and 56 together with Tables 29 and 30.

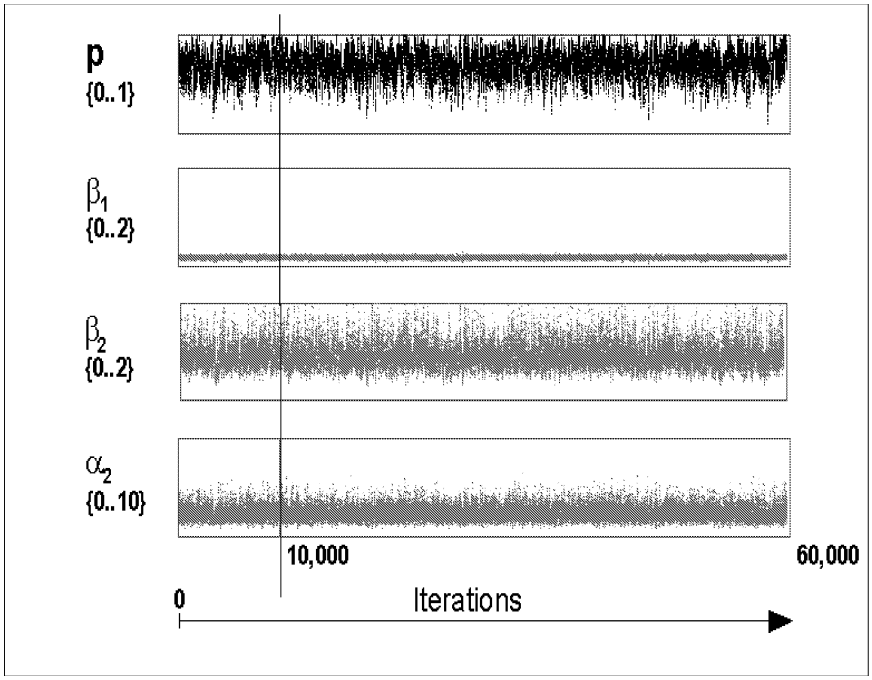


Figure 54: Raw outputs for the Gamma Exponential Model

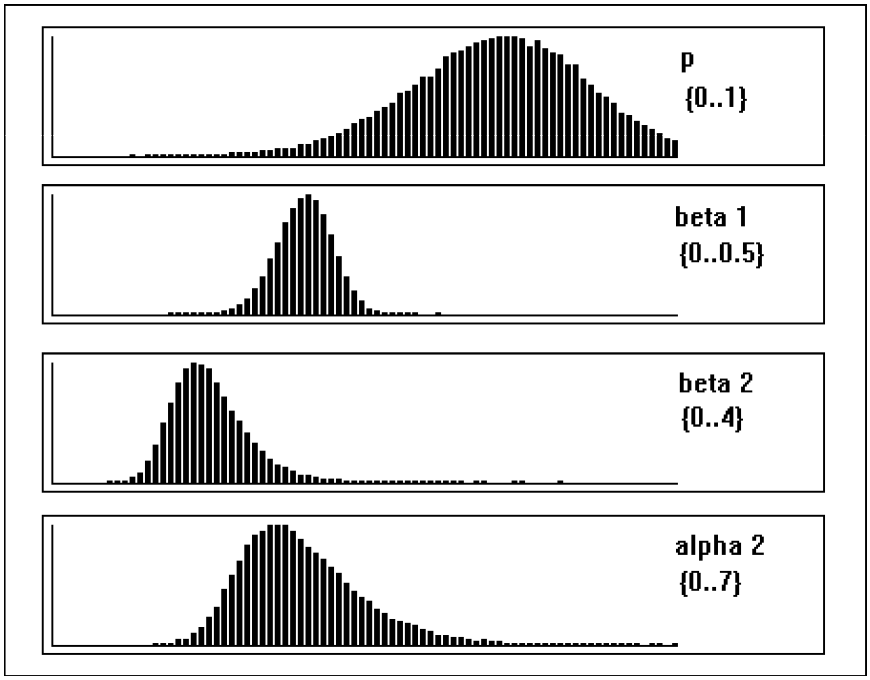


Figure 55: Marginal posterior distributions for the Gamma Exponential distribution

It can be seen from an examination of Figures 54, 55 and 56 together with Tables 29 and 30 that this run can be deemed a success and from this the inference can be drawn that the methodology is appropriate. By way of checking this a second run of the Gibbs sampler was carried out, identical to the previous one but using File 3

Parameter	Posterior Mean	Posterior Variance
p	7.3492E-1	1.6472E-2
β_1	1.9690E-1	4.7319E-4
β_2	1.0851	8.7516E-2
α_2	3.1374	7.8262E-2

Table 29: Posterior means and variances for the Gamma Exponential distribution

Parameter	Raw Output	Thinned output
p	1.0788E-2	2.0945E-2
β_1	0.61886E-2	1.2017-2
β_2	1.3795E-2	1.2748E-2
α_2	0.72714E-2	1.8598E-2
Critical value	1.3238E-2	2.9603E-2

Table 30: K-S values for the Gamma Exponential Distribution

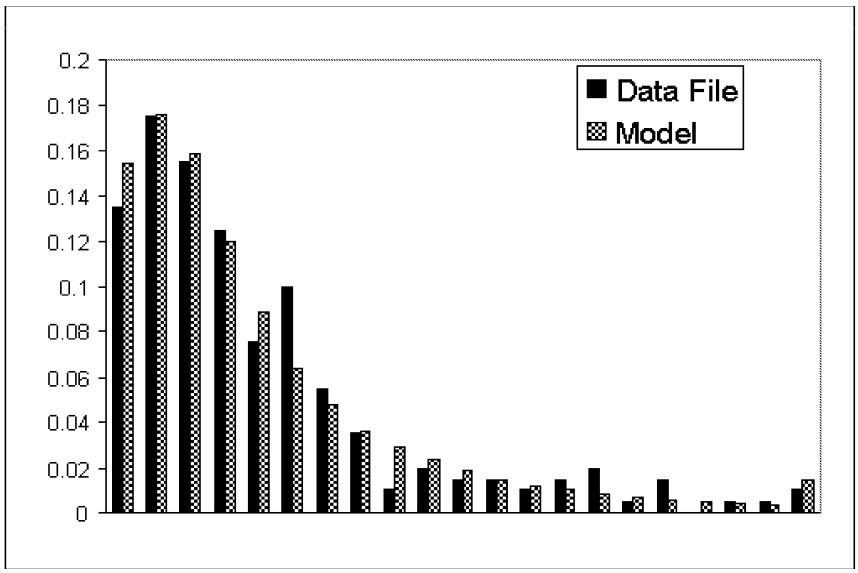


Figure 56: Model fit diagram

as the data file. This file represents a more congested stream of traffic than File 2.

7.9.5 Run outcome using File 3

A further run of the program was done using File 3 as the data file and details are tabulated in Table 31. Also, for the sake of brevity, only the posterior histograms and the model fit diagram will be shown here.

Model	Gamma Exponential
Constraint	$E(C_1) > E(C_2)$
Prior type	Mildly informative
Blocking	Second component only
R.U.G.S.	First component only
Data file	File 3

Table 31: Details of run using File 3

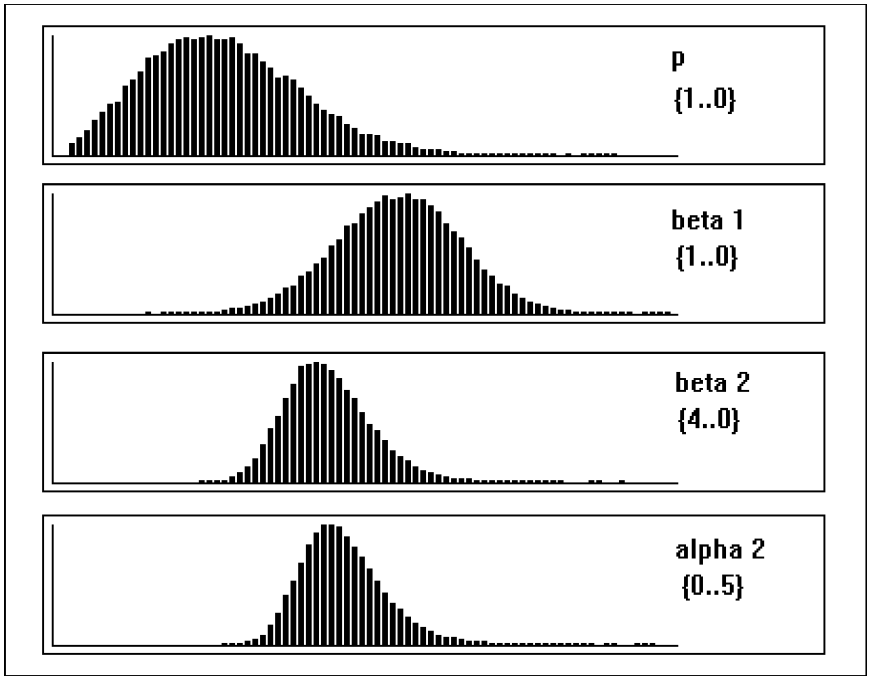


Figure 57: Marginal posterior distributions for the Gamma Exponential distribution

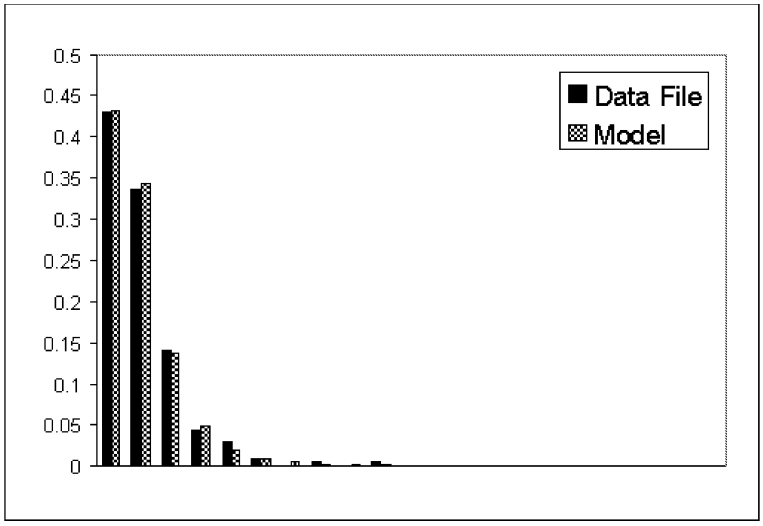


Figure 58: Model fit diagram

There are two interesting points that arise from this particular run.

1. Thinning did not prove to be necessary in this case since convergence was achieved using raw output chains.

2. Although the posterior distributions are not perfectly symmetrical, their modes were sufficiently close to their means so that these two quantities could be interchanged with very little discernable difference to the model fit diagram.

7.10 Summary

The three models have been subjected to a high degree of scrutiny in this section and, as a result, it is possible to draw firm conclusions regarding their suitability for modelling headways. But before each model is considered separately, it has to be stated that the Bayesian paradigm, and Gibbs sampling in particular, have proved highly successful tools which can give valuable insights into model behaviour.

- **The Griffiths & Hunt model** would appear to be the wrong model for this particular application. The usual problems of identifiability and slow mixing can be overcome but model fit remains unacceptable. It was also shown how Griffiths & Hunt obtained satisfactory results when using this model but their method is, in the author's opinion, unacceptable.

- **The Schuhl model**, although capable of reflecting the data in a much better way than the Griffiths & Hunt model, has difficulties of its own which arise from the discontinuity caused by the shift parameter, k, whose marginal posterior distribution can never be unimodal as required. Although the model may be adequate for modelling light traffic flow its general use is not recommended.

- **The Gamma \ Exponential distribution** proposed by the author, proved a more successful model than those previously considered. Identifiability and slow mixing were overcome and good model fits were achieved both for light and heavy traffic flows.

It is now possible state the successful algorithm in diagrammatic form, as shown in Figure 59.

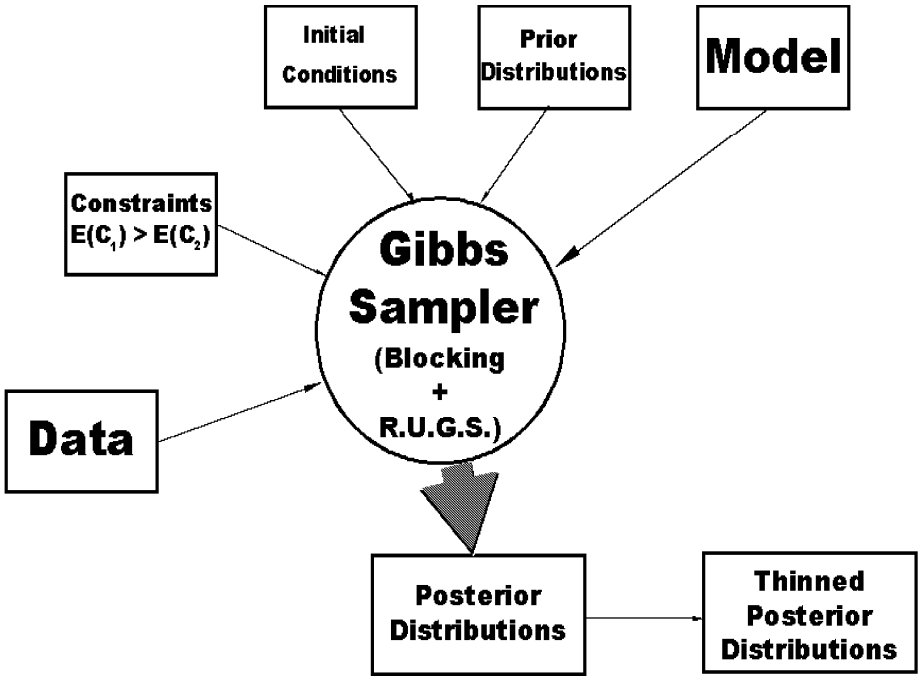


Figure 59: The successful algorithm

Further description of each element of the diagram is given below :-

The model is a two component mixture of Gamma distributions with the first component parameterised as an exponential distribution (i.e. $\alpha_1 = 1$ and the second such that $\alpha_2 \geq 1$.

Prior distributions are mildly informative except in the case of the parameter p whose prior distribution is $\beta(1, 1)$

Initial conditions, or starting values, were found to have no effect on the posterior distribution.

Constraints. The only constraint used was that $E(C_1) > E(C_2)$. Such a constraint actually forms part of our prior beliefs concerning the model.

The data consited of two files, one representing fairly light traffic and the other for more congested traffic. The model performed well with either.

We conclude, therfore, that the Gamma \ Exponential model is suitable for modelling headways in both light and heavy traffic on dual carriageways.

8 An examination of model behaviour via Bayesian deviance

Introduction

In Section 7 the problems commonly associated with the use of mixture models were described and, in the case of the Gamma / Exponential distibution, working solutions that would enable the model to be used for modelling vehicle headways were offered. These issues of identifiability, correlation and convergence have been detailed and in this section an attempt will be made to detect and explain the underlying cause of these phenomena.

In this section it will be shown that over parameterisation is an important factor in mixture model behaviour and that dealing with this issue is relatively straightforward under the Bayesian paradigm.

The method by which this concept will be explored will be that of Bayesian deviance.

8.1 Bayesian deviance

Let Θ be a vector of parameters, in the case of the Gamma/Exponential model, i.e.,

$$\Theta = [p, \beta_1, \beta_2, \alpha_2]^T$$

and suppose that the model, denoted $p(t|\Theta)$ is the subject of Gibbs sampling where there are m observations, i.e., $\{t_1, t_2, \ldots, t_i, \ldots, t_{m-1}, t_m\}$, and N iterations after the initial burn-in. Now, let the quantity D be defined by :-

$$D = -2 \sum_{j=1}^{j=m} \log p(t_i|\Theta) + 2 \log f(t)$$

where f(t) is a term that depends only on the data. Also let

$$\bar{D} = \sum_{i=1}^{i=N} \left\{ -2 \sum_{j=1}^{j=m} \log p(t_j|\Theta_i) + 2 \log f(t) \right\} \cdot \frac{1}{N}$$

where Θ_i is the vector of parameters, Θ, sampled at the ith iteration of the Gibbs sampler. This gives :-

$$\bar{D} = \frac{1}{N} \cdot \left\{ \sum_{i=1}^{i=N} -2 \sum_{j=1}^{j=m} \log p(t_j|\Theta_i) \right\} + 2 \log f(t)$$

Let $\bar{\Theta}$ be the vector of parameter posterior means, i.e.,

$$\bar{\Theta} = [\bar{p}, \bar{\beta_1}, \bar{\beta_1}, \bar{\alpha_2}]^T$$

where, for example,

$$\bar{p} = \frac{1}{N} \cdot \sum_{i=1}^{i=N} p_i$$

and let

$$D(\bar{\Theta}) = -2 \sum_{j=1}^{m} \log p(t_j|\bar{\Theta}) + 2 \log f(t)$$

Spiegelhalter, Best and Carlin (1998) defined p_D, the *effective* number of parameters, as

$$p_D = \bar{D} - D(\bar{\Theta})$$

$$\begin{aligned} p_D &= \frac{1}{N} \cdot \left\{ \sum_{i=1}^{i=N} -2 \sum_{j=1}^{j=m} \log p(t_j|\Theta_i) \right\} + 2 \log f(t) \\ &+ 2 \sum_{j=1}^{m} \log p(t_j|\bar{\Theta}) - 2 \log f(t) \\ &= \frac{1}{N} \cdot \left\{ \sum_{i=1}^{i=N} -2 \sum_{j=1}^{j=m} \log p(t_j|\Theta_i) \right\} + 2 \sum_{j=1}^{m} \log p(t_j|\bar{\Theta}) \end{aligned}$$

and, hence, the effective number of parameters can be calculated without the need to evaluate the term $2 \log f(t)$ (known as the *saturated* deviance). See also Spiegel-

halter, Best, Carlin and van der Linde (2002). Also the quantity $\bar{D_0}$ can be defined as

$$\bar{D_0} = \frac{1}{N} \cdot \left\{ \sum_{i=1}^{i=N} -2 \sum_{j=1}^{j=m} \log p(t_j|\Theta_i) \right\}$$

and similarly,

$$D_0(\bar{\Theta}) = -2 \sum_{j=1}^{m} \log p(t_j|\bar{\Theta})$$

The following equation is now arrived at :-

$$p_D = \bar{D_0} - D_0(\bar{\Theta})$$

The significance of the terms in the above equation is explained below :-

p_D	This has already been described as the *effective* number of parameters and is, therefore, a measure of model complexity.
$\bar{D_0}$	This term, clearly linked to the likelihood, is a measure of model fit. It is the expectation of the null deviance.
$D_0(\bar{\Theta})$	This is the null deviance evaluated at the posterior parameter means.

The principles described above can be demonstrated by carrying out a number of runs of the Gibbs sampler, using the Gamma / Exponential model, with different prior variances for the parameter α_2. The parameters β_1 and β_2 are given informative priors throughout the sequence of runs and this eliminates the need for blocking and the use of R.U.G.S. More importantly though, the use of informative prior distributions gives rise to approximately normal posterior ditributions. This property of the posterior distributions is essential for the above theory to apply.

In the following runs of the Gibbs sampler, the posterior distributions are assumed to be close enough to being normal for these principles to hold although it is acknowledged that cases can quite easily arise where this is not the case. This assumption of normality would not be possible if, for example, the posterior mean of the weighting parameter, p was close to either 1 or 0 and the sample size was relatively small. For this reason, the data set File 2 was used in the runs of the Gibbs sampler that follow. The prior distributions for the model parameters are set out in Table 32

Parameter	Prior Distribution
p	Uniform(0,1)
β_1	$\propto \beta_1^{\gamma-1} e^{-\delta\beta_1}$
β_2	$\propto \beta_2^{\theta-1} e^{-\nu\beta_2}$
α_2	$\propto \alpha_2^{\omega-1} e^{-\kappa\beta_2}$

Table 32: Parameter prior distributions

Table 33 shows the values of the parameters of these parameters.

Run No.	Prior parameter	Value
a	ω	3
a	κ	1
b	ω	9
b	κ	3
c	ω	24
c	κ	8
d	ω	75
d	κ	25
e	ω	240
e	κ	80
f	ω	300
f	κ	100

Table 33: Prior parameter values

Inspection of the above table reveals that the prior mean of α_2 remains constant at 3 for each run, but its prior variance is successively decreased. These are shown in Table 34

Run No.	Prior variance of α_2, $V_0(\alpha_2)$
a	3
b	1
c	0.375
d	0.12
e	0.0375
f	0.0030

Table 34: Prior variance of α_2

The parameters β_1 and β_2 were given idential prior distributions and $\gamma = \theta = 5$ and $\nu = \omega = 20$.

Six runs of the Gibbs sampler were carried out using the Gamma/Exponential distribution in a similar manner to that of the previous section but with the following important difference :-

- The only deviation from the standard algorithm was the constraint that $E(C_1) > E(C_2)$. Blocking and R.U.G.S. were not used and their effect was compensated for by the use of highly informative prior distributions.

The starting values, burn-in and sample size for the runs are the same as those used in Section 7.

8.2 Run Outcomes

Posterior histograms relating to the first run of the Gibbs sampler (Run a) are shown in Figure 60.

Whilst it can be seen that these posterior distributions are not exactly normal, the results that follow do seem to indicate that the assumption is valid.

Quantities of interest for the six runs of the Gibbs sampler are tabulated in Figure 35 , with the prior variance of α_2 being denoted by $V_0(\alpha_2)$.

Inspection of the above table reveals that the effective number of parameters, p_D, is a real number and not an integer as might reasonably be expected. This is due to the "effectiveness" of three of the parameters being reduced by constraints

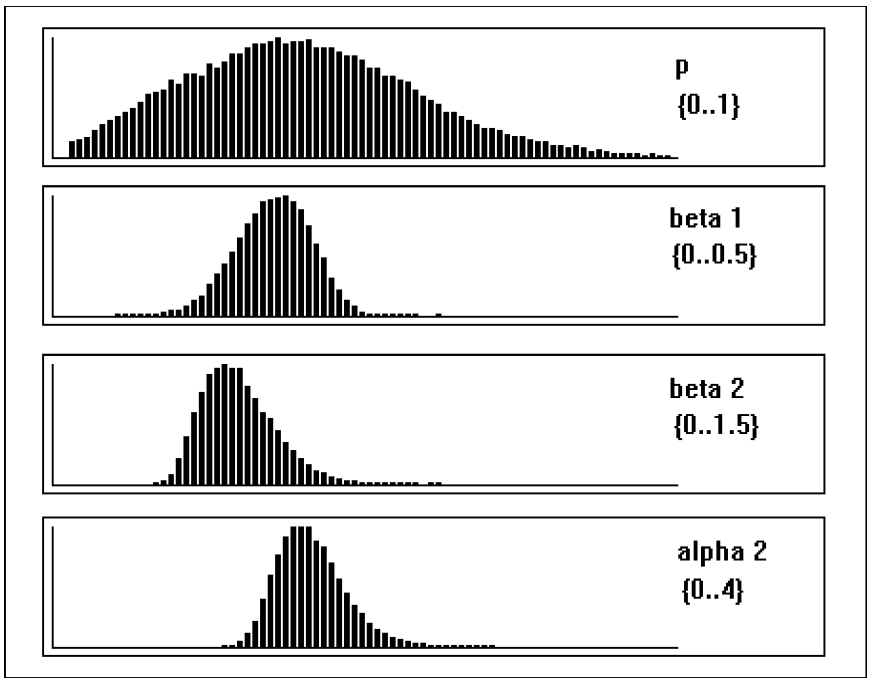


Figure 60: Posterior histograms for the Gamma Exponential distribution

Run	$V_0(\alpha_2)$	$\bar{D}_0$	p_D
a	3.0	1006.43	3.77
b	1.0	1005.99	3.68
c	0.375	1005.77	3.29
d	0.12	1006.91	2.73
e	0.0375	1008.18	2.44
f	0.03	1008.29	2.39

Table 35: Quantities of interest

imposed, mainly in the form of prior distributions, and so the effective number of parameters is less than the actual number. In this case there are four parameters, which we can denote by n_p and so it might be useful to define the *degree of influence* of Θ as being

$$d_i(\Theta) = \frac{p_D}{n_p}$$

Nevertheless, in Figures 61 and 62 that follow p_D will be used.

Figure 61 shows, perhaps unsurprisingly, that p_D decreases as does the prior variance of α_2. The relationship is not linear and p_D assymptotically approaches a value four although this value may never be reached due to the prior distributions

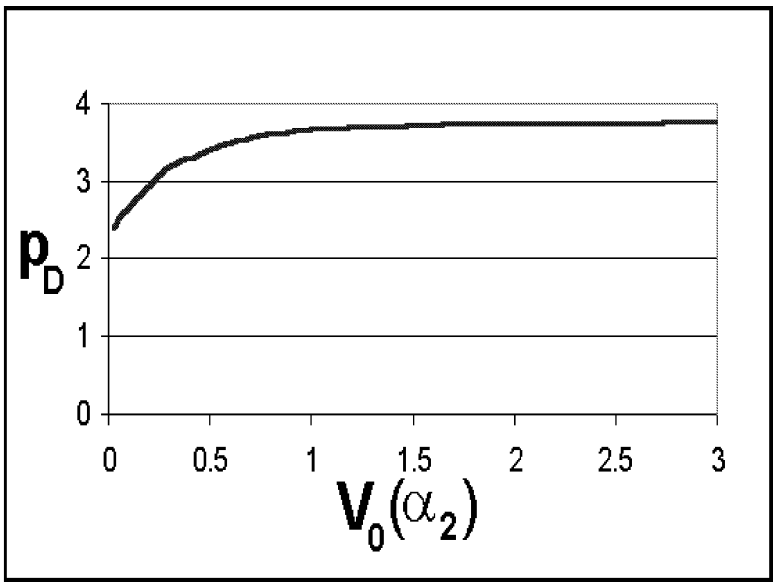


Figure 61: Graph of $V_0(\alpha_2)$ v p_D

given to β_1 and β_2. Whilst this graph contains, perhaps, few surprises the following plot of p_D against $\bar{D}$ is more revealing.

It is clear from Figure 62 that there exists a value of P_D for which there also exists a minimum value of $\bar{D}$. Since P_D is related to the prior variance of α_2, there is also a value of $V_0(\alpha_2)$ which minimises $\bar{D}$. Since this latter quantity is a measure of model fit it is clear that it is possible to constrain parameters beyond the point of optimum model fit.

8.3 Summary

In this section Bayesian deviance has been used to demonstrate a link between the effective number of parameters and model fit. It has been found that optimum model fit occurred when the effective number of parameters took a non-integer value below that of the actual number of model parameters. This appears to suggest that instead of using a model with fewer parameters, it would be better to use informative prior distributions and constrain the existing model which could be slightly over parameterised. One drawback of the technique is, however, that autocorrelation is stronger and the thinning factor needs to be greater for the sampler to pass the

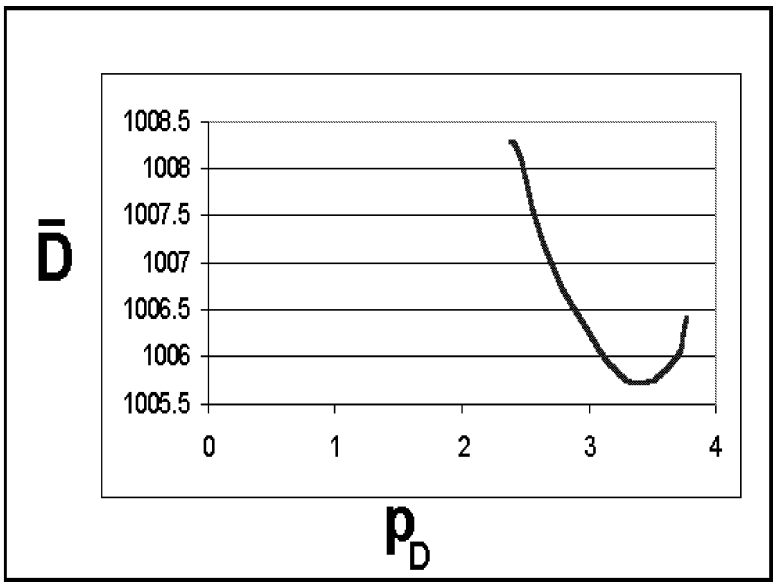


Figure 62: Graph of p_D v $\bar{D}$

convergence test.

In the next section another Bayesian method of checking model behaviour will be proposed and examined.

9 Bayesian model comparison using posterior predictive datasets and Mahalanobis distance

Introduction

"Checking the model is crucial to statistical analysis". This point, made by Carlin, Rubin, Gelman & Stern (1995) would find acceptance with statisticians of both persuasions but the controversy begins when the simple question "How?" is asked. Under the Bayesian paradigm the question "Is the model true or false?" is not asked. That is to say, no test is applied to the model such that, as a result of this test, the model is deemed to either fit the data or not. A well known example of this method is the χ^2 test where observed and expected values are compared and, as a result of the test, the model is either rejected or not.

Bayesian model checking, however, does not function in this absolute manner. Rather, models or formulations of models and prior distributions are compared with one another. A common example of this approach is the calculation of Bayes factors or Likelihood ratios which can be explained as follows.

Suppose we have two fully specified models, M_1 and M_2 with parameter vectors Θ_1 and Θ_2 respectively and a dataset, D. Now, let π_1 be the prior probability that model 1 is correct and let π_2 be similarly defined. We require the posterior probability that model 1 is correct given the dataset, i.e. $p(M_1|D)$. By Bayes Theorem :-

$$p(M_1|D) = \frac{p(D|M_1).p(M_1)}{p(D)}$$

where $p(D|M_1)$ is the likelihood under model 1 which will be denoted by L_1, $p(M_1)$ is the prior probability of model 1, i.e. π_1 and $p(D)$ is, by the theorem of total probability, $p(D|M_1).p(M_1) + p(D|M_2).p(M_2)$ which can be reduced to $\pi_1 L_1 + \pi_2 L_2$. If we now let the posterior probability of model 1 be denoted by p_1, the following

can be written :-

$$p_1 = \frac{\pi_1 L_1}{\pi_1 L_1 + \pi_2 L_2}$$

By similar reasoning the following relationship is also true :-

$$p_2 = \frac{\pi_2 L_2}{\pi_1 L_1 + \pi_2 L_2}$$

If we now define the posterior odds in favour of model 1 as the ratio of p_1 to p_2 then, after some algebra, can be written :-

$$\frac{p_1}{p_2} = \frac{\pi_1}{\pi_2} \times \frac{L_1}{L_2}$$

The term $\frac{\pi_1}{\pi_2}$ is known as the prior odds (in favour of model 1) and $\frac{L_1}{L_2}$ is referred to as the Likelihood ratio or Bayes factor.

The above explanation presupposes that the two models have already been estimated. It is possible, however, to incorporate the calculation of Bayes factors into Gibbs sampling but at this point computational difficulties arise.

Further descriptions of Bayes factor methodology are given by Kass & Raftery (1995), Gelfand (1996) and Raftery (1996). In this project an alternative method is proposed which fully exploits the posterior predictive distribution resulting from the use of Gibbs sampling and is, therefore, fully Bayesian. The proposed method is also computationally straightforward.

9.1 Background

The proposed method is for use where there is no alternative model depends upon two points which are

1. The Gibbs sampler generates **all** credible parameter vectors for the model, given the data. That is each $f(t|\Theta_i, t_1, t_2, \ldots, t_j)$ where, in this case, $i \in \{1..50,000\}$ is a credible pd.f. given the data.

2. The data is a sample, of size j, from a p.d.f. with unknown parameter vector Θ^*.

Suppose, for each $i \in \{1..50,000\}$, a sample of size j were drawn from $f(t|\Theta_i, t_1, t_2, \ldots, t_j)$ it would then be possible to compare features of the data with the same features of the samples. These samples, known as posterior predictive datasets, represent all credible samples, of size j, that can be drawn from the model given the data. This comparison would, in effect, show the extent to which the data differs from all those samples that the model is capable of generating and so would be a measure of model fit. This approach is highly flexible since any feature of the data can be chosen for comparison depending on the wishes of the modeller. For example, comparing extreme values would demonstrate the models ability to accommodate outlying values. Alternatively, specific intervals could be compared.

9.2 Implementation

In this project the following features will be used for comparison

1. The mean
2. The standard deviation
3. The coefficient of skew
4. The coefficient of pointedness or kurtosis

where the skew is defined by

$$E\left(\frac{x-\mu}{\sigma}\right)^3$$

and the kurtosis by

$$E\left(\frac{x-\mu}{\sigma}\right)^4$$

using standard notation.

The first four moments are used as features of comparison since this approach provides a general method that could be used in any modelling situation.

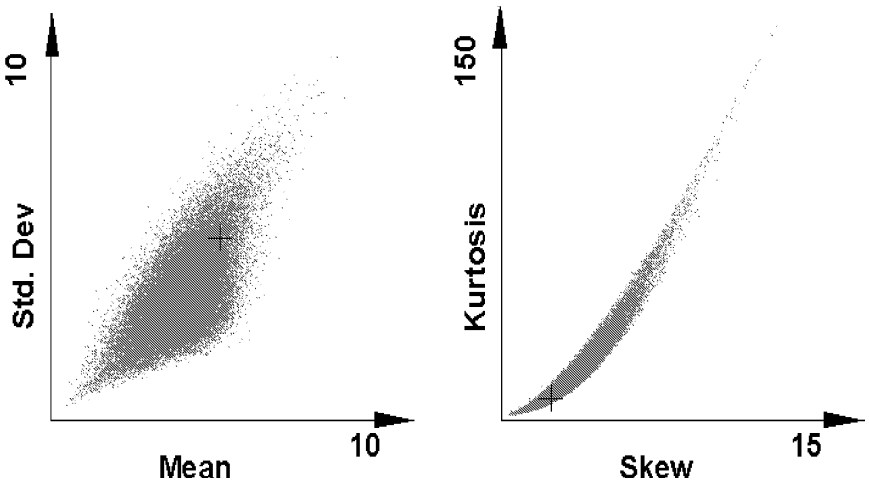


Figure 63: Posterior predictive datasets for Run 1

To illustrate this technique the actual dataset will be compared to posterior predictive datasets resulting from two separate runs of the Gibbs sampler. Details of the two runs are shown in Table 36.

Run	1	2
Model	Gamma Exponential	Gamma Exponential
Constraint	$E(C_1) > E(C_2)$	$E(C_1) > E(C_2)$
Prior type	Mildly informative	Highly informative
Blocking	Second component only	Not used
R.U.G.S.	First component only	Not used
Data file	File 2	File 2

Table 36: Details of Runs 1 and 2

Run 1 is detailed in Section 7 and Run 2 is referred to as Run C in Section 8.

For each run, the dataset mean, standard deviation, skew and kurtosis are compared to those features of the 50,000 posterior predictive datasets initially by means of two graphs. One graph is a plot of mean versus standard deviation for each posterior predictive dataset and the other is the corresponding plot of skew versus kurtosis.

Figure 63 refers to Run 1. The grey area is made up of 50,000 data points from the posterior predictive datasets and the black cross represents the data point of the actual dataset. It can be seen that the actual dataset point falls, on each graph,

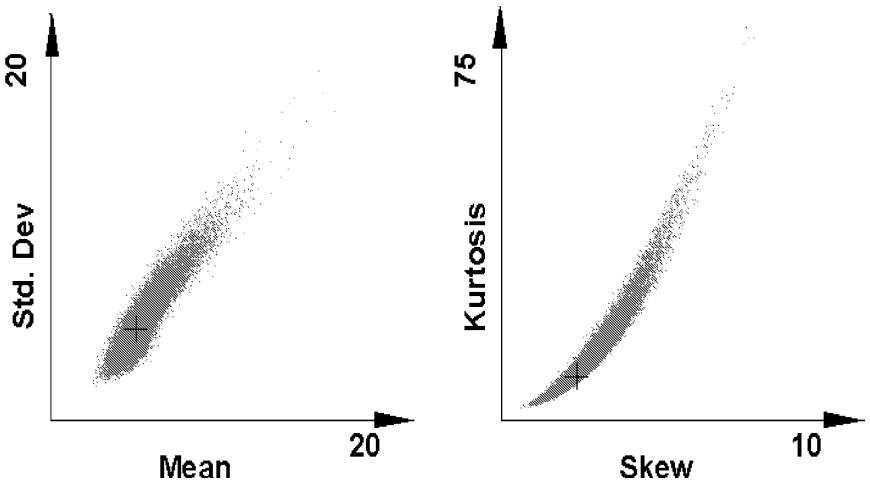


Figure 64: Posterior predictive datasets for Run 2

within the limits of the density formed by the posterior predictive datasets. This seems to indicate that whilst the model fit is acceptable, it should be possible to improve it.

Figure 64 refers to Run 2.

Here, it is clearly visible that the actual dataset point is closer to the centre of the probability density in the case of the mean versus standard deviation graph. There appears to be little, if any, difference in the skew versus kurtosis plot. From this the inference could be drawn that Run 2 produced better model fit than Run 1. Having compared the two runs by examining graphs the obvious question to ask would be *" Is there a numerical method by which we can quantify the differences that have been observed?"*

The answer is that the distance from the actual dataset point to the mean of the joint posterior probability density of the mean, standard deviation, skew and kurtosis of the posterior predictive datasets can be measured. This is a distance in four dimensional space but the covariances and variances of the parameters are taken into account by means of their covariance matrix and so the space in non-Euclidian. What is, in fact, measured is the square of this distance which is called Mahalanobis distance, D^2 (Mahalanobis, 1936).

9.3 Mahalanobis distance (D^2)

Notation

The following notation will be used in this section :-

Let S_i be a sample of size j drawn from $f(t|\Theta_i, t_1, t_2, \ldots, t_j)$ where $f(t|\Theta, t_1, t_2, \ldots, t_j)$ is the Gamma Exponential Distribution and $i \in \{1..50,000\}$. Further notation is shown in Table 37.

m_0 = data mean
d_0 = data standard deviation
s_0 = data skew
k_0 = data kurtosis
m_i = the mean of the i^{th} sample
$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}$, i.e. the mean of all the m_i's.
d_i = the standard deviation of the i^{th} sample
$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$, i.e. the mean of all the d_i's.
s_i = the skew of the i^{th} sample
$\bar{s} = \frac{\sum_{i=1}^{n} s_i}{n}$, i.e. the mean of all the s_i's.
k_i = the kurtosis of the i^{th} sample
$\bar{k} = \frac{\sum_{i=1}^{n} k_i}{n}$, i.e. the mean of all the k_i's.

Table 37: Further notation

where, in this case, n = 50,000.

Mahalanobis distance, D^2, is defined by

$$D^2 = X^T V^{-1} X \tag{20}$$

where

$$X = \begin{bmatrix} m_0 - \bar{m} \\ d_0 - \bar{d} \\ s_0 - \bar{s} \\ k_0 - \bar{k} \end{bmatrix}$$

and V^{-1} is the inverse of the covariance matrix.

For each of the two runs, Mahalanobis distance was computed and the results are shown in Table 38.

Run	1	2
D^2	3.3975	0.53260

Table 38: The value of D^2 for Runs 1 & 2

A much smaller value of D^2 in the case of Run 2 confirms what was suggested by comparison of the graphs.

9.4 Two important caveats

9.4.1 The use of D^2 in isolation

One particular run of the Gibbs sampler gave a value of 1.7373 for D^2. This, taken on its own would place the run between Runs 1 and 2 in terms of model fit. However, the run in question was the base run for the G.E.D. model which, as shown in Section 7, suffered from identifiability problems which manifested themselves in the output of the parameter marginal posterior distributions which "jumped state" several times during the course of sampling. This gave rise to bimodal distributions and, if we examine the mean / standard deviation / skew / kurtosis graphs for this run, bimodality can again be observed.

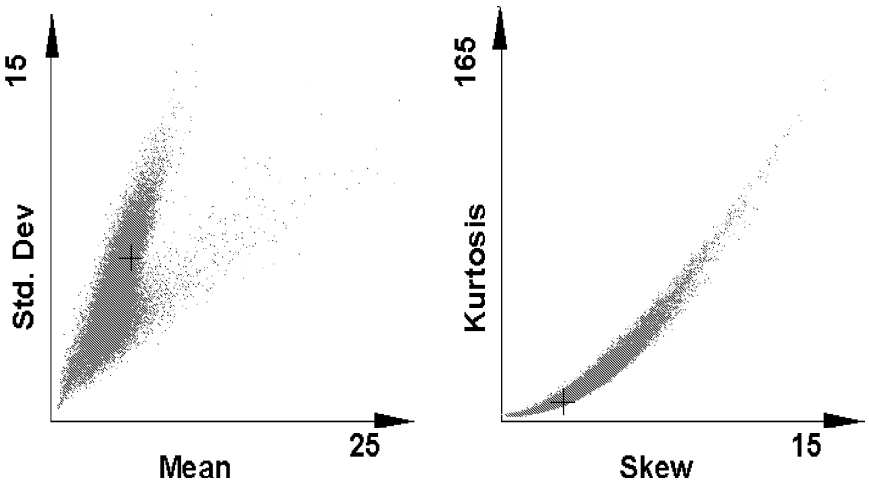


Figure 65: Posterior predictive datasets where bimodality is present

It is, therefore, recommended that this method should not be used as a single tool for model analysis without reference to other methods. It can, however, be used to compare competing modelling stratagies which are known to meet basic criteria such as those set out in Section 7.

9.4.2 The sampling distribution of D^2

It should be remembered that the χ^2 distribution of D^2 is dependent on the variates involved having normal distributions. This is not always the case as can be seen from Figure 66 which shows a histogram of the skews of 50,000 posterior predictive datasets.

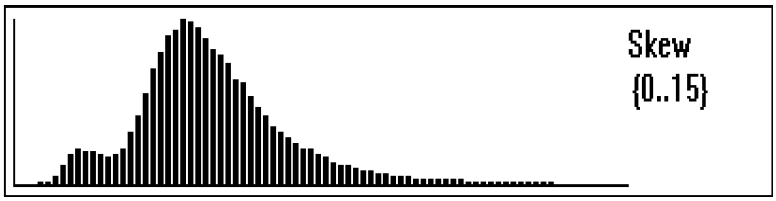


Figure 66: An example of a non-normal marginal posterior distribution

If, however, the distributions had been normal in this case it would have been appropriate to use the χ^2 distribution with 4 degrees of freedom as a test. The upper critical value at 95% is 9.488 so it can be seen that the values of Mahalanobis distance tabulated in Table 38 fall within this range.

9.5 Summary

It has been shown that posterior predictive datasets can be useful for model comparison in two ways. Firstly, they can be used to construct graphs that give a clear visual indication of model performance. Secondly they can provide, via Mahalanobis distance, a numerical measure of model behaviour that is relatively straightforward to implement and is not, to use the words of Wlodzimierz Bryc (1995) a "numerical nuisance". It must be remembered, however, that interpretation of the value of D^2 is only only approximate due to the non-normality of the posterior distribution.

10 Conclusion

10.1 Research questions

It was stated, in the Introduction to this thesis, that the primary area of interest was the application of the Bayesian paradigm to vehicle headway modelling. This has been a success, the extent of which can be judged from the answers to the research questions which are set out below:-

10.1.1 The Bayesian paradigm

The first research question asked in Section 1 was *" Can we usefully apply the Bayesian paradigm to inferences about these models?"* Here, the object was to determine not only if the Bayesian paradigm could be applied but also if there was any advantage to this approach. It is now clear that these models can be made the subject of Bayesian inference and that there are definite advantages to be gained by using this methodology. Among these are

- Explicit use of prior belief
- Proper handling of uncertainty
- Usefulness of posterior predictive datasets

One interesting advantage afforded by the methods used is what might be called *"transparency"*. The nature of Gibbs sampling is such that, for example, actual output traces of the model parameters can be plotted and examined as can marginal posterior distribution histograms. It is also possible and, in fact, necessary to plot the output of one parameter against another to look for correlation or against itself (with lag) to detect autocorrelation. In effect, every step of the process of analysis is open to scrutiny with the only limit being the desire of the individual statistician. This property will be utilised in one of the suggestions for further research.

10.1.2 Model suitability

Secondly, the question was asked *"Are these models appropriate?"* or, more accurately, *"Are these models appropriate for the datasets considered in this thesis?"* In the case of the two models found in the literature the answer is *"No"*. The double displaced exponential distribution, as proposed and analysed by Griffiths & Hunt (1991), is only successful when observed headways less than half a second are ignored, with their presence being attributed to *"trigger happy"* observers. Under the Bayesian paradigm model fit appears to be a problem for which no remedy can be found, the most likely reason for this being incorrect model choice given the data.

Whilst Salter (1974) has achieved some success with the double exponential headway distribution model, under the Bayesian paradigm the behaviour of the shift parameter, k, is highly problematic in that its marginal posterior distribution was always multimodal. Since the property of having a unimodal marginal posterior distribution was laid down as one of the criteria for a successful run of the Gibbs sampler this model cannot be used. This criteria would also exclude any other model that contained a discontinutity.

The Gamma Exponential model, proposed by the author, has been shown to be useful for modelling headways. The usual difficulties associated with Gibbs sampling are encountered (see 10.1.3 below) but can be overcome.

10.1.3 Problems associated with the Bayesian paradigm

The third research question was *" What problems arise when the Bayesian paradigm is applied ?"* and in Section 2 it was shown how computational difficulties can arise in Bayesian statistics. Having chosen Gibbs sampling as the computational algorithm, there are still difficulties to overcome. The problem of identifiability, ever present in the case of mixture models, can be overcome by placing constraints on the model. These constraints form part of the prior beliefs concerning the model and its parameters.

The issues of correlation and convergence are far from independent of each other. Whilst mixture models are notoriously slow in mixing, i.e. the sampler is slow to move thoroughout the support of the posterior distribution, autocorrelation has the effect of misleading the experimenter into believeing that output chains have not converged when, in fact, the opposite is sometimes true. The technique of thinning counters autocorrelation and has the advantage of great simplicity over reparameterisation methods which, by means of some appropriate algebraic transform, change the shape of the posterior distribution.

10.1.4 The Bayesian paradigm and highway engineers

The final question, *" Is the routine use of these models feasible in highway engineering?"* has only been answered in part by this thesis. The gamma exponential model has been shown to be a suitable model. It fits the data, is flexible, and a reliable method of fitting has been developed. There still remains, however, work to be done in terms of producing a methodology or modelling algorithm that could be conveniently used by highway engineers.

10.2 Further research

There are four areas which, if taken further, would benefit mixture modelling in general and headway modelling in particular.

1. Model refinement
2. A methodology for highway engineers
3. Further work on mixture models
4. A convergence / correlation analysis tool

10.2.1 Model refinement

There are two areas for consideration here :-

1. A model for single carriageway use

2. The non-independence of headways

The model proposed by Griffiths & Hunt has been found to be unsuitable for modelling headways on single carriageway roads. Since this model was the only one examined for use in this way a suitable alternative must be sought. One important point to be considered in proposing an alternative is the fact that very short headways can be very rare in single streams of traffic, e.g. when only one direction of flow is being considered on a single carriageway road. Having observed the behaviour of the shift parameter, k, in the double displaced headway model it is not considered appropriate to use any mixture distribution that has such a shift parameter. This points towards the use of either a single component model or a two component model where one component is capable of handling these very short headways.

The work done in this thesis has been carried out under the assumption that headways are independent of each other. However, the more congestion present in a stream of traffic the more this assumption could be called into question. One avenue of research that should be followed is to determine the suitablity of using a Hidden Markov Model. Here a Gamma / Exponential distribution is used as before but the modelling process is complicated the fact the allocation of a particular observation, t_j, at a given sweep of the Gibbs sampler, say i, is dependendent on the allocation of the previous observation, t_{j-1} and on the next one, t_{j+1}, which will have been allocated to a component at the previous sweep. This seems, at first, to be counterintuitive but if an example is considered from the investigation of a genetic disorder, say haemophilia, the apparent problem can be resolved.

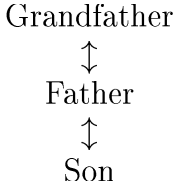


Figure 67: Direction of inferences in a genetic investigation

Suppose we wish to make inferences concerning the individual labelled "Father" in Figure 67. If data were available for the Grandfather, then we would be able to draw inferences concerning the Father. The same is true if data were available concerning only the Son. If, however, data were available concerning both Grandfather and Son, then inferences concerning the Father would be more robust since there would be more information upon which to base them. The same principle holds when we come to allocate the observation t_j and we can redraw Figure 67 to illustrate this.

$$t_{j-1} \updownarrow t_j \updownarrow t_{j+1}$$

Figure 68: Direction of inferences in observation allocation

Given the reasoning above and by reference to Figure 68 it can be seen that, if there is data concerning the allocation of observations t_{j-1} and t_{j+1}, then inferences concerning the allocation of t_j will be more accurate. In practice, this gives rise to much more complicated calculation of allocation probabilities and there is also a transition matrix to be dealt with. All of this means that the application of a Hidden Markov Model is a considerable project in itself.

10.2.2 A methodolgy for highway engineers

Given that a suitable model can be found for use on single carriageway roads, the modelling process still needs to be modified in such a way that it is readily accessible to highway engineers. In practice this means that certain areas will need to be "hidden", .i.e. it would not be required of the highway engineer to specify prior distributions or their parameter values from scratch without help. This could be done indirectly by asking the highway engineer to specify likely rates of traffic flow. The same principle applies to convergence diagnosis and it is envisaged that routines for the assessment of convergence could be appended onto the Gibbs sampling software.

That is to say, convergence diagnosis could be automated with appropriate messages being displayed if problens were encountered. There is clearly much to be done in this area, mainly in the field of elicitation, but its important cannot overemphasised.

10.2.3 Further work on mixture models

An important point to come out of this project is the crucial role of the allocation step in the Gibbs sampling algorithm. One possible area of research would be to look closer at this step and, possibly, see if there is any benefit from "tracking" an observation (or group of observations) to monitor which component it is allocated to as sampling progresses. From a computational point of view this would not be difficult and it may be possible to numerically investigate the effect of the well documented "trapping states" that occur. This could lead to improved convergence diagnostics.

10.2.4 A convergence / correlation analysis tool

As the work involved with this thesis progressed it became necessary to write various computer programs. The first requirement was for a Gibbs sampler. It then became clear that a convergence diagnostic was required and after that the need for a thinning program emerged. All these programs proved invaluable but they were all written separately as the need arose. It would be advantageous, to practitioners of Gibbs sampling, if a program were available that could readily assess the convergence and correlation properties of output chains both numerically and graphically. Such a program could be written without the addition any new routines. The necessary procedures and functions already exist but in different programs. The task would be more of a collation exercise than writing a completely new program. The program would have two main advantages over currently available diagnostics :-

1. It would be a stand-alone program. There would be no need use the program in conjunction with any other program or package.

2. The convergence diagnostic can be used on any form of marginal posterior distribution.

Although much has been achieved in this project there remains enormous scope for further research and it would be appropriate to end by quoting Thompson (1995) who said *"Research never ends"*.

11 References

Adams, W.F.. (1936) Road traffic Considered as a Random Series. *The Journal of the Institution of Civil Engineers.* **Nov 1936** pp.121 - 132

Ashton, W. (1971) Distributions for gaps in Road Traffic. *Journal of the Institution of Mathematics Applications.* **7** pp.37 - 46

Barnard, G.A. (1958) Studies in the History of Probability and Statistics. *Biometrika.* **Volume 45, Parts 3 and 4** pp.293 - 315

Baras, J.S., Dorsey, A.J., Levine, W.S. (1979) Estimation of traffic platoon structure from headway statistics. *IEEE Transactions on Automatic Control.* **AC - 24(4)** pp.553 - 559

Berger, J.O. (2000) Bayesian Analysis : A Look at Today and Thoughts of Tomorrow. *Journal of the American Staistical Society.* **95** pp.1269 - 1276

Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M. (Eds.) (1998), *Invited Papers* The Sixth Valencia International Meeting on Bayesian Statistics, Alcossebre, Spain.

Brooks, S. (1998) Markov chain Monte Carlo method and its application. *The Statistician.* **Part 1** pp.69 - 100

Bryc, W. http:math.uc.edu/~brycw
The term "numerical nuisance" was used by Professor Bryc on a web page that contained lecture notes. Unfortunately, the page has has been removed and the reference given is that of Professor Bryc's current home page. The page referred to was accessed in September 2002.

Calabria, R., Pulcini, G. (1994) Bayes Credibility Intervals for the Left Truncated Exponential Distribution. *Microelectronics and Reliability.* **Vol. 34, No. 12** pp.1897 - 1907

Copas, J.B., Heydari, F. (1997) Estimating the Risk of Reoffending by using Exponential Mixture Models. *Journal of the Royal Statistical Society, Series A.* **Vol. 160, Part 2** pp.237 - 252

Dey, D.K., Kuo, L., Sahu, S.K. (1995) A Bayesian predictive approach to determining the number of components in a mixture distribution. *Statistics and Computing.* **Vol. 5** pp.297 - 305

Diebolt, J., Robert, C.P. (1994) Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society, Series B.* **Vol. 56, No. 2** pp.363 - 376

Gelman, A. (1996) Inference and monitoring convergence. *Markov chain Monte Carlo in practice (eds, Gilks, W., Richardson, S., Spiegelhalter, D.J.).* **London : Chapman & Hall** pp.131 - 144

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995) Bayesian Data Analysis. *London.* **Chapman and Hall** pp.161 - 183

Geman, S., Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Trans. Patt. Anal. Mach. Intel..* **6** pp.721 - 741

Gilks, W. (1996) Full conditional distributions. *Markov chain Monte Carlo in practice (eds. Gilks, W., Richardson, S., Spiegelhalter, D.J).* **London : Chapman & Hall** pp.77 - 88

Gilks, W., Roberts, G.O. (1996) Strategies for improving McMC. *Markov chain Monte Chain in practice. (eds. Gilks, W., Richardson, S., Spiegelhalter, D.J.).* **London : Chapman & Hall** pp.89 - 144

Gilks, W.R., Wild, P. (1992) Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics.* **41, No.2** pp.337 - 348

Griffiths, J.D., Hunt, J.G. (1991) Vehicle Headways in Urban Areas. *Traffic Engineering and Control.* **Oct. 1991** pp.458 - 462

Gullberg, J. (1997) Mathematics from the Birth of Numbers. *London, W.W. Norton & Company Ltd.* pp.963-967

Hammersley, J.M., Handscomb, D.C. (1964) Monte Carlo Methods. *London, Methuen & Company Ltd..* pp.25 - 42

Hastings, W.K.. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrica.* **57, 1** pp.97 - 109

Kalos, M.H., Whitlock, P.A. (1986) Monte Carlo Methods. Volume 1 : Basics. *Chichester, John Wiley & Sons.* pp.39 - 88

Kass, R., Raftery, A. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass..* **90** pp. 773 - 795

Katti, B.K., Pathak, R.H. (1986) A study on headway distribution models for the urban road sections under mixed traffic condition. *Highway Research Bulletin (New Dehli).* **26** pp.1 - 39

Kloej, T., Van Dijk, H.K. (1978) Bayesian Estimates of Equation System Parameters : An Application of Integration by Monte Carlo. *Econometrica.* **Volume 6, No. 1** pp.1 - 19

Krishnan, M.S. Ramaswamy, V., Meyer, M.C. Damien, P. (1999) Customer Satisfaction for Financial Services : The Role of Products, Services and Information Technology. *Management Science.* **Vol. 45, No. 9** pp.1194 - 1209

Lau, E.W., Pathamanathan, R.K., André, N.G., Cooper, J., Shekan, J.D., Griffith, M.J . (2000) The Bayesian approach improves the electrographic diagnosis of broad complex tachycardia. *Pacing and clinical electrophysiology.* **23, 10** pp.1518 - 1526

Madi, T.M., Leonard, T. (1996) Bayesian estimation for shifted exponential

distributions. *Journal of Statistical Planning and Inference.* **55** pp.345 - 351

Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proc. Natl. Institute of Science of India.* **12** pp.49 - 55

Mengersen, K.L., Robert, C.P., Guihenneuc-Joyaux, C. (1998) McMC Convergence Diagnostics : A revieWWW. *Bayesian Statitsics 6. (eds. Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M..* pp.399 - 432

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics.* **11, 6** pp.1087 - 1092

Miller, A.J. (1961) A Queuing Model for Road Traffic Flow. *Journal of the Royal Statistical Society, Series B.* **23** pp.64 - 90

Norton, P. (1987) Inside the IBM PC. . **USA : Prentice Hall** pp. 302

O'Hagan, A. (1987) Monte Carlo is fundamentally unsound. *The Statistician.* **36** pp.247 - 249

Ohno, K., Mine, H. (1972) Traffic light queues with dependent arrivals as a generalisation to queuing theory. *Journal of applied probability.* **Vol. 19, No. 3** pp.630 - 641

Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. London.* **A 185** pp.71 - 110

Open University, The. (1984) Unit 1 : Chance. *M245 : Probability and Statistics.* pp. 20 - 32 (This booklet forms part of an undergraduate course in Statistics)

Polus, A. (1979) An analysis of the headway distribution of an urban bus service. *Traffic Engineering and Control.* **Aug/Sept. 1979** pp.419 - 421

Raftery, A.E., Lewis, S.M. (1996) Implementing McMC. *Markov chain Monte Carlo in practice. (eds. Gilks, W., Richardson, S., Spiegelhalter, D.J.).* **London :**

Chapman & Hall pp.115 - 130

Richardson, S., Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B.* **59** pp.731 - 792

Robert, C.P. (1996) Mixtures of distributions : inference and estimation. *Markov chain Monte Carlo in practice. (eds. Gilks, W., Richardson, S., Spiegelhalter, D.J).* **London : Chapman & Hall** pp.441 - 464

Robert, C.P., Mengersen, K. L. (1999) Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms. *Computational Statistics and Data Analysis.* **29** pp.325 - 343

Robert, C.P., Ryden, T., Titterington, D.M. (1999) Convergence controls for McMC algorithms with application to hidden Markov models. *J. Statist. Comput. Simul..* **Vol. 64** pp.327 - 355

Roberts, G.O., Sahu, S.K. (1997) Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *J. Roy. Statist. Soc..* **B 59** pp.291 - 317

Salter, R.J. (1974) Highway Traffic Analysis and Design. *London, Macmillan.* pp.107 - 124

Schuhl, A. (1955) The probability theory applied to distribution of vehicles on two-lane highways. *Poisson and Traffic.* **The Eno Foundation** pp.59 - 75

Smith, A.F.M., Roberts, G.O. (1993) Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B.* **1** pp.3 - 33

Spigelhater, D.J., Best, N.G. & Carlin, B. (1998) Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *McMC*

Pre-print service, Bristol University.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. (2002) Bayesian measures of complexity and model fit. *J.R.Statistic. Soc.* **B, 64, Part 4** pp. 583 - 639

Spiegelhalter, D.J., Thomas, A., Best, N.G. & Gilks, W. R. (1994) BUGS : Bayesian inference using Gibbs sampling, version 0.30. *Cambridge : Medical Research Council, Biostatistics Unit.* pp.1 - 59

Stephens, M. (2001), Private Communication

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B.* **4** pp.795 - 809

Stewart, L.T., Johnson, J.D. (1971) An example of the use of Monte Carlo integration in Bayesian decision problems. *10th Annual Reliability and Maintainability Conference, Anaheim, CA.* pp.19 - 23

Swartz, Tim., Evans, Michael. (1995) Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration problems. *Statistical Science.* **Vol. 10, No. 3** pp.254-272

Tamura., Y., Chishaki, T. (1983) Time Headway Distribution Model Based on the Composition of Free and Constrained Flowing Vehicles. *Transaction of the Japanese Society of Civil Engineers.* **15** pp.448 - 451

Thompson, J. (1995), Private Communication

Thompson, T.J., Smith, P.J., Boyle, J.P. (1998) Finite mixture models with concomitant information : assessing diagnostic crieria for diabetes. *Applied Statistics.* **Vol. 47, Part 3** pp.393 - 404

Titterinton, D.M., Smith, A.F.M., & Makov, U.E. (1985) Statistical analysis of finite mixture distributions. . **New York : Wiley** pp.various

Trader, R.L. (1985) Bayesian inference for truncated exponential distributions. *Communications in Statistics : Theory and Methods.* **14(3)** pp.585 - 592

Troutbeck, R.J. (1986) Average Delay at an Unsignalised Intersection with Two major Streams Each Having a Dichotomised Headway Distribution. *Transportation Science.* **Vol. 20, No. 4** pp.272 - 286

Troutbeck, R.J., Kako, S. (1999) Limited priority merge at unsignallised intersections. *Transportation Research (Part A : Policy & Research).* **Vol. 33A, No. 3/4** pp.291 - 304

Vounatsou, P., Smith, T., Smith, A.F.M. (1998) Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Applied Statistician.* **Vol. 47, Part 4** pp.575 - 587

Wasielewski, P. (1979) Car following headways on freeways interpreted by the semi-Poisson headway distribution model. *Transportation Science.* **13, 1 Feb 79** pp.36-55

12 Appendix

12.1 1 Source code listing : headw1.c : data collection program

The program listed below was used to collected headway data and is referred to in Section 6.

```
#include "c:\pcc\headers\stdio.h"

char data, *buffer;
long ticks;
int n;
FILE *fp;

int main()
{
   n = 0;
   scr_setup();
   puts("Enter name of data file to be created >>>");
   gets(buffer);
   fp = fopen(buffer, "w");

   while( (data = scr_csts()) != 32)
      ;

   puts("Timing has started\n");
   while( (data = scr_csts()) != 17)
   {
      tone(0, 1);
```

```
      ticks++;
      if ( data == 32)
      {
         n++;
         printf("%6.2f secs : total headways read %d\n", ticks/18.5, n);
         fprintf(fp, "%6.2f\n", ticks/18.5);
         ticks = 0;
      }

   }
   fclose(fp);
   exit(0);
}
```

12.2 Source code listing : A typical sampling routine

The Pascal function listed below returns a single value from the Beta distribution with parameters α and β.

```
function betasim(alpha, beta: extended):Real;
{************************************************************************}
{This function returns one value simulated from Beta(alpha, beta,)}
{************************************************************************}

type
    ordinates = record
        x     : extended;
        y     : extended;
        grad  : extended;
    end;
```

```
type
    envelopes = record
        x1    : extended;
        x2    : extended;
        m     : extended;
        c     : extended;
        area  : extended;
    cumarea   : extended;
    end;
var
   totarea, I : extended;
   p, q, r, s, t, u : extended;
   w1, w2, w3, xrexp, k : extended;
   a, b, c : extended;
   ordarray : array[1..3] of ordinates;
   envarray : array[1..3] of envelopes;
   l_accept : Boolean;
   counter1, counter2, x, y : integer;

begin {The simulation}
{initialise & check where necessary, variables}

   a := alpha;
   b := beta;

   c := min(lptbeta(0.999, a, b, 0.0), lptbeta(1.0/100.0, a, b, 0.0));
   c := abs(c)  ;
```

```
totarea := 0.0;
ordarray[1].x := 0.5*(a - 1.0)/(a + b - 2.0);
ordarray[2].x := (a - 1.0)/(a + b - 2.0);
ordarray[3].x := 0.5*(1 + (a - 1.0)/(a + b - 2.0));

{initialise remainder of ORDARRAY}
 for counter2 := 1 to 3 do
 begin
    ordarray[counter2].y := lptbeta(ordarray[counter2].x,
                                    a, b, c);
    ordarray[counter2].grad := grlptbet(ordarray[counter2].x,
                                a, b);
end;
ordarray[2].grad := 0.0;
{ORDARRAY initialised}
{initialise ENVARRAY}
envarray[1].x1 := 0.0;
envarray[3].x2 := 1.0;

for counter2 := 2 to 3 do
begin
   p := ordarray[counter2].y;
   q := ordarray[counter2 - 1].y;
   r := ordarray[counter2].x;
   s := ordarray[counter2].grad;
   t := ordarray[counter2 - 1].x;
   u := ordarray[counter2 - 1].grad;
   envarray[counter2 - 1].x2 := ((p-q) - (r*s) + (t*u))/(u - s);
```

```
    envarray[counter2].x1 := envarray[counter2 - 1].x2;
 end;

 for counter2 := 1 to 3 do
 begin
       envarray[counter2].m := ordarray[counter2].grad;
       envarray[counter2].c := ordarray[counter2].y
                             - ordarray[counter2].x
                             *ordarray[counter2].grad;

    envarray[counter2].area := envarea(envarray[counter2].x1,
                               envarray[counter2].x2,
                               envarray[counter2].m,
                               envarray[counter2].c);

    if counter2 < 2 then
       envarray[counter2].cumarea := envarray[counter2].area
    else
       envarray[counter2].cumarea := envarray[counter2].area
                                   + envarray[counter2 - 1].cumarea;
end;

totarea := envarray[3].cumarea;
l_accept := False;
while not l_accept do
begin {while not l_accept}
   w1 := random;
   w2 := w1;
```

```
 w1 := w1*totarea;
 counter2 := 1;
 while envarray[counter2].cumarea < w1 do
      counter2 := counter2 + 1;

 if counter2 = 1 then
    I := w1
 else
    I := w1 - envarray[counter2 - 1].cumarea;

if envarray[counter2].m = 0.0 then
begin
   p := exp(envarray[counter2].c);
   q := envarray[counter2].x1;
   xrexp := (I/p)+q;
end
else
begin
   p := exp(envarray[counter2].c);
   q := envarray[counter2].m;
   r := envarray[counter2].x1;
   s := r*q;
   t := exp(s);
   xrexp := (1.0/q)*ln((q*I/p) + t);
end;

p := lptbeta(xrexp,a, b, c);
q := envarray[counter2].m;
```

```
    r := envarray[counter2].c;
    s := q*xrexp + r;
    t := exp(p - s);

    w3 := random;
    if w3 <= t then
    begin
        l_accept := true;
        betasim := xrexp;
    end;
    end; {while not l_accept}

end; {The betasim function}
```

The function was used for sampling a value from the marginal posterior distribution of the weighting parameter, p, a typical example of which can be found in Section 5.5.1. The following line of Pascal source code shows how the function is called :-

```
psimv := betasim((n1 + phi), (n2 + psi));
```

12.3 Additional runs of the Gibbs sampler

Those runs of the Gibbs sampler that are documented in Section 7 are obviously considered essential to the main objective of this thesis. Other runs were, however, carried out and some are documented here. Because of what might be termed their "secondary" importance the details presented here are only intended to give an outline of model performance under the given conditions.

12.3.1 Making the Griffiths & Hunt model "work"

This run of the sampler was exactly the same as the final run for this model as in Section 7 but in this case every observation less than 2 seconds was deleted from the

data file used. Also, more informative priors for β_1 and β_1 were used. The purpose of this was to mimic the strategy of Griffiths & Hunt and see if a better model fit could be achieved. Figure 69 shows the marginal posterior distributions for the model parameters for this run.

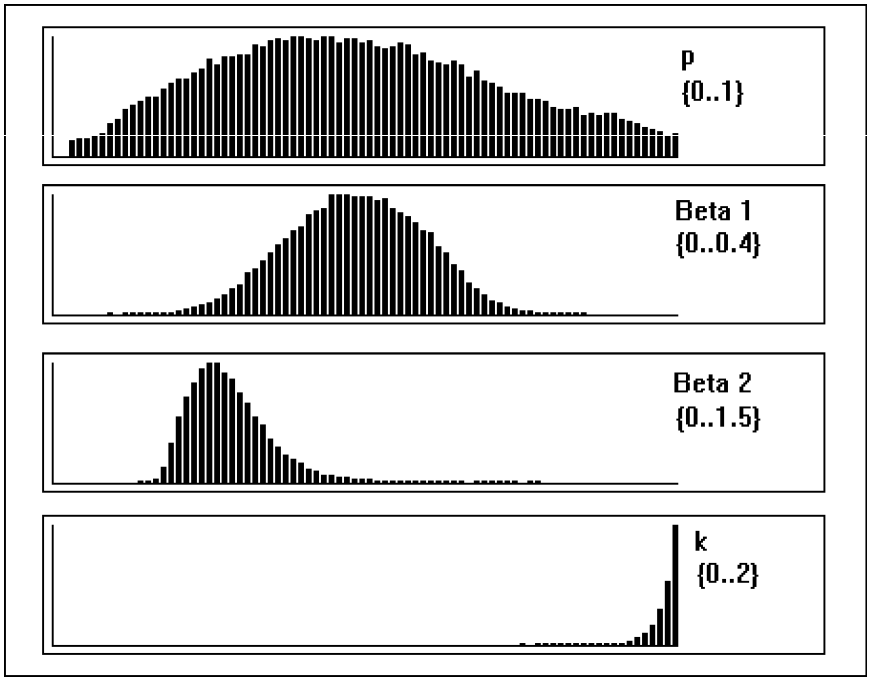


Figure 69: marginal posterior distributions for the Griffths & Hunt model

From Figure 69 it can be observed that all marginal posterior distributions are unimodal. However, the use of informative priors merely served to increase autocorrelation with a corresponding reduction in the rate of mixing. Model fit, though, was noticeably improved. This is shown in Figure 70 although it must be pointed out that there is still some way to go before the model/data fit could be described as good. It is, therefore, believed that this run of the Gibbs sampler confirms the explanation given in Section 7 of the apparent success of the Griffiths & Hunt model.

12.3.2 Simulated data : Run 1 : Data simulated from an exponential distribution

In order to assess how the Gamma/Exponential distribution would perform in conditions of totally free-flowing traffic, a file containing 200 observations was used with

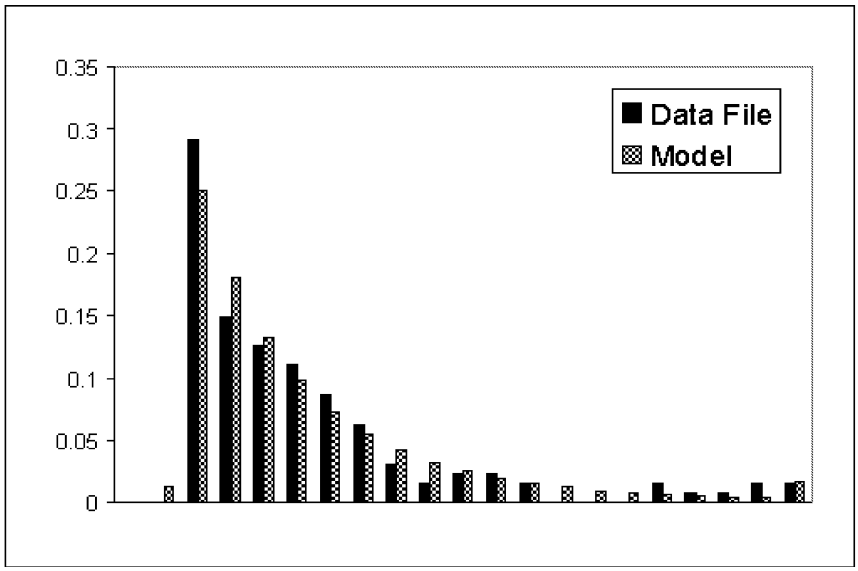


Figure 70: Model fit diagram for the Griffths & Hunt model

each observation being simulated from an exponential distribution whose parameter was 0.25. Prior distributions etc were the same as those used for the last run of the Gibbs sampler for this model in Section 7.

Figure 71 shows the marginal posterior distributions which can all be seen to be unimodal. Also, it can be observed from the parameter, p, that the sampler has coped well with data that would always be expected to be allocated to the first component, with the modal value of p being 0.982 (3 d.p.). This is encouraging.

The run passed the test for convergence after thinning was applied and model fit is quite satifactory as shown in Figure 72 As already stated, marginal posterior means were used in model fit diagram. If posterior modes were used then model fit is slightly better. This gives rise to the question as to whether posterior modes or posterior means should be used in such circumstances. This raises other questions that are fairly central to Bayesian statistics and the choice is, perhaps, best left to the individual practitioner although, as is often the case, practicality rather than philosophy may be the deciding factor.

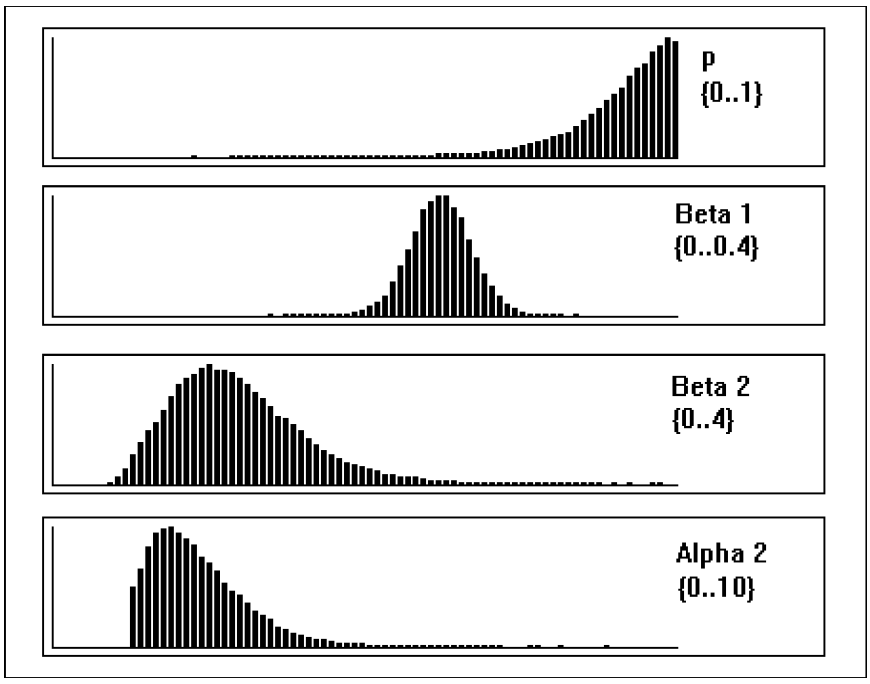


Figure 71: marginal posterior distributions for the Gamma Exponential distribution

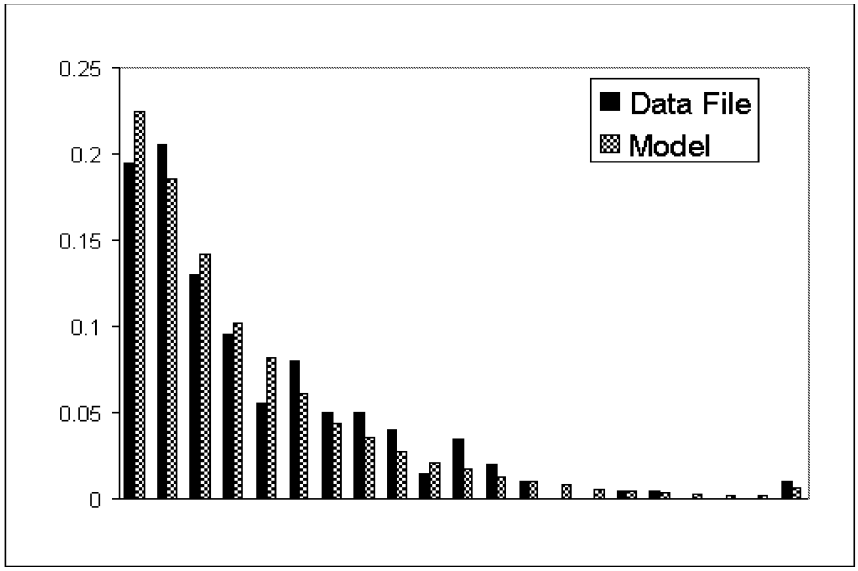


Figure 72: Model fit diagram for the Gamma Exponential distribution

12.3.3 Simulated data : Run 2 : Data simulated from an gamma distribution

Here, it required to examine model behaviour under conditions of heavily congested flow and so a file of observations simulated from a gamma distribution has been created. In this case we would expect a posterior value of p to be slightly greater

than zero and this is, in fact, the case with the posterior mean for p being 0.0366 (4 d.p.). Figure 73 shows the marginal posterior distributions for this run and all are unimodal.

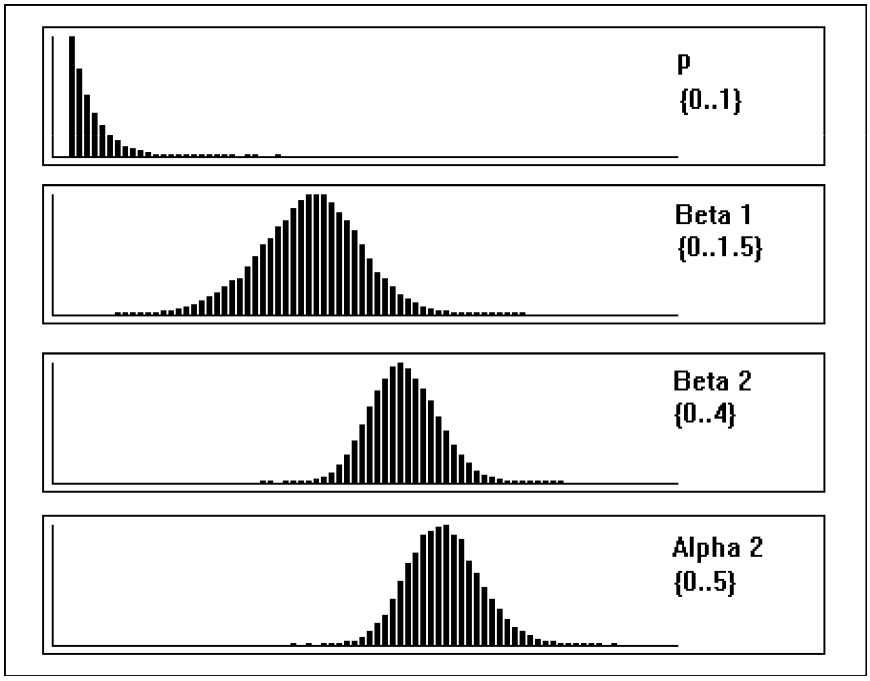


Figure 73: marginal posterior distributions for the Gamma Exponential distribution

Convergence was satisfactory but it is noted that less thinning was required in order for the convergence test to be passed. So far, it appears that the convergence properties of the model are better when more congested traffic is modelled. This may be because, under such conditions, more observations are allocated to component 2 but to date this phenomenom has not been rigourously examined. Figure 74 shows the model fit diagram for this run.

Again, model fit can be seen to be satisfactory. Given this, and the other results from this run, we see that the model has performed well with data corresponding to congested traffic.

When both the above runs are considered, along those where real data was used, it can be seen that results are very encouraging. The Gamma / Exponential distribution has emerged as a suitable model for use with vehicle headways on dual carriageway roads.

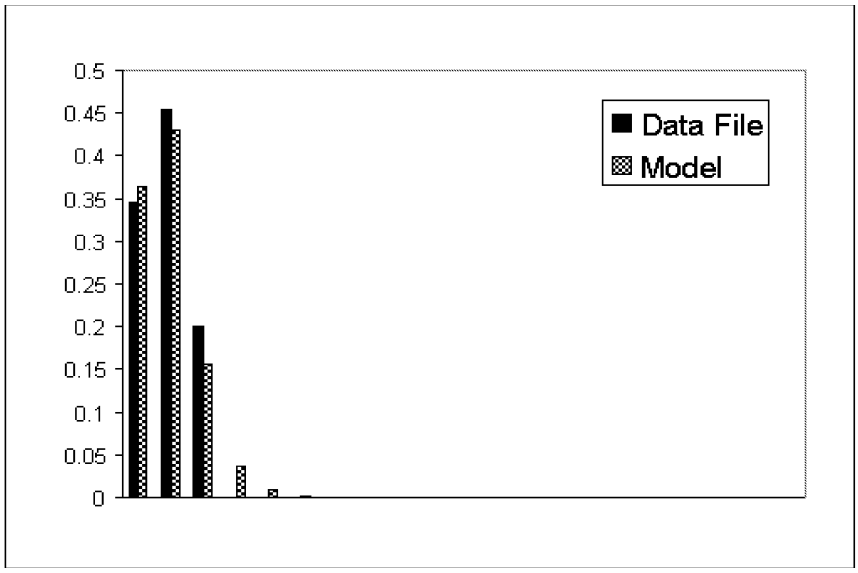


Figure 74: Model fit diagram for the Gamma Exponential distribution