Bayesian Models for PCR Stutter

Madappuli Arachchige Chaminda Sri Sampath Fernando

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Statistics,

the University of Auckland,

2017.

Abstract

For several decades, increasing attention has been given to the improvement of the quality of DNA mixture evidence interpretation, because of its importance in resolving problems in criminal investigations. Peaks at different positions along a molecular weight axis in an electropherogram (epg) are known to correspond to the alleles in a DNA sample, and these alleles can be used to describe differences between individuals. The process of DNA mixture interpretation largely involves probability and statistical models. Continuous probabilistic models ensure relatively greater objectivity and consistency between analysts than do other types of models. Implementing such models requires statistical models for PCR phenomena such as stutter. A peak at an allelic position, generally one repeat unit lower than a 'parental' peak, and usually with a height lower than the parent peak height, is called a 'stutter'. The presence of stutters in an epg presents a problem in DNA mixture evidence interpretation. Therefore, practitioners search for sophisticated methodologies to model the contributions of stutters and real alleles to the peak height in order to make the interpretation more accurate. In modelling PCR stutter, the stutter ratio (SR) which represents the proportion between the observed stutter peak height and the parent allelic peak height is generally used.

This research reviews the existing models for SR and develops new, advanced, Bayesian models for increased accuracy in predicting stutter. The developed models include non-hierarchical, hierarchical, and infinite mixture models. In these models, the longest uninterrupted sequence (LUS) of an allele was used as the key covariate in explaining the behaviour of SR. For hierarchical and non-hierarchical models, standard model evaluation techniques, including information criteria such as AIC, BIC, DIC, and WAIC, cross-validation measures, and Bayesian p-values, were used considering their limitations and appropriateness under different modelling conditions. Initially, eleven non-hierarchical models, including six new models and five models developed in previous studies for predicting SR, were evaluated. Next, hierarchical models corresponding to seven of these models were investigated. Finally, the study used an algorithm based on the collapsed Gibbs sampling that uses the Chinese restaurant process as a non-parametric Dirichlet process prior, for fitting an infinite mixture of simple linear regression models for SR using LUS as the predictor. The overall contribution includes improvements in the prediction of PCR stutter through various Bayesian modelling techniques, an extension of infinite mixtures to the linear regression case, and advances to the collapsed Gibbs sampling algorithm that uses CRP as a non-parametric Dirichlet prior.

Dedication

To
my beloved Mother
&
late Father

Acknowledgements

I take this opportunity to thank all those who helped me in numerous ways to complete this thesis, and it is a great pleasure to mention the following key persons here.

I would like to express my most important acknowledgement and gratitude to my supervisor Prof. James M. Curran for being a tremendous mentor for me who comes from a solely statistics background and acquired knowledge on biology and DNA analysis with his guidance. His intellectual vigour, generous support, and advice on my research have been priceless. The great freedom he offered allowed me to significantly improve my skills and allowed me to grow as a research scientist. Also, I am especially grateful to Prof. Renate Mayer, my co-supervisor, for her expert advice, valuable assistance, suggestions, and encouragement throughout my study.

I would like to offer my deepest sense of gratitude to the New Zealand government and Education New Zealand for granting me the New Zealand International Doctoral Research Scholarship (NZIRDS); and this work would not have materialised without this financial support.

I am indebted to the University of Auckland for granting me the Auckland University Doctoral Scholarship for six months after my previous scholarship and for providing me all the facilities required for conducting my study. Special thanks to the staff at the New Zealand eScience Infrastructure (NeSI) pan cluster for their supportiveness.

I extend my sincere gratitude to the Department of Statistics for kindly providing a graduate teaching assistantship position and travel funds to attend conferences.

I am greatly indebted to Sabaragamuwa University of Sri Lanka for providing me with study leave to complete this PhD without any disruption.

My heartfelt thanks to the Dr Jo-Anne Bright, Science Leader at the Institute of Environmental Science and Research (ESR) who helped me as a fellow student in obtaining DNA profile data for my study, arranging a visit for me to the ESR, and explaining the technical details I needed.

In addition, I would like to thank Mr. Richard Gyde, the Managing Editor at Editwrite English text editing service, for his professional, untiring support and patience in proof-reading my thesis.

I would like to thank my colleagues and the staff at the Department of Statistics in the University of Auckland, who have always been extremely helpful and welcoming.

I owe a great deal of appreciation to my wife Chathurani for all her efforts in reading this thesis and suggesting improvements for my writing. Without her continuous unconditional support and motivation, this PhD would have not been completed.

Finally, I would like to thank my beloved son Vikum and daughter Mindulie for their patience and making me forget all my burdens in this PhD for most of the time.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 Introduction

Forensic DNA analysis has received much public attention over the last thirty years because of its incredible usefulness in criminal investigations. It has become an extraordinarily powerful technique in forensic science. It has also had considerable scientific scrutiny, mainly in response to changes in science and legal challenges. The field of statistics is paramount in DNA evidence interpretation because of the intrinsic probabilistic nature of the problem. Statistical DNA interpretation is one of the most mature fields in forensic science, which uses knowledge from the fields of: statistics, population genetics, and molecular biology [146]. This has been in existence longer than any of the current technologies used for typing the evidence, because the same basic ideas generally apply, regardless of how the evidence is typed. Consequently, there exists a large body of literature devoted to this subject, which reflects its importance in the legal and scientific community.

In a typical criminal case, biological materials such as blood, semen, saliva, or other body tissues, may be recovered. These materials may have been exposed to a range of surface or environmental conditions, and this can affect their usefulness to the investigator. The materials are taken to a forensic laboratory where a scientist will attempt to extract DNA using reagents specially designed for this task. The amount of template DNA extracted is often very small, in the range of 50 to 100 picograms (10^{-12} g). There-

fore, forensic biologists amplify the template DNA using the *polymerase chain reaction* (PCR) process. This amplification allows length variants in the DNA, called *short tandem repeats* (STRs), to be detected by measuring relative fluorescence when the sample is exposed to laser light. The resulting signal is collected by a photomultiplier and displayed graphically as an *electropherogram* (epg).

The epg consists of a trace signal displayed on a molecular weight axis, which is mostly flat with peaks in various locations. The presence of a peak corresponds to the alleles present in the DNA sample. Crudely, alleles are variants or *polymorphisms* of a gene, which can be used to describe differences between individuals. The heights of the peaks are approximately proportional to the amount of template DNA present. This quantitative information (as opposed to the discrete allele information) can greatly enhance the interpretation process.

A genotype at a locus consists of two alleles, each inherited from the donor's biological mother and father. The alleles for each locus are usually denoted by integer, or occasionally decimal values in the epg. If the pair of alleles is identical at a given locus, then the individual is said to be *homozygous* at that locus. In contrast, if they are different, the person is said to be *heterozygous* at that locus. Modern forensic labs are well-equipped with various commercial multiplexes and each multiplex tests a distinctive collection of STR loci.

1.1.1 Characteristics of an Electropherogram (epg)

A hypothetical example of an epg (Figure 1.1) is used to illustrate some key characteristics of an ideal DNA profile. The epg illustrates peak height information of three loci corresponding to a biological sample originated from a single-source. Each allele possesses an 800 rfu (relative florescence units) signal. The individual is homozygous at the Locus B and heterozygous at the other two loci. As a consequence of allele masking at the Locus B, the peak height is almost double compared to other peaks of the heterozygous loci. Either *peak height ratio* (PHR) or *heterozygote balance* (or *imbalance*) is used to describe the degree of balance (or imbalance) between two peaks at a locus of an epg. The heterozygote balance is regarded as a PCR artefact. The possible reasons for this

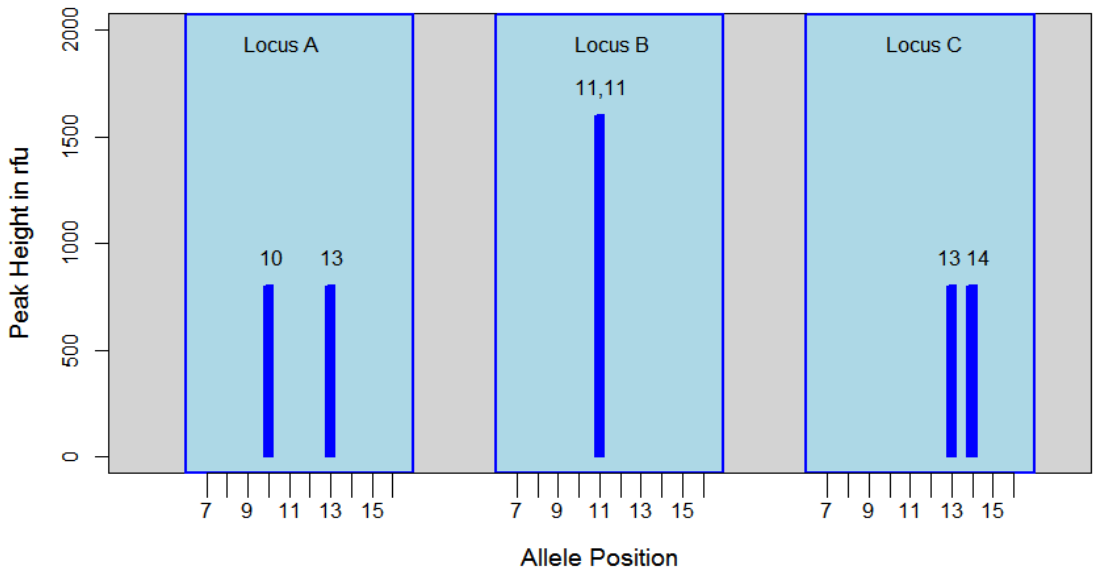

Figure 1.1: A hypothetical example of an epg

are unequal sampling of starting templates, and unequal amplification of two alleles of a heterozygote due to natural variation in the PCR process [31]. The peak height of one particular allele is compared with the corresponding height(s) of sister allele(s) in defining both measures. The PHR of a locus is also called intra-locus balance and, for this example, it is 100% at each locus. The presence of an epg with the above characteristics makes the interpretation much easier and provides a greater confidence that a profile originated from a single source. However, there are some critical issues that arise from epgs which originate from real-world biological samples, and they often create severe challenges, difficulties and uncertainties in the interpretation of DNA profiles. Figure 1.2 is a part of an epg, which consists of peak height information for two loci (D10S1248 and vWA), and it illustrates some practical problems associated with real world data. There are two major peaks in each locus and the vWA locus exhibits a relatively high PHR compared to the D10S1248 locus. In the D10S1248 locus there are two minor peaks located at allele positions 11 and 14. Another minor peak is located at allele position 15 in the vWA locus. If it is assumed that the questioned biological sample corresponding to this epg contains DNA from a single donor, then these three minor peaks are categorised as ‘stutters’. A detailed description on stutter and stutter mechanism is provided in section 1.1.3.

Biological samples containing very low quantities of template DNA are designated as

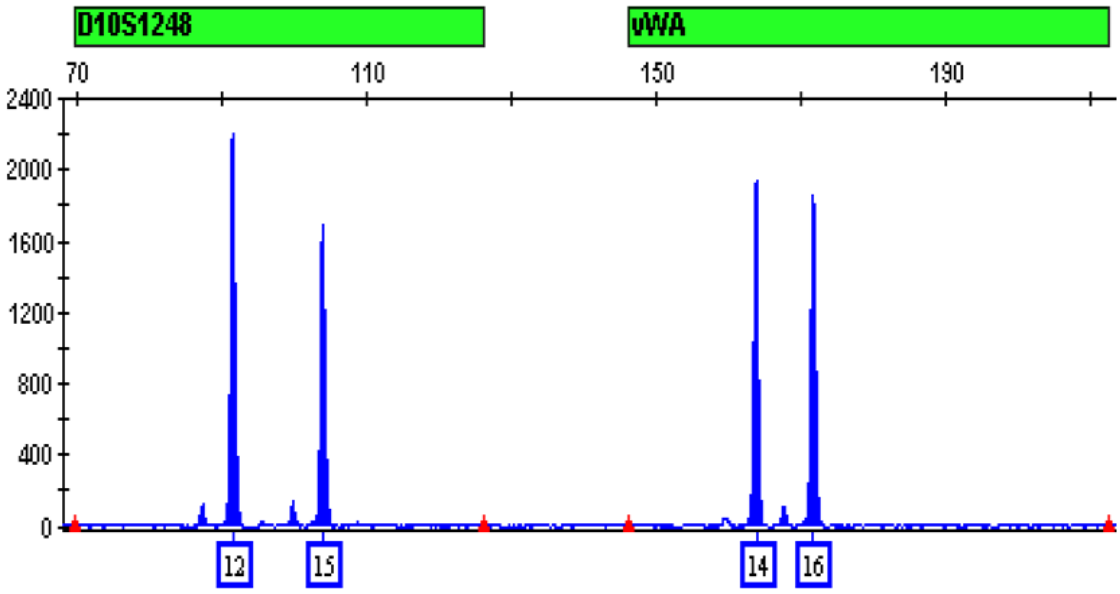

Figure 1.2: An example of an epg (peak heights in rfu)

Source: Torben Tvedebrink [167]

LCN (low copy number) or LtDNA (low template DNA) samples [30]. Modern forensic laboratories can retrieve DNA from items where the DNA has been transferred only by casual handling of evidential items. Many of these 'touch DNA' samples can also be classified as LCN samples. The analysis of these is very challenging due to low quantity of DNA available. The LtDNA profiles tend to be severally affected by profiling artefacts. A higher incidence of heterozygote imbalance is an example for profiling artefacts, which often creates problems in standard profiling of donors with heterozygote loci. Both *allelic drop-out* and *allelic drop-in* are also regarded as profiling artefacts.

The condition where an allele cannot be visualised is known as allelic drop-out. Mostly, when one allele of a heterozygote cannot be visualized, the drop-outs are observed. The stochastic variation in the PCR process has been identified as the reason for this non-visualisation [81]. The appearance of extra alleles in a profile in addition to the alleles of donors is called drop-in. Very small quantities of contaminant DNA is the potential cause of this. However, the presence of allelic drop-in is an occasional incidence even with LtDNA.

1.1.2 Analytical and Stochastic Thresholds

The background noise (observation of tiny peaks) in the epg is another issue which makes some difficulties in the interpretation of complex biological samples. The uncertainty in the classification between real alleles and background noise is expected to be minimised with the use of minimum signal thresholds. In many applied scientific fields, thresholds are often used for the practical convenience in discriminating between two states [28]. The transition between two states is gradual. Therefore, theoretical or empirical thresholds are often applied to delineate the two states. However, strictly defined thresholds can cause problems in decision making especially when an observation is close to the threshold. In the field of forensic genetics, laboratory defined thresholds are used to designate alleles and stutters based on peak height information derived from the epg. These thresholds can be very useful as they may simplify interpretation, but the simplification comes at some cost. Binary decisions, such as these, may be wrong and consequently can have drastic effects on interpretation.

An *analytical threshold* is defined as the minimum height requirement of an epg to differentiate real allelic peaks from background noise (artifactual peaks). They are conditionally defined based on the sensitivity of the genotyping instruments; hence, each laboratory establishes their own thresholds based on an analysis of internally derived signal-to-noise data. Most of those empirically derived thresholds vary in a range from 30 rfu to 50 rfu.

Stochastic thresholds are used with the presence of a relatively low single peak at a specific STR locus. The observed peak height is compared with a threshold and designated as either a homozygous or a heterozygous genotype. Generally, empirically derived stochastic thresholds vary between 150 rfu to 300 rfu. A single peak appearing between the analytical and stochastic thresholds is regarded as a potential heterozygous genotype whose sister allele is believed to have dropped out. In contrast, a single peak locating above the stochastic threshold would be designated as a homozygous genotype because drop-out is considered as fairly unlikely in this situation.

There are a number of proposed methods for the derivation of thresholds in the literature. Gill et al. [85] proposed a method based on a logistic regression, which provides a

way to evaluate the risk of a false designation of a heterozygote as a homozygous genotype. The risk levels were determined based on an experimental dataset of low-template-DNA, which includes extreme drop-outs. A graph of extreme drop-out probabilities (calculated upon the fitted logistic regression model) against the height of an existing sister allele visualises how likely a sister allele of an observed single allele having dropped out. Tvedebrink et al. [168] also employed a model based on logistic regression, to estimate the allelic drop-out probabilities of STR alleles and have shown the locus-specific dependency of these probabilities. In addition, the variations in the drop-out probabilities due to different typing kits for profiling and the use of diverse machinery even within the same laboratory have been highlighted and recommended to use machine-specific estimates of the parameters for the logistic regression model.

1.1.3 Stuttering and the Stutter Mechanism

A peak at an allelic position, generally one repeat unit lower than a parental peak, and usually with a height lower than the parent peak height, is called a *stutter*. Slippage (shadow band) [94] or miscopying [27] during the PCR process is presumed to be the reason for it. The magnitude of stutter product is often measured in terms of a ratio or a proportion. The ratio of observed peak height relative to the corresponding parent allele height is defined as the stutter ratio (*SR*). The stutter proportion, in contrast, is defined as the proportion of observed height compared to the sum of the stutter and allele peak heights. However, stutter proportions are not as frequently used as the *SR*. There are several modes of stuttering, the most common of these being back stutter. The modes of stuttering are described with names corresponding to their positions relative to the parent allele location. A stutter peak in a position that is one repeat unit lower than the parent peak is called a *back stutter*, *negative stutter*, or *reverse stutter* [31]. This is the most commonly observed type of stuttering. On rare occasions, stutters may occur at positions that are two repeat units lower or one repeat unit higher than the corresponding parent allele. These two uncommon situations are labelled as *double stutter* (*double back stutter*) and *over stutter* (*forward stutter* or *positive stutter*) respectively [31, 157]. The stutter ratios (or heights) of double back stutter and forward stutter are always smaller

than the back stutter for a given allele. However, interpretation of mixtures with any of these two types of stutters is as complicated as with back stutter.

Traditionally forensic laboratories have used rule-based approaches to designate stutters. Stutter thresholds are not directly defined for observed peak heights, unlike analytical and stochastic thresholds. Rather, stutter thresholds are defined in relation to stutter ratios. The stutter threshold applied for a 28-cycle PCR is generally 10% [27] while universal stutter thresholds used in laboratories can vary from 10% to 20%, with the consideration of outliers [31]. However, stutter thresholds within this range are quite generous and most of the stutter peaks are less than 5% of the parent allele height. The mean stutter ratio usually remains below 15% even for LCN profiles. However, stochastic effects mean that even larger values are possible. Locus-specific stutter thresholds (upper-limit) are empirically defined as either three standard deviations (SD) above the mean *SR* or the largest observed *SR*. Rarely, the value that is three standard deviations beyond the empirical maximum of stutter ratios is also used as the threshold. The estimated locus-specific thresholds vary across different Applied Biosystems STR Kits. Table 1.1 summarises the estimated stutter thresholds recommended in user manuals for the D18S51 STR locus. According to the information available in the table, the inherent properties of STR kit and the statistical method (mean stutter + 3 SD, the empirical maximum, or the largest observed stutter + 3 SD) can be identified as the key factors affecting locus-specific stutter thresholds. If a suspected peak has a lower percentage peak height (relative to the height of the parent peak) than the given threshold, then the peak is classified as a stutter.

1.1.4 DNA Mixtures

A biological specimen being tested containing DNA contributions from two or more individuals is called a mixture profile. Vaginal swab collected from a rape victim, body fluid (e.g. saliva and blood) of a perpetrator recovered from the surface of victim's skin, and contaminated single source crime samples are examples of possible mixture profiles. The unintentional introduction of exogenous DNA into a biological sample or PCR is known as contamination (SWGDAM 2012) [160]. Such detection of exogenous DNA originated from reagents, consumables, operator and/or laboratory environment can be evaluated

Table 1.1: Stutter thresholds recommended for the D18S51 locus across different Applied Biosystems STR Kits

STR Kit	Stutter threshold (%)	Method of calculation
Identifiler Direct	12.9	mean stutter + 3 SD (N = 669)
Profiler Plus	< 13	highest observed stutter
Identifiler Plus	13.7	mean stutter + 3 SD (N = 500)
NGM SElect Express	13.8	mean stutter + 3 SD (N = 668)
NGM SElect	13.8	mean stutter + 3 SD (N = 1080)
NGM	13.9	mean stutter + 3 SD (N = 996)
NGM Plus	16.0	highest observed stutter + 3 SD
SEfiler Plus	16.4	highest observed stutter
Identifiler	17.0	highest observed stutter
MiniFiler	18.0	mean stutter + 3 SD (N = 668)

Source: Advanced topics in forensic DNA typing: Interpretation [31]

using both known and control samples. Contamination assessment is an important requirement of internal validation processes. The police or other individuals who access the scene of a crime may be other potential sources of contamination.

The key characteristics of a mixture profile are: the appearance of more than two alleles at multiple loci and the presence of several loci with only one pair of alleles that exhibit fairly extreme peak imbalance [31]. Amelogenin, the sex-typing marker, provides some decisive information to distinguish between male and female contributors in a mixture profile. A severe imbalance in the amelogenin X and Y alleles recommends a male-female mixture profile with the female as the major contributor.

The analysis of complex mixtures like this has always been problematic in evidence interpretation. Models that rely solely on expert judgement are susceptible to analyst's specific biases or misinterpretation. In contrast, the use of continuous probabilistic models which assign a non-zero weight (probability) for each possible genotype significantly enhances the use of continuous peak height information and makes it possible to produce very efficient reliable interpretations.

A study conducted by Dror and Hampikian [54] described the subjectivity and the bias in forensic DNA mixture interpretation. The study selected a DNA mixture profile from an actual adjudicated criminal incident related to a gang rape case. According to the study, the same evidence had been presented to an independent group of 17 expert DNA

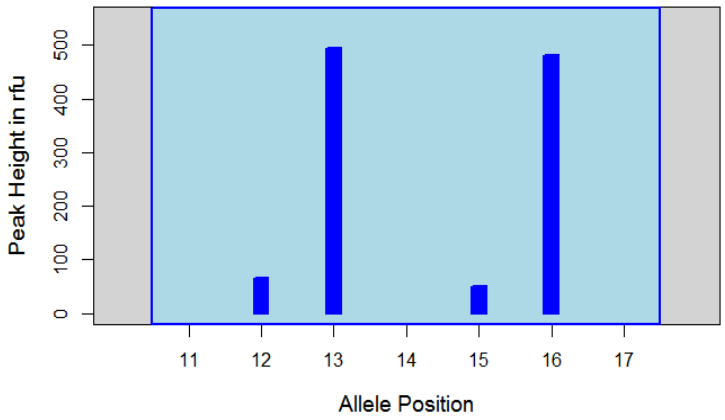

Figure 1.3: A hypothetical epg with unknown contributors

analysts (including 2 PhD, 12 M.Sc. and 2 B.Sc. qualified experts) who were working in forensic casework in a reputed forensic laboratory in North America. These examiners independently analysed the same source of information used by the original forensic experts, and this included the DNA mixture profile of the questioned biological sample recovered from the rape. Only one of them came up with a conclusion that was similar to the original experts' conclusion. The inconsistencies between the original decision and the decisions of the expert panel clearly emphasise the risk of using experts' judgement methods in evidence interpretation.

1.1.4.1 How do stutters complicate mixture interpretation?

Analysis of mixed stains can be seriously affected by stutter peaks in the epg and in particular when the contributions of DNA are very unequal [176]. A complex mixture containing many contributors enhances the possibility of allele sharing, and increases the uncertainty of the associating genotypes of contributors. Thresholds become very difficult to apply in complex mixture cases.

As previously mentioned, stuttering can seriously complicate the interpretation of DNA mixtures. In order to illustrate this issue, a small hypothetical example is presented. Figure 1.3 is an idealisation of an epg without any knowledge of the contributors. Some of the possible explanations for the epg are listed below.

1. A mixture of two heterozygous contributors – a (13, 16) major and a (12, 15) minor contributors

2. A single heterozygous contributor – a (13, 16) contributor with two stutter peaks at 12 and 15

3. A mixture of two heterozygous contributors (subject to a severe heterozygous imbalance) – (12, 13) and (15, 16) contributors or (12, 16) and (13, 15) contributors

4. A three-person mixture with two homozygous major contributors – (13, 13), (16, 16) major and a (12, 15) minor contributors

5. A two-person mixture of two homozygous contributors – (13, 13) and (16, 16) contributors with two stutter peaks at 12 and 15

In some recent publications [11, 47, 78, 84, 86, 152], stutters have been discussed in the context of DNA mixtures. Analysis of evidence is always problematic with the presence of mixtures. The analysis of a crime scene stain which is assumed to be contributed by only two individuals is the simplest situation of mixtures. However, matching of information related to the stain is very important in evidence interpretation. In a two-person mixture, where major and minor contributors can often be distinguished and the minor component matches with the suspect (e.g. analysis of vaginal swab recovered in a rape case), the ambiguity of minor peaks may severely affect the interpretation of evidence. In contrast, in some situations, a match between the minor peaks and the victim (e.g. analysis of the evidence recovered from finger nail tip of a suspect in a rape case) may also be very complicated. Therefore, the ambiguity of a match between the minor peaks and the suspect or the victim is conditionally dependent on the circumstances of the crime.

If we assume that this is a two person mixture, and both contributors are heterozygous at this locus and there is no allele sharing, then all the four peaks represent real alleles. However, if a single contributor situation is assumed, then both minor peaks are interpreted as stutters corresponding to their parent alleles.

Assume that there are two typed individuals who are alleged to have contributed to the crime stain. Furthermore, one is a victim and his/her contribution is not disputed by either the prosecution or the defence and is of type 13, 16. The defence may argue that their client (of type 12, 15) was not a contributor and that 12, 15 are stutters or 12, 15 belongs to another random individual. In this particular situation, a match between the minor peaks

and the suspect makes an ambiguity of minor peaks and it may severely impact on the interpretation of evidence.

Forensic practitioners believe that heterozygous balance varying between 0.6 and 1.66 is fairly acceptable. Therefore, it is obvious that the ratio of peak heights between a pair of heterozygote alleles (heterozygote balance) cannot be far away from unity. Therefore, in this situation, where the minor component matches the suspect, the most credible combination of major and minor contributors is (13, 16) and (12, 15) respectively. Under these circumstances, the major and minor contributors are the victim (V) and the suspect (S) respectively. When this case is taken to a court, the prosecution would propose that both suspect and victim share the evidence. However, the defence would claim that (12, 15) are not real alleles but stutters.

The following pair of propositions summarises the issue in the court.

Prosecution Proposition (H_P) - The evidence contains the victim's and suspect's DNA (both peaks at 12 and 15 are real alleles)

Defence Proposition (H_D) - The evidence contains the victim's and some other person's DNA or both minor peaks at 12 and 15 are stutters and the evidence contains only the victim's DNA

In reality, the truth lies between the two extremes that both 12 and 15 peaks are either allelic peaks or stutters. Therefore, a sound methodology to distinguish between stutters and real alleles is essential for the accuracy of the interpretation. Sensibly, any such method has to be able to focus on modelling stutter peaks.

1.2 The Statistical Evaluation of DNA Evidence

Generally, in forensic genetics, each individual's DNA is believed to be unique with the exception of identical twins. The standard DNA profile largely depends on the resulting post-amplification product of PCR process, which uses a tiny sample of donor's entire DNA in amplification. Hence, irrespective of the uniqueness of an individual's DNA, there could be a random match between a particular crime scene profile and a random

person who does not have any relationship with the crime. Therefore, the forensic DNA evidence interpretation is explicitly probabilistic and the role of statistics is vital.

The classical profile probability and the *likelihood ratio* (LR) are the two approaches used to report DNA profiles [83]. The probability of evidential DNA profile under a specific given hypothesis is conveyed in the profile probability approach. The likelihood ratio approach, the one which is focused on in this study, is (now) the favoured method for presenting forensic evidence in the court in many jurisdictions. It links the evidence related to two hypotheses (propositions): prosecution (H_p) and defence (H_d). The prosecution hypothesis claims that the accused is the donor of recovered DNA from the crime scene. The defence hypotheses, in contrast, claims that an unknown person who is not blood-related to the accused is the donor of DNA. Then the ratio of the probabilities of the prosecution hypothesis to the defence hypothesis is defined as *LR*.

There have been many different approaches suggested for the evaluation of forensic evidence. However, this thesis believes that correct approach is a Bayesian approach. Bayes' theorem is a statistical representation of the logical process of updating one's beliefs on the basis of evidence. It is not hard to see the connection of this theory to the legal process. Forensic statistics uses the odds form of Bayes' theorem. That is,

$$\underbrace{\frac{\Pr(H_p|Evidence)}{\Pr(H_d|Evidence)}}_{\text{Posterior Odds}} = \underbrace{\frac{\Pr(Evidence|H_p)}{\Pr(Evidence|H_d)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{\Pr(H_p)}{\Pr(H_d)}}_{\text{Prior Odds}}$$

Conditional probabilities along with the Bayes' theorem are used to calculate the *LR*, in presence of more information and/or conditions. The third principle of evidence interpretation ("Scientific interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated") proposed by Evett and Weir [59] is used in calculating *LR* whenever the non DNA background information related to the crime scene is available. The DNA evidence is used to infer the questioned genotype, and to compare it with a suspect genotype relatively to a reference genotype population to assess the strength of match [135]. This assessment is performed using *LR*, and it is also used to discriminate two hypotheses with regard to DNA evidence. It determines which hypothesis is more likely to be true under the given evidence.

Likelihood ratio often appears in different forms of mathematical formulation and scientific interpretation. Perlin et al. [134] examined the following forms of *LR* and proved mathematical equivalence among them.

1. *Hypothesis form* calculates the information gain in hypothesis as the ratio between the posterior odds and the prior odds of the hypothesis.

2. *Likelihood form* considers both identification and alternative hypotheses, and calculates the information gain in likelihood as the ratio between the likelihood of data given the identification hypothesis and the same likelihood given the alternative hypothesis.

3. *Genotype form* calculates the information gain at the suspect's genotype as the ratio between the probability of evidence genotype and the probability of coincidental genotype.

4. *Match form* calculates the information gain in match as the ratio between the probability of evidence match and the probability of coincidental match.

The presence of mixtures, close relatives or an involvement of partial profiles create more complicated evidence, and the use of *LR* to communicate the strength of evidence becomes more important [146]. In the presence of a mixture with different numbers of contributors, the genotype probabilities can only be compared with *LR*.

In a criminal investigation, a forensic scientist or an expert witness is invited to give the court an assessment of the weight, or value, of the evidence. This is reduced to the following form to simplify the evaluation of DNA evidence

$$\sum_j w_j \Pr(S_j|H_i) \tag{1.1}$$

where

H_i – the of hypothesis of interest (e.g. only the victim and suspect have contributed to the stain),

S_j – a set of possible genotypes, and

w_j – the weight representing the possibility that S_j can explain the evidence.

There are four competing models for the interpretation of DNA evidence. These are usually referred to as: *classical*, *binary*, *semi-continuous*, and *continuous*. They essentially differ in their definitions of the weights w_j. The classical model takes no account of peak height information. It therefore regards every *feasible* genotype set S_j as having a weight of $w_j = 1$. Feasibility is defined in terms of the genotype combinations that **completely** describe the alleles observed in the crime scene stain. The classical model cannot deal with PCR artefacts such as drop-in, drop-out, stutter, or uncertainty in input parameters. The relative advancements in binary, semi-continuous, and continuous models are reviewed with more details in the following section.

1.3 Models for DNA Interpretation

There are two Bayesian models used for the interpretation of DNA evidence, namely, binary and continuous. However, in the recent past, models that are able to overcome some of the problems in binary model have been identified as semi-continuous models [104]. The following sections review background literature on these three models and discuss their relative advancement in terms of the ability to use the epg information for *LR* calculation.

1.3.1 Binary Models

A binary model combines the experience of experts and sets of empirically derived rules during the probability assignment as weights in Equation 1.1, and considers a much wider set of genotype combinations than the classical models. Based on empirical guidelines that take peak height information into account, an expert decides whether any of all possible genotype combinations at a locus can be excluded. The model assigns the values 0 or 1 to the unknown probability $\Pr(O|S_j)$ (the probability of observed crime stain O given genotype combination S_j), based on whether the analyst's judgement on possibility and selecting them as either included (weight = 1) or excluded (weight = 0). The binary model primarily assumes: [27]:

- approximately equal locus-specific mixture proportions,

- proportionality between the quantity of DNA peak area, and

- that the total contribution of two individuals equals to the area of shared peaks.

Since the binary model combines the experience of experts and a set of empirical guidelines when assigning the weights, it has been classified as a manual method which can be used for the resolution of two-person mixtures. That means application of a binary model is very difficult with higher order mixtures due to the extensive involvement of mathematics. The assignment of discrete weights (0 and 1) will lead to increasing the risk of less accurate decisions because it does not assign weights for every genotype combination. This problem is critical in complex mixtures and LtDNA (low-template DNA) where there exist stochastic factors that make interpretation more complex. Failure to consider the value of peak height and inability to deal with PCR artefacts are the major shortfalls in the binary model.

1.3.2 Semi-continuous Models

Semi-continuous models explicitly model stochastic phenomena such as drop-out and drop-in (contamination), and assign probabilistic weights [11, 168, 169]. This leads to partially resolving the problems in binary models discussed above. However, the differences in capabilities of both PCR amplifications and light intensity measuring systems for the epg, result in variations in the drop-out probabilities. The variations can be seen across laboratories, machineries, and typing kits used for profiling. Treating probability of drop-out as a random variable and integrating it out can be a solution for these problems. Although it is difficult because the probability distribution of drop-out is not fixed across cases [104], implementations of discrete models incorporating these probabilities are recently found [141]. In summary, even semi-continuous models are not capable of assigning different weights for all genotype combinations in complex mixtures and have significant problems in using information available with peak heights.

1.3.3 Continuous Models

Binary and semi-continuous models, assign zero probability (weights) for many genotype sets S_j. Hence, compared to these two models, continuous models may consider more genotype sets as they allow the associated weights in Equation 1.1 to vary from zero to one $(0 < w_j < 1)$. Since the weights can be any continuous value between 0 and 1, the model is known as the continuous model. Using a suitable model for peak heights for all the peaks in the profile, a fully continuous model assigns a value to the probability of the observed peak, given any particular genotype combination [104]. A "good" explanation gets a weight close to one and a "poor" explanation receives a weight close to zero. Since the weights can assume any value from 0 to 1, continuous models ensure high reliability in evaluation of DNA evidence. However, statistical models for PCR phenomena (and computer software) are required to implement the continuous models. Diagnosis of true alleles from epg peaks, decision of possible allelic combinations, and stutter allocations are some problematic scientific decisions involved in developing these models. The essence of continuous models is the use of epg peak heights as a source of quantitative information to determine the probability of peak heights given all possible genotypes.

Two types of continuous models: normal approximation based methods and MCMC (Markov Chain Monte Carlo) methods have been discussed by Buckleton et al [27]. A program called "BETAMIX" was developed based on the normal approximation-based method. However, the effect of stuttering was not assessed and heterozygous balance was not modelled appropriately in BETAMIX. Buckleton et al. [27] considered MCMC approaches as superior against normal approximation based approaches as they do not address PCR artefacts. MCMC simulation methods can probabilistically deal with complex interaction of preferential amplification, stuttering, and other PCR artefacts or sampling effects such as drop-out and drop-in [27]. Following are the key assumptions of MCMC models:

- approximately equal locus-specific pre-amplification mixture proportions
- proportionality between peak area and the quantity of DNA
- the total contribution of two individuals equals to the area of shared peaks

- the contributions from other sources and stuttering can be combined together

However, some practical problems of using MCMC methods for DNA evidence interpretation could be highlighted. These methods drastically increase the complexity of the method of DNA evidence interpretation. In the courts point of view MCMC methods can be termed as "black box" type methods. The ability of lawyers, forensic scientists, juries, and judges to understand them is relatively low. MCMC methods introduce an additional problem, which is not a problem for the statisticians but the court. It could be difficult to understand why the answers are not the same from run to run. MCMC procedures do not give exactly the same result every time unless the use of a random number seed.

It is hard to know whether the MCMC methods do or do not explore the full genotype space. However, genetic calculators consider all possible genotype combinations that are indicated by the electropherogram. The number of genotype combinations that need to be considered within each locus is massively large when drop-in and drop-out are taken into account. Genetic calculators like STRmix^{TM} and Cybergenetics TrueAllele® consider every possible genotype combination for different contributors at each locus. Even though it considers a large number of combinations as a Bayesian system, only a few of them will be plausible and informative. All the unrealistic genotype combinations are rated only with extremely low probabilities. The combinations that exhibit a good explanation between them are obviously justified with higher genotype probabilities. The locus-specific posterior probabilities of every possible genotype are calculated for each contributor and used them to calculate the likelihood ratio of any interested genotype.

Inefficiencies of MCMC methods in evaluating higher order mixtures and dealing with PCR artefacts have been highlighted in literature [46, 47]. Buckleton and Gill [26] have obtained a patent for their MCMC method that uses a distribution to allow more variation in heterozygous balance when the peak areas are low. Evett has used this methodology to develop a prototype named "Mixtures Full Monte" (MFM). A heterozygous balance model, an assumed stutter model, and a known mixture proportion (M_X) with no stochasticity were used when creating MFM. All these models or methods provide sufficient evidences for the ability of MCMC methods to use all types of continuous epg information. Therefore, these methods can be identified as the most prominent recent trend in quanti-

tative DNA mixture interpretation models that use either peak height or area information [47].

1.4 Statistical Modelling of Stutter

The presence of stutters in an epg has been known for a long time as a problem in DNA evidence interpretation. As it always makes the mixture interpretations complicated, practitioners have started to use some ad-hoc approaches though these do not work all the times. As mentioned previously, a peak in a stutter position with a height not exceeding a laboratory defined threshold value, and not masking as an allele of a potential contributor is suspected as a stutter peak. The ad-hoc rules are defined based on the ratio between the height of the suspected peak and the height of related parent peak. Many of these ratios are lower than 0.05, while a very generous value like 0.15 is used in the ad-hoc rules, as the threshold.

There have been various efforts in recent publications to deal with the issue of stutter peaks with quantitative information. Buckleton et al. [27] emphasised the importance of using peak height or peak areas in analysing DNA mixtures. Evett et al. [58], Gill et al. [82], and Buckleton et al. [25] used quantitative information from peak heights or peak areas to improve interpretation. Neither binary nor semi-continuous models can handle PCR stutter. However, most recent research has focused on continuous models because they avoid "black and white" binary decisions. Buckleton et al. [27] emphasised the importance of understanding the behaviour of non-mixtures before interpreting a potential mixture. According to their recommendation, a dataset consisting of single source profiles is the best to study the behaviour of stutters. In addition, they highlighted the usefulness of analysing stutter in terms of stutter ratio. Although back stutter is the most common phenomenon, forward and double back stutter also have some complicating influence on forensic DNA mixture interpretation.

Cowell et al. [45] presented a methodology based on a gamma distribution, for identification and separation of DNA mixtures. Even though the methodology used the peak area values of epg, it was not capable in taking into account the PCR artefacts including

stutters. Extending the methodology adopted in this study, these researchers have presented another coherent probabilistic framework [46]. Based on the continuous peak area values obtained from the epg, the new probabilistic expert system (PES) addresses the issue of stutters, silent alleles, and allelic drop-outs. According to this PES, the observed peak height of an allele at allele position a is assumed to be affected by two possible effects of stuttering. If there is a back stutter at allele position $a-1$, the peak height of the allele at a can be reduced. However, in case of a back stutter in relation to a parent allele at allele position $a+1$, the peak height of the allele at position a can be increased. Considering these two facts, the observed allele heights at each allele position are revised prior to the likelihood calculations.

Recently Bleka et al. [17] adopted a Bayesian network method that incorporates both allele drop-out and stutters developed by Cowel et al. [44] to model peak height information. In the presence of DNA mixtures of several unknown contributors, this method estimates a number of unknown parameters using maximum likelihood method. In this method, an observed peak height at allele position a is classified as: a stutter, if none of the contributors has allele a, and as a drop-out if at least one contributor has it. This information is included in the network as a binary variable. A revised mixture analysis is then performed conditionally on the pre-classification of alleles as stutters or drop-outs. Even though this is an improved method which is capable in incorporating PCR artefacts including stutters, it still relies on pre-classification of stutters based on threshold values. Therefore, a more robust probabilistic approach for predicting stutters would be advantageous.

Bright et al. [21] investigated the performance of five different statistical models for predicting stutter. Each model was evaluated by applying to three sets of known single source DNA profiles. This study has been reviewed with more details in Chapter 2. In addition, Bright et al. discussed the variance of stutter ratio [20], allelic and stutter peak height models in the context of continuous DNA interpretation methods [23], and the relationship between stutter product and the longest uninterrupted sequence [24]. From theoretical considerations, Weusten and Herbergs [177] recommended that the variance in stutter ratio is inversely proportional to the amount of template DNA. This is often ob-

served and predicted by intuition. Findings of the above four studies are further reviewed and used in Chapter 2.

Forensic computer systems that use MCMC-based statistical models for continuous peak height data to effectively interpret complex DNA evidence have recently been developed (e.g. STRmix™ and Cybergenetics TrueAllele®). Perlin et al. [135] emphasised the possibility of incorporating additional model variables into the TrueAllele® genetic calculator to facilitate the problems associated with PCR process such as stutter. STRmix™ already incorporates statistical models for both forward and backward stutter peaks [22]. Therefore, developing more improved Bayesian probabilistic models for predicting stutter, which can be implemented in these software products will add a significant value to the evidence interpretation.

The main concern of study is stutters and their statistical behaviour. Bright et al. [21] have investigated the performance of five statistical models for predicting the observed behaviour of stutter. Recognising the strengths and weaknesses of them, this study proposes an additional set of Bayesian models including hierarchical and two-component mixture models. There are numerous approaches available for evaluating the performance of Bayesian models. This study reviews different criteria for model evaluation and uses them appropriately for evaluating the existing and proposed Bayesian models for predicting stutters. Finite and infinite mixture models are continuously receiving increasing attention in various fields of science as they provide a greater flexibility in modelling complex data. In this study, infinite mixture modelling approach is used to model stutter ratio with improved accuracy.

1.5 Summary

This chapter discusses the background of the problems associated with interpretation of DNA evidence and reviews important literature related to stutter prediction. The approximate proportional relationship between the amount of template DNA and the peak height information in the epg, the effects of PCR artefacts, the applicability of analytical and stochastic thresholds, and the use of Bayesian approach in the field of evidence interpre-

tation are mainly highlighted. The importance of a sound methodology to incorporate the behaviour of stutters for accuracy of the interpretation is emphasized.

1.6 Organisation of Chapters

The thesis consists of eight chapters. The initial part of the thesis, chapters 2 to 5, includes a discussion of eleven non-hierarchical and seven hierarchical models for predicting stutter, a review of Bayesian model assessment measures, and an evaluation of the performance of the discussed models. The middle part, chapters 6 and 7, explores theoretical aspects of the Bayesian multiple linear regression model assuming a fully conjugate prior, and evaluates the performance of infinite mixtures of linear regression models for stutter prediction. Finally, Chapter 8 combines the results and finding of the study and provides useful suggestions for future research in the context of stutter modelling and infinite mixture models. The specific content of each chapter can be summarized as follows.

Chapter 2 (Beyond the Log-normal) discusses five existing models developed by Bright et al. [21] for predicting stutter and extends this knowledge by introducing six new models. The existing five models were: two log-normal, two gamma, and a two-component log-normal mixture. The new models include two normal, two non-standardized Student's t, and two two-component mixtures of normal and non-standardised Student's t distributions. Before developing the new models the chapter reviews previous research attempts at modelling stutter ratio. Subsequently, it provides relevant technical details about the two sets of data used (NGM SElect^{TM} and Identifiler^{TM}) and explains the basis for introducing improved models. Finally, the chapter examines the variations and relationships among the estimated slopes and intercepts of mean models, and standard deviation parameters of similar models.

Chapter 3 (Measures of Model Assessment) reviews measures available for assessing Bayesian statistical models. Several information criteria, cross-validations and their approximations, and Bayesian p-values are discussed as measures of predictive model accuracy. The usefulness and shortfalls of these measures and the conditions to be satisfied for their use are reviewed, expecting the selection of appropriate measures for evaluating

the models developed in Chapter 2 for predicting stutter.

Chapter 4 (Assessment of Models) tests the performance of the five models developed by Bright et al. [21] and the six new models developed by the present study in Chapter 2. First, the distributional assumptions of the normal and log-normal models are graphically assessed. Second, the models fitted to the two datasets are compared using BIC (Bayesian Information Criterion) while showing the unacceptability of other information criteria. The chapter presents the estimated values of the WAIC (Watanabe Akaike Information Criterion) which is the most widely applicable for any type of model (hierarchical, mixture etc.), and compares the variations of them along with the posterior variances of log-predictive densities. Showing that none of the models fitted in the study satisfy the conditions for using the WAIC, the leave one out cross-validation (LOO-CV) which is the most suitable measure for model evaluation, is approximated as the computational cost of the exact LOO-CV is unaffordable. IS (importance sampling), TIS (truncated importance sampling), and PSIS (Pareto smoothed importance sampling) are the LOO-CV approximations considered.

Chapter 5 (Investigation and Assessment of Hierarchical Models) introduces hierarchical models for four locus-specific variance models (gamma, normal, log-normal, and non-standardised Student's t) and three two-component mixture models (normal, log-normal, and non-standardised Student's t). These seven are non-hierarchical models that the results in Chapter 4 indicated better performance relative to the profile-wide variance models. Providing a detailed introduction to hierarchical models, the chapter discusses the shrinkage of mean model parameters and inverse variance parameters of these models when modelling stutter ratio. Subsequently, the shrinkage of the parameters is examined using credible intervals, and the goodness-of-fits of the parameters to the posterior inferred distributions of the respective parameters are tested. Finally, the empirical cumulative distributions of the parameters, the log-likelihoods, and the log predictive densities are compared across the hierarchical and non-hierarchical models to determine the effects of hierarchical modelling.

Chapter 6 (Bayesian Multiple Linear Regression with a Conjugate Prior Distribution) explains the analytical process of the Bayesian version of multiple linear regres-

sion involving a fully conjugate prior distribution. First, the chapter presents the derivation of the likelihood function and discusses the selection of conjugate prior distribution and combination of them to derive the posterior distribution of model parameters. Second, it derives analytical relationships between the prior information, the observed data, and the parameters of posterior predictive distribution of the data. These theoretical outcomes are useful in the next stage of the study to develop infinite mixtures of linear regression models for stutter prediction.

Chapter 7 (Infinite Mixtures of Linear Regression Models) develops infinite mixtures of linear regression models for predicting stutter ratio using LUS as the explanatory variable. Initially, the theoretical background of finite mixtures with normal densities is reviewed due to their relevance in infinite mixture modelling. Next, the chapter discusses some representations of Dirichlet process, which can be used as non-parametric prior distributions in building infinite mixture models, and introduces some improvements to collapsed Gibbs sampling that uses one of these representations (the Chinese restaurant process) expecting increased accuracy. Finally, it presents the infinite mixture models fitted to the D2S1338 locus in the NGM SElect^{TM} dataset, and evaluates their performance in order to select the best model.

Chapter 8 (Conclusions and Future Work) summarises conclusions and contribution of the overall study that focused on developing Bayesian models for predicting stutter ratio. Providing a brief background to the interpretation of DNA mixture evidence, first this chapter recapitulates the need for investigaing improved models for stutter prediction. Second, it presents the models developed in each research phase following different modelling approaches and the key findings on their performance evaluations. Finally, the chapter suggests possible extensions to the study in relation to stutter prediction as well as the methodologies adopted.

Chapter 2

Beyond the Log-normal

2.1 Introduction

t Statistical modelling of stutters is an indispensable process in continuous DNA interpretation. In this process, a continuous weight is assigned to any suspicious peak that can be treated as a stutter in an epg. Stutters have been discussed in more detail in Chapter 1. Bright et al. [21] investigated the performance of five different statistical models for predicting stutter ratios. The present study extends their work with alternative models and evaluates the performance of both the existing and the proposed models. First, this chapter discusses the methodology used by Bright et al. After providing a brief description of the data used for testing the original five models, six new models are introduced. Finally, the variations of the estimated parameters and their relationships, especially for the similar models, are also discussed.

2.2 Existing Models for Stutter Ratio (SR)

Bright et al. [21] were interested in modelling the observed behaviour of stutter in terms of stutter ratio (SR), which is defined as,

$$SR = \frac{\text{Observed height of the stutter peak}}{\text{The height of the parent allelic peak}}.$$

The linear relationship between SR and the longest uninterrupted sequence (LUS) in an allele has been described in earlier works [10, 136]. LUS has been used as the key determinant in explaining the behaviour of SR, considering the strong evidence available in the literature [23, 24]. When models are derived for a profile-wide relationship of mean SR, the relevant LUS information can be amalgamated as an informative covariate. However, the empirical evidence strongly indicates significant differences in the result with locus-specific mean models. Hence, the mean of the i^{th} stutter ratio at locus l, $\mu_{li} = \mathrm{E}(SR_{li})$, of the relevant model has been modelled in the following form:

$$\mu_{li} = \beta_{0l} + \beta_{1l} LUS_{li},$$

where the intercept and slope parameters of the model for locus l are denoted by β_{0l} and β_{1l} respectively. The LUS value corresponding to the i^{th} observed SR for locus l (i.e. SR_{li}) is denoted by LUS_{li}. Following the method used by Brooks et al [24], the LUS of an allele, which is the longest stretch of basic repeat motifs in it, was determined. The short tandem repeat DNA internet database (STRBase) was used as the source to decide the LUS corresponding to each allele [32, 147]. In a case where multiple LUS values for the reported variants are available, the average of them is used. The alleles whose LUS values are not available have been removed from the datasets. The two datasets that have been used in this study do not have many examples in relation to multiple LUS values. The collection of additional data with multiple LUS values is also impossible as the biological experimentation related to the data generation process is beyond the control of the researcher of this study. However, if there are enough data with multiple LUS values, in a perfect context, a latent class model that assumes LUS as a missing covariate can be used.

Weusten and Herbergs [177] developed a stochastically simple but mathematically complex model to describe the technical events in the background of the PCR amplification process. They assumed three possible outcomes during each PCR cycle: successful amplification, no amplification, and an amplification with a slipped strand. The slipped strand mechanism is a kind of folding of the original strand which causes the introduction of a stutter sequence. The model used a recursive mathematical approach employing a

multinomial distribution to accommodate the three outcomes. It assumes fixed probabilities for the three events over the PCR cycles. The model derived recursive relationships for the expectations and covariances on the number of amplicons with the presence of up to two stutters. The inversely proportional relationship between the coefficient of variation (relative error) and the square root of the expected number of DNA strands entering the amplification is one of the results revealed in this study. In another study, Bright et al. [20], also empirically confirmed the inversely proportional relationship of the variance of stutter height to the template DNA. Employing these theoretical considerations for each model, the variance of *SR* is modelled inversely proportional to the amount of template DNA. Bright et al. [21] proposed two types of models: one with profile-wide variance σ^2 and the other with locus-specific (l) variance σ_l^2. The profile-wide and locus-specific variances of i^{th} observation are denoted by σ_i^2 and σ_{li}^2 respectively. Employing the inversely proportional relationship between the variance of stutter height and the template DNA, these two types of variances are defined as

$$\sigma_i^2 = \frac{\sigma^2}{O_i} \qquad \text{and} \qquad \sigma_{li}^2 = \frac{\sigma_l^2}{O_{li}},$$

where O_i denotes the observed peak height of the i^{th} stutter ratio (SR_i) and O_{li} denotes the locus-specific (l) observed peak height of the i^{th} stutter ratio (SR_{li}). In the standard practice of model fitting, it is essential to assume a suitable family of distributions to describe the behaviour of data. Subsequently, the parameters of the model, usually the mean and variance, are estimated following a standard estimation method. Maximum likelihood and Bayesian methods are frequently used for parameter estimation.

In another study by Bright et al. [23], the observed *SR* has been treated as a positively skewed random variable with long tails. In this study they have used a log-normal distribution to model the behaviour of observed *SR*. Based on this study, two of the five models have been proposed for their later study [21] along with log-normal distributions: one with profile-wide variance (LN_0) and the other with locus-specific variance (LN_1). In describing the mean, both models have assumed locus-specific mean models. The specific parametrisation of these two models are given in the Table: 2.1. Log-normal distribution can only be used with positive quantities. Therefore, the use of log-normal distribution in

SR model fitting fully guarantees the strictly positive characteristic in predictions.

The family of gamma distributions also provides much flexibility in modelling heavy tailed positively skewed distributions. Hence, two gamma models have been introduced as competing alternatives to the log-normal models. As it does with the log-normal, the profile-wide variance model in gamma distribution, G_0 is distinguished from its locus-specific variance model G_1 by different parametrisations in variance terms. The natural logarithm is a useful link function for gamma generalised linear models. Hence, the logarithm of the mean is modelled as a linear function of LUS. Consequently, it confirms a strictly positive mean for the model. The parametrisation of G_0 and G_1 models are also given in Table: 2.1.

Bright et al. [21] have proposed a two-component log-normal mixture model (MLN_1) for *SR* as the fifth alternative. It can also be regarded as a two-component normal model for the logarithm of *SR*. It has been introduced in order to provide more robust modelling when there exist outlier type stutter ratios as well. The mean of each component has assumed a common locus-specific mean model. As showing in Table: 2.1, these mixture models assume two different locus-specific variances for the two components. A majority of the stutter ratio values are expected to be modelled by the component with smaller variance. The rest of the stutter ratio values which are a long way from the mean are modelled with the component having larger variance. Mixtures of normal distributions with unique means but different variances are generally known as normal scale mixtures [123]. The family of two-component normal mixtures is the simplest class of normal scale mixtures. In 1960, Tukey has also discussed the two-component heteroscedastic mixtures of normal densities with equal means under the family of contaminated normal distributions [166]. In this family, the component with higher variance is connected to the model with a lower weight than the other.

The parameter estimation of gamma models is slightly more complex compared to the log-normal models because they require a translation from a shape and scale parametrisation to a mean-variance parametrisation. The family of gamma distributions can be expressed with two parameters: shape parameter α and scale parameter θ (or rate parameter $\beta = \frac{1}{\theta}$). Then the mean and variance of the gamma distribution are $\alpha\theta$ and $\alpha\theta^2$

respectively. Both of these parameters can be empirically calculated in terms of estimated mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ as,

$$\hat{\alpha} = \frac{\hat{\mu}^2}{\hat{\sigma}^2} \qquad \text{and} \qquad \hat{\theta} = \frac{\hat{\sigma}^2}{\hat{\mu}}.$$

This therefore suggests a natural mean-variance relationship.

Table 2.1: Descriptions of existing models

Model	Distribution	Mean	Variance
LN_0	$ln(SR_{li}) \sim N(\mu_{li}, \sigma_i^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_i^2 = \frac{\sigma^2}{O_{ai}}$
LN_1	$ln(SR_{li}) \sim N(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
G_0	$SR_{li} \sim Gamma(\alpha_{li}, \theta_{li})$	$\mu_{li} = exp(\beta_{0li} + \beta_{1li} LUS_{li})$	$\sigma_i^2 = \frac{\sigma^2}{O_{ai}}$
G_1	$SR_{li} \sim Gamma(\alpha_{li}, \theta_{li})$	$\mu_{li} = exp(\beta_{0li} + \beta_{1li} LUS_{li})$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
MLN_1	$ln(SR_{li}) \sim \pi N(\mu_{li}, \sigma_{0li}^2) +$ $(1-\pi) N(\mu_{li}, \sigma_{1li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{qli}}$ $\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$

2.3 The Data used for Testing the Models

The performance and consistency of the five models proposed by Bright et al. [21] were tested with three datasets. The present study uses two of these three datasets to compare the performance of the same five models and six new models proposed in the next section based on some additional performance criteria. The technical description of the two datasets: NGM SElect™ and Identifiler™ provided by Bright et al. [21] is given in the following paragraph.

Single source saliva stains on FTA® Elute cards (Whatman, Maidstone, UK) were extracted using an automated elute method [131]. A target amount of 1 ng DNA was amplified using the Applied Biosystems NGM SElect™ (N = 291) and Identifiler™ (N = 341) multiplexes (Life Technologies, Carlsbad, CA) and run on a 9700 silver block following the manufacturer's recommended protocols. All Quantifiler™ and STR amplifications were set up on a Hamilton Nimbus (Hamilton, Reno, NV, USA) liquid handling

robot. All samples were run on an ABI PRISM 3130xl capillary electrophoresis instrument and analysed using GeneMapper ID v 3.2.1, with a 30 rfu analysis threshold. The analytical threshold of 30 rfu used for data analysis is lower than that used normally for casework to avoid bias. In addition, only samples with allele heights greater than or equal to 500 rfu were selected for the creation of the models. This was to remove any bias towards alleles more likely to stutter.

The two sets of multiplexes NGM SElect^{TM} and Identifiler^{TM} return DNA profiles for 16 and 15 loci respectively. Ten of these loci: D16S539, D18S51, D19S433, D21S11, D2S1338, D81179, FGA, TH01, vWA and D3S1358 are common for both sets. The DNA profiles of NGM SElect^{TM} set contain additional six loci: D10S1248, D22S1045, D2S441, D1S1656, D12S391, and SE33. The additional loci of DNA profiles in the Identifiler^{TM} set are: CSF1PO, D13S317, D5S818, D7S820, and TPOX. The peak heights of homozygous alleles are approximately twice the heights of the corresponding heterozygous alleles. Hence, the stutter ratios related to homozygous alleles are typically smaller than that of heterozygous alleles. Therefore, it will be more beneficial to build statistical models for stutter ratios related to homozygous and heterozygous alleles separately. The scope of this study is limited only to the stutter ratios related to heterozygous alleles. Hence, the datasets are stutter ratios of alleles from only the loci where the individuals are heterozygous. Setting of 30 rfu as the analytical threshold and removal of allele peaks whose heights are less than 500 rfu were the two filtering criteria that ensure the observed peaks as stutters.

2.4 The Basis for New Models

Usually, the magnitude of stutter ratio is expected to below 0.15 [31] and in fact is usually below 0.05. However, higher stutter ratios can also be occasionally expected. Therefore, a heavy-tailed right skewed distribution such as gamma and log-normal would be theoretically more appropriate in modelling *SR*. The family of Gaussian distributions is the most popular statistical distribution family in modelling continuous data. This distribution is capable of accommodating the bulk of the data around its mean. However, relatively large

or extreme observations can also be captured. Normal distribution assigns relatively low tail probabilities on extreme values than alternative heavy-tailed right-skewed counterparts. Therefore, in case of infrequent large observations of *SR*, a better goodness-of-fit can be expected even from the Gaussian distributions.

When the occurrence of large *SR* values is more frequent than expected, a non-standardised Student's t distribution can be used as a potential option than a Gaussian family distribution. The non-standardised Student's t distribution is defined with three parameters: location (μ), scale ($\sigma > 0$), and degrees of freedom ($\nu > 0$). The additional parameter, degrees of freedom, enables an extensive flexibility in robust modelling of data. The well-known Student's t distribution is symmetrically distributed around its mean $\mu = 0$, while having a scale parameter, $\sigma = 1$. The non-standardised Student's t distribution is sometimes discussed as the general form of it [1]. Let us assume that T is a random variable that has a non-standardised Student's t distribution with location, scale, and degrees of freedom parameters μ, σ, and ν respectively. Then the density function of T, $f_T(t)$, is in the form,

$$f_T(t) = \frac{1}{\sigma\sqrt{\nu}B\left(\frac{\nu}{2},\frac{1}{2}\right)}\left[1+\frac{1}{\nu}\left(\frac{t-\mu}{\sigma}\right)^2\right]^{-\frac{(1+\nu)}{2}},$$
$$-\infty < t < \infty, \quad \nu > 0, \quad \sigma > 0,$$

and $B(\cdot,\cdot)$ denotes the complete beta function. The mean of the distribution is equal to its location parameter (i.e. $\mathrm{E}(T) = \mu$) for $\nu > 1$. When $\nu = 1$, the distribution reduces to the Cauchy distribution. The variance of distribution ($\frac{\nu\sigma^2}{\nu-2}$) exists only when $\nu > 2$.

Simple linear regression models were fitted to the means of the three log-normal models, taking *LUS* as the predictor. Meanwhile, a log link function was used in the two gamma models. Hence, both types of models assume a simple linear regression model on logarithm of *SR* against *LUS*. In contrast, literature [23, 24] suggests a linear relationship between *LUS* and *SR*. Therefore, the strength of the relationship between *LUS* and *LR* with and without logarithmic transformation would be interesting. For the NGM SElect$^{\mathrm{TM}}$ dataset, 69.5% of the total variation in *SR* can be explained using *LUS* as a predictor. However, only 58.9% of the total variation in $\ln(SR)$ can be explained with *LUS*.

For the Identifiler^{TM} dataset, 64.5% of the total variation in SR and 61.7% in $\ln(SR)$ can be explained with LUS. Therefore, a better performance can be expected with normal and non-standardised Student's t distributions over log-normal and gamma distributions.

Considering all these facts and expecting an improvement in stutter prediction, two models with normal distributions and two models with non-standardised Student's t distributions are proposed for modelling SR. Bright et al. [21] have observed a better performance of two-component mixture models in modelling SR. Therefore, two two-component mixtures: normal and non-standardised Student's t are also proposed. The parametrisation of the proposed six models are summarised in Table 2.2.

Table 2.2: Descriptions of proposed models

Model	Distribution	Mean	Variance
N_0	$SR_{li} \sim N(\mu_{li}, \sigma_i^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_i^2 = \frac{\sigma^2}{O_{aj}}$
N_1	$SR_{li} \sim N(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{alj}}$
T_0	$SR_{li} \sim t(\mu_{li}, \sigma_i^2, \nu)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_i^2 = \frac{\sigma^2}{O_{aj}}$
T_1	$SR_{li} \sim t(\mu_{li}, \sigma_{li}^2, \nu_l)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{alj}}$
MN_1	$SR_{li} \sim \pi N(\mu_{li}, \sigma_{0li}^2) +$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{alj}}$
	$(1-\pi) N(\mu_{li}, \sigma_{1li}^2)$		$\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$
MT_1	$SR_{li} \sim \pi t(\mu_{li}, \sigma_{0li}^2, \nu_{1l}) +$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{alj}}$
	$(1-\pi) t(\mu_{li}, \sigma_{1li}^2, \nu_{2l})$		$\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$

Note: ν is profile-wide and ν_l, ν_{1l}, and ν_{2l} are the locus-specific (l) degrees of freedom of the t distributions.

2.5 Model Fitting

The models were fitted using a Bayesian approach along with Markov Chain Monte Carlo (MCMC) techniques. The package "rjags" (version 4.2.0) was used since it provides an interface from statistical package R (version 3.2.2) for Bayesian data analysis. In Bayesian model fitting, it is required to assume suitable prior distributions for the model parameters. Since the prior information related to the model parameters were not available, this study attempted to use vague prior distributions. However, the effect of vague

prior distributions is minimised against the size of datasets used in this study. Normal vague prior distributions were assumed for the slope and intercept parameters of the simple linear regression models related to the mean of each model. Vague inverse gamma prior distributions were assumed for the variance parameters in each model. The degrees of freedom parameters in the respective Student's t models were modelled with log-uniform prior distributions. Mixing proportion of each mixture model is modelled with a uniform prior. After 50000 burn-in steps, each model ran for another 50000 iterations with a thinning interval of 25. Finally, the parameters of the models were estimated over 2000 posterior draws.

2.6 Variations and Relationships among the Parameters of Similar Models

For both log-normal and gamma models, the locus-specific mean of the stutter ratio measured in a logarithmic scale is modelled as a simple linear regression of LUS. Therefore, slopes and intercepts of these models are theoretically comparable. As shown in Figures: 2.1 to 2.4, only a moderate concordance can be seen in the estimates of slope and intercept parameters between log-normal and gamma models for each dataset. The slope parameters of these models are varying within an approximate range of [-7, -3]. Similarly, the intercepts are varying within an approximate range of [0.05, 0.40]. Significant differences among both slopes and intercepts across different loci for both types of models for both datasets can be observed. However, any locus-specific significant difference within each dataset cannot be detected in either slope or intercept of log-normal models except for TPOX locus for the Identifiler^{TM} dataset. For this particular locus, both slope and intercept of the mean model show significant differences in mixture model MLN_1 compared to the other two log-normal models. Similarly, for the gamma models, only the TH01 locus for the NGM SElect^{TM} dataset shows a significant difference between two models for both parameters.

In both normal and Student's t models, unlike log-normal and gamma models, the mean stutter ratio is directly modelled as a simple linear regression of LUS. Figures 2.5

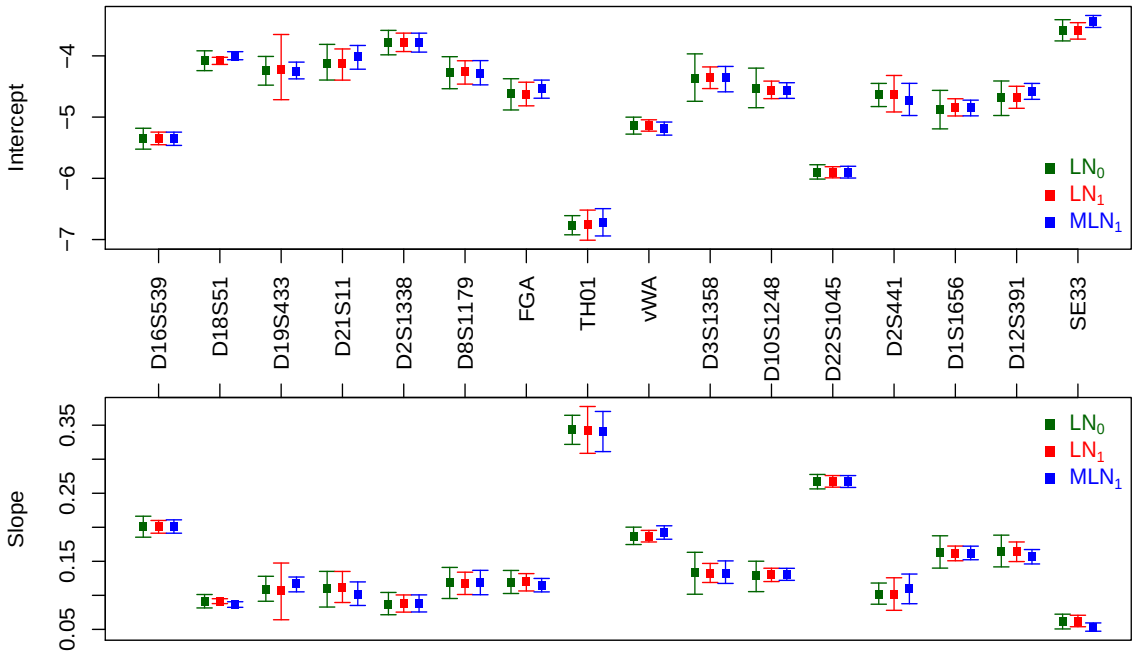

Figure 2.1: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the log-normal models for the NGM SElect^{TM} dataset.

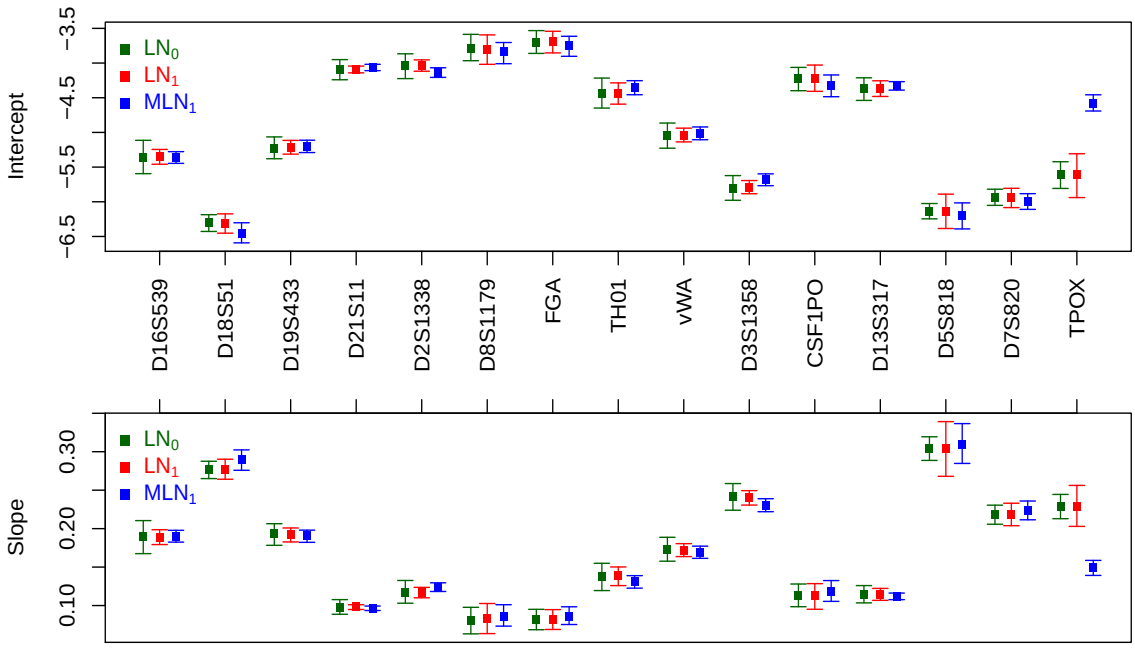

Figure 2.2: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the log-normal models for the Identifiler^{TM} dataset.

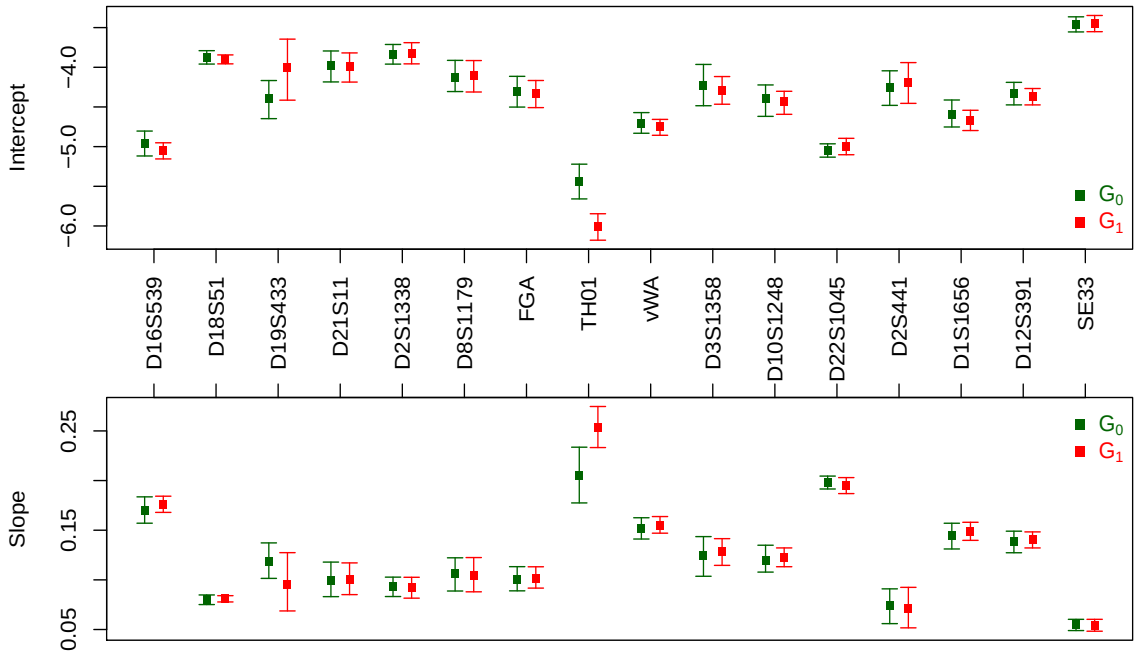

Figure 2.3: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the gamma models for the NGM SElect^{TM} dataset.

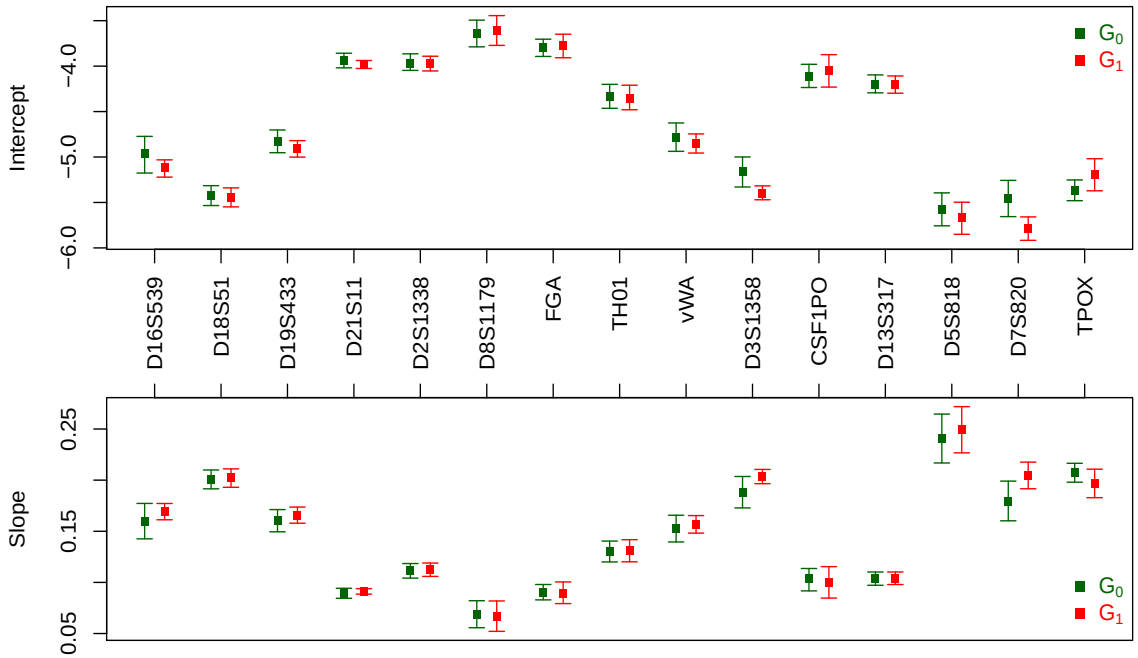

Figure 2.4: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the gamma models for the Identifiler^{TM} dataset.

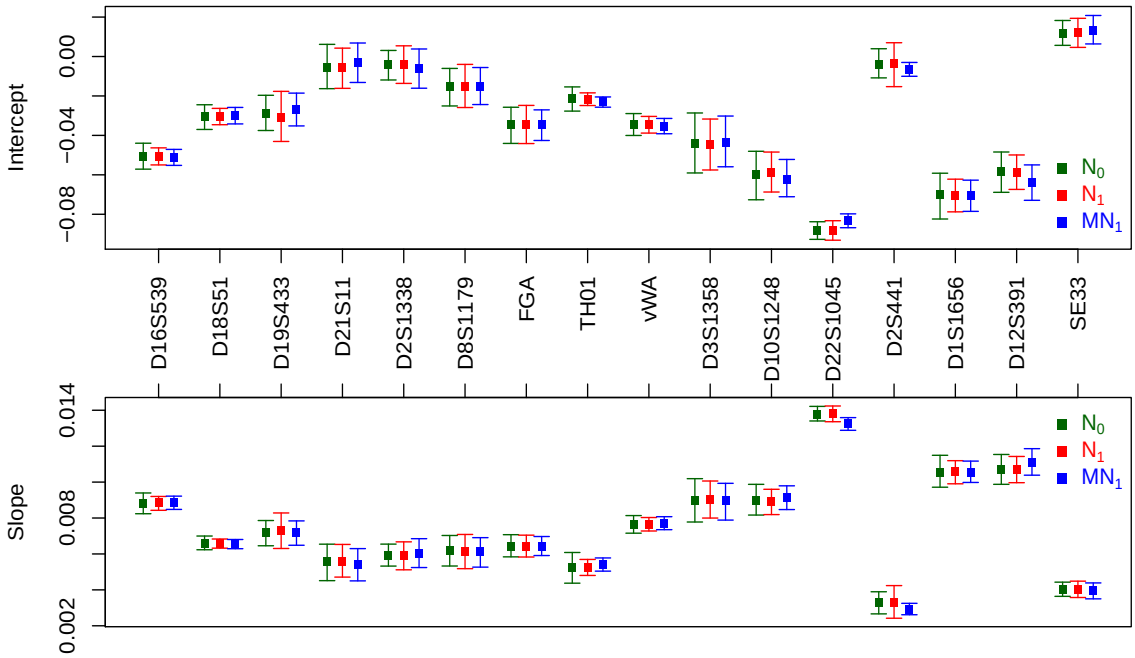

Figure 2.5: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the normal models for the NGM SElect^{TM} dataset.

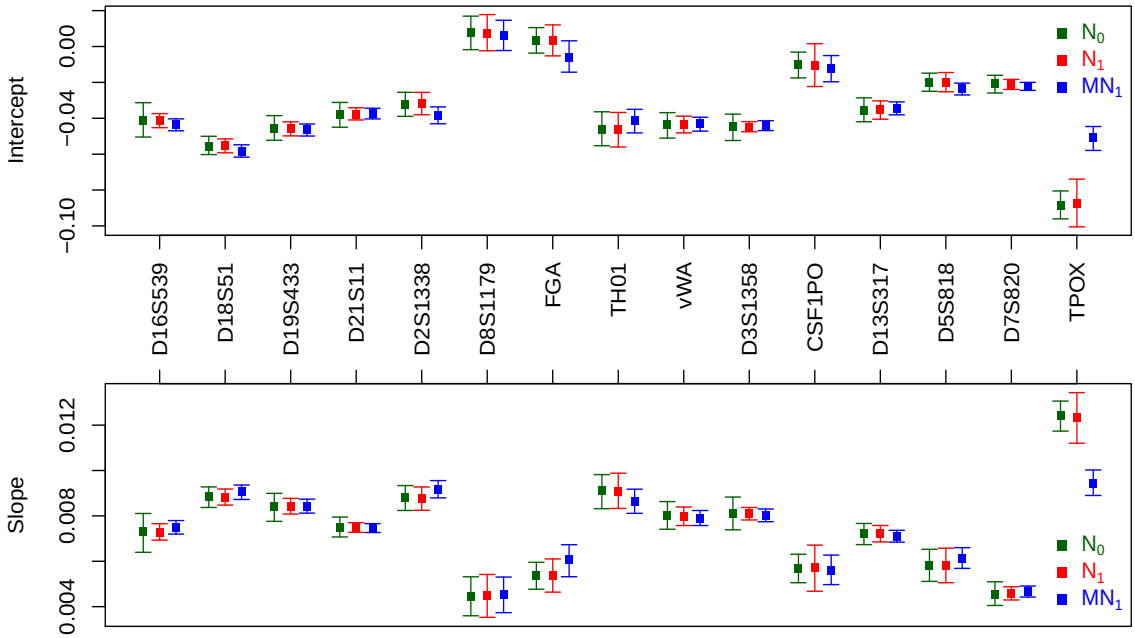

Figure 2.6: The locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) under normal models for the Identifiler^{TM} dataset.

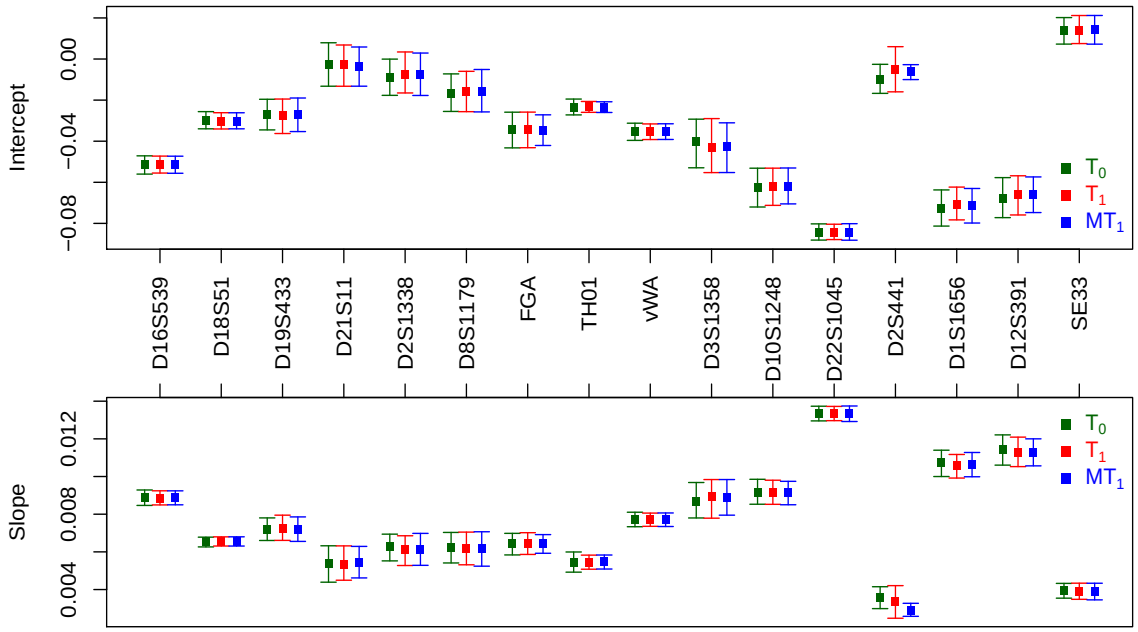

Figure 2.7: Locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) of the Student's t models for the NGM SElect^{TM} dataset.

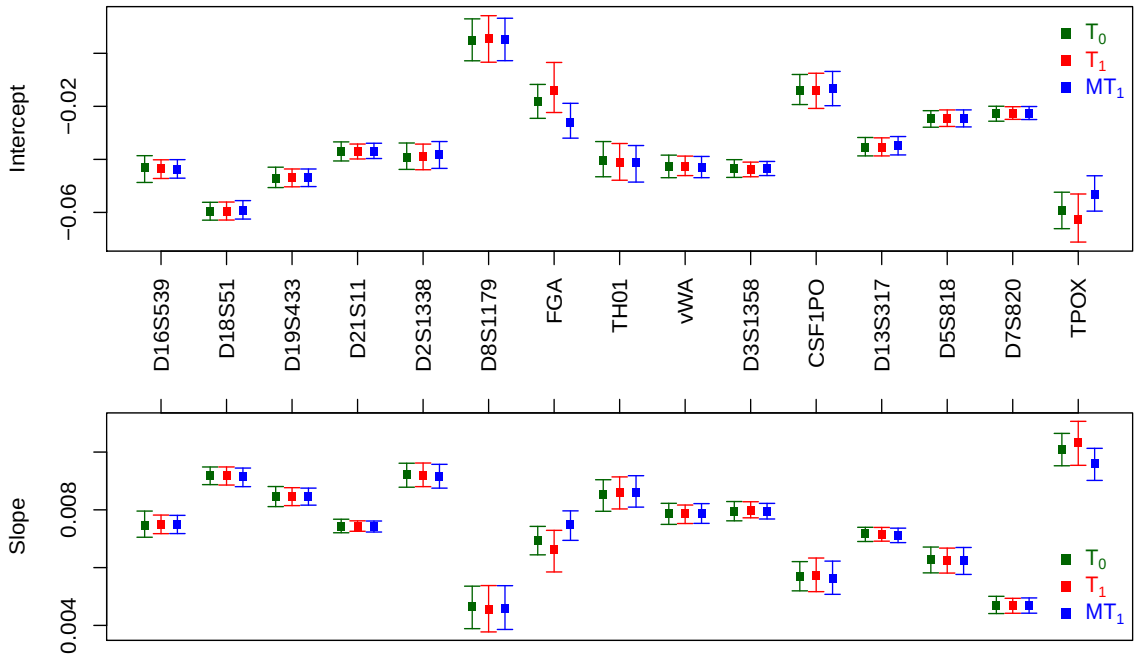

Figure 2.8: The locus-specific variation (95% credible interval with posterior median) in mean model parameters (slope and intercept) under Student's t models for the Identifiler^{TM} dataset.

to 2.8 clearly show a greater concordance in the estimates of slope and intercept parameters between normal and non-standardised Student's t models for each dataset. This is expected in theoretical point of view as both normal and Student's t distributions are symmetric around their location parameters. The intercept and slope parameters calculated for both datasets under each model vary approximately in regions of [-0.10, 0.02] and [0.002, 0.014] respectively. For both models, as it does with log-normal and gamma models, there are significant differences among locus-specific slopes and intercepts over both datasets. Any significance difference cannot be expected in either slopes or intercepts in normal models or Student's t models for both datasets except TPOX locus for the Identifiler^{TM} dataset. The normal mixture model, like log-normal mixture for the Identifiler^{TM} dataset also exhibits a significantly different slope and intercept parameters for TPOX locus compared to the other two normal models.

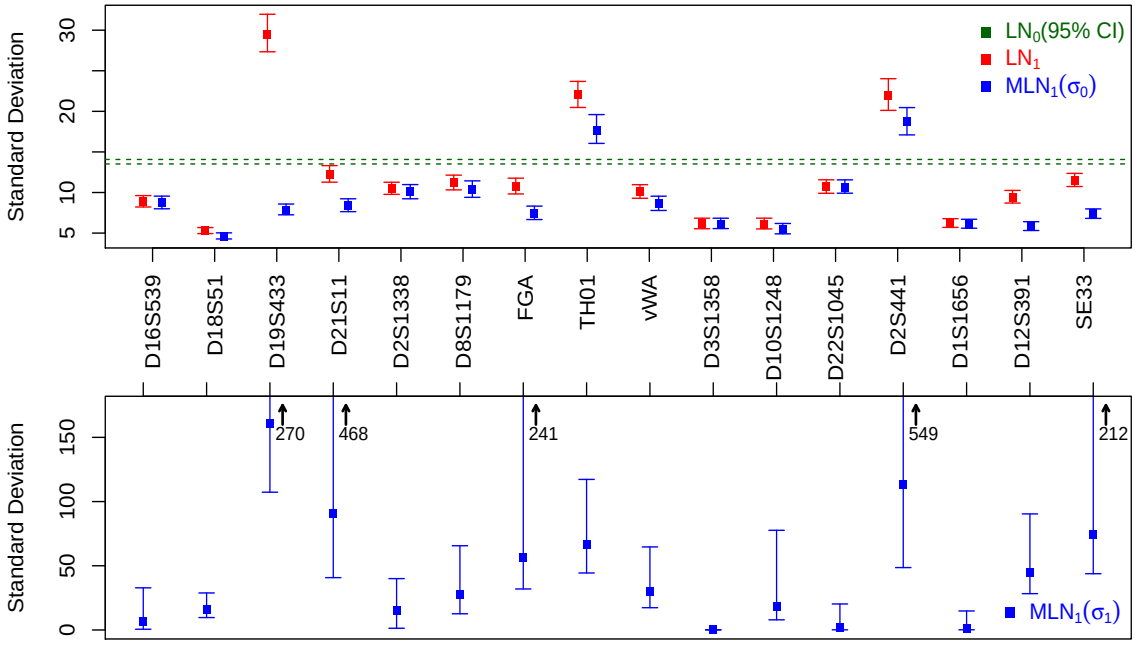

Figure 2.9: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the log-normal models for the NGM SElect^{TM} dataset.

Figures: 2.9 to 2.16 clearly demonstrate outstanding deviations of the locus-specific standard deviations from the profile-wide standard deviation for all the four models for both datasets. The second component of each mixture model has been introduced to capture the statistical behaviour of the data points that are largely deviated from their

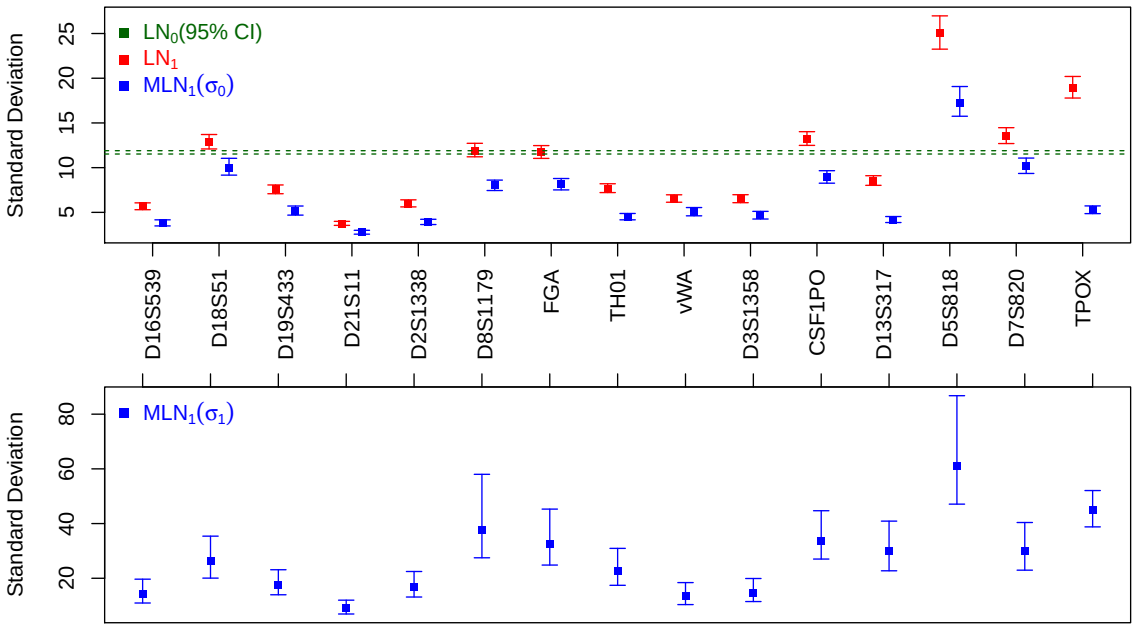

Figure 2.10: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the log-normal models for the Identifiler^{TM} dataset.

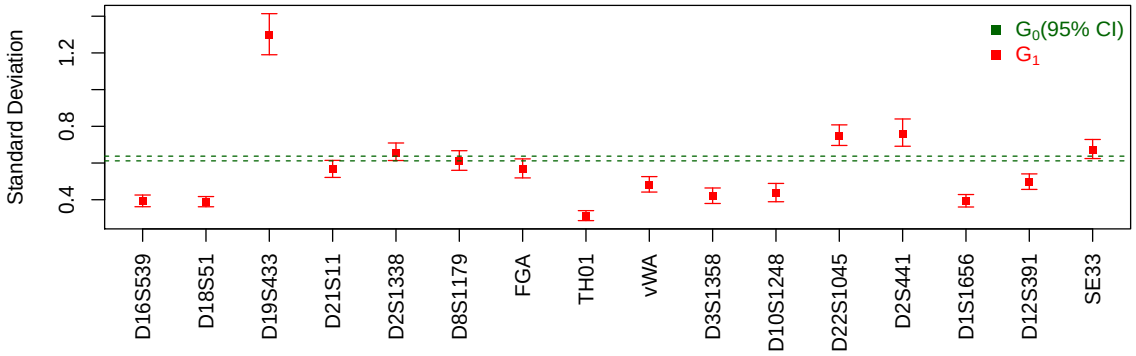

Figure 2.11: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the gamma models for the NGM SElect^{TM} dataset.

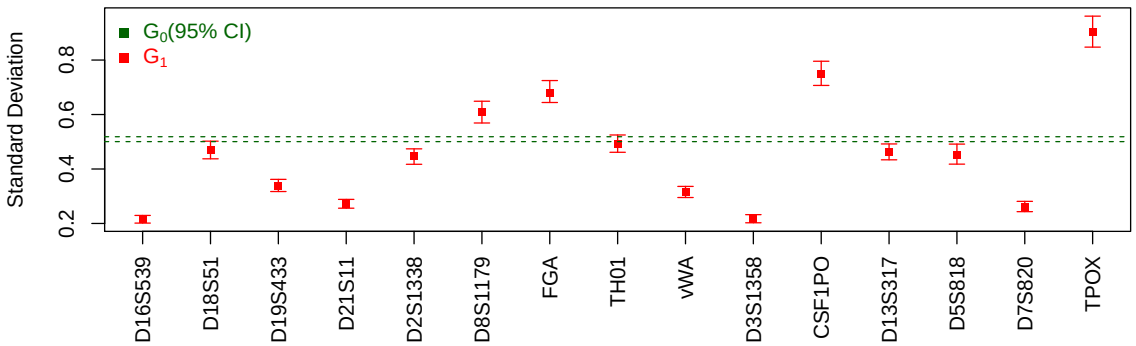

Figure 2.12: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the gamma models for the Identifiler^{TM} dataset.

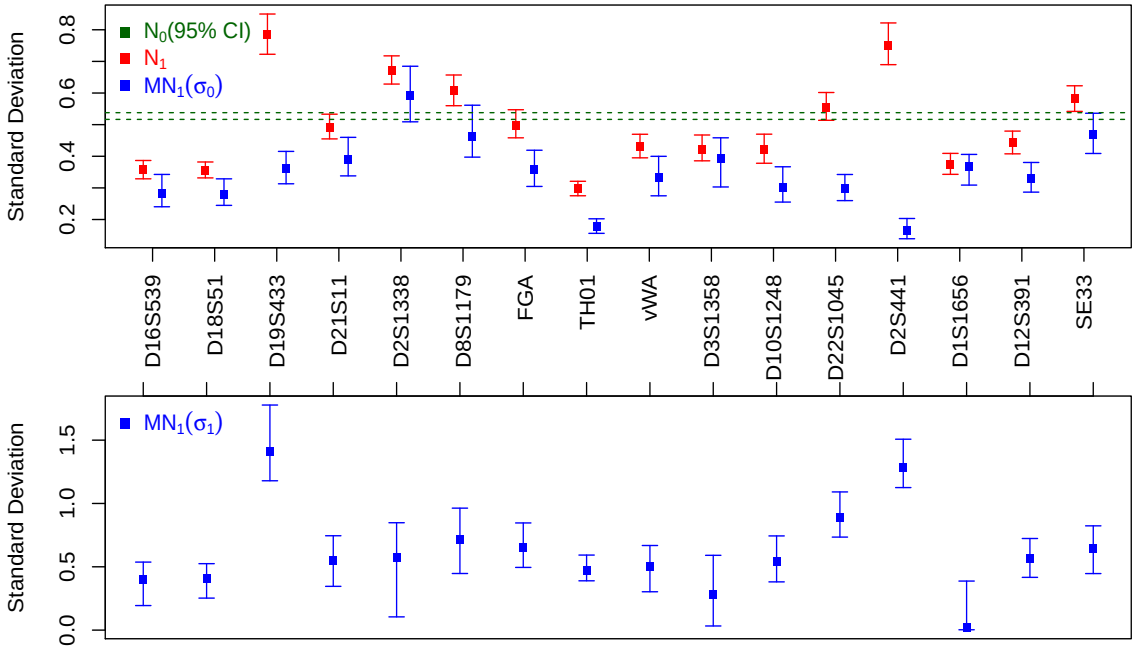

Figure 2.13: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the log-normal models for the NGM SElect^{TM} dataset.

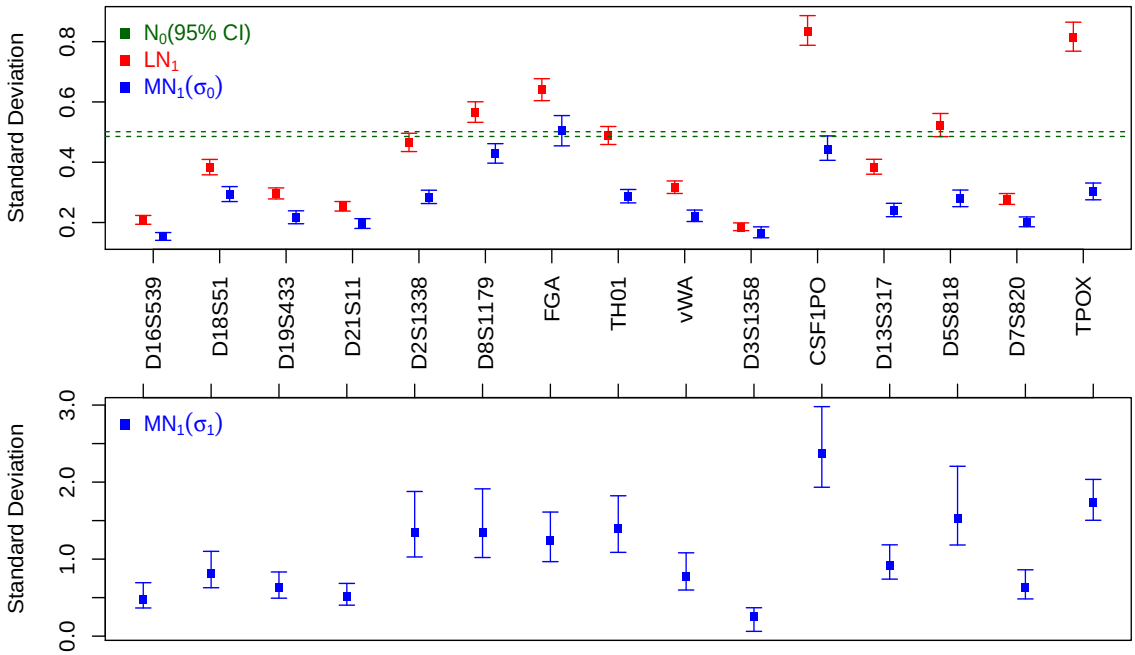

Figure 2.14: The locus-specific variation (95% credible interval with posterior median) in standard deviation parameters under log-normal models for the Identifiler^{TM} dataset.

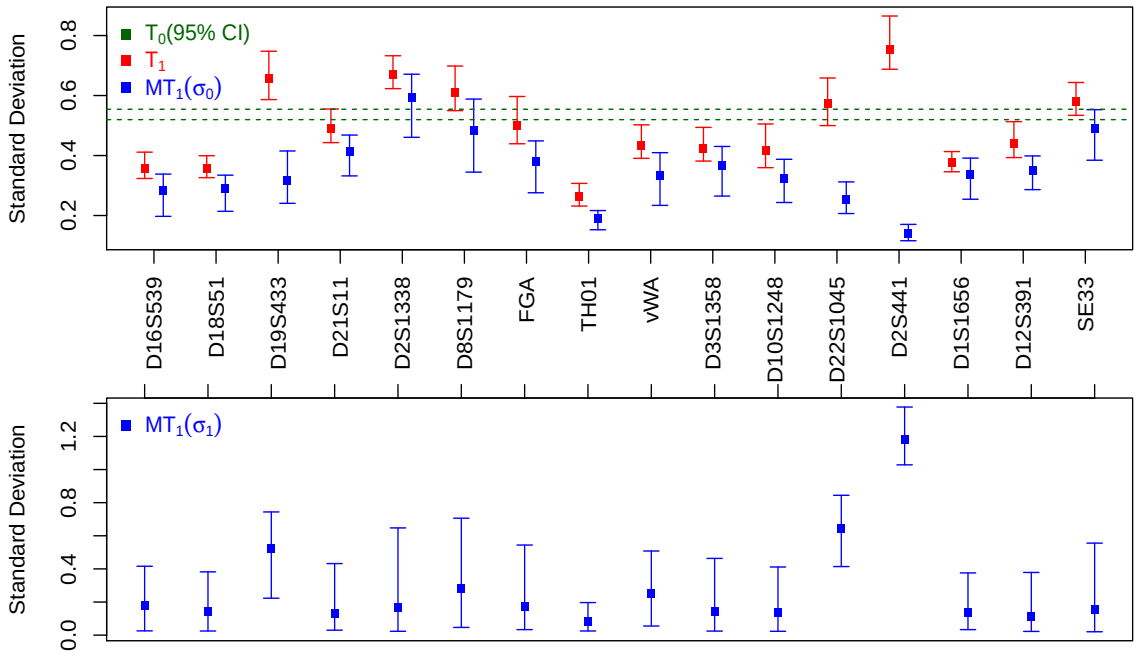

Figure 2.15: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the log-normal models for the NGM SElect^{TM} dataset.

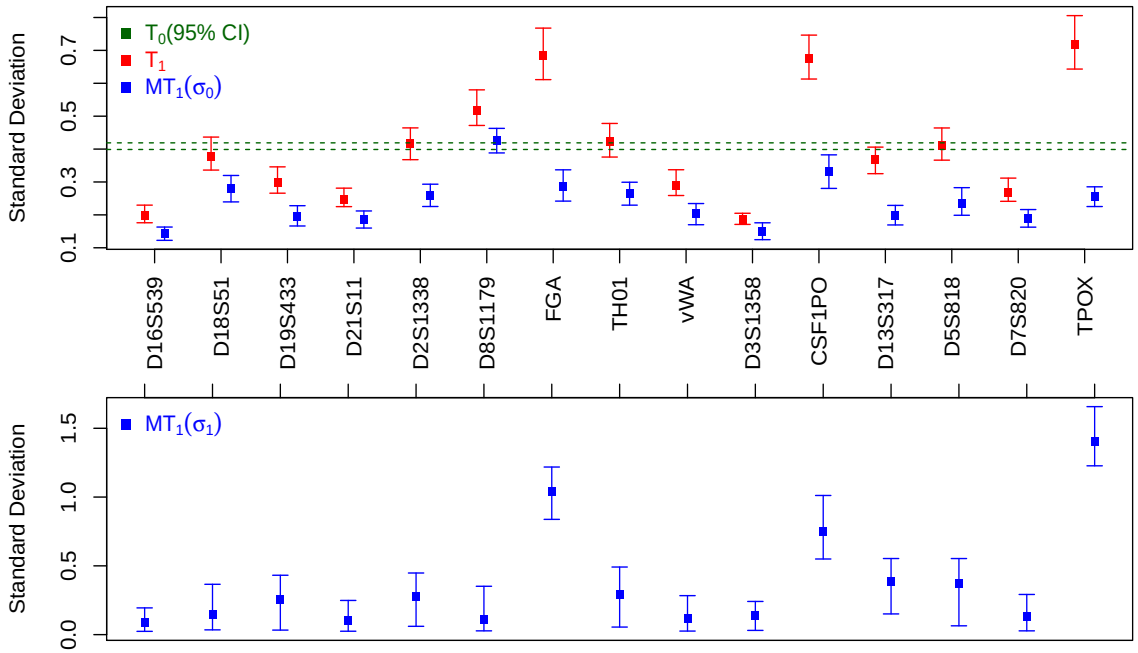

Figure 2.16: Locus-specific variation (95% credible interval with posterior median) in standard deviation parameters of the log-normal models for the Identifiler^{TM} dataset.

Table 2.3: Mixing percentages of the mixture distributions.

Mixture Model	Dataset NGM SElect™	Identifiler™
Log-normal (MLN_1)	2.7 [1.8, 3.8]	9.1 [7.8, 10.5]
Normal (MN_1)	29.0 [24.1, 34.4]	10.2 [8.5, 12.0]
Student's t (MT_1)	48.2 [40.8, 57.1]	31.1 [24.8, 37.4]

Note: Percentage (with 95% credible interval) of points modelled by the component with larger variance in each mixture model is given.

mean. The standard deviation of the component with low variability is denoted by σ_1 and the other by $\sqrt{\sigma_1^2+\sigma_2^2}$ for all the three mixture models. Hence, the large estimates of σ_2 clearly indicate the presence of highly deviated values from the mean of the respective model. The log-normal model exhibits larger credible intervals for standard deviations for the second component under some loci for the NGM SElect™ dataset.

Table: 2.3 presents the percentage of stutter ratios explained with the component of larger variance. It is relatively higher for the Student's t mixture model for both datasets. For each dataset, it approximately captures additional 20% of the stutter ratios over the normal mixture models .

The variations of the degrees of freedom parameters in the Student's t models are presented in Figure 2.17, Figure 2.18, and Table 2.4. The profile-wide variance models exhibit heavy-tailed behaviours as their degrees of freedom parameters are consistently smaller for all the loci for both datasets. In general, locus-specific variance models fitted to the Identifiler™ dataset demonstrate more heavy-tailed behaviour than that of the NGM SElect™ dataset. Even in mixture distributions, both Student's t components generally exhibit more heavy-tailed behaviour than normal distributions with similar scale and location parameters, as their degrees of freedom parameters are smaller than 30 in many situations. However, larger upper bounds of the credible intervals calculated for both degrees of freedom parameters illustrate the possibility of having approximately normal like tail behaviours.

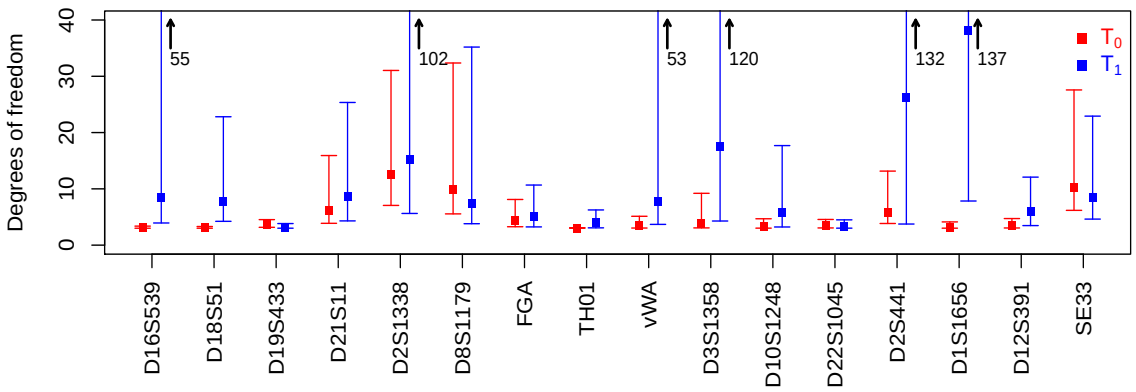

Figure 2.17: Locus-specific variation (95% credible interval with posterior median) in degrees of freedom parameters of the Student's t non-mixture models for the NGM SElect^{TM} dataset.

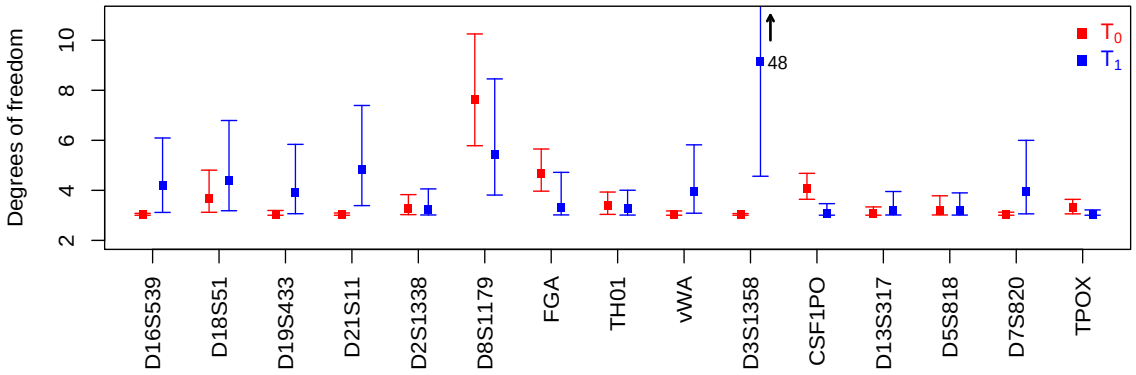

Figure 2.18: Locus-specific variation (95% credible interval with posterior median) in degrees of freedom parameters of the Student's t non-mixture models for the Identifiler^{TM} dataset.

2.7 Summary

This chapter evaluated five existing models (log-normal and gamma) developed by Bright et al.[21] for predicting stutter ratio and introduced six new models (normal and non-standardised Student's t) expecting improved performance. Based on the characteristics of variance modelling, these 11 models were classified into three categories: profile-wide variance, locus-specific variance, and two-component mixture with heteroscedastic variances. Log-normal, normal, and non-standardised Student's t distributions were used with all the three categories. However, gamma distribution was used only with locus-specific and profile-wide variance categories. The two sets of data: NGM SElect^{TM} and Identifiler^{TM} that include stutter peak information related to 4646 and 6949 heterozygous loci respectively, were utilized in the analysis.

Table 2.4: Variation in degrees of freedom of non-standardised Student's t mixture models

	NGM SElect™				Identifiler™			
Locus	ν_1		ν_2		Locus	ν_1		ν_2	
D16S539	12	[3, 127]	17	[4, 128]	D16S539	17	[3, 131]	4	[3, 112]
D18S51	10	[3, 123]	14	[4, 132]	D18S51	14	[3, 127]	5	[3, 118]
D19S433	22	[3, 134]	3	[3, 5]	D19S433	19	[3, 138]	6	[3, 95]
D21S11	12	[3, 124]	13	[4, 129]	D21S11	7	[3, 126]	6	[3, 122]
D2S1338	17	[4, 128]	23	[5, 131]	D2S1338	38	[3, 139]	3	[3, 96]
D8S1179	14	[3, 131]	16	[4, 126]	D8S1179	65	[12, 144]	4	[3, 6]
FGA	8	[3, 121]	8	[3, 117]	FGA	9	[3, 120]	40	[7, 138]
TH01	8	[3, 125]	5	[3, 129]	TH01	34	[3, 135]	3	[3, 110]
vWA	12	[3, 132]	7	[4, 131]	vWA	5	[3, 122]	5	[3, 112]
D3S1358	23	[4, 134]	26	[4, 135]	D3S1358	16	[4, 124]	26	[4, 139]
D10S1248	12	[3, 135]	8	[3, 127]	CSF1PO	12	[3, 128]	3	[3, 5]
D22S1045	38	[5, 137]	11	[3, 119]	D13S317	27	[3, 138]	4	[3, 11]
D2S441	34	[4, 137]	63	[13, 142]	D5S818	27	[3, 139]	3	[3, 60]
D1S1656	28	[5, 139]	39	[6, 140]	D7S820	18	[3, 136]	4	[3, 114]
D12S391	9	[3, 129]	8	[3, 127]	TPOX	44	[7, 141]	59	[12, 142]
SE33	11	[3, 131]	12	[4, 136]					

Cell contents: Posterior median and 95% credible interval

All the normal and Student's t models clearly show a higher concordance in slopes and intercept parameters for each dataset. However, only a moderate concordance in these parameters has been observed between log-normal and gamma models for each dataset. There are significant differences among locus-specific slopes and intercepts estimated within each of the 11 models for the two datasets. Locus-specific slopes and intercepts of mean models fitted based on each distribution do not show significant differences among them except TPOX locus of the Identifiler™ dataset for normal and log-normal models and TH01 locus of the NGM SElect™ dataset for gamma models. With regard to variability parameters of non-mixture models, the locus-specific standard deviations demonstrate outstanding deviations from the profile-wide standard deviation for all the four models for both datasets. The standard deviations observed for the components with larger variances are relatively higher than that was expected for all the mixture models fitted to both datasets. Percentage of points modelled by the component with larger variance in each mixture model was examined and it was found that log-normal, normal, and non-standardised Student's t mixture models capture approximately 3%, 29%, and

48% respectively from the NGM SElect^{TM} dataset. The respective percentages for the Identifiler^{TM} dataset were 9%, 10%, and 31%. The Student's t mixture model approximately captures additional 20% of the stutter ratios over the normal mixture models. For all the loci of both datasets, the profile-wide variance non-standardised Student's t models exhibit heavy-tailed behaviours as their degrees of freedom parameters are smaller. In the locus-specific variance non-standardised Student's t models, more heavy-tailed behaviour was observed for the Identifiler^{TM} dataset. Although both components of the Student's t mixture model exhibit more heavy-tailed behaviour than normal distributions with similar scale and location parameters, larger upper bounds of the credible intervals for both degrees of freedom parameters indicate the possibility of having approximately normal like tail behaviours.

Chapter 3

Measures of Model Assessment

3.1 Introduction

This chapter describes the methods that will be used in Chapter 4 and 5 to assess the performance of the Bayesian models developed in Chapter 2 (non-hierarchical models including two-component mixtures) and Chapter 5 (hierarchical models including two-component mixtures) for predicting stutter ratio. The theoretical background, benefits, and limitations of various performance measures are reviewed in order to identify appropriate measures for evaluating the Bayesian statistical models presented.

Statistical models, in general, are developed based on few fundamental assumptions. The distributional assumption on the data, for instance, plays a key role in the plausibility of the inference that is based on the fitted model. Models incorporated in Bayesian data analysis are also subjected to key assumptions. Hence, a model that exhibits poor plausibility tends to produce misleading inferences. Therefore, an assessment of these assumptions is always a good practice. Generally, it is essential to check the statistical capability of a model in order to produce a realistic summary of the data at hand [75]. In the classical (frequentists') approach, comparisons between the observations and the predictions (expected results under the model) are used as the basis of goodness-of-fit tests that quantify the inconsistency in terms of a probability value (p-value). In the context of Bayesian data analysis, the posterior predictive distribution, which describes the characteristics and statistical behaviour of unobserved future observations conditioned on the

observed real data at hand, is used to answer the prediction problems [174].

In general, a statistical model is a probabilistic system that involves a probability distribution or a finite/infinite mixture of distributions. These models are widely used in explanation, prediction, or making inferences on some real-world phenomena. It is possible to approximate a given phenomenon with more than one model. Accordingly, the complexity of a model can vary from simple to very complex. Very complex models may include very large number of parameters. A fully non-parametric model, for example, may consist of enormous number of parameters. Models where the number of parameters can grow with the size of training data set are more appropriately referred to as non-parametric. In 1976, George E. P. Box stated, "**all models are wrong, but some are useful**" [19]. This is a widely believed fact in modelling and hence, no single statistical model is able to capture the real mechanism behind naturally generated data [174]. However, a model that is rich enough to approximate the behaviour of data including essential uncertainties is generally accepted as a good model. Usually, it is more convenient to build different models based on one particular distribution (e.g. regression models with normal distribution). A set of such models can be easily compared using an appropriate criterion. However, in situations where the models have originated from different distributions, the comparisons are quite interesting. This becomes further complicated with the use of different models adopting various modeling concepts. For example, a situation that requires selecting one out of a set including hierarchical, mixture, and hierarchical mixture models will be very complicated in practice.

Assessing Bayesian models can involve evaluation of the fit of a model to data and comparisons of several candidate models for predictive accuracy and for improvements. The methods available for assessing the model fit are of three types [74, 75]:

1. posterior predictive checks
2. prior predictive checks
3. mixed checks.

Prior predictive checks are used to evaluate replications with different parameter values whereas mixed checks are used for evaluating hierarchical models. In posterior predic-

tive checks, data simulated under the fitted model are compared with the actual data [73]. Therefore, it examines whether there are systematic differences between the actual and replicated data [72]. Predictive model accuracy is estimated using information criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), and Watanabe-Akaike (or widely available) information criterion (WAIC), and cross-validation (CV). In addition, Bayesian p-values calculated based on the discrepancy measures (test quantities) can also be used as tools for posterior predictive checks, especially in the contexts of model improvement. The goal of information criteria is to obtain an unbiased measure of out-of-sample prediction error [74, 173]. Since posterior checks use the data twice; once for model estimation and once testing, a penalty constant or bias correction is applied to these criteria. Although, these criteria are unable to reflect the goodness-of-fit in an absolute sense, the differences (in the information theoretic criterion of choice between competing models) can measure the relative performance of the models of interest. However, the use of some of these measures is only valid under certain circumstances. The computation cost is also another problem. Calculation of predictive accuracy measures should not take a long time relative to the model fitting and obtaining initial posterior draws.

Any particular model may provide an adequate fit to the data. However, there may be some plausible alternative models that are also capable in producing a fairly similar fit. Therefore, in the contexts of posterior inferences, where the model at hand differs from the others, posterior predictive checks are very informative. Any discrepancy that can be observed as a result of this self-consistency assessment is considered as consequences of either model misfit or chance or both.

3.1.1 Bayesian p-values

Tail area probabilities can be used as they are in classical statistics, to obtain p-values [72]. The posterior distribution of the unknown model parameters is used to answer the Bayesian inferential problems. Hence, a test quantity which represents the level of discrepancy between the fitted Bayesian model and data, is a function of both data and the unknown model parameters.

Let us assume that the observed data and all the parameters of the fitted model are denoted $\mathbf{y}$ and θ respectively where all the hyper parameters in a hierarchical model are also included in parameter vector θ. The simulated data drawn from the posterior distribution and the future observable data are denoted by $\mathbf{y}^{rep}$ and $\tilde{\mathbf{y}}$ respectively. Then the posterior predictive distribution of $\tilde{\mathbf{y}}$ or the distribution of $\mathbf{y}^{rep}$ is defined with the posterior of unknown parameter θ as

$$p(\mathbf{y}^{rep}|\mathbf{y}) = \int_{\theta} p(\mathbf{y}^{rep}|\theta)p(\theta|\mathbf{y})d\theta.$$

$T(\mathbf{y},\theta)$ denotes a discrepancy measure that summarises parameters and data into scalars and $T(\mathbf{y})$ is a test statistic that depends only on data. The tail-area probability is calculated based on the posterior simulations of $(\theta,\mathbf{y}^{rep})$ and used to measure the goodness-of-fit of the data with respect to the posterior predictive distribution. The extremeness of the simulated data in comparison with the observed data is calculated as a probability p_B, the Bayesian p-value:

$$p_B = Pr\big(T(\mathbf{y}^{rep},\theta) \geq T(\mathbf{y},\theta)|\mathbf{y}\big).$$

The posterior density of θ, $p(\theta|\mathbf{y})$, posterior predictive density of $\mathbf{y}^{rep}$, $p(\mathbf{y}^{rep}|\theta,\mathbf{y}) = p(\mathbf{y}^{rep}|\theta)$, and the indicator function $I_{T(\mathbf{y}^{rep},\theta)\geq T(\mathbf{y},\theta)}$ are used to calculate the posterior predictive p-value, p_B as follows.

$$p_B = \int_{\theta}\int_{\mathbf{y}} I_{T(\mathbf{y}^{rep},\theta)\geq T(\mathbf{y},\theta)} p(\mathbf{y}^{rep}|\theta)p(\theta|\mathbf{y})d\mathbf{y}^{rep}d\theta. \tag{3.1}$$

Let us consider the observed data $\mathbf{y}$ consisting of n observations and assume that there are S draws from the posterior distribution of θ. Then $\mathbf{y}_s^{rep}$ also consists of n replicated values for each parameter value θ_s (where $s = 1,2,\ldots,S$). Then the posterior predictive probability p_B, which is defined as in Equation: 3.1 can be approximated based on the above replicated samples. In the context of Bayesian p-values, the realised and predictive test quantities denoted by $T(\mathbf{y},\theta_s)$ and $T(\mathbf{y}_s^{rep},\theta_s)$ respectively are compared over S replicated draws to perform the posterior predictive checks. The proportion of predictive test

quantities $T(\mathbf{y}_s^{rep}, \theta_s)$ which are not less than corresponding realised value $T(\mathbf{y}, \theta_s)$ is an estimate of the Bayesian p-value. Mathematically, it can be written in the following way.

$$p_B = \frac{1}{S}\sum_{s=1}^{S} I_{T(\mathbf{y}^{rep},\theta_s)\geq T(\mathbf{y},\theta_s)},$$

where,

$$I_{T(\mathbf{y}^{rep},\theta_s)\geq T(\mathbf{y},\theta_s)} = \begin{cases} 1, & \text{if } T(\mathbf{y}^{rep},\theta_s) \geq T(\mathbf{y},\theta_s) \\ 0, & \text{otherwise.} \end{cases}$$

Since various aspects of the model can be tested using the concept of posterior predictive p-value, the selection of the test quantity is very important. The inferential aspect that is expected to be assessed by the test quantity must be in line with the practical purpose of the model.

The interpretation of posterior predictive p-values is more interesting compared to classical p-values. In the classical approach, a p-value that is close to zero implies a greater disagreement between the data and statistical concept being tested while a value that is close to one evidences a greater agreement between them. An extreme posterior predictive p-value (close to 0 or 1) in Bayesian approach, in contrast, implies a greater discrepancy between the data and model. However these extreme p-values can be omitted in the situations where the misfits of the model are practically very small in comparison with the variation within the model. In general, extreme p-values can be used to identify the possible departures of the test quantities from the model rather than rejecting the model. These are very important in practice to identify unusual observations and provide appropriate suggestions to improve the model and data. A p-value close to 0.5 in a posterior predictive check exhibits a better adequacy of the model to data, except in some misleading situations. Since the sample variance is always a sufficient statistic, a test quantity that is a function of sample variance may not be capable in assessing the quality of a posterior predictive distribution. Such discrepancy measures generally produce p-values close to 0.5 and are misleading. A scatter plot of $T(\mathbf{y}, \theta_s)$ vs $T(\mathbf{y}^{rep}, \theta_s)$ or a histogram of $T(\mathbf{y}, \theta_s) - T(\mathbf{y}^{rep}, \theta_s)$ can also be used to display the discrepancy between the data and

the model. The scatter plot should be symmetric around $T(\mathbf{y},\theta_s) = T(\mathbf{y}^{rep},\theta_s)$ line and the value zero must be in the middle of the histogram for a better fit.

Gelman et al.[71] suggested the following discrepancy quantity, which corresponds to the chi-squared goodness-of-fit measure, as an omnibus goodness-of-fit test where the model parameter θ is known.

$$D(\mathbf{y},\theta) = \sum_{i=1}^{n} \frac{\Big(y_i - \mathrm{E}(y_i|\theta)\Big)^2}{\mathrm{Var}(y_i|\theta)},$$

This can be calculated for both observed data $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$ and unobserved future data $\tilde{\mathbf{y}}^T = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)$ as $D(\mathbf{y},\theta)$ and $D(\tilde{\mathbf{y}},\theta)$ respectively [34]. In the Bayesian context where the posterior distribution of θ represents its behaviour, a p-value can be defined in the following way to evaluate the extremeness of future observations.

$$p_D = p\big[D(\tilde{\mathbf{y}},\theta) \geq D(\mathbf{y},\theta)\big] = \int_\theta p\big[D(\tilde{\mathbf{y}},\theta) \geq D(\mathbf{y},\theta)\big]\, p(\theta|\mathbf{y})\, d\theta.$$

As it does in the other posterior checks, p_D can be estimated over the posterior predictive simulations as below.

$$\widehat{p}_D = \frac{1}{S}\sum_{s=1}^{S} I_{D(\tilde{\mathbf{y}},\theta_s)\geq D(\mathbf{y},\theta_s)}, \tag{3.2}$$

where, θ_s $(s = 1,2,\ldots,S)$ are the posterior draws of model parameters θ and

$$I_{D(\tilde{\mathbf{y}},\theta_s)\geq D(\mathbf{y},\theta_s)} = \begin{cases} 1, & \text{if } D(\tilde{\mathbf{y}},\theta_s) \geq D(\mathbf{y},\theta_s) \\ 0, & \text{otherwise.} \end{cases}$$

3.1.2 Marginal Predictive Checks

Marginal predictive distributions are calculated for each observation y_i of $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$ in the observed data and used for overall model calibration or to find possible outliers [72]. Let us assume that y_i^{rep} denotes the replicated values of the i^{th} observation in the data. Then the tail area probability p_i corresponding to each observation y_i is calculated

as

$$p_i = Pr\big(T(y_i^{rep}) \leq T(y_i)|\mathbf{y}\big).$$

A natural discrepancy measure $T(y_i)$ is defined as $T(y_i) = y_i$, when y_i is continuous. In this case the tail-area probability reduces to the computation of

$$p_i = Pr(y_i^{rep} \leq y_i|\mathbf{y}).$$

Similar to the way that the Bayesian p-value was calculated in the previous section, p_i can be estimated as

$$\widehat{p}_i = \frac{1}{S}\sum_{s=1}^{S} I_{y_{is}^{rep} \leq y_i}$$

where, y_{is}^{rep} $(s = 1, 2, \ldots, S)$ are the replicated data of y_i and

$$I_{y_{is}^{rep} \leq y_i} = \begin{cases} 1, & y_{is}^{rep} \leq y_i \\ 0, & \text{otherwise.} \end{cases}$$

It is important to perform a combined check by pooling these marginal predictive p-values into a single figure. Therefore, this study derives the following p-value $\widehat{p}_{\mathrm{M}}$ to estimate the overall average of the marginal predictive p-values.

$$\widehat{p}_{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_i = \frac{1}{nS}\sum_{i=1}^{n}\sum_{s=1}^{S} I_{y_{is}^{rep} \leq y_i}, \tag{3.3}$$

In addition, the overall variability in marginal predictive p-values can be represented by their standard deviation.

The cross-validation predictive p-value is an alternative approach that can be used for posterior model check. However, p-values calculated based on marginal and posterior predictive checks generally reveal different behaviours. Here, the marginal distribution of y_i is calculated based on all the other observations except y_i (i.e. $\mathbf{y}_{-i}$). Consequently, the

cross-validation p-value for y_i is defined as

$$p_i = Pr(y_i^{rep} \leq y_i | \mathbf{y}_{-i}).$$

Replicated data can be used to estimate this as it is calculated in marginal predictive p-values. Since the cross-validation predictive p-values involve additional computations, its computational cost has to be particularly considered in practice. However, in the situations where new observations under exactly similar conditions of the model predictors are possible, the gap between cross-validation and full Bayesian predictive check can be fulfilled. This is regarded as mixed predictive check in Bayesian data analysis.

3.2 Predictive Accuracy

Measuring the accuracy of predictions made by a model is a common way of evaluating models. In model assessment, various measures can be discussed. For instance, the scoring function is a method for measuring the predictive accuracy of a point prediction [72]. A value replicated using the model fitted for an observed value, which represents the future observation under similar circumstances corresponding to the observed value is regarded as a point prediction. Mean squared error, mean absolute error, and mean absolute percentage error of predictions are examples of simple scoring functions that can be used to evaluate the predictive accuracy of a model that is close to a normal distribution.

3.2.1 Log-likelihood

Predictive accuracy of probabilistic predictions is evaluated using scoring rules such as quadratic, logarithmic, and zero-one scores [72]. The logarithmic score is a widely used scoring rule in probabilistic predictions and in selecting models [74, 174]. Let us consider a model with parameter θ, that is expected to fit on data $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$. Assuming the independence of data, the likelihood function $p(\mathbf{y}|\theta)$ of the model is defined as

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

Then the log-likelihood is defined as

$$\log p(\mathbf{y}|\theta) = \log \prod_{i=1}^{n} p(y_i|\theta) = \sum_{i=1}^{n} \log p(y_i|\theta).$$

3.2.2 Kullback-Leibler Information

The log density of the unobserved future data given the model parameters and observed data is generally referred to as log predictive density. It is a well-known summary measure of predictive fit [74]. For normal models with constant variance, the log predictive density is proportional to the mean squared error. In statistical model comparison the log predictive density involved in a decisive role as it connected to the Kullback-Leibler information measure. Especially for large samples, expected log predictive density, Kullback-Leibler information, and posterior probabilities are greatly inter-connected. The model that produces the lowest Kullback-Leibler information leads to produce highest expected log predictive density, which will have the highest posterior probability compared to the other models. Hence, the expected log predictive density is used to measure the overall model fit.

The relationship between log predictive density and Kullback-Leibler information measure has been discussed in literature related to information theory (e.g. [7, 29, 74, 151, 156, 170]). The idea of measuring the conceptual distance between two models (or densities) as a directed divergence was originally introduced in 1951 by Solomom Kullback and Richard A. Liebler [110, 111]. The Kulback-Leibler (K-L) information measures the quality of approximation or information loss $I(f;g)$ [9, 101]. In a situation where one approximates the true density $f(x)$ by $g(x)$, where x is a $q \times 1$ random vector, K-L information is defined as

$$I(f;g) = \mathrm{E}\Big\{\log\Big[\frac{f(x)}{g(x)}\Big]\Big\} = \int_{R^q} \log\Big[\frac{f(x)}{g(x)}\Big] f(x)dx.$$

$I(f;g)$ is always non-negative and is zero when $f(x) = g(x)$. In model selection, the true function $f(x)$ is treated as fixed, however, unknown. The function $g(x)$ with parameter

vector θ (i.e. $g(x|\theta)$) is used to approximate $f(x)$. Then $I(f;g)$ becomes [29],

$$I(f;g) = \int_{R^q} \log\Big[\frac{f(x)}{g(x|\theta)}\Big] f(x)dx.$$

The logarithmic term of the above equation can be further expanded into a difference of two logarithmic terms,

$$I(f;g) = \int_{R^q} \log\big[f(x)\big] f(x)dx - \int_{R^q} \log\big[g(x|\theta)\big] f(x)dx.$$

Both integrals of the above equation are in the form of statistical expectations with respect to the true function f, hence, $I(f;g)$ can be expressed equivalently as

$$I(f;g) = \mathrm{E}_f\Big[\log\big[f(x)\big]\Big] - \mathrm{E}_f\Big[\log\big[g(x|\theta)\big]\Big].$$

As the true function $f(x)$ is fixed and unknown, the expectation $\mathrm{E}_f\Big[\log\big[f(x)\big]\Big]$, which depends only on $f(x)$ is also an unknown constant (say k). Finally,

$$I(f;g) = k - \mathrm{E}_f\Big[\log\big[g(x|\theta)\big]\Big].$$

As the constant k is unknown, the absolute information loss cannot be calculated, hence, the appropriateness of $g(x|\theta)$ in approximating $f(x)$ cannot be evaluated. Fortunately, the selection of the best candidate model among two or more alternative models in the context of information loss is obvious. It is known that the inferential aspects that are used in the calculation of information criteria, is highly conditional on the data. Hence, model comparisons cannot be accomplished across different datasets and completely restricted for a fixed given dataset. However, two or more models fitted to a fixed dataset can be compared. Let us assume that $g_1(x|\theta_1)$ and $g_2(x|\theta_2)$ are two models that used to approximate $f(x)$. As the Kullback-Leibler information $I(f;g_i)$, where $i = 1,2$ measures the information loss or the closeness between the true and fitted models, the one that corresponds to the lowest information loss is the best relative to the other. Hence, the model $g_1(x|\theta_1)$ is better than $g_2(x|\theta_2)$ in approximating $f(x)$, if $I(f;g_1) < I(f;g_2)$. This

inequality can be further simplified as

$$
\begin{aligned}
I(f;g_1) &< I(f;g_2) \\
k - \mathrm{E}_f\Big[\log\big[g_1(x|\theta_1)\big]\Big] &< k - \mathrm{E}_f\Big[\log\big[g_2(x|\theta_2)\big]\Big] \\
-\mathrm{E}_f\Big[\log\big[g_1(x|\theta_1)\big]\Big] &< -\mathrm{E}_f\Big[\log\big[g_2(x|\theta_2)\big]\Big].
\end{aligned}
$$

Finally, this expression reveals the statistical basis of the use of the expected log predictive density as the key quantity in model comparison. The model that produces the highest expected log predictive density, especially for large samples, ensures uppermost posterior probability compared to the alternative candidates.

3.2.3 Out-of-sample Predictive Accuracy Measures Using Posterior Simulations

The predictive accuracy of a model can be evaluated by measuring the out-of-sample predictive performance for a new data point generated from the true data-generating process [72]. A dataset that could be seen in future under the true process f is denoted by $\tilde{\mathbf{y}}^T = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_i, \ldots, \tilde{y}_n)$. The posterior distribution of the unknown parameter θ is denoted by $p_{\text{post}}(\theta)$. The probabilities and expectations calculated as averages over the posterior distribution of θ are denoted by p_{post} and E_{post} respectively. Then, as stated by Gelman et al. [74], the log predictive density of a new data point $\tilde{y}_i$, (i.e. $\log p_{\text{post}}(\tilde{y}_i)$), can be calculated using the posterior density of θ as

$$
\log p_{\text{post}}(\tilde{y}_i) = \log \mathrm{E}_{\text{post}}[p(\tilde{y}_i|\theta)] = \log \int_\theta p(\tilde{y}_i|\theta) p_{\text{post}}(\theta) d\theta.
$$

Since this is practically impossible to calculate as the future data are unknown, the **expected** (out-of-sample) **log predictive density** (**elpd**) for a new datum $\tilde{y}_i$ is used. Sometimes the elpd is denoted as the mean log predictive density and defined as,

$$
\text{elpd} = \mathrm{E}_f[\log p_{\text{post}}(\tilde{y}_i)] = \int_{\tilde{\mathbf{y}}} [\log p_{\text{post}}(\tilde{y}_i)] f(\tilde{y}_i) d\tilde{\mathbf{y}}.
$$

Although the data distribution f is always unknown, a posterior distribution for the parameters of a model assumed for a given problem can be obtained. Using a plug-in estimator of f is a convenient way of estimating the elpd. However, these estimates generally raise some problems due the risk of overfitting.

Defining the **expected log point-wise predictive density** (**elppd**) as a point-wise measure of predictive accuracy is an alternative way of handling this problem in the Bayesian context. The point-wise measure, elppd, for a new dataset can be defined as

$$\text{elppd} = \sum_{i=1}^{n} \text{E}_f[\log\, p_{\text{post}}(\tilde{y}_i)].$$

The advantage in point-wise measure of predictive density over the joint posterior predictive distribution $p_{\text{post}}(\tilde{\mathbf{y}})$ is its connection with cross-validation. Cross-validation is a well-known method that uses data at hand for approximating out-of-sample fit. Sometimes, the calculation of predictive accuracy at a given point estimate of θ, (say $\hat{\theta}$) is a useful method. For instance, for the models with the presence of independent data conditional on model parameters, the expected log predictive density calculated on $\hat{\theta}$, $\text{E}_f[\log\, p(\tilde{y}|\hat{\theta})]$, can be expressed in terms of point-wise prediction as

$$\text{E}_f[\log\, p(\tilde{y}|\hat{\theta})] = \text{E}_f[\log \prod_{i=1}^{n} p(\tilde{y}_i|\hat{\theta})] = \sum_{i=1}^{n} \text{E}_f[\log\, p(\tilde{y}_i|\hat{\theta})].$$

As the parameter θ is unknown, it is impossible to derive the log predictive density $\log p(\mathbf{y}|\theta)$. However, the predictive accuracy of the fitted model can be summarised based on the posterior distribution of θ, $p_{\text{post}}(\theta) = p(\theta|\mathbf{y})$. As previously discussed, the joint prediction at a given point estimate $\hat{\theta}$ can be expressed in terms of point-wise predictions subjected to the independence of the data given parameters. Consequently, the **log point wise predictive density** (**lppd**) can be evaluated as

$$\text{lppd} = \log \prod_{i=1}^{n} p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log\, p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log \int_{\theta} p(y_i|\theta) p_{\text{post}}(\theta) d\theta.$$

Let us consider S draws from the posterior distribution of θ, $p_{\text{post}}(\theta)$. The s^{th} value of the parameter θ among posterior simulations is denoted by θ_s, where, $s = 1, 2, \ldots, S$. Based

on the posterior replicated values, the predictive density of y_i can be estimated as

$$\int_{\theta} p(y_i|\theta)p_{\text{post}}(\theta)d\theta = \frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_s).$$

Consequently, the **computed** or estimated value of **log pointwise predictive density**, **clppd** can be obtained as

$$\text{clppd} = \sum_{i=1}^{n} \log\left[\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_s)\right].$$

MCMC is a well-known numerical approximation technique that calculates unknown quantities averaging over posterior draws. Practically, Monte Carlo sample size S must be sufficiently large to capture the important features of a posterior distribution and such samples arbitrarily improve the accuracy of estimates [41]. However, the inherent autocorrelation between consecutive posterior draws must also be taken in to account. Thinning in MCMC; which refers to the selection of the first draw from every consecutive sub group that consists of a certain number (thinning interval) of draws, is routinely used by many Bayesian practitioners expecting a reduction in the effects of autocorrelation.

A majority of the Bayesian practitioners have encouraged the use of thinning due to one or both of the following reasons.

1. In the presence of high level of autocorrelations in posterior draws, it is required longer runs for the convergence of parameters [73, 100, 127]. The tendency of over-representation of some values while having under-representation of other values due to the clumpy behaviour is another issue in autocorrelated MCMC chains [109].

2. The MCMC algorithms with slow mixing properties require extremely large numbers of posterior draws to derive precise estimates of the features of posterior density [100]. A group of consecutive posterior draws generated under a slow mixing algorithm usually contributes with a little more information compared to the total amount of information provided by a single observation in the group.

In MCMC simulations, posterior draws are assumed to be independent samples from the target distribution, which is regarded as the gold standard of Monte Carlo simulations

[95, 100]. Both slow-mixing chains and highly autocorrelated posterior draws are interconnected in many occasions and create a clear violation of the gold standard. The use of thinning with a suitable interval is the most frequently used technique that helps to keep the gold standard. Model re-parameterisation is an alternative and practical technique that can be used to get rid of strong autocorrelations [72]. However, some literature that discourages thinning even with moderate autocorrelations, and recommends long MCMC runs instead are available [40].

Some Bayesian practitioners regarded thinning as a useful method for the following reasons.

1. It gives a conservative measure of precision of the estimated quantities over unthinned posterior draws [114, 115].

2. It effectively manages the limited storage resources in the post-chain processing of posterior draws [79, 114].

3. Gigantic models containing thousands of parameters, hierarchical or latent variable models for example, usually require a huge number of draws for the convergence and store massive arrays of MCMC outputs in Random Access Memory (RAM) of computers [100].

Longer runs and extra time are the costs or the consequences associated with thinning. However, even with these, thinning is useful especially with the limited resources available in many laptop and desktop computers.

3.3 Information Criteria

Measures of predictive accuracy are generally known as information criteria [72]. The probability of data conditional on estimated model parameters is often expressed in a logarithmic scale and labelled as the log-likelihood. In 1972, Nelder and Wedderburn proposed the idea of deviance to assess the goodness-of-fit of models against their increasing complexity [129]. It is defined by multiplying the log-likelihood by a factor of

-2 [73]. Deviance plays an important role in model comparisons as it is connected with Kullback-Leibler and other information criteria.

Performance evaluation of a given model and comparison of different models are recognised as the key uses of prediction accuracy [72]. In fact, there are two ways to calculate predictive accuracy measures: within-sample and out-of-sample; however, out-of-sample predictions will always be less accurate compared to the other. The log predictive density calculated over the observed data is a naïve estimate of the expected log predictive density of future data that is believed to be drawn under the same data generating process. The computed log point-wise predictive density (clppd) is relatively more straightforward in terms of estimation and comparison. However, the clppd always overestimates the expected log point-wise density (elppd) as it uses data twice for model fitting and evaluation. As clppd introduces a bias in estimating elppd, it can be statistically treated by introducing a suitable bias correction. This provides the basis for many information criteria and some of them are discussed in this section with details.

Cross-validation (CV) is another method that can be used for out-of-sample prediction. Even though many versions of cross-validations are available for evaluating prediction error, they all essentially take the same form. Basically, in any version of the CV, the data are divided into two parts, a training set and a test (or validation) set. The predictive accuracy of the model fitted on the training set is evaluated over the analogous test set. CV is a useful technique that can be used to avoid overfitting. However, it is computationally expensive as it may use many partitions and may fit a large number of models. For example, in k-fold cross-validation, the data are partitioned randomly into k roughly equal-sized subgroups [87]. Consequently, k separate models are fitted treating each of $k-1$ subgroups as a training set. For each model, the remaining set is used as the test set. Leave-one-out cross-validation (LOO-CV) is the most extreme case of k-fold CV, which considers each observation as a subgroup. LOO-CV is the most computationally expensive variant among other CV methods, as it fits as many models as the number of observations. However, some modified versions of importance sampling techniques have been combined with LOO to get rid of expensive computational issues [173].

3.3.1 Akaike Information Criterion (AIC)

Model comparison procedures generally consist of two measures [156]:

1. Goodness-of-fit – frequently measured in terms of deviance statistic, and
2. Model complexity – commonly measured in terms the number of free (or effective) parameters.

A model with high complexity often leads to a better fit but will be of little predictive utility because of over-fitting. Hence, a balance between these two is essential in model building and evaluation processes.

Kulback-Leibler (K-L) and maximum likelihood are two paradigms that dominate the fields of information theory and statistics respectively [29]. In 1973, Akaike information criterion (AIC) was introduced [2] as an estimator for expected K-L information, in terms of a bias corrected maximised log-likelihood value. In fact, the estimated expected relative K-L information was derived as the difference between the log-likelihood and the number of parameters (k) in the model as the bias correction. That is

$$\text{estimated expected relative K-L information} = \text{log-likelihood} - k.$$

Subsequently, the difference between the log-likelihood and k is multiplied by a factor of -2, expecting comparability with the definition of deviance which measures the lack of fit of a fitted model. When the data and maximum likelihood estimators of model parameters are denoted by $\mathbf{y}$ and $\hat{\theta}_{mle}$ respectively, the AIC is formulated as

$$\text{AIC} = -2 \log p(\mathbf{y}|\hat{\theta}_{mle}) + 2k.$$

The model with the lowest AIC value is treated as the best relative to the alternative candidate models. AIC is recognised as a useful tool in model comparison, especially for variable selection, as it does not depend on the order in which the models are computed [122]. Overfitting can be avoided by employing the models that involve hierarchical structures or informative priors [72]. AIC cannot be used to compare them as the numbers of free parameters of these models are unknown. The actual (or effective) number of parameters

in both types of models is strongly influenced by the variance of group-level parameters. Some extensions of AIC with adjustments related to k are: the TIC (Takeuchi information criterion), the RIC (regularized information criterion, and the NIC (network information criterion). These criterion are not widely used because the estimate's variance increases due to stability problems and computational difficulties [174].

AIC exhibits a poor performance when the size of sample is not large enough in connection with the number of parameters in the model. As AIC was derived to correct the asymptotic bias of maximum likelihood, corrected AIC (AIC_c) was introduced by Sugiura in 1978, considering the exact bias of four practical problems including a regression problem [159]. In 1989, Hurvich and Tsai confirmed the superior performance of AIC_c over AIC with the presence of small samples and recommended its use in the context of regression and autoregressive models [96]. Let us consider a model fitted with k parameters over a sample of size n. Then AIC_c is calculated as follows:

$$AIC_c = -2\log p(\mathbf{y}|\hat{\theta}_{mle}) + 2k\frac{n}{n-k-1} = -2\log p(\mathbf{y}|\hat{\theta}_{mle}) + 2k + \frac{2k(k+1)}{n-k-1}; n > k+2.$$

According to Burnham and Anderson, the use of AIC_c is strictly recommended instead of AIC if $n < 40k$ as it gives better results even for larger samples [29].

3.3.2 Bayesian Information Criterion (BIC)

In 1978, Schwarz proposed Bayesian information criterion (BIC) as a competitive measure for AIC in model comparison 148. The derivation of BIC is linked to Bayesian procedures involving asymptotic behaviour of Bayes factor of a comparison between two models. Replacing the factor 2 of the penalty parameter $2k$ in AIC by the natural log of the sample size ($\log n$), BIC is defined as

$$BIC = -2\log p(\mathbf{y}|\hat{\theta}_{mle}) + k\log n.$$

The BIC, unlike AIC, adjusts for the number of fitted parameters with a penalty that increases with respect to the sample size. The factor 2 in the penalty parameter of AIC, replaces by the (natural) log value of the sample size n in BIC. As the natural log of eight

greater than two (i.e. $\log 8 > 2$), the penalty per parameter in BIC is steeper for bigger datasets that has more than seven observations. As BIC heavily penalises complex models, it favours smaller models (in terms of number of parameters) than AIC. As discussed under AIC, the calculation of number of free parameters under the models that involves hierarchical structures or informative priors are always misleading. As a consequence of this, BIC cannot be used to compare these models.

3.3.3 Deviance Information Criterion (DIC)

Complex hierarchical models are not easily compared as the number of parameters in these models is not well-defined. Rapid expansion of MCMC methods and the development of super computers enhance the possibilities of exploring real world phenomena with highly complex models. Hence, a method that can cope with comparisons of such massive models along with MCMC approaches was essential, and the Deviance information criterion (DIC) introduced by Spiegelhalter et al. [156] fulfils this requirement.

The DIC is considered to be the Bayesian analogue of AIC [72] as it blends the frequentist approach of AIC and Bayesian thinking in its derivation process [29]. The DIC replaces the maximum likelihood estimate of θ in the AIC with its posterior mean $\hat{\theta}_{Bayes}$ and the penalty parameter k with a data-based bias correction p_{DIC} which represents the **effective number of parameters** in the model [72, 74]. Consequently, the expected log predictive density is estimated as

$$\widehat{\text{elpd}}_{\text{DIC}} = \log p(\mathbf{y}|\hat{\theta}_{Bayes}) - p_{\text{DIC}},$$

where p_{DIC} is defined as twice as the difference between log predictive density of the Bayes estimator of θ and posterior expectation of the log predictive density. That is,

$$p_{\text{DIC}} = 2\big\{\log p(\mathbf{y}|\hat{\theta}_{Bayes}) - \text{E}_{\text{post}}\big[\log p(\mathbf{y}|\theta)\big]\big\}.$$

The posterior expectation of the log predictive density calculated over the posterior reali-

sations of θ, estimates the value of p_{DIC} as follows.

$$\widehat{p}_{\text{DIC}} = 2\Big[\log\, p(\mathbf{y}|\hat{\theta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log\, p(\mathbf{y}|\theta_s)\Big].$$

The log predictive density is maximised when the posterior mean and mode of the parameter θ are identical, and thus their difference is crucial in estimating p_{DIC}. Since larger differences tend to produce a negative value for the effective number of parameters, an alternative measure $p_{\text{DIC alt}}$ was introduced.

$$p_{\text{DIC alt}} = 2\text{Var}_{\text{post}}\big[\log\, p(\mathbf{y}|\theta)\big].$$

Although $p_{\text{DIC alt}}$ defined as twice as posterior variance of the log predictive densities, is certainly positive, p_{DIC} is numerically more stable than $p_{\text{DIC alt}}$. The two forms of the effective number of parameters formulate DIC as follows:

$$\text{DIC} = -2\,\log\, p(\mathbf{y}|\hat{\theta}_{Bayes}) + 2\; p_{\text{DIC}}$$

$$\text{DIC}_{\text{alt}} = -2\,\log\, p(\mathbf{y}|\hat{\theta}_{Bayes}) + 2\; p_{\text{DIC alt}}$$

The use of DIC in model comparison has some practical limitations in relation to missing data models such as mixture and random effect models [36]. The DIC exhibits some inconsistency in the results of mixture models as posterior estimates of means are quite delicate under these models. Overall, the poor performance of posterior means in estimating model parameters is a key problem of DIC with respect to mixture models.

3.3.4 Widely Available Information Criterion (WAIC)

A learning machine or a statistical model is described as regular if its Fisher information matrix is positive definite and if the map taking parameters to probability distributions is one-to-one [175]. Many machine learning methods including normal mixtures, artificial neural networks, Bayes networks, and hidden Markov models do not have this property,

hence they are known as singular. Non-representativeness of the plug-in estimates of posterior parameters and lack of convergence in the distribution of deviance to a chi-square distribution are the key problems associated with singular models [74]. Singular learning theory is required for models with a hierarchical structure or hidden variables. Maximum likelihood estimators under singular models are not asymptotically normal, diverge or increase the generalisation error. Hence, maximum likelihood estimation is not applicable for singular models. In contrast, Bayesian estimation reduces the generalization error of models with singularities. However, in both regular and singular models, average cross-validation is equal to the average generalization error.

The Widely available or the Watanabe-Akaike information criterion (WAIC) was introduced in 2010 [175] as a fully Bayesian method for estimating the out-of-sample expectation. In addition, Watanabe showed the asymptotic equivalence of Bayesian leave-one-out cross-validation (LOO-CV) to WAIC. Since Bayesian cross-validations are always applicable for both singular and non-singular models, the asymptotic equivalence implies the validity of WAIC even with singular models. WAIC is defined based on the computed log pointwise predictive density (clppd), which is used as an estimate of expected log pointwise predictive density (elppd) [72, 74, 171]. The bias correction term p_{WAIC} is used to estimate the effective number of parameters against over-fitting. Considering the deviance form of an information criterion, the WAIC is defined as

$$\text{WAIC} = -2\,\widehat{\text{elpd}}_{\text{WAIC}} = -2\,\Big(\text{clppd} - p_{\text{WAIC}}\Big).$$

Similar to DIC, WAIC also has two types of penalty terms p_{WAIC} and $p_{\text{WAIC alt}}$. The first version p_{WAIC} reflects the sum of the differences between posterior expectation of point-wise predictive densities, calculated in logarithmic scale and the posterior expectation of point-wise log predictive densities. That is

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\Big\{\log \text{E}_{\text{post}} p(y_i|\theta) - \text{E}_{\text{post}}\big[\log\,\big(p(y_i|\theta)\big]\Big\}.$$

The expectation terms in the above expression are replaced with the corresponding aver-

ages calculated over posterior draws. Consequently, p_{WAIC} is estimated as

$$\widehat{p}_{\text{WAIC}} = 2\sum_{i=1}^{n}\Big\{\log\Big[\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta_s)\Big] - \frac{1}{S}\sum_{s=1}^{S}\log p(y_i|\theta)\Big\}.$$

The second version, $p_{\text{WAIC alt}}$ is defined as the sum of posterior variances of the point-wise log predictive densities. That is

$$p_{\text{WAIC alt}} = \sum_{i=1}^{n}\text{Var}_{\text{post}}\big[\log p(y_i|\theta)\big].$$

Let us assume that $d_{is} = \log p(y_i|\theta_s)$ is the log density of y_i given θ_s. Then the mean log density of y_i (say $\bar{d}_i$) is computed over S posterior draws as $\bar{d}_i = \frac{1}{S}\sum_{s=1}^{S} d_{is}$. The notation $V_{s=1}^{S}$ is used to define the mathematical operation in relation to the calculation of sample variance of the quantity d_{is}, such that, $V_{s=1}^{S} d_{is} = \frac{1}{S-1}\sum_{s=1}^{S}(d_{is} - \bar{d}_i)^2$. Subsequently, $p_{\text{WAIC alt}}$ is estimated as

$$\widehat{p}_{\text{WAIC alt}} = \sum_{i=1}^{n}\text{Var}_{\text{post}}\big[\log p(y_i|\theta)\big] = \sum_{i=1}^{n} V_{s=1}^{S} d_{is} = \sum_{i=1}^{n} V_{s=1}^{S}\log p(y_i|\theta).$$

The second penalty term, $p_{\text{WAIC alt}}$ was recommended for practical use based on two reasons.

1. The closer relationship between leave-one-out cross-validation (LOO-CV) and $p_{\text{WAIC alt}}$, in series expansions and

2. Both LOO-CV and $p_{\text{WAIC alt}}$ tend to produce approximately similar results.

The WAIC takes the average over the posterior distribution, unlike AIC and DIC that condition on a point estimate. Therefore, WAIC is more popular than AIC and DIC in Bayesian contexts. Although the two approaches in WAIC have similarities with p_{DIC} and $p_{\text{DIC alt}}$, $p_{\text{WAIC alt}}$ is more stable compared to $p_{\text{DIC alt}}$ since it takes the sum of separately calculated variances for each data point. In addition, both AIC and BIC fail in evaluating hierarchical models, while DIC fails in mixture model evaluations. WAIC however performs successfully for hierarchical and mixture models where the number of parameters increases with the sample size and point estimates are often delicate and misleading. Once a set of candidate models consists of both types of models, WAIC can be

effectively used to rank them. However, performance of WAIC also has some practical problems associated with the posterior variances of point-wise log predictive densities calculated over posterior draws. Simulation studies highlight some reliability issues in WAIC with the presence of any observation whose posterior variance exceeds 0.4 (i.e. $\text{Var}_{\text{post}}\big[\log\, p(y_i|\theta)\big] > 0.4$ for any i) [173]. Therefore, 0.4 is used as the standard for the maximum posterior variance of log predictive density of any observation. When WAIC is calculated with R package **loo**, it displays a warning message together with the percentage of observations that violate this standard.

3.4 Leave-one-out Cross-validation (LOO-CV)

Bayesian cross-validation like regular cross validation, involves repeated partitioning of data into **training** and **holdout** sets denoted by y_{train} and $y_{holdout}$ respectively [72, 74]. Posterior distributions of model parameters (θ) are obtained by fitting models to the training sets. The posterior distribution of θ calculated on the training set is specified as

$$p_{\text{train}}(\theta) = p(\theta|y_{\text{train}}).$$

An estimate of the log predictive density of the holdout dataset, $\log\, p_{\text{train}}(y_{\text{holdout}})$ defined below, is used to evaluate the model fit.

$$\log\, p_{\text{train}}(y_{\text{holdout}}) = \log \int_{\theta} p_{\text{pred}}(y_{\text{holdout}}|\theta) p_{\text{train}}(\theta) d\theta.$$

It is assumed that the posterior distribution is summarized by a sufficiently large number of simulation draws. Assuming S posterior draws ($\theta_s;\ s = 1,2,\ldots,S$), the log predictive density is calculated as

$$\log \Big[\frac{1}{S}\sum_{s=1}^{S} p(\text{holdout}|\theta_s)\Big].$$

When each holdout set includes a single data point y_i where $i = 1,2,\ldots,n$, then it generates n different posterior densities. Let us assume that $p_{\text{post}(-i)}$ where $(-i)$ indicates that

the i^{th} observation is being omitted, is the posterior density calculated for y_i over S posterior draws θ_{is}. Then the estimated out-of-sample predictive fit in the context of Bayesian LOO-CV is defined as

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i).$$

The value of $\text{lppd}_{\text{loo-cv}}$ is calculated over posterior draws as defined below.

$$\text{computed lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log \left[\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_{is})\right].$$

The risk of underestimation needs to be considered as it uses only $n-1$ data points in each predictive fit. However, since the bigger the dataset gets the smaller the difference, considerable differences cannot be anticipated in the fits to large samples. The computation cost associated with large datasets is another problem, and alternatively, k-fold cross-validations are used to reduce massive computations associated with LOO-CV. In practice, a first order bias correction b is used to evaluate the quality of predictions. It is defined as

$$b = \text{lppd} - \overline{\text{lppd}}_{-i},$$

where

$$\overline{\text{lppd}}_{-i} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \log p_{\text{post}(-i)}(y_i).$$

Assuming S posterior draws, $\overline{\text{lppd}}_{-i}$ is calculated as

$$\text{computed } \overline{\text{lppd}}_{-i} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \log \left[\frac{1}{S}\sum_{s=1}^{S} p(y_j|\theta_{is})\right].$$

Consequently, the bias corrected Bayesian LOO-CV, $\text{lppd}_{\text{cloo-cv}}$ is calculated as

$$\text{lppd}_{\text{cloo-cv}} = \text{lppd}_{\text{loo-cv}} + b.$$

The impact of b is negligible due to its low magnitude, and therefore, the bias corrected LOO-CV is rarely used. Subsequent to the calculation of log point-wise predictive density (lppd) of the fitted model over its posterior simulations, the effective number of parameters in LOO-CV, $p_{\text{loo-cv}}$ is estimated by

$$p_{\text{loo-cv}} = \text{lppd - lppd}_{\text{loo-cv}}.$$

Similarly, the effective number of parameters in the bias-corrected LOO-CV, $p_{\text{cloo-cv}}$ is estimated by

$$p_{\text{cloo-cv}} = \text{lppd - lppd}_{\text{cloo-cv}} = \overline{\text{lppd}}_{-i} - \text{lppd}_{\text{loo-cv}}.$$

Approximating the posterior distribution by re-fitting the model a number of times associates with a high computational cost. However, it performs well with singular models. In addition, the following approximations of other information criteria to LOO-CV are advantageous.

1. AIC is asymptotically equal to LOO-CV
2. Regularised information criteria (RIC), a variant of DIC, is asymptotically equal to LOO-CV
3. WAIC is asymptotically equal to Bayesian LOO-CV.

3.5 Importance Sampling (IS) for Calculating Leave-one-out Cross-validation (LOO-CV)

As it mentioned in section 3.3.4, the performance of WAIC begins to decrease when any of the posterior variances of the log predictive densities increases beyond 0.4. The use of exact cross-validation is also expensive since it requires as many re-fits as the number of observations in the dataset. However, the problem of involving massive computations can be avoided by using important sampling (IS) techniques. Applicability and the computational convenience of IS in approximating LOO-CV has been widely discussed in

literature [15, 67, 173]. In the Bayesian context, the LOO-CV out-of-sample predictive fit is estimated by Vehtari et al. [173]

$$\text{lpd}_{\text{loo}} = \sum_{i=1}^{n} \log\, p(y_i|y_{-i}),$$

where $p(y_i|y_{-i})$ denotes the leave-one-out predictive density of the i^{th} observation y_i given the remaining observations and is defined as

$$p(y_i|y_{-i}) = \int_{\theta} p(y_i|\theta)p(\theta|y_{-i})d\theta.$$

Assuming the conditional independence among data, the importance ratio of the i^{th} observation over the s^{th} posterior draw (θ_{is}), r_{is} is calculated as follows based on the findings of Gelfand et al. [68]

$$r_{is} = \frac{1}{p(y_i|\theta_{is})} \propto \frac{p(\theta_{is}|y_{-i})}{p(\theta_{is}|y)}.$$

Then the importance sampling leave-one-out (IS-LOO) predictive distribution is calculated as a weighted average over S posterior draws treating importance ratios (weights) r_{is} as weights. Therefore, IS-LOO predictive distribution of y_i is

$$p(y_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} r_{is}p(y_i|\theta_{is})}{\sum_{s=1}^{S} r_{is}} = \frac{\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}p(y_i|\theta_{is})}{\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}} = \frac{S}{\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}}.$$

It can also be written as,

$$p(y_i|y_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}}.$$

As this result implies, IS-LOO predictive density of y_i is the harmonic mean of S posterior densities of y_i, calculated conditionally on the simulated parameter values θ_{is}. Hence, in IS-LOO, the estimated log predictive density of the point y_i can be written in the form

$$\log\, p(y_i|y_{-i}) \approx \log\left(\frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}}\right) = -\log\left(\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\theta_{is})}\right).$$

Therefore, the expected log point-wise predictive density with IS-LOO can be estimated by

$$\widehat{\text{elpd}}_{is-loo} = -\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S}\frac{1}{p(y_i|\theta_{is})}\right).$$

However, the resulting estimate can be noisy due to large or infinitely large variances of the important weights [56].

3.5.1 Truncated Importance Sampling (TIS) for Calculating Leave-one-out Cross-validation (LOO)

The distribution of importance weights is highly positively skewed. In addition, large or infinite variance in the tails tends to produce instability in the results [171]. Ionides [99] has discussed some important aspects of importance sampling along with useful suggestions with special reference to MCMC methods [173]. Following these recommendations and expecting more stabilised weights, raw importance weights r_{is} are replaced by truncated weights w_{is}. The new weights are in the form:

$$w_{is} = \min\left(r_{is}, \sqrt{S}\bar{r}_i\right),$$

where $\bar{r}_i$ is the simple arithmetic mean of the raw importance weights calculated over S posterior draws. That is

$$\bar{r}_i = \frac{1}{S}\sum_{s=1}^{S} r_{is}.$$

Considering the truncated importance sampling weights, the TIS-LOO estimate of the expected point-wise predictive density is calculated as

$$\widehat{\text{elpd}}_{tis-loo} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_{is}p(y_i|\theta_{is})}{\sum_{s=1}^{S} w_{is}}\right).$$

According to Ionides [99], the finite variance of truncated importance sampling weights is ensured under this method (TIS-LOO). However, possible biases introduced as a con-

sequence of truncation in weights is a remarkable drawback of this method.

3.5.2 Pareto-smoothed Importance Sampling (PSIS) for Calculating Leave-one-out Cross-validation (LOO-CV)

Considering the weakness associated with IS-LOO and TIS-LOO, in 2016, Vehtari et al. introduced a new method [173] that employs a special smoothing technique to the importance weights based on the Pareto distribution. Pareto-smoothed importance sampling (PSIS) [172] is a new approach for regularising importance weights and is an efficient way of calculating leave-one-out cross-validation (PSIS-LOO).

It is known that the distribution of importance weights is positively skewed with long tails. The Generalised Pareto distribution (GPD) introduced by Pickands in 1975 [137] is a well-known distribution to model the exceedances over a threshold [35]. The density function of three-parameter GPD is in the following form [49, 172].

$$f(x|\mu,\sigma,k) = \begin{cases} \frac{1}{\sigma}\Big[1+k\big(\frac{x-\mu}{\sigma}\big)\Big]^{-\frac{1}{k}-1}, & k \neq 0 \\ \frac{1}{\sigma}exp\big(-\frac{x-\mu}{\sigma}\big), & k = 0, \end{cases}$$

where σ (>0) and k $(-\infty < k < \infty)$ denote the scale and shape parameters respectively. The continuous location parameter can assume any real value (i.e. $-\infty < \mu < \infty$). For non-negative values of k, the range is $\mu \leq x < \infty$. However, it has a finite upper bound $\mu - \frac{\sigma}{k}$ for negative values of k (i.e $\mu \leq x \leq \mu - \frac{\sigma}{k}$ for $k < 0$). The existence of mean and variance of the distribution strongly depends on the value of shape parameter k as shown below.

$$E(X) = \frac{\sigma}{1-k},\ k < 1$$
$$Var(X) = \frac{\sigma^2}{(1-k)^2+(1+2k)},\ k < \frac{1}{2}$$

The shape parameter k plays an important role in describing valuable properties of the GPD and especially, the thickness of tail of the fitted GPD is characterised by the magnitude of k. As it mentioned above, finite variance of the distribution is guaranteed if k is

less than 0.5 whereas the variance is infinite and the mean is still finite, if k lies between 0.5 and 1. Both mean and variance of the GPD do not exist whenever the shape parameter exceeds or equals one.

Considering many of these vital properties associated with GPD, it has been used to model the upper tail of the weights distribution. In addition, the method proposed by Zhang and Stephens [180] is used to calculate the empirical Bayes estimates of the parameters of the GPD. This method has an increased efficiency and a lesser bias relative to the maximum likelihood estimate. The methodology can be summarised as below [172]. Initially, the parameters of the GPD are redefined as (b, k) so that $b = \frac{k}{\sigma}$. The profile likelihood of k, which maximises the conditional likelihood given b, is selected instead of its maximum likelihood to eliminate high correlation between b and k. Subsequently, the posterior mean of b $(\hat{b})$ is numerically computed combining the estimated profile likelihood, assuming a weakly informative prior on it. In the last step, estimating σ $(\hat{\sigma})$ as $\frac{\hat{k}}{\hat{b}}$, the estimated value of k $(\hat{k})$ is calculated by maximising the likelihood conditional on $\hat{b}$. Even though the estimate associates with a small bias, it is highly efficient and computationally fast compared to fully Bayesian approaches that provide better estimates.

The smoothing technique adopted in PSIS basically depends on larger values of the calculated importance weights [172, 173]. Subsequent to the calculation of importance weights r_{is} for each data point y_i and for each posterior simulation s (where, $s = 1, 2, \ldots, S$), it is necessary to select the largest M weights for each held-out data point y_i. Generally, $M = 0.2S$ representing the largest 20% of weights. However, this number can be further reduced for larger posterior simulations. Depending on the number of posterior samples S, the ratio between M and S can vary between 10% and 20%. In general, a larger value of S requires a smaller percentage. The uncertainty in estimates decreases with larger M, however, the bias increases with larger percentages of $\frac{M}{S}$. After selecting a suitable M, the remaining steps associated with PSIS can be summarised as below.

1. Consider $\left(\frac{S-M}{S}100\right)^{th}$ percentile of the calculated importance weights r_{is} as μ. For example, when $M = 0.2S$, then $\mu = 80^{th}$ percentile of r_{is} is considered.

2. Following the method proposed by Zhang and Stephens described above, estimate the remaining two parameters k and σ, fitting the GPD to the selected largest M

importance weights.

3. Calculate the expected order statistics of the fitted GPD as below.

   The inverse distribution function for any given probability p (0 < p < 1) is

$$F^{-1}(p|\mu,\sigma,k) = \begin{cases} \mu + \frac{\sigma}{k}\Big[(1-p)^{-k} - 1\Big], & k \neq 0 \\ \mu - \sigma ln(1-p), & k = 0, \end{cases}$$

   where $p = \frac{m-0.5}{M}$ for $m = 1,2,\ldots,M$.

   Replace the selected M largest importance ratios with these M estimated order statistics and label the new weights as $m_{is},\ s = 1,2,\ldots,S$. Now m_i is a distinct vector of length S for observation y_i. For more clarity, the first $S-M$ elements of m_i are replaced with the smallest $S-M$ values of the vector r_{is} while the remaining M are replaced with the estimated order statistics of fitted GPD.

4. Truncate each vector of these estimated importance weights m_i at $S^{\frac{3}{4}}\overline{m}_i$, where $\overline{m}_i$ is the average of all the S elements m_{is} in m_i. That is, the new importance weights $\widetilde{w}_{is}$ are defined using truncated weights m_{is}. This ensures a finite variance for the estimate. Then the new weights are in the form

$$\widetilde{w}_{is} = \min\Big(m_{is}, S^{\frac{3}{4}}\overline{m}_i\Big).$$

5. Repeat the above four steps for all the data points y_i, $i = 1,2,\ldots,n$.

Finally, combining all the results, the PSIS-LOO estimate of the expected point-wise predictive density is calculated as

$$\widehat{\text{elpd}}_{psis-loo} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} \widetilde{w}_{is}\, p(y_i|\theta_{is})}{\sum_{s=1}^{S} \widetilde{w}_{is}}\right).$$

It is already known that the importance of shape parameter k is determining the existence of moments in GPD. Consequently, the reliability of estimates can be assessed based on the estimated shape parameter $\hat{k}$ as below.

1. When $\hat{k}$ is less than 0.5, then the estimate is not sensitive to influential (larger) importance weights, hence, the finite variance of the raw importance weights is guaranteed. In addition, a rapid convergence in the estimate can be observed.

2. When $\hat{k}$ lies between 0.5 and 1, a slow convergence in the PSIS estimate can be seen. The raw importance weights are ended up with an infinite variance, but a finite mean. A large variance in the PSIS estimate is obvious under these circumstances.

3. When $\hat{k}$ exceeds 1, the mean and variance of the fitted distribution of importance weights do not exist and consequently, a larger variance in the PSIS estimate cannot be avoided.

When there are influential observations or weak prior information involved in finite models, PSIS-LOO is more robust than WAIC. It is possible to rely on LOO-CV approximation even for alarming values of $\hat{k}$ (i.e. $0.5 < \hat{k} < 1$). However, IS-LOO fails when $\hat{k}$ goes beyond 1 while PSIS-LOO and TIS-LOO measures remain with finite variances.

The approximate standard error calculated for predictive errors under LOO can be highlighted as a small extension of respective calculations [173]. Let us assume that $\text{lpd}_{\text{loo,i}}$ denotes the LOO log predictive density calculated for the observation y_i. As n observations in the data are independent, it can assume that the estimate $\widehat{\text{elpd}}_{\text{loo}}$ consists of n independent components $\widehat{\text{elpd}}_{\text{loo,i}}$. Then the approximated variance of $\widehat{\text{elpd}}_{\text{loo,i}}$ can be calculated as the sample variance of n values of $\widehat{\text{elpd}}_{\text{loo,i}}$. That is

$$Var\big(\widehat{\text{elpd}}_{\text{loo,i}}\big) = V_{i=1}^{n}\widehat{\text{elpd}}_{\text{loo,i}},$$

where the operator $V_{i=1}^{n}a_i$ denotes the sample variance of any variable a with n values $a_1, a_2, \ldots, a_n$. Since $\widehat{\text{elpd}}_{\text{loo}}$ is estimated as the sum of independent components $\widehat{\text{elpd}}_{\text{loo,i}}$, the variance of $\widehat{\text{elpd}}_{\text{loo}}$ is calculated as

$$Var\big(\widehat{\text{elpd}}_{\text{loo}}\big) = n\, V_{i=1}^{n}\widehat{\text{elpd}}_{\text{loo,i}}.$$

The same method can be adopted to calculate the variances of $\widehat{\text{elpd}}$ under IS-LOO, TIS-LOO and PSIS-LOO methods.

In model comparisons, the estimated expected log predictive densities calculated under the two models for the same dataset are compared. Therefore, variance of the difference between two measures can be used to calibrate the precision of comparison. Let us assume that $\widehat{\text{elpd}}^{A}_{\text{loo}}$ and $\widehat{\text{elpd}}^{B}_{\text{loo}}$ denote the estimated LOO expected log predictive densities calculated for model A and B respectively. Point-wise log predictive densities of the two models for the i^{th} observation y_i are denoted by $\widehat{\text{elpd}}^{A}_{\text{loo,i}}$ and $\widehat{\text{elpd}}^{B}_{\text{loo,i}}$ respectively. These two point estimates are paired as they are calculated for the same datum. Hence, the variance of any paired difference can be approximated by the variance of the differences as given below.

$$Var\left(\widehat{\text{elpd}}^{A}_{\text{loo,i}} - \widehat{\text{elpd}}^{B}_{\text{loo,i}}\right) = V_{i=1}^{n}\left(\widehat{\text{elpd}}^{A}_{\text{loo,i}} - \widehat{\text{elpd}}^{B}_{\text{loo,i}}\right).$$

Consequently, the overall variance of the difference between two measures used for comparison can be calculated as

$$Var\left(\widehat{\text{elpd}}^{A}_{\text{loo}} - \widehat{\text{elpd}}^{B}_{\text{loo}}\right) = n\, V_{i=1}^{n}\left(\widehat{\text{elpd}}^{A}_{\text{loo,i}} - \widehat{\text{elpd}}^{B}_{\text{loo,i}}\right).$$

This calculation can be extended to assess the precision of model comparisons using IS-LOO, TIS-LOO and PSIS-LOO methods. The R package **loo** provides necessary facilities for calculating WAIC, LOO and PSIS-LOO. In addition, it provides various standard errors in relation to WAIC and LOO approaches.

3.6 L-Measure

The L-measure is another Bayesian model assessment criterion, which is defined on the posterior predictive distribution of the data. The L-measures calculated for large class of plausible models can be compared to select the best model. It was first introduced by Ibrahim & Laud [98] and Laud & Ibrahim [112], only for the linear models. Then a more general version of L-measure was introduced in 2001 by Ibrahim, Chen & Sinha [97].

Let us assume that $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$ denotes a vector of observed data and $\mathbf{z}^T = (z_1, z_2, \ldots, z_n)$ is a vector of unobserved future values generated from the same model m

from which the observed data **y** are coming. Then the more general version of L-measure (L_m) is defined as

$$L_m^2 = L_m^2(\mathbf{y}, \nu) = \sum_{i=1}^{n} \Big[\mathrm{Var}(z_i) + \nu \big(\mathrm{E}(z_i) - y_i \big)^2 \Big],$$

where, $0 \leq \nu \leq 1$. In the original version of L-measure, $\nu = 1$, and it gives equal priority to both the variance component and the squared bias term. As ν is allowed to vary between zero and one, it provides a great flexibility in assigning weighting for the bias and the variance. The L-measure is calculated as $L_m = \sqrt{L_m^2}$, expecting a comparability with the observed data in terms of measurement unit.

3.7 Summary

Assessing the validity of the underlying assumptions and measuring the accuracy of predictions are essential in any Bayesian statistical model. This chapter reviews measures available for Bayesian model assessment and discusses the usefulness and shortfalls of them and the conditions to be satisfied for their use. This review provides important suggestions for selecting appropriate measures for evaluating the models presented in Chapter 2, and hence supports the identification of better non-hierarchical models in Chapter 4 and hierarchical models in Chapter 5, for predicting stutter ratio. Bayesian p-values, information criteria such as AIC, BIC, DIC, and WAIC, and cross-validations are the discussed measures of predictive model accuracy. However, these methods take different aspects of predictive accuracy into consideration and provide varied benefits in model evaluations. Bayesian p-values measure the discrepancy between the data and the fitted model by evaluating the extremeness of future observations. However, they are not suitable for evaluating test quantities that are functions of sample variance since they always produce p-values close to the desired value 0.5 regardless of the model fit. Both AIC and BIC perform better for Bayesian models with flat priors, fitted to large samples. However, BIC is better than AIC as it penalises the model complexity more than AIC. Both AIC and BIC are not suitable for evaluating hierarchical models whereas DIC is suitable for these. However, DIC is not suitable for mixture model comparisons. WAIC is the

best option with both singular (e.g. mixture models) and non-singular models as it has been developed based on singular learning theory. However, the posterior variances of log pointwise predictive densities for all the observations in a dataset must be below 0.4 for WAIC to be valid. Exact (leave-one-out) cross-validation is one of the best approaches available for evaluating out-of-sample predictive fit. However, it is computationally expensive and time consuming with large datasets. Therefore, importance sampling, truncated importance sampling, and Pareto-smoothed importance sampling techniques are used as approximations to the exact cross-validation. Furthermore, L-measure is identified as another Bayesian model assessment criterion which evaluates a weighted average of the variance and squared bias of predictions.

Chapter 4

Assessment of Models

4.1 Introduction

In this chapter, the methodology discussed in Chapter 3 is applied to compare the models described in Chapter 2. These models include five proposed by Bright et al. [21] and another six models proposed in this study to explain the behaviour of PCR stutter ratio (*SR*). The five models proposed by Bright et al. modelled *SR* with a right-skewed heavy-tailed distribution and log-normal and gamma distributions were used. The six models proposed in this study, in contrast, modelled *SR* as a symmetrically distributed random variable and the distributions proposed were the non-standardised Student's t and the normal. The mean of each model was modelled as a locus-specific simple linear regression that used longest uninterrupted sequence (*LUS*) as the predictor. The gamma models assumed two versions of variance models: profile-wide and locus-specific variances. The normal, log-normal, and non-standardised Student's t models assumed an additional variance structure that provides the basis for two-component mixtures. Chapter 2 provided all the details of the 11 models that are examined in this chapter. In relation to variance modelling, the models can be classified into three categories:

1. Models assuming profile-wide variance – There are four models that assume log-normal (LN_0), gamma (G_0), normal (N_0), and non-standardised Student's t (T_0) distributions in stutter modelling.

2. Models assuming locus-specific variance – The models LN_1, G_1, N_1, and T_1 are

fitted based on the log-normal, gamma, normal, and non-standardised Student's t distributions respectively to model the behaviour of *SR*.

3. Two-component mixture models – Log-normal, normal, and non-standardised Student's t distributions are assumed for the two components in each model and these are denoted by MLN_1, MN_1, and MT_1respectively.

4.2 Graphical Assessment of Distributional Assumptions

Assessment of the distributional assumptions or evaluating the validity of the proposed distributions is always very important in the model building process. The use of probability plots for testing the goodness-of-fit is very common in classical statistics. Quantile-quantile (Q-Q) and percent-percent (P-P) plots are the most commonly used probability plots [165]. Q-Q plots always provide a greater emphasis to the tails [66]. P-P plots, in contrast, highlight the differences in the middle of the distribution as the cumulative probabilities are rapidly changing in the regions of higher probabilities. Therefore, both P-P and Q-Q plots jointly describe the goodness-of-fit of the fitted distribution.

According to their parametrisation, the means and variances of all the models used in this study vary from observation to observation . The variance of *SR* (or log(*SR*)), for example, is a function of the observed allele height which varies from observation to observation. The log-normal model for *SR* can also be interpreted as a normal model for log(*SR*). Hence, the observed value of each log(*SR*) and *SR* can be standardised based on the estimated means and standard deviations of respective normal and log-normal models. However, even within one family either normal or log-normal, the mean and variance of the distribution for each observation are fixed only for the particular observation. Therefore, the standardisation of the observations are not subjected to the mean and variance of a single distribution as it is in usual practice. However, subjected to the validity of zero mean and unit variance of the standardised values, they can be assumed to be realisations of a normally distributed random variable. Even though the standardisation of the observations is possible under the normal mixture model, one cannot expect any normality from the standardised values, as mixtures of normal densities are no longer normal. Simi-

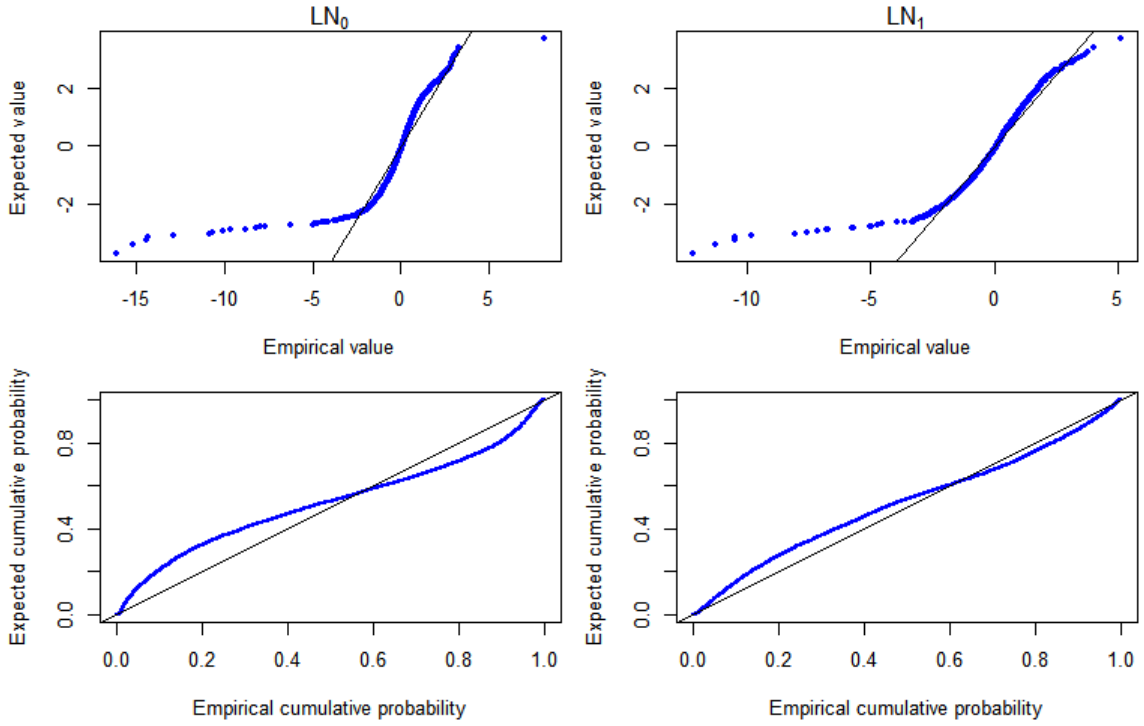

Figure 4.1: Log-normal Q-Q and P-P plots for the NGM SElect^{TM} dataset.

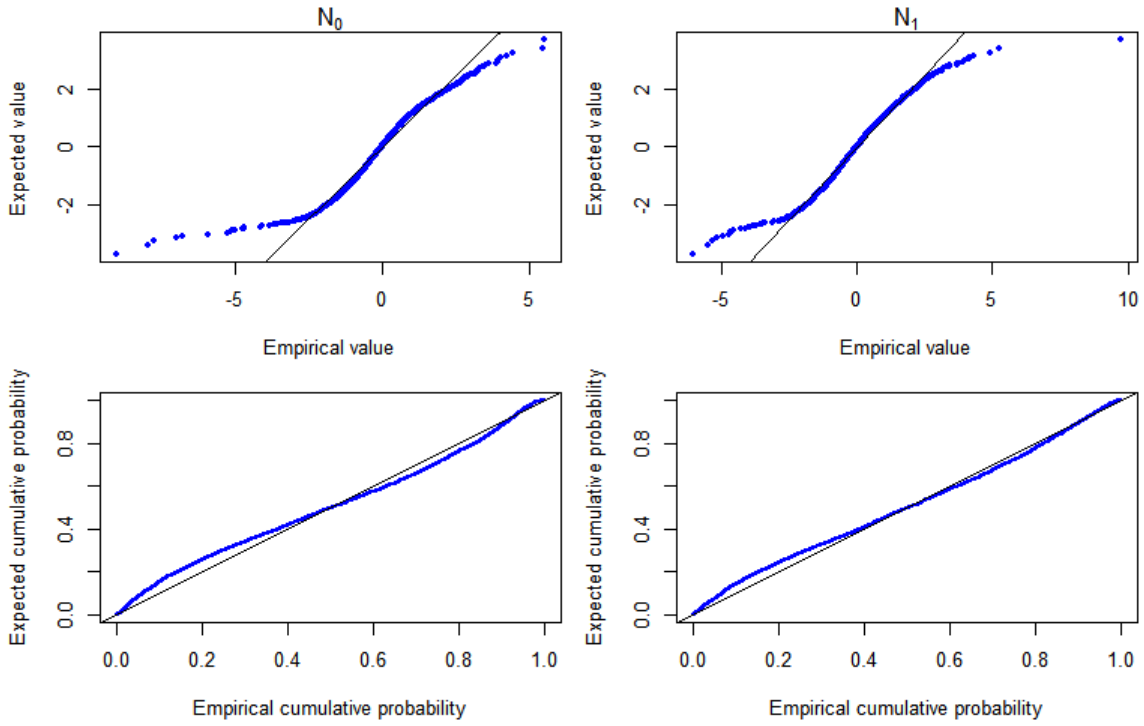

Figure 4.2: Normal Q-Q and P-P plots for the NGM SElect^{TM} dataset.

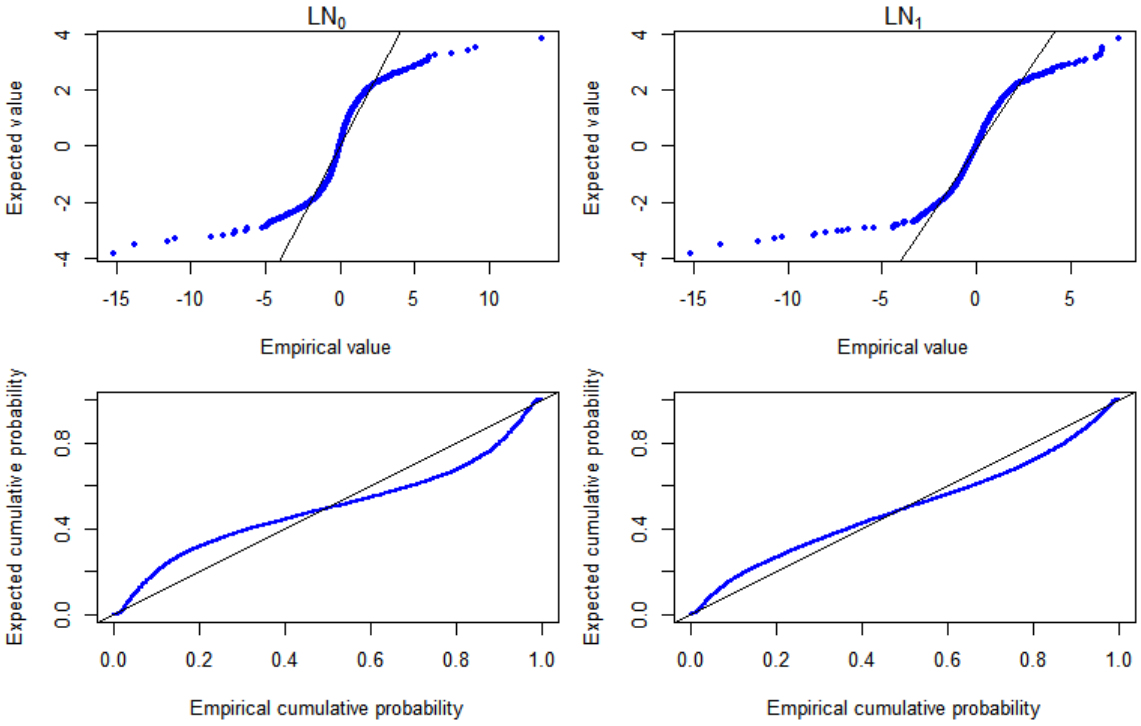

Figure 4.3: Log-normal Q-Q and P-P plots for the Identifiler™ dataset.

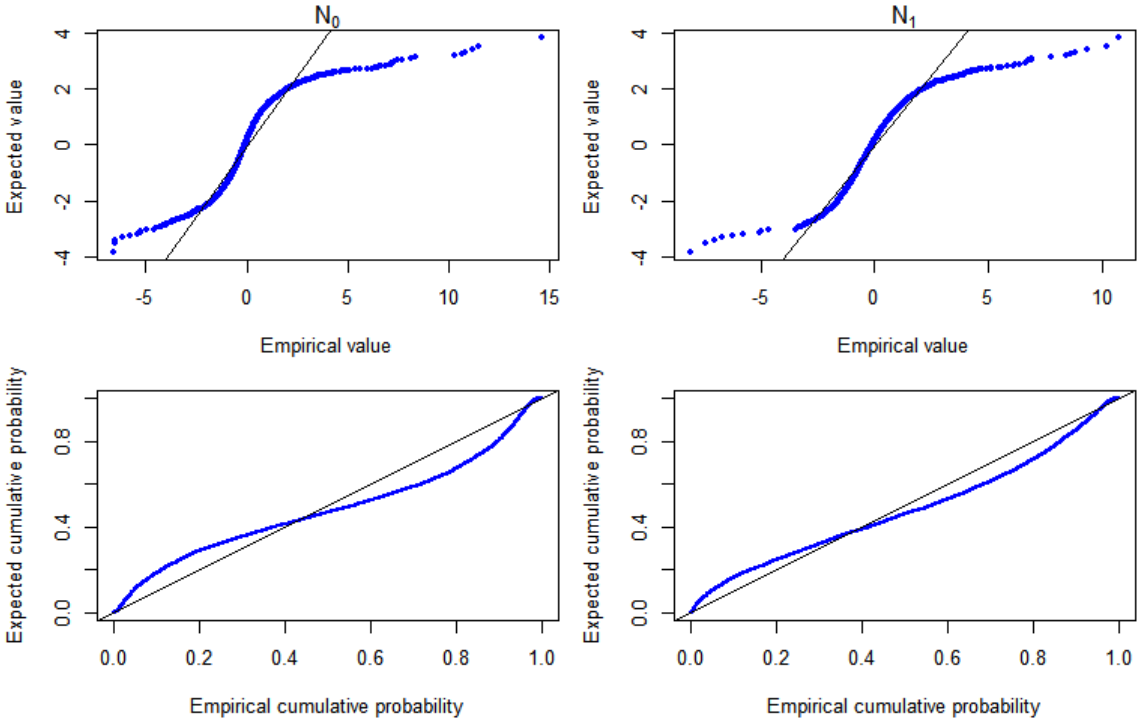

Figure 4.4: Normal Q-Q and P-P plots for the Identifiler™ dataset.

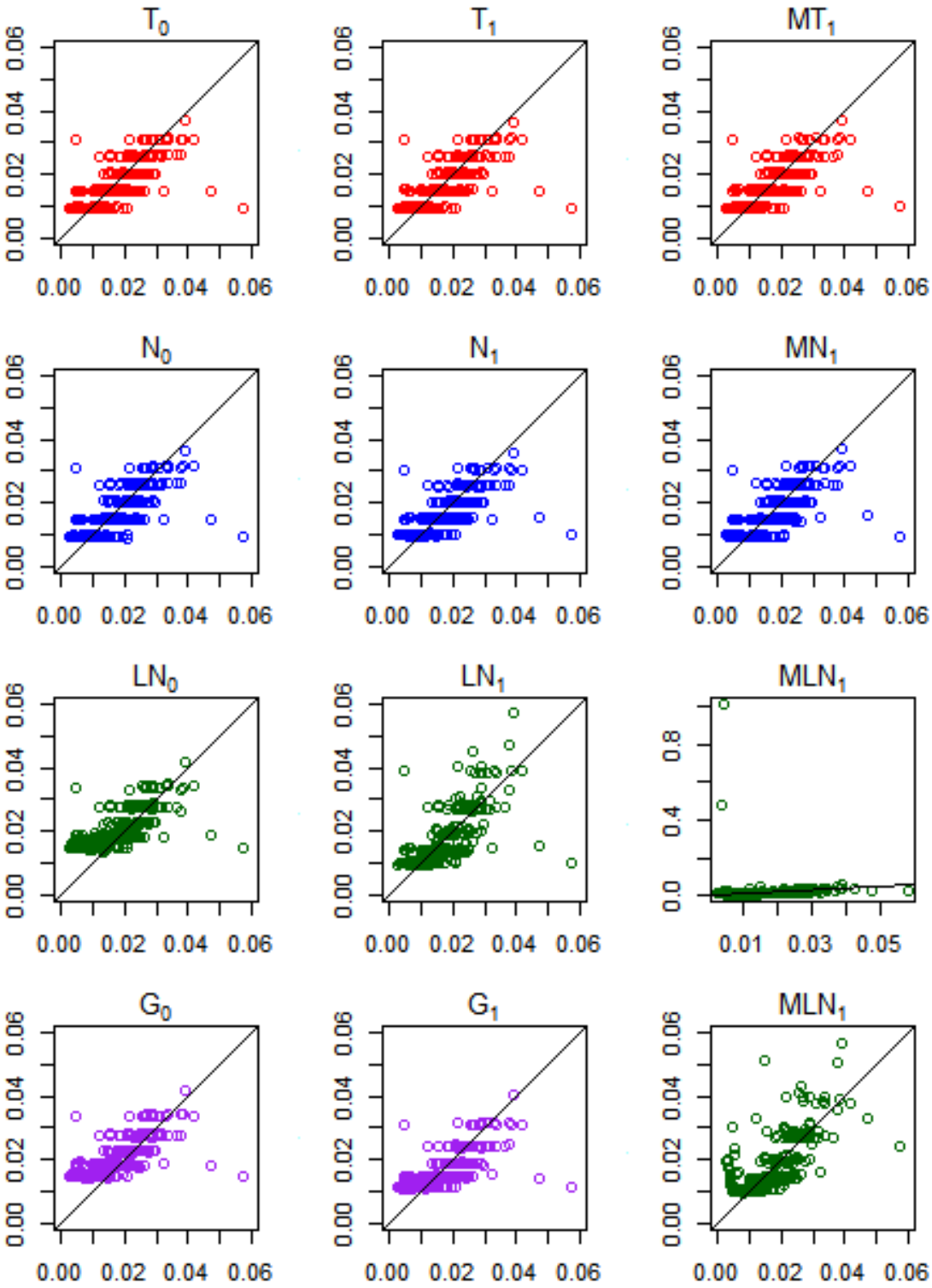

Figure 4.5: Plots of predicted versus observed *SR* for THO1 locus in the NGM SElect^{TM} dataset.

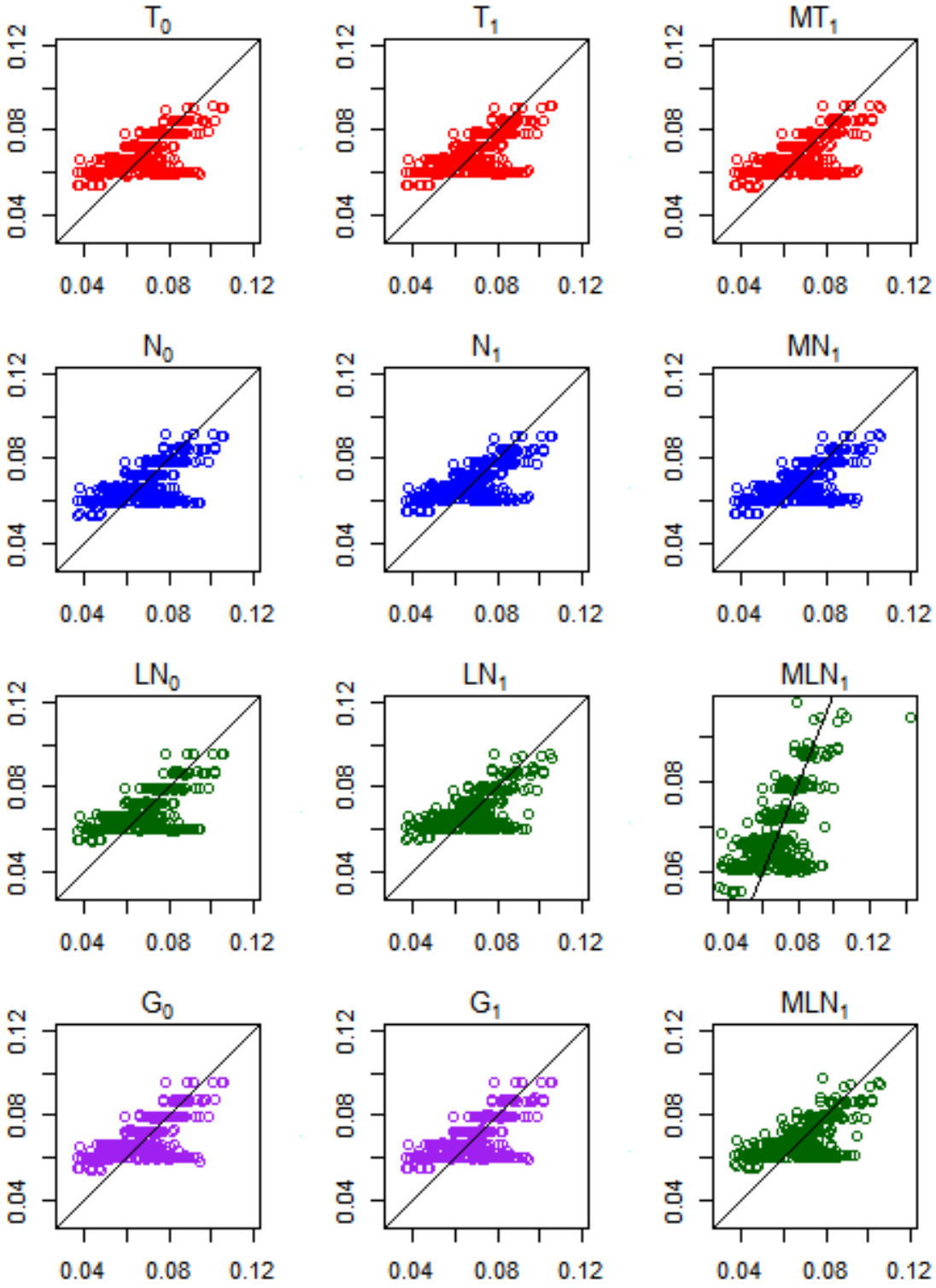

Figure 4.6: Plots of predicted versus observed *SR* for D2S1338 locus in the NGM SElect^{TM} dataset.

larly, a mixture of log-normal densities is no longer a log-normal density. Construction of both P-P and Q-Q plots for gamma and Student's t models are always impossible as they are not simplified into a unique distribution like the standard normal distribution.

Bright et al. [21] also used log-normal Q-Q plots to evaluate the goodness-of-fit of the log-normal models. Figure: 4.1 and Figure: 4.2 graphically illustrate to the goodness-of-fit of both normal and log-normal models fitted for the NGM SElect^{TM} dataset. The Q-Q plots clearly indicate severe lack-of-fit problems in the lower tails of both log-normal and normal models. Normal models exhibit goodness-of-fit problems in the upper tail too. The Q-Q plots illustrate comparatively weak fitting in the middle of the distribution for log-normal models over the normal models. According to Figure: 4.3 and Figure: 4.4, both normal and log-normal models fitted for the Identifiler^{TM} dataset show serious goodness-of-fit problems in both tails of their respective distributions. In addition, both models exhibit considerable departures from the theoretically expected behaviour even in the middle of their distributions. As indicated by the P-P plots, irrespective of the dataset, both types of models reveal thin-tail behaviours in the respective distributions. Lack-of goodness-of-fit problems associated with both normal and log-normal models are possibly reduced with the use of two-component normal and log-normal mixture models. However, the goodness-of-fit of the mixture models cannot be graphically examined.

Figures: 4.5 and 4.6 present the scatter plots of predicted versus observed *SR* for THO1 and D2S1338 loci respectively in the NGM SElect^{TM} dataset. There are no major differences among the relationships between predicted and observed *SR* for both loci. However, Figure: 4.5 shows two extremely large predicted *SR* for THO1 locus for the log-normal mixture model (MLN_1).

4.3 Comparison of Existing and Proposed Models

Various information criteria are available for model comparisons. However, the limitations associated with these criteria usually restrict their applicability. Even though, AIC (the Akaike information criterion) and BIC (the Bayesian information criterion) are well-known in model comparisons, they require the calculation of the log-likehood under the

maximum likelihood estimators (MLEs) of the model parameters. However, it is well-known that the Bayesian estimates calculated with the presence of large samples or flat (especially uniform) priors tend to produce similar estimates to MLEs. In this study, the Bayesian estimates of model parameters were calculated using flat priors, with large samples. Therefore, the 11 models can be compared with either AIC or BIC, replacing MLEs with corresponding Bayesian estimates.

Mixture models are generally treated as singular models. Asymptotic behavioural problems and non-representativeness of the plug-in estimates of posterior parameters of singular models have been highlighted in literature [74]. The comparison of mixture models is recommended with information criteria developed based on singular learning theory. However, this study does not reveal any unusual result in using either AIC or BIC for comparing three two-component simple mixture models. Bright et al. [21] also used AIC to compare the performance of their five models including a log-normal mixture. However, since BIC penalises the model complexity more than AIC for large samples, the 11 models were compared with BIC.

Table 4.1: The differences of BIC values for the NGM SElect^{TM} dataset

Model	LN_0	G_0	LN_1	G_1	N_0	MLN_1	N_1	T_1	T_0	MN_1
G_0	1723									
LN_1	2100	377								
G_1	2727	1004	627							
N_0	3053	1330	953	326						
MLN_1	3584	1861	1484	857	531					
N_1	3640	1917	1540	912	586	55				
T_1	3744	2021	1644	1017	691	160	105			
T_0	3823	2100	1723	1095	769	238	183	79		
MN_1	3979	2256	1879	1251	925	394	339	156	235	
MT_1	4635	2912	2535	1907	1581	1050	995	812	890	656

The differences of BIC values for all possible combinations of the models for the NGM SElect^{TM} and the Identifiler^{TM} datasets are presented in Table 4.1 and Table 4.2 respectively. The models are arranged in increasing order of their performances. The magnitude of the differences reflects the extent that the models in the rows are better than the corresponding models in the columns, with respect to BIC. For example, the first value

Table 4.2: The differences of BIC values for the Identifiler^{TM} dataset

Model	LN_0	N_0	G_0	LN_1	G_1	N_1	MLN_1	T_0	T_1	MN_1
N_0	1687									
G_0	1739	52								
LN_1	3019	1332	1280							
G_1	3901	2214	2162	882						
N_1	4201	2514	2462	1182	300					
MLN_1	5566	3878	3827	2547	1664	1365				
T_0	5626	3938	3887	2607	1724	1425	60			
T_1	6100	4412	4361	3081	2198	1899	534	474		
MN_1	6267	4580	4528	3248	2366	2066	702	642	168	
MT_1	8043	6355	6304	5024	4141	3842	2477	2417	1943	1775

1723 in Table 4.1 represents, BIC(LN_0) - BIC(G_0) = 1723.

Regardless of the dataset, the two component non-standardised Student's t mixture (MT_1) and the two component normal mixture (MN_1) have been selected as the best and the second best model respectively. For both datasets, non-standardised Student's t models, one with locus-specific variance (T_1) and the other with profile-wide variance (T_0) outperform over all the non-mixture models. For both datasets, normal models perform better than log-normal models and non-standardised Student's t models perform consistently better than both normal and log-normal models among all the three modelling categories: profile-wide variance, locus-specific variance, and mixture models. Performance of the gamma models are consistently greater than that of the log-normal models and mostly lower than the normal models, in both locus-specific and profile-wide variance modelling categories. The normal model with locus-specific variance (N_1) is the best and the most convenient option for the forensic practitioners who are not comfortable with advanced modelling techniques such as mixture models or rarely used statistical distributions such as non-standardised Student's t.

4.4 Model Comparison Beyond AIC and BIC

The use of Bayesian estimates in place of MLEs of model parameters is a key assumption in the use of AIC and BIC in Bayesian model comparisons. On the contrary, the infor-

mation criteria established with singular learning theory are recommended for the comparison of mixture models. DIC (the deviance information criterion) is regarded as the Bayesian version of AIC and it provides a convenient way of performance evaluation even with very large complex models. The problem of estimating MLEs for Bayesian models can be avoided with the use of DIC that calculates log-likelihoods based on the Bayesian estimates of model parameters. However, DIC cannot be used for model comparison as the posterior estimates of means are quite delicate under the mixture models. WAIC (the widely available or Watanabe-Akaike information criterion) has been developed based on singular learning theory and is recommended for comparison of both singular and non-singular models. However, the use of WAIC also has a practical limitation in relation to the posterior variance of log predictive distribution. When the posterior variance of any observation, calculated over MCMC simulations exceeds 0.4, WAIC starts to exhibit its inability in model comparison and performance evaluation. Thus, it is vital to explore the behaviour of log densities of the observations under each model when WAIC is used.

Figure 4.7 and Figure 4.8 clearly indicates that a substantial majority of the variances of log predictive densities exceed the standard 0.4 limit in both datasets. In particular, for each model, at least 99.7% of data points in the Identifiler^{TM} dataset and 95.5% of data points in the NGM SElect^{TM} dataset exceed the 0.4 margin. More interestingly, the models associated with normal distribution in each of the three model categories for both datasets, provide on average the lowest posterior variance of log predictive density along with the lowest variability compared to the other models. In contrast, the models related to the non-standardised Student's t distribution indicate the largest posterior variance among all the modelling categories consistently for both datasets.

All the models fitted to the two datasets do not fulfil the posterior variance requirement of log predictive densities and hence, the validity of WAIC is controversial. Leave-one-out cross-validation (LOO-CV) is the next alternative approach available for comparing the performance of the models. However, LOO-CV is computationally very expensive and time consuming. Importance sampling LOO-CV (IS-LOO), truncated importance sampling LOO-CV (TIS-LOO), and Pareto-smoothed importance sampling LOO-CV (PSIS-LOO) are three variants of importance sampling discussed in Chapter 3 as the potential

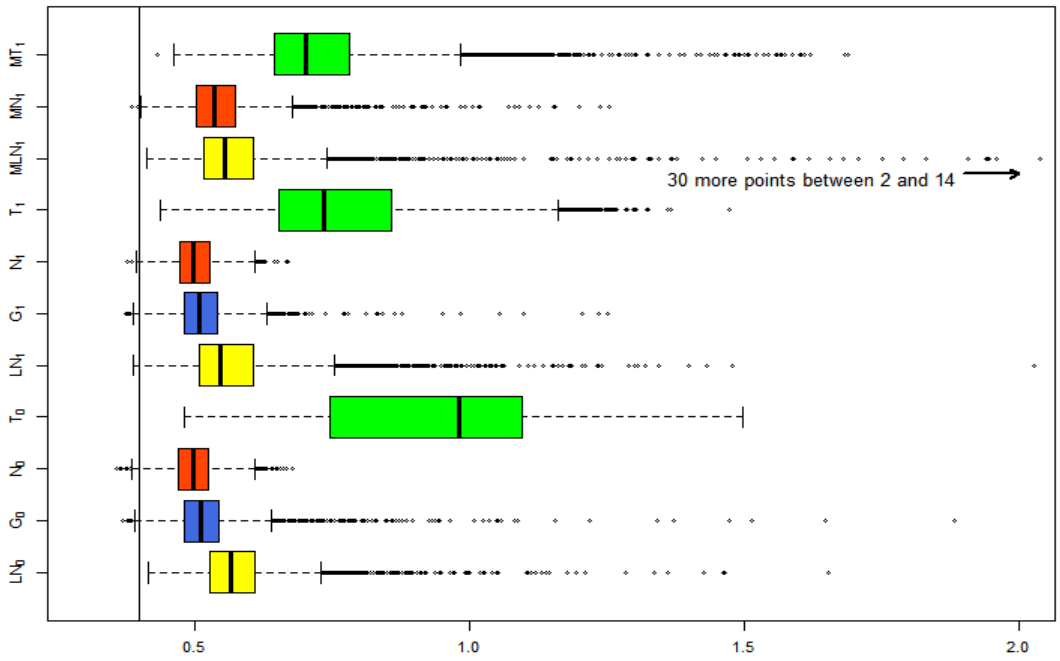

Figure 4.7: Posterior variances of log predictive densities of the models for the NGM SElect^{TM} dataset.

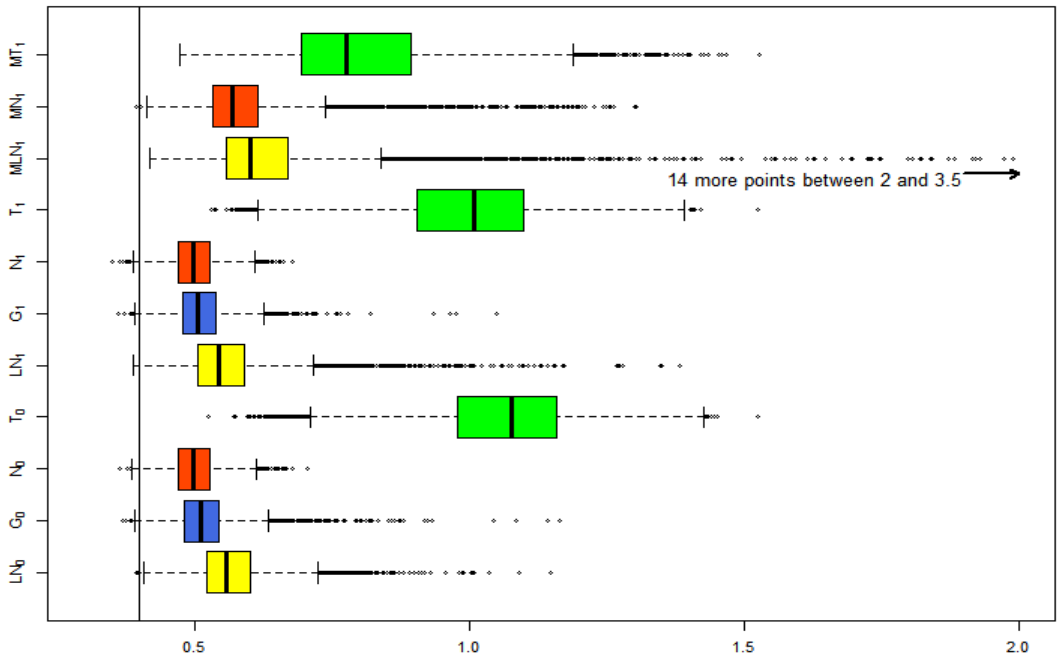

Figure 4.8: Posterior variances of log predictive densities of the models for the Identifiler^{TM} dataset.

approximations to LOO-CV. This study used and compared the three LOO-CV approximations together with the computed log pointwise predictive density (clppd) and expected log pointwise predictive density calculated under the two versions of WAIC (i.e.; WAIC and WAIC_{alt}).

All the posterior predictive measures presented in Figure: 4.9 and Figure: 4.10 are estimated based on the individual posterior densities calculated at each observed value in the respective datasets. These individual density estimates represent a summary measure of a sample of 2000 posterior draws of densities. The simple arithmetic mean is used as the summary measure in estimating the log point-wise density (clppd) of each model. IS-LOO is calculated based the harmonic mean of sample values. The simple arithmetic mean gives equal priority to all the observations whereas harmonic mean gives high priority to smaller values and low priority to larger values. Therefore, harmonic mean calculated for any distinct dataset is always less than its simple arithmetic mean. As a consequence of this relationship between simple arithmetic mean and harmonic mean, it is intuitive to expect a smaller value for IS-LOO compared to the corresponding clppd. In TIS-LOO, a weighted average of sample values is used as the summary measure. The raw importance weights (the reciprocals of respective posterior densities) are used as the weights. However, the raw importance weights that are larger than the truncation point (the mean of importance weights multiplied by the square root of the number of MCMC draws) are replaced by the truncation value itself prior to the calculations. The larger importance weights are corresponding to the smaller predictive densities by its definition and as a result, smaller predictive densities receive relatively low weights. Hence, the relative significance (percentage weight) of the larger densities corresponding to non-truncated weights certainly increases. Consequently, TIS-LOO yields a larger lppd than IS-LOO. However, with the presence of a huge sample of posterior draws, practically a small or no impact from the truncation can be expected as the value of truncation point is proportional to the square root of the size of posterior draws.

The lower values of TIS-LOO and IS-LOO in comparison with clppd can be explained in relation to the behaviour of the variances of log predictive densities. When the variance of log predictive densities of an observation calculated based on the MCMC sample

Table 4.3: Calculated log predictive densities of the models for the NGM SElect^{TM} dataset.

Model	IS	TIS	PSIS	clppd	WAIC	WAICalt	pWAIC	pWAICalt
LN_0	9002	10447	9722	14188	12575	11492	1613	2696
LN_1	10207	11576	10855	15295	13680	12600	1616	2695
MLN_1	9834	11871	11617	16475	14411	13636	2065	2839
G_0	10509	11649	11048	14961	13470	12529	1491	2432
G_1	11153	12259	11677	15533	14052	13132	1482	2401
N_0	11383	12427	11870	15573	14149	13253	1424	2920
N_1	11873	12784	12208	15926	14492	13596	1434	2330
MN_1	12053	13133	12609	16501	14903	13961	1598	2540
T_0	6334	9650	8370	16408	14038	12050	2370	4358
T_1	8469	11105	10065	16488	14458	12889	2029	3598
MT_1	9751	11979	11147	16821	14734	13395	2088	3427

Note: Calculated penalty constants of WAIC are given in the last two columns. WAIC and WAIC alt excluding the factor -2 are reported.

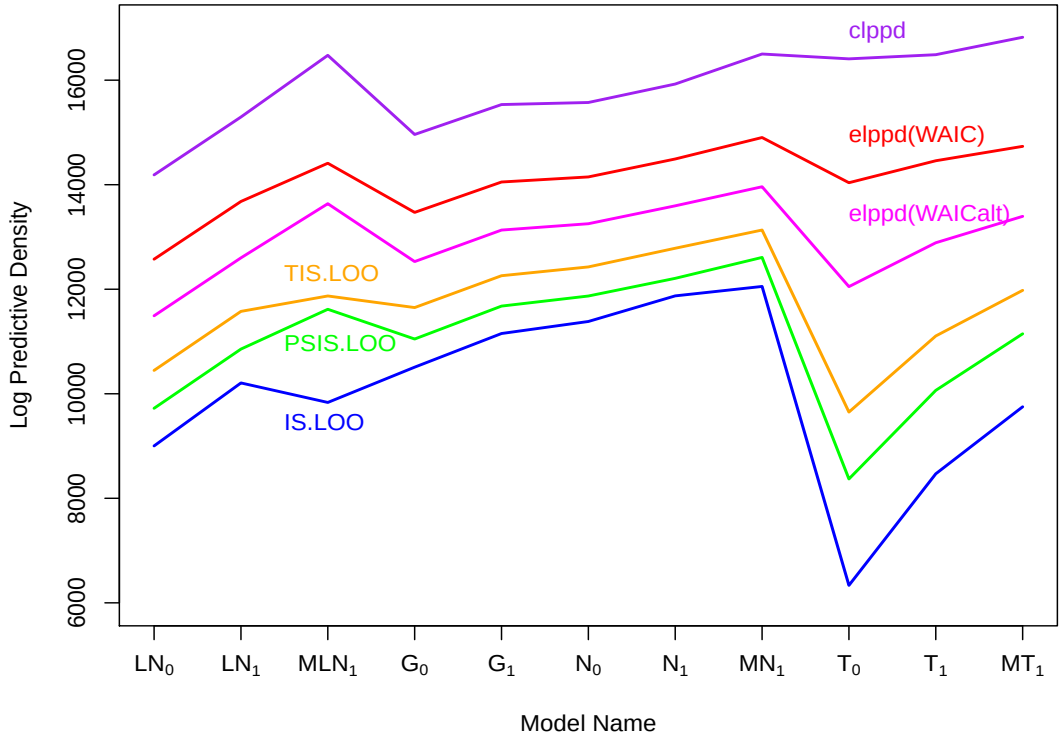

Figure 4.9: Calculated log predictive density profiles of the models for the NGM SElect^{TM} dataset.

is larger, then the variance of predictive densities must be very large. Therefore an ob-

Table 4.4: Calculated log predictive densities of the models for the Identifiler^{TM} dataset.

Model	IS	TIS	PSIS	clppd	WAIC	WAICalt	pWAIC	pWAICalt
LN_0	14902	16962	15929	22452	20079	18498	2373	3953
LN_1	16601	18627	17600	24000	21652	20101	2348	3900
MLN_1	17974	20283	19363	26400	23668	21920	2733	4480
G_0	16640	18335	17448	23226	21038	19639	2187	3587
G_1	17807	19536	18687	24386	22211	20832	2175	3554
N_0	16971	18527	17682	23244	21111	19767	2133	3477
N_1	18283	19846	19022	24563	22425	21085	2138	3478
MN_1	19761	21378	20753	26725	24127	22623	2599	4102
T_0	8749	14706	12547	26416	22512	19065	3904	7351
T_1	10187	15718	13675	26607	22880	19679	3727	6928
MT_1	14935	18876	17518	26984	23638	21369	3346	5615

Note: Calculated penalty constants of WAIC are given in the last two columns. WAIC and WAIC alt excluding the factor -2 are reported.

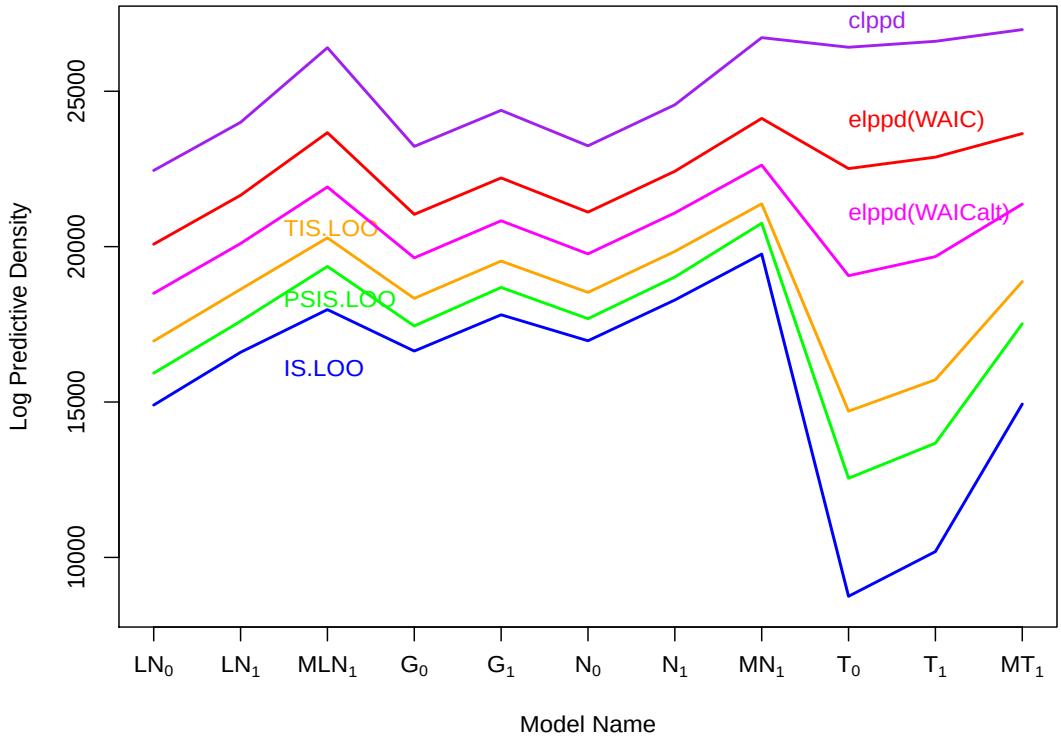

Figure 4.10: Calculated log predictive density profiles of the models for the Identifiler^{TM} dataset.

servation with a relatively larger posterior variance of log predictive densities must have a larger spread than another similar observation with a relatively low variance. The har-

monic mean of any distinct set of observations is always smaller than its simple arithmetic mean. Therefore, in general, it is intuitive to expect a comparatively bigger difference between the harmonic mean and its simple arithmetic mean for an observation with a larger posterior variance in log predictive densities. Consequently, the models with relatively large posterior variances of log predictive densities produce larger reductions in IS-LOO and TIS-LOO in comparison with clppd. Figure: 4.9 and Figure: 4.10 clearly exhibit parallel log-likelihood profiles for both datasets except for the models associated with the Student's t distribution. The model MLN_1 fitted to the NGM SElect^{TM} dataset also shows a slightly larger reduction in IS-LOO and TIS-LOO.

Figure 4.7 shows very large posterior variances of log predictive densities for all the models fitted to the NGM SElect^{TM} dataset based on the non-standardised Student's t distribution. Model T_0 corresponds to the largest variance, while T_1 and MT_1 correspond to the second and the third largest variances. The variance corresponding to model T_0 is substantially larger than that of both T_1 and MT_1 models. The variance corresponding to model T_1 is slightly larger than that of MT_1. According to Figure 4.9, the largest reduction in the estimates of IS-LOO and TIS-LOO can be observed in model T_0. The second and the third largest reductions correspond to the models T_1 and MT_1 respectively. In addition, model T1 has little larger reductions in both IS-LOO and TIS-LOO compared to those of model MT_1. The model MLN_1 fitted to the NGM SElect^{TM} dataset produces a longer upper tail in the distribution of the posterior variances of log predictive densities than all the other models fitted to the same dataset. However, the median of the posterior variances is approximately similar in all the models except the non-standardised Student's t models. The unusual reduction in both IS-LOO and TIS-LOO estimates of MLN_1 model is probably a consequence of the extended upper tail of the distribution of posterior variances.

The Identifiler^{TM} dataset also produces larger posterior variances of log predictive densities for all the models that are based on the non-standardised Student's t distribution. The variances corresponding to the models T_0 and T_1 (Figure 4.8) are substantially larger than that of model MT_1. Even though the model T_0 corresponds to the largest posterior variance, the difference of the posterior variances between T_0 and T_1 is much smaller

than the difference between T_1 and MT_1. Figure 4.10 clearly indicates the biggest and the second biggest reductions in IS-LOO and TIS-LOO compared to clppd of the models T_0 and T_1 respectively. The reduction associated with the model T_1 is slightly lower in magnitude compared to that of the model T_0. The upper tail of the distribution of posterior variances of log predictive densities corresponding to the model MLN_1 is slightly longer than that of all the other models. However, the length of the tail is longer for the NGM SElect^{TM} dataset in comparison with that of the Identifiler^{TM} dataset. It results in considerably unusual reductions in IS-LOO and TIS-LOO of the model MLN_1 compared to those of the other models except the non-standardised Student's t models. The relationship between the posterior variances of log predictive densities and the reductions in IS-LOO and TIS-LOO compared to the corresponding value of lppd was consistently established by the models fitted to the NGM SElect^{TM} and the Identifiler^{TM} datasets. In particular, the reductions are larger for the larger variances and smaller for the smaller variances.

The calculations associated with PSIS-LOO of an observation are rather complex than the similar calculations in TIS-LOO. Initially, the largest 20% of the raw weights are considered and replaced by the estimated order statistics fitting a Pareto distribution to them. Subsequently, the method adopted in TIS-LOO is used to obtain the truncated weights. The truncation point under the PSIS-LOO is defined as the multiplication between the average of importance weights (including revised weights) and the three fourth power of the posterior sample size. Although there are some similarities between PSIS-LOO and TIS-LOO, it is very difficult to compare the values that will be inferred by these two methods as their weight calculations are not directly comparable. However, Figure 4.9 and Figure 4.10 consistently demonstrate that the estimated PSIS-LOO is approximately the average of IS-LOO and TIS-LOO for all the models with fairly similar posterior variances of log predictive densities. In case of larger posterior variances, in contrast, the estimated PSIS-LOO is closer to TIS-LOO than IS-LOO. Vehtari et al. [172] strongly recommended to report the cases whose estimated shape parameter of the generalised Pareto distribution ($\hat{k}$) exceeds 0.5. PSIS-LOO offers an improved accuracy when $\hat{k} > \frac{1}{2}$. However, the estimated predictive density under this method tends to be more biased with high variance

Table 4.5: The distribution of estimated shape parameters ($\hat{k}$) under each model for the NGM SElect^{TM} (NGM) and the Identifiler^{TM} (Idn) datasets.

Model	LN_0	LN_1	MLN_1	G_0	G_1	N_0	N_1	MN_1	T_0	T_1	MT_1
NGM											
$\frac{1}{2} < \hat{k} < 1$	69.4	70.6	68.9	82.3	84.5	86.0	86.6	87.6	13.6	27.6	43.1
$\hat{k} \geq 1$	30.6	29.4	31.1	17.7	15.5	14.0	13.4	12.4	86.4	72.4	56.7
Idn											
$\frac{1}{2} < \hat{k} < 1$	71.2	73.5	72.9	83.0	84.7	86.0	86.6	91.4	3.6	5.1	29.4
$\hat{k} \geq 1$	28.8	26.5	27.1	17.0	15.3	14.0	13.4	8.6	96.4	94.9	70.6

Note: The percentage (%) of observations are tabulated according to their estimated $\hat{k}$ value.

when $\hat{k}$ is not less than one (i.e. $\hat{k} \geq 1$). The shape parameters of the generalised Pareto distribution fitted to the largest 20% of the raw weights for each of observation under each model for both datasets are very informative in relation to the accuracy in PSIS-LOO. Table 4.5 summarises the percentage of observations which produced larger $\hat{k}$ values under each model for both datasets. An extensive majority of the $\hat{k}$ values associated with all the non-standardised Student's t models fitted to both datasets exceed unity. Therefore the bias and the variance of log predictive density (lppd) estimated under PSIS-LOO are very high for these models. PSIS-LOO estimated under the three log-normal models also demonstrates similar problems as their percentages of $\hat{k}$ values that exceed unity are moderately large. The three models based on normal distribution indicate the minimum risk of bias and variance as they produced relatively lower percentages of $\hat{k}$ values greater than unity.

Assuming the validity of Watanabe-Akaike (or widely available) information criterion (WAIC), one can discuss the effect of posterior variances of log predictive densities in estimating the penalty constant. WAIC is defined with two versions as it uses two different types of bias correction (or penalty) terms against over-fitting. The first penalty term p_{WAIC} is defined as twice as the sum of differences between the log of the posterior mean of predictive densities and the posterior mean of log predictive densities estimated over posterior draws for each observed value in the dataset. Any logical relationship between p_{WAIC} and the posterior variances of log predictive densities is not identified. However,

the models that show very large posterior variances (T_0, T_1, and MT_1) result in larger penalty terms than others consistently for both datasets (see Table 4.3 and Table 4.4). The alternative version of WAIC is highly recommended for practical use. The penalty term of this version ($p_{\text{WAIC alt}}$) is defined as the sum of posterior variances of log predictive densities. Therefore, posterior variances of log predictive densities and $p_{\text{WAIC alt}}$ are highly related. Hence, it is not surprising to see larger penalty terms for non-standardised Student's t models as they provided larger posterior variances. In addition to these models, the other two mixture models: MN_1 and MLN_1 also result in slightly larger values for both penalty terms. The alternative version WAIC heavily penalises all the models compared to the usual method.

The penalty parameter of the Bayesian information criterion (BIC) is defined as the product between the logarithmic value of the sample size and the number of parameters in the model. The number of parameters in the models varies between 33 to 97 for the NGM SElect^{TM} dataset and 31 to 91 for the Identifiler^{TM} dataset. The two datasets consist of 4646 and 6949 observations respectively. Therefore, the penalty constant defined under the BIC ranges from 279 to 819 for the NGM SElect^{TM} dataset and from 274 to 805 for the Identifiler^{TM} dataset. In BIC, the sum of negative log-likelihoods is multiplied by a factor of -2 and added to the penalty constant. Therefore, the actual penalty constant without the factor -2 is only a half of the constant (hence the maximum is 410). Table 4.3 and Table 4.4 present the expected log predictive densities (elppd) estimated under WAIC excluding the factor -2. Therefore, p_{WAIC} was at least 5 to 11 times larger than the penalty constant estimated under BIC for the NGM SElect^{TM} dataset. For the Identifiler^{TM} dataset, it was at least 8 to 19 times. Similarly, $p_{\text{WAIC alt}}$ was at least 8 to 21 times lager for the NGM SElect^{TM} dataset and 13 to 36 times larger for the Identifiler^{TM} dataset.

As observed in Figure: 4.9 and Figure: 4.10, the estimated log predictive density of each model has been heavily penalised by IS-LOO, TIS-LOO, and PSIS-lOO for both datasets. Therefore, none of them can be treated as a good approximation to the exact LOO-CV. The limitation associated with the posterior variances of log predictive densities restricts the use of WAIC. If it is possible to calculate the exact LOO-CV for these models, then it would be possible to explore the real picture and test the reliability of the methods

in evaluating model performance. However, it will be really computationally expensive as the NGM SElect^{TM} and the Identifiler^{TM} datasets consist of 4646 and 6949 observations respectively. The *k*-fold (e.g. 10-fold) cross-validation is an alternative method that could be used to reduce the computational cost associated with LOO-CV. However, that method was not used in this study.

Regardless of the degrees of freedom parameter in the non-standardised Student's t distribution, the number of parameters of the models within each modelling category: profile-wide variance, locus-specific variance, and two-component mixtures is equal for a given dataset. Hence it is possible to assume equal complexities for the models within each category. This assumption provides a basis on which to compare the performance of these models within each modelling category. Although, the overall lppds estimated under each method for each model are unable to reflect the goodness-of-fit in absolute sense, the differences of them evaluated under each method can be utilized to measure the relative performance of the models.

A comparison of these models within each modelling category considering all the methods specified in Table 4.3 and Table 4.4 is very informative. In addition, the consistency of them over both datasets is also vital. According to IS-LOO, TIS-LOO, PSIS-LOO, and WAIC alt, when profile-wide and locus-specific variance models are considered separately, the models based on normal distribution (N_0 and N_1) outperform over the other three models in the respective modelling categories, consistently for both datasets. In contrast, WAIC does not exhibit any consistency across the datasets. When the models are ranked based on WAIC, N_0 and N_1 for the NGM SElect^{TM} dataset and T_0 and T_1 for the Identifiler^{TM} dataset are selected as the best models. However, the difference in the WAICs for normal and non-Standardised Student's t models is very small. Considering the complexity of Student's t models against normal models, the models related to normal distribution can be identified as the best.

Obviously, a much better performance is expected with two-component mixture models relative to their univariate counterparts. Again, the model developed on normal distribution (MN_1) has been identified as the best two-component mixture and MT_1 as the second best model based on each performance measure, consistently for both datasets.

Table 4.6: Bayesian p-values based on marginal predictive distributions (p_M) and chi-squared discrepancy measure (p_D) for the NGM SElect^{TM} (NGM) and the Identifiler^{TM} (Idn) datasets.

Model	LN_0	LN_1	MLN_1	G_0	G_1	N_0	N_1	MN_1	T_0	T_1	MT_1
NGM											
p_M	0.48	0.48	0.49	0.49	0.49	0.50	0.50	0.50	0.49	0.49	0.49
p_D	0.52	0.51	0.50	1.00	1.00	0.49	0.49	0.51	0.52	0.16	0.87
Idn											
p_M	0.51	0.51	0.51	0.52	0.51	0.52	0.52	0.50	0.49	0.49	0.49
p_D	0.51	0.49	0.51	0.36	0.49	0.49	0.49	0.51	0.01	0.02	0.60

Note: The standard deviations of marginal predictive probabilities calculated under these models ranges from 0.0071 to 0.0079 for the NGM SElect^{TM} dataset and from 0.0058 to 0.0065 and the Identifiler^{TM} (Idn) dataset.

Therefore, the models related to normal distribution can be recommended as the best within each modelling category. The simplicity and familiarity of normal models are additional advantages, especially for the forensic practitioners who may not be greatly familiar with statistics. Finally, considering both the performance and simplicity, the model based on normal distribution with locus-specific variance (N_1) is recommended as the best univariate model. Similarly, the two-component normal mixture (MN_1) is recommended as the best among mixture models.

4.5 Bayesian p-values and L-measure for Model Comparison

Performance of the models can also be compared with the Bayesian version of p-values. Table 4.6 summarises the p-values calculated upon marginal predictive distributions of the data (p_M) and chi-squared discrepancy measure (p_D). The two p-values: p_M and p_D are calculated based on Equation 3.3 and Equation 3.2 respectively. As shown in the table, the p-values that represent marginal predictive distribution do not show any problem with any of the models as they are extremely close to the typical value of 0.5 for both datasets. Furthermore, the considerably low standard deviations of the marginal predictive p-values for

both datasets (see the note under Table 4.6) imply the low-variability or closeness of the individual p-values around their overall means (p_M). Hence, future observations generated by each of these fitted models will greatly exhibit unbiased characteristic around the observed values for all the models. However, the p-values that are based on chi-squared discrepancy measure indicate serious problems in both gamma models (G_0 and G_1) for the NGM SElect^{TM} dataset and the non-standardised Student's t models with profile-wide and locus-specific variances (T_0 and T_1) for the Identifiler^{TM} dataset (i.e the extreme p-values that are close to 0 or 1). The larger p-values of both gamma models under the NGM SElect^{TM} dataset indicate larger deviations of the predictions than actual observations, from the estimated mean of the respective distributions. The smaller p-values of T_0 and T_1 models fitted for the Identifiler^{TM} dataset, in contrast, indicate relatively smaller deviations of the predictions than actual observations, from the estimated mean of the respective distributions. However, both smaller and larger deviations of the predictions compared to the actual observations highlight potential discrepancies between actual observations and predictions under the models.

The L-measure is calculated as a weighted sum of the variance and squared bias of the future observations generated based on the fitted model, and it can be informatively used in model comparisons and performance evaluations. Table 4.7 and Table 4.8 summarise mean L-measures and their standard deviations calculated over 100 repetitions of each model fitting. The two tables clearly exhibit comparatively large L-measures for all the log-normal models fitted to both datasets. Two component log-normal mixture model exhibits the largest mean value for the Identifiler^{TM} dataset along with a large standard deviation. Surprisingly, L-measures calculated under the same model for the NGM SElect^{TM} dataset produced infinitely large summary values. Hence, MLN_1 model was not included in Table 4.7. Since all the observed stutter ratios in both datasets are less than 0.17, larger predictions are not expected under any reliable model. A careful in-depth analysis of the model revealed the existence of unbelievably large predictions for some points in the NGM SElect^{TM} dataset, which is the root cause of the large summaries observed under MLN_1 model. For example, on average, there were more than 10 observed values that create a bias of more than 100. Similarly, at least, on average, 18, 38, and 72

Table 4.7: Means and standard deviations of L-measures of the models for the NGM SElect^{TM} dataset.

ν value	0.0	0.2	0.4	0.5	0.6	0.8	1.0
LN_0	1.4756	1.5176	1.5583	1.5783	1.5981	1.6369	1.6747
	0.0602	0.0594	0.0586	0.0582	0.0579	0.0572	0.0566
LN_1	2.0901	2.1265	2.1623	2.1799	2.1974	2.2318	2.2657
	0.8226	0.8183	0.8143	0.8124	0.8105	0.8068	0.8034
G_0	0.8756	0.9328	0.9867	1.0126	1.0378	1.0865	1.1331
	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0016
G_1	0.8805	0.9355	0.9873	1.0123	1.0366	1.0837	1.1287
	0.0038	0.0036	0.0034	0.0033	0.0033	0.0032	0.0031
N_0	0.7402	0.7987	0.8532	0.8792	0.9044	0.9529	0.9990
	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
N_1	0.7373	0.7960	0.8507	0.8767	0.9020	0.9506	0.9968
	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0005
MN_1	0.7343	0.7939	0.8494	0.8758	0.9015	0.9507	0.9975
	0.0069	0.0072	0.0075	0.0077	0.0078	0.0081	0.0084
T_0	0.7530	0.8113	0.8657	0.8916	0.9169	0.9653	1.0115
	0.0026	0.0024	0.0023	0.0022	0.0021	0.0021	0.0020
T_1	0.7266	0.7864	0.8420	0.8684	0.8941	0.9434	0.9902
	0.0018	0.0017	0.0016	0.0015	0.0015	0.0015	0.0014
MT_1	0.7466	0.8070	0.8632	0.8900	0.9160	0.9659	1.0133
	0.0393	0.0371	0.0353	0.0345	0.0337	0.0323	0.0310

Cell contents: Means and standard deviations of L-measures

observed stutter ratios are corresponding to a minimum bias of 10, 1, and 0.1 respectively.

Normal and non-standardised Student's t distributions based models perform equally better than gamma-based models consistently for both datasets. The variance of the predictions is much bigger than the squared bias for all the models for both datasets. Consequently, the changes in weightage (ν) have not been effective in adding useful information to the results.

4.6 Summary

Bright et al. [21] developed five models to explain the behaviour of PCR stutter ratio (*SR*). They modelled *SR* as a right-skewed heavy-tailed distribution using log-normal and gamma models. This chapter compares the performance of these five models and

Table 4.8: Means and standard deviations of L-measures of the models for the Identifiler^{TM} dataset.

ν value	0.0	0.2	0.4	0.5	0.6	0.8	1.0
LN_0	1.4615	1.5055	1.5482	1.5691	1.5898	1.6303	1.6698
	0.0095	0.0095	0.0094	0.0094	0.0094	0.0094	0.0094
LN_1	1.1866	1.2380	1.2874	1.3113	1.3349	1.3808	1.4252
	0.0102	0.0100	0.0098	0.0097	0.0096	0.0095	0.0094
MLN_1	3.1550	3.1888	3.2220	3.2384	3.2546	3.2865	3.3179
	2.2954	2.2858	2.2765	2.2721	2.2677	2.2592	2.2510
G_0	0.9531	1.0103	1.0645	1.0905	1.1160	1.1652	1.2124
	0.0014	0.0014	0.0013	0.0013	0.0013	0.0013	0.0013
G_1	0.9147	0.9731	1.0282	1.0547	1.0805	1.1304	1.1781
	0.0020	0.0019	0.0019	0.0019	0.0019	0.0019	0.0019
N_0	0.9238	0.9780	1.0293	1.0541	1.0782	1.1250	1.1699
	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
N_1	0.8765	0.9334	0.9871	1.0129	1.0380	1.0865	1.1330
	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
MN_1	0.8358	0.8961	0.9526	0.9796	1.0059	1.0565	1.1048
	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
T_0	0.7639	0.8297	0.8906	0.9196	0.9477	1.0015	1.0525
	0.0027	0.0025	0.0023	0.0022	0.0022	0.0020	0.0020
T_1	0.8070	0.8692	0.9272	0.9549	0.9819	1.0336	1.0829
	0.0032	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025
MT_1	0.8262	0.8882	0.9461	0.9738	1.0007	1.0525	1.1018
	0.0027	0.0026	0.0025	0.0024	0.0024	0.0023	0.0023

Cell contents: Means and standard deviations of L-measures

the six new models proposed in this study for predicting *SR*. Providing relevant theoretical justifications in Chapter 2, in these six models, *SR* was modelled as a symmetrically distributed random variable, and the distributions proposed were non-standardised Student's t and normal. In each model, the mean *SR* was modelled as a locus-specific simple linear regression of longest uninterrupted sequence (*LUS*). In variance modelling, the models were classified into three categories, namely, profile-wide variance, locus-specific variance, and two-component mixture models.

The selected model evaluation criteria include graphical evaluation methods (Q-Q and P-P plots), BIC, WAIC, and LOO-CV approximations. The usability of these criteria under different modelling conditions were discussed in detail and the best criteria

for the models were selected accordingly. The validity of distributional assumptions of log-normal and normal models was examined using Q-Q and P-P plots which indicated several lack-of-fit problems in these models. The mixture models were initially evaluated using BIC and WAIC whereas WAIC which uses posterior predictive distribution instead of point estimates of parameters in calculating likelihood is relatively better. However, WAIC is valid only when the posterior variance of log-predictive densities calculated for each observation is not more than 0.4. In this study, for all the models within each dataset, more than 95% of the observations exceed 0.4 limit. Therefore, leave-one-out cross-validation (LOO-CV), the best measure of predictive model accuracy was approximated using IS (importance sampling), TIS (truncated importance sampling), and PSIS (Pareto smoothed importance sampling) as the computational cost associated with exact LOO-CV is very high for large datasets. However, these measures also had some issues in evaluating models considered in the study. The three measures are smaller than the corresponding calculated log pointwise predictive densities (clppd) of all the models within each dataset, and the degree of the reduction is higher for the models with larger posterior variances of lppd. However, all of them are estimates of unobserved log-likelihood of the model for new data. Furthermore, for each model, there is a substantial proportion of observations whose shape parameters of the generalised Pareto distributions exceeds one, indicating a greater bias in PSIS-LOO measures, and this confirms lower values of PSIS-LOO. When the posterior variances of lppd are not largely varied, all the criteria (WAIC, WAIC_{alt}, IS-LOO, TIS-LOO, and PSIS-LOO) provided parallel lppd profiles indicating approximately similar results in performance evaluation of the models. When the model complexity in terms of the number of parameters is concerned, all the models within each modelling category: profile-wide variance, locus-specific variance, and two-component mixture are similar except the degrees of freedom parameter in the Student's t models. Despite the limitations in WAIC and LOO-CV measures, they all confirm that the models based on normal distribution outperform in all the modelling categories, for both datasets.

Bayesian p-values that represent the marginal predictive distributions are close to the desired value 0.5 and hence, do not reveal any problem in any model fitted to the datasets. However, the p-values that are based on the chi-squared discrepancy measure indicate

problems in predictions of the gamma models fitted to the NGM SElect^{TM} dataset and both profile-wide and locus-specific non-standardised Student's t models fitted to the Identifiler^{TM} dataset. The deviations of predictions in comparison with actual observations, from the estimated mean of the respective distribution were larger in the gamma models and smaller in the non-standardised Student's t models. A few unbelievably large predicted values produced by the log-normal mixture model result in large L-measures, and hence indicate larger variations in the predictions.

In summary, when BIC is considered as the model selection criteria, two-component non-standardised Student's t mixture (MT_1) and the two-component normal mixture (MN_1) models are selected as the best and the second best models. The first and second best non-mixture models respectively are non-standardised Student's t (T_1) and normal (N_1) models with locus-specific variance. According to the LOO-CV approximations (IS-LOO, TIS-LOO, and PSIS-LOO), two-component normal mixture model (MN_1) performed better than all the other models consistently for both datasets. Among the non-mixture models, the normal model with locus-specific variance (N_1) is selected as the best. The models related to normal distributions are easy to understand, easy to implement, and most likely are not computationally expensive. Therefore, considering the computational complexity associated with non-standardised Student's t distribution, this study recommends to use the two-component normal mixture model (MN_1) for the caseworks related to PCR stutter (stutter ratio) problems. However, when the complexity of mixture models cannot be ignored, the normal model with locus-specific variance (N_1) is recommended.

Chapter 5

Investigation and Assessment of Hierarchical models

5.1 Introduction

The performance of locus-specific variance models (LN_1, G_1, N_1, and T_1) and three two-component mixture models (MLN_1, MN_1, and MT_1) have been discussed as non-hierarchical models in chapters 2 and 4. Since these models have been identified as better performing than the profile-wide variance models, this chapter introduces the hierarchical models of these seven models and evaluate their performance. The relevant performance measures have already been discussed in Chapter 3 with their limitations.

Hierarchical modeling can be basically regarded as a generalised version of regression methods [70]. It uses for the purposes of prediction, data reduction, and casual inference in both observational and experimental studies. Hierarchical modelling is essential in prediction, very useful in data reduction, and helpful in causal inference. Even though the level of prominence varies based on the purpose, hierarchical models provide an overall improvement over the use of regression models. Usually, multilevel models are used in modelling observed data conditionally on a particular set of parameters that also come from certain probability distributions with a further set of parameters known as hyper-parameters [71]. A multilevel model can be regarded as a hierarchical model for two reasons: the structure of data and its own hierarchy, where the group level parameters are

controlled by the parameters of the upper-level model or models [73]. In many studies, potential factors are considered as a part of the designing of data collection. A reflection of these factors in a model makes it highly informative with regard to the inferential aspects of the study. The cluster indicators at each level of the data design can be easily incorporated in multilevel modelling.

Multilevel models can also be discussed in the context of data pooling. Analysis of the data ignoring its inherent hierarchical structure is called complete pooling, and it suppresses variations in the data, which may affect the overall objective of the study [73]. No pooling, in contrast, performs a separate analysis for each source of the data and tends to give misleading inferences. Both group-level and individual models are simultaneously incorporated in multilevel models, and they enable partial (or semi) pooling of the data [72, 100]. It generates results that are more informative than their extreme alternatives. However, both complete and no-pooling models are useful in a preliminary analysis.

Assume the observed data y_{ij} that represent j^{th} $(j = 1,2,\ldots,n_i)$ observation within $i^{th}(i = 1,2,\ldots,k)$ group can be used to estimate population parameters θ_is which are not actually observed. A scenario that assumes independence among the groups often performs individual independent analysis for each group. All the groups can be combined into a single collection assuming identical model parameters for them, and a single analysis can be performed. The models involved in these two types of analyses are called non-hierarchical. In the Bayesian context, per-group parameters, or the parameters of the combined group in a non-hierarchical model, assume appropriate prior distributions with fixed hyper-parameters.

Joint probability models for statistical applications involving multiple parameters need to reflect the interdependence of parameters [72]. In the context of observed variables, presumably there may be some real-world dependencies among these groups due to geographical, environmental, socio-economical,or any other potential predictors. The scenarios that indicate such dependencies typically correspond to hierarchically structured data. In hierarchical modelling, the groups are treated as different but related sub populations [14]. At the first stage of any hierarchical model, the uncertainty in the observables within each group is separately modelled with parametric distributions. Then the model

parameters $\theta_1, \theta_2, \ldots, \theta_k$ are themselves modelled within each group with suitable prior distributions. Generally, these take the same functional form across the groups but with different unknown (hyper) parameters. The interrelatedness among the groups are then considered and incorporated in the model, introducing the third stage hierarchy which binds hyper-parameters with a second-level distribution called hyper-prior. Parameters of the hyper-prior distribution can be decided based on the historical data and/or experience. Otherwise, considering the theoretical aspects of hyper-parameters, suitable vague (flat) hyper-prior distributions are used. However, complex Bayesian models typically consist of more than three levels, and hence the parameters of hyper-prior distributions are again statistically treated with higher-level prior distributions. Priors of these multilevel models are hierarchical as they are specified in layers. Mathematically, there is no restriction to the number of layers in hierarchical prior distributions [80].

The use of simple non-hierarchical models with the presence of hierarchically structured data always throws-out useful information in the data. Models with a smaller number of parameters fail to accurately fit large sets of data, while those using a large number of parameters are likely to overfit [72]. Generally, over-parametrised simple models tend to result in overfitting including the random noise of the data, though they fail in generating good predictions. Hierarchical models, in contrast, can be more effectively utilised in many complex real-world problems expecting good predictions. Fully Bayesian approaches are largely adopted along with modern computational methods such as MCMC, in estimating general and specific parameters that reflect population characteristics and group-level profiles respectively [140].

In non-hierarchical models, the group level parameters of each group are independently estimated based on the observed data. Hierarchical models, in contrast, accommodate the possible interdependencies among groups through first-level prior distributions, whose parameters are interconnected through the hyper-prior distribution. Therefore, it is obvious to expect shrinkage in the group level parameters towards their average across groups. Hierarchical methods can also be treated as smoothing techniques as they are forced to shrink group level parameters [43]. Generally, hierarchical models provide more precise estimates as they share information across groups. However, this results

in a risk of unwanted bias among the estimates. The dilemma of increasing the precision along with additional bias in the estimates is termed as bias-variance trade-off. Hierarchical models play an important role as a smoothing technique, especially in rare events or studies that are based on small geographical areas. The risk of having unstable estimates due to small sample sizes and/or rare events can be minimised with hierarchical models which pool the strength of data over different groups [120, 144].

5.2 Investigation of Hierarchical models for Stutter Ratio

Hierarchical models are more appropriate for modelling stutter ratios as they demonstrate hierarchical structure in the data. Different brands of PCR instruments usually have their own techniques to optimise the sensitivity, efficiency, and precision of PCR methodology. Therefore, technical settings of each brand could have varied effects on the performance of statistical modelling of peak heights. Stutter peaks in an EPG are classified as artefacts. Possible relationships between manufacturer specifications, and thus, the variability in artefacts cannot be neglected. Irrespective of whether these relationships are known or unknown, they can create some interdependencies in the behaviour of the observed stutter peaks across different loci for a given PCR instrument. Hierarchical models could be employed to take these interdependencies into account. Therefore, the seven locus-specific models are extended to appropriate hierarchical models with suitable hyper-prior distributions. The specifications of these hierarchical models are given in Table 5.1. Furthermore, the nature of hierarchy of the models is graphically illustrated in Figure 5.1 for the locus-specific hierarchical normal model (N_2).

Normal vague hyper-prior distributions were assumed for hyper-parameters of the normal prior distributions of slope and intercept parameters of each model. Gamma vague hyper-prior distributions were assumed for hyper-parameters of the inverse gamma prior distributions of variance parameters in each model. The degrees of freedom parameters of the respective non-standardised Student's t models were modelled with log-uniform prior distributions without any hyper-prior distribution. Profile-wide mixing proportion of each of the mixture model is modelled with a uniform prior.

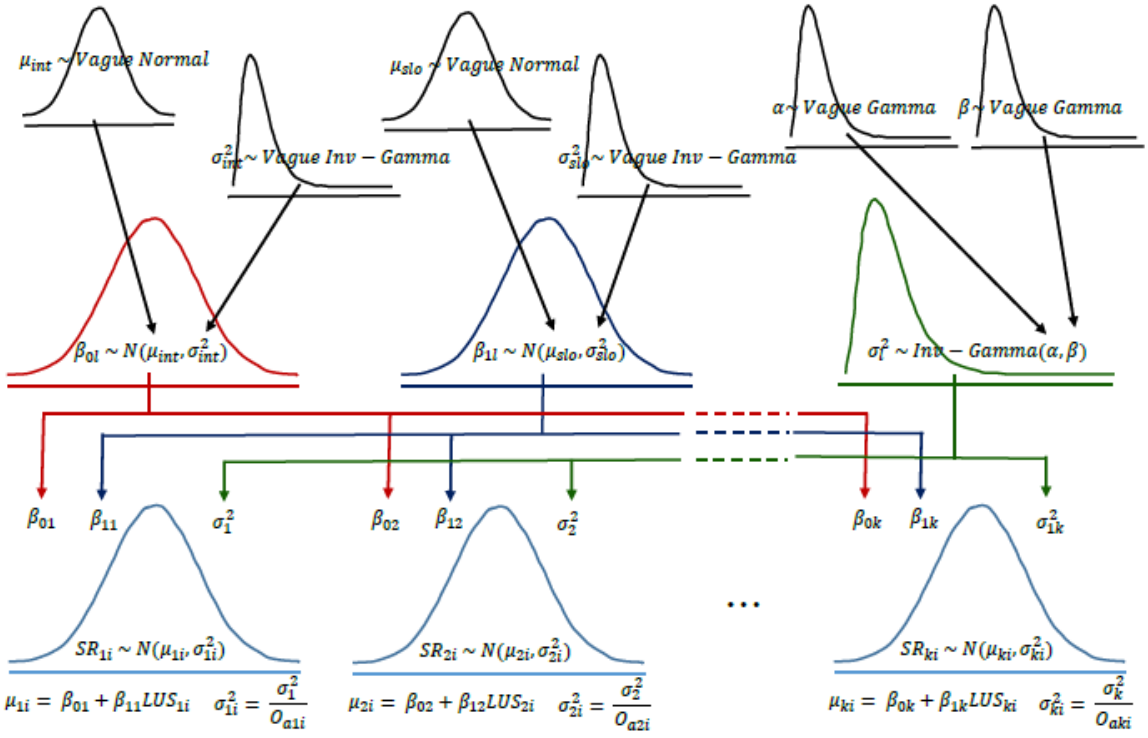

Figure 5.1: Hierarchical dependencies of locus-specific normal model (N_2) for *SR*. $k = 16$ for the NGM SElect^{TM} dataset and 15 for the Identifiler^{TM} dataset.

5.3 Evaluation of Hierarchical Models

As was done with the non-hierarchical models, the proposed hierarchical models were also fitted using MCMC techniques. Each model was run for 50000 iterations after 50000 burn-in steps. A thinning interval of 25 was applied in order to reduce possible inter-correlations among posterior draws.

The hierarchical models presented in this chapter are compared with the corresponding non-hierarchical models presented in Chapter 4, using credible intervals of the estimated mean model parameters (slope and intercept) and the standard deviation parameters of these models. In hierarchical models these three parameters were calculated for each locus (16 for the NGM SElect^{TM} and 15 for the Identifiler^{TM} datasets) assuming relevant hyper prior distributions. According to the results presented in sections 5.3.1 and section 5.3.2, the locus-specific estimates of the three parameters are not significantly different across hierarchical and non-hierarchical models. Therefore, based on the estimated parameters of hyper-prior distributions, the goodness-of-fit of the parameters was

Table 5.1: Descriptions of the proposed hierarchical models

Model	Distribution	Mean	Variance
LN_2	$ln(SR_{li}) \sim N(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
G_2	$SR_{li} \sim Gamma(\alpha_{li}, \theta_{li})$	$\mu_{li} = exp(\beta_{0li} + \beta_{1li} LUS_{li})$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
N_2	$SR_{li} \sim N(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
T_2	$SR_{li} \sim t(\mu_{li}, \sigma_{li}^2, \nu_l)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$
MLN_2	$ln(SR_{li}) \sim \pi N(\mu_{li}, \sigma_{0li}^2) +$ $(1-\pi) N(\mu_{li}, \sigma_{1li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{qli}}$ $\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$
MN_2	$SR_{li} \sim \pi N(\mu_{li}, \sigma_{0li}^2) +$ $(1-\pi) N(\mu_{li}, \sigma_{1li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{ali}}$ $\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$
MT_2	$SR_{li} \sim \pi t(\mu_{li}, \sigma_{0li}^2, \nu_{1l}) +$ $(1-\pi) t(\mu_{li}, \sigma_{1li}^2, \nu_{2l})$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma_{0l}^2}{O_{qli}}$ $\sigma_{1li}^2 = \frac{\sigma_{0l}^2 + \sigma_{1l}^2}{O_{ali}}$

Note: ν_l, ν_{1l}, and ν_{2l} are the locus-specific (l) degrees of freedom of the t distributions.

tested by performing the Kolmogorov-Smirnov test on relevant estimates of the parameters of both hierarchical and non-hierarchical models. In Kolmogorov-Smirnov test, the goodness-of-fit of the observed data in relation to the proposed (theoretical) distribution is assessed by evaluating the maximum absolute difference between theoretical and empirical cumulative distributions [165].

5.3.1 Mean Model Parameters of Hierarchical Models

Locus-specific variation represented by 95% credible intervals and medians of the mean model parameters (slope β_0 and intercept β_1) of the hierarchical and non-hierarchical models of all the types (log-normal, gamma, normal, non-standardised Student's t, and their mixtures) fitted to both datasets, do not show any significant difference (i.e. the intervals presented in Figures A.1 to A.14 in Appendix A are almost completely overlapping except only a few with extremely small differences).

The empirical cumulative distributions of the posterior means of slope and intercept parameters are presented in Figures 5.2 and 5.4 for the NGM SElect^{TM} dataset and Fig-

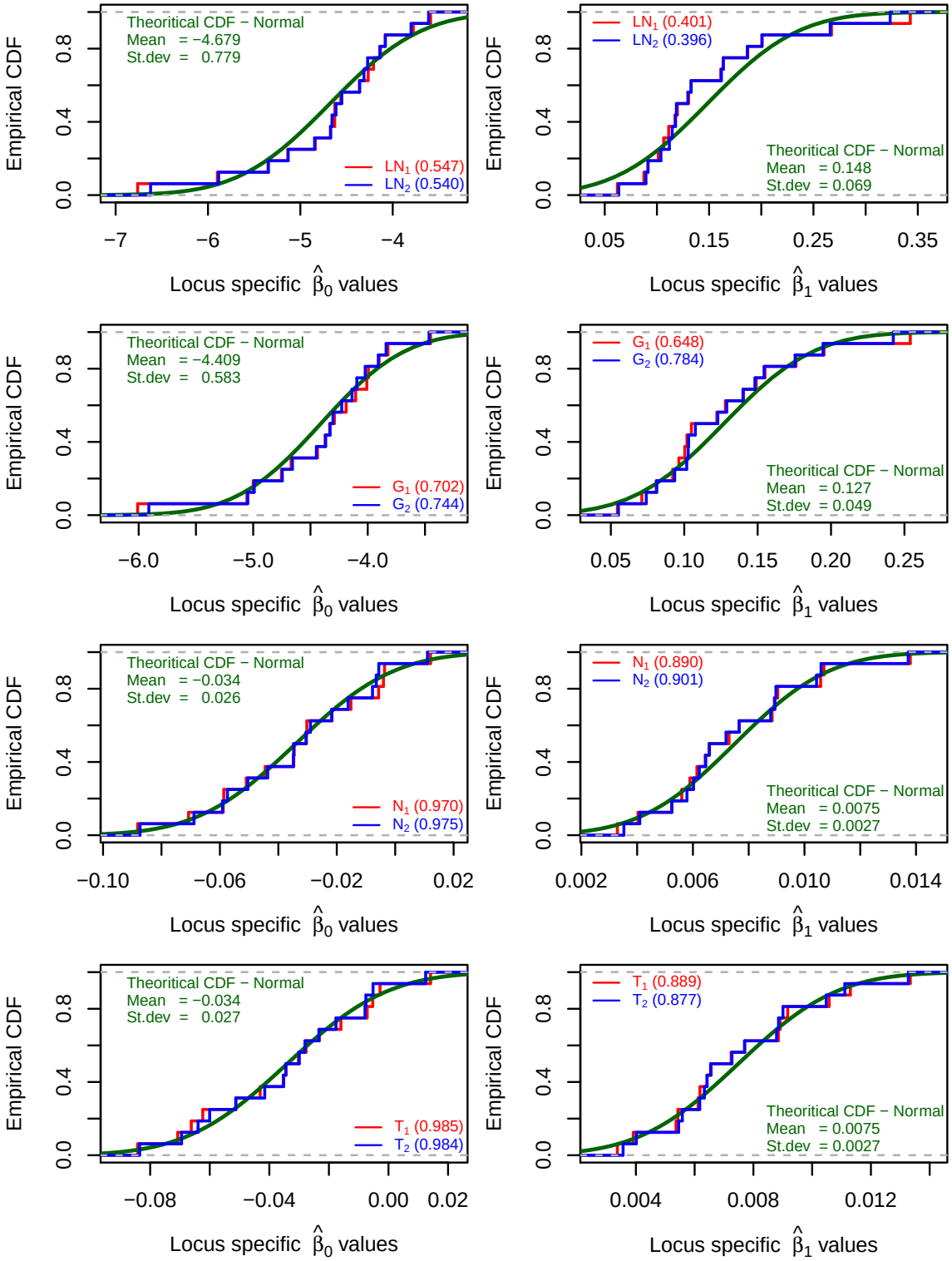

Figure 5.2: Inferred distributions of mean model parameters (slope β_0 and intercept β_1) of the locus-specific variance non-mixture models for the NGM SElect^{TM} dataset

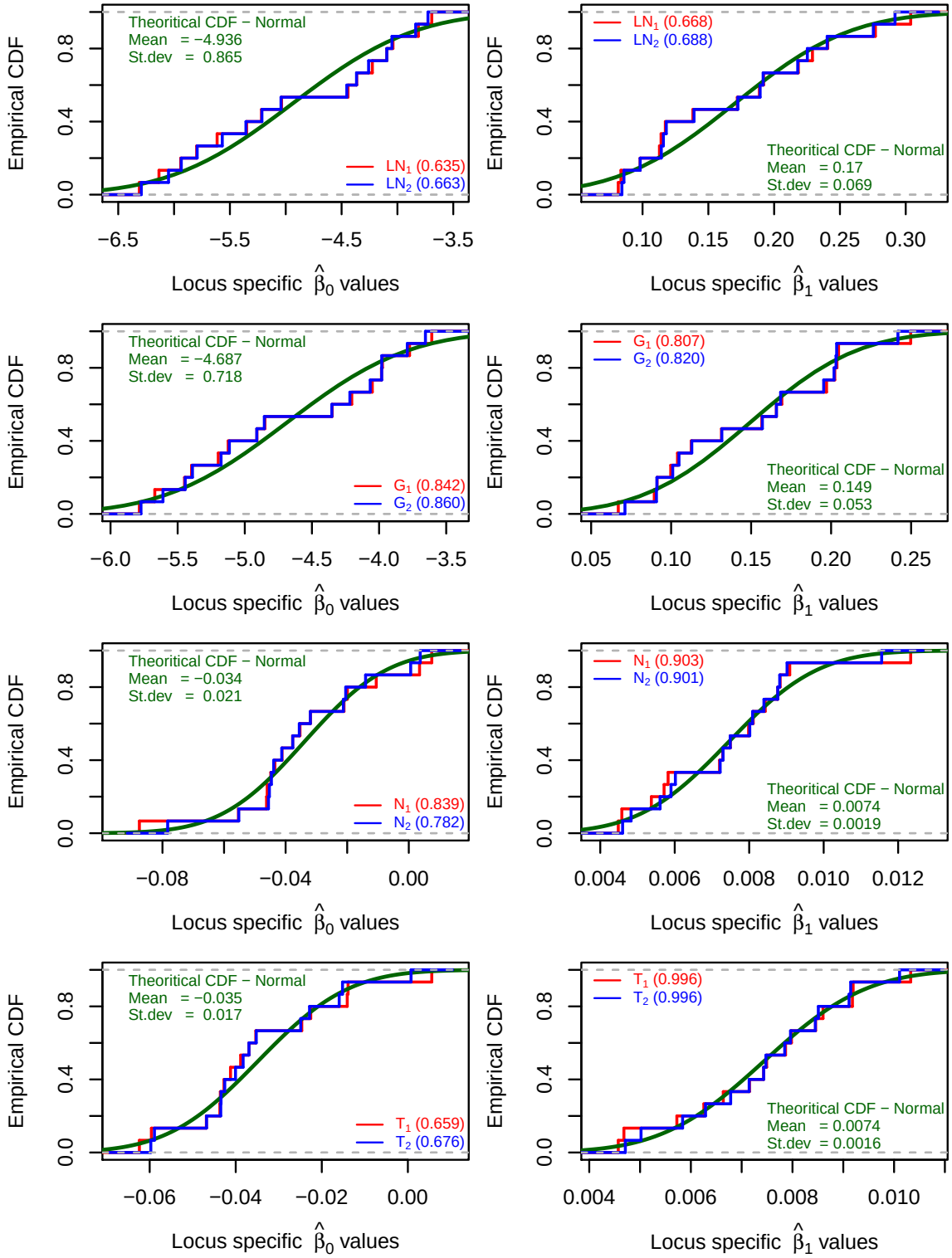

Figure 5.3: Inferred distributions of mean model parameters (slope β_0 and intercept β_1) of the locus-specific variance non-mixture models for the Identifiler^{TM} dataset

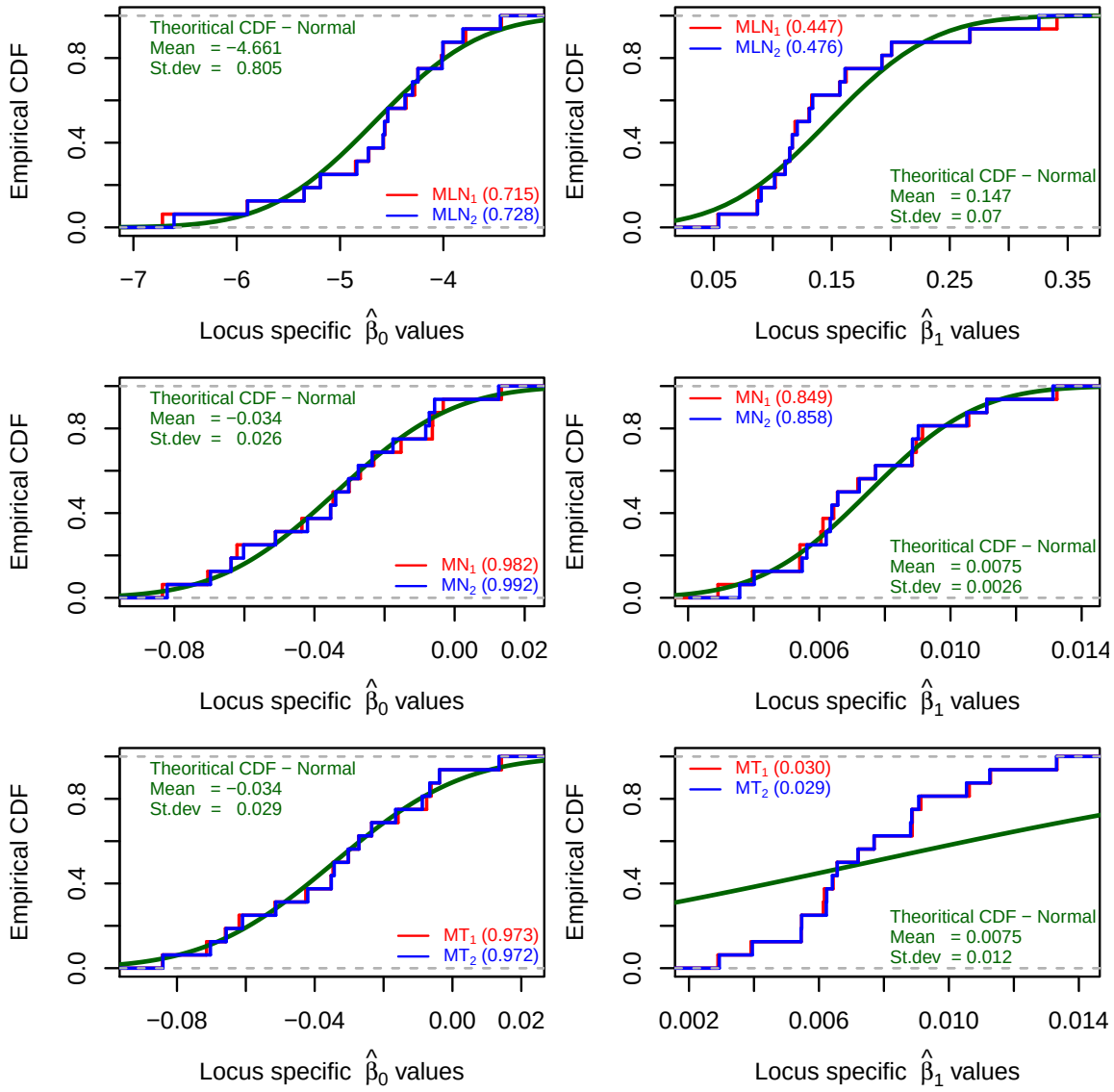

Figure 5.4: Inferred distributions of mean model parameters (slope β_0 and intercept β_1) of the mixture models for the NGM SElect^{TM} dataset

ures 5.3 and 5.5 for the Identifiler^{TM} dataset. These figures do not show any significant discrepancy between two cumulative distributions. All the non-mixture models that assumed locus-specific variances reveal very high goodness-of-fit with the inferred distributions of both slope and intercept parameters irrespective of the dataset. Both log-normal and normal mixture models consistently indicate a high goodness-of-fit for these two parameters in relation to their inferred distributions. The non-standardised Student's t model, in contrast, shows significant deviations in empirical distributions of locus-specific slope parameters from the normal inferred distribution. In fact the theoretical and empiri-

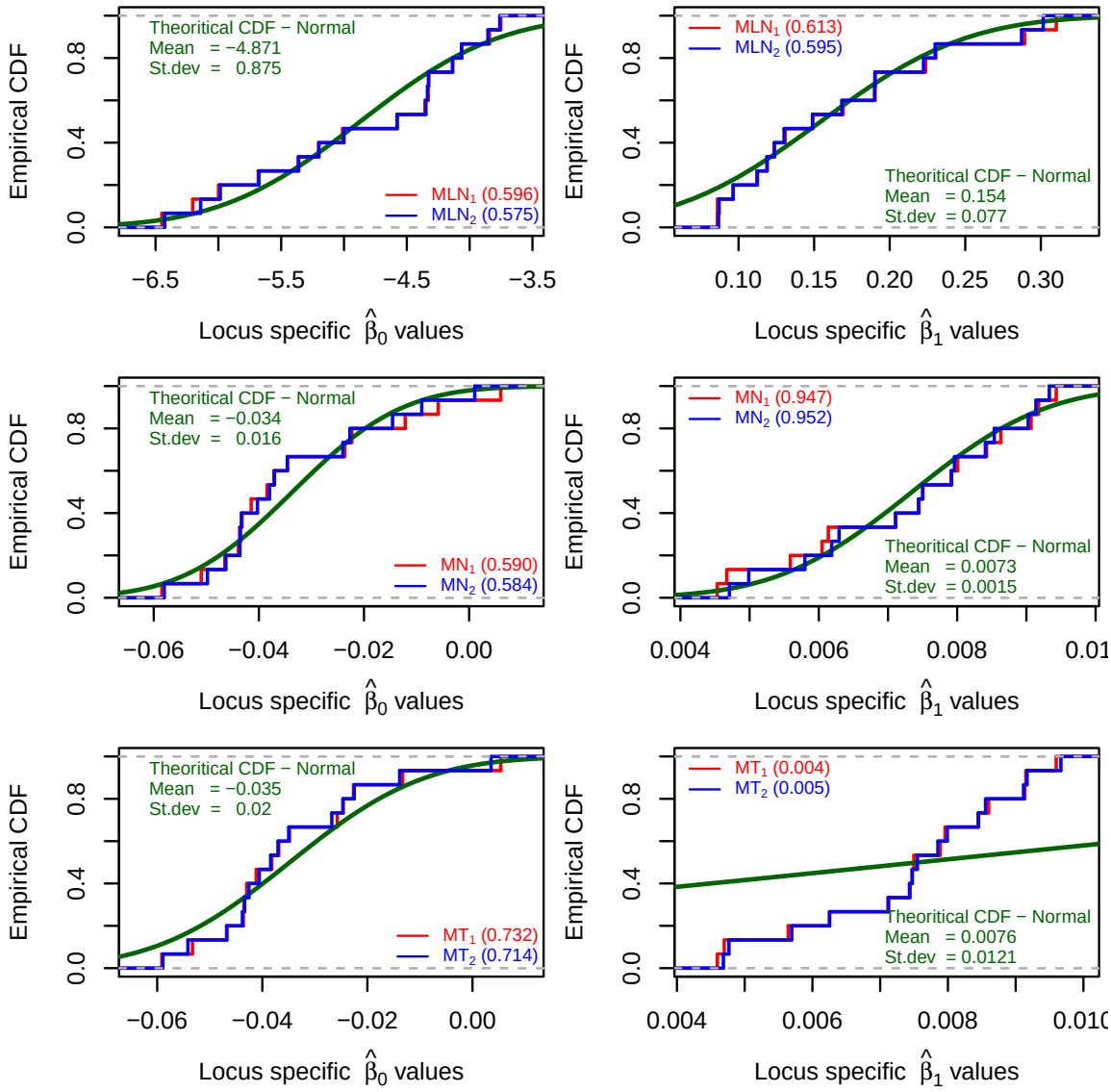

Figure 5.5: Inferred distributions of mean model parameters (slope β_0 and intercept β_1) of the mixture models for the Identifiler^{TM} dataset

cal means are approximately similar. However, for both datasets, the variance of theoretical inferred distribution was much larger than the corresponding variances of empirical distributions.

The similarities in the credible interval plots and empirical cumulative distributions drawn for each parameter of both hierarchical and non-hierarchical models fitted to each dataset graphically reveal the absence of pooling in locus-specific slope and intercept parameters. Similar goodness-of-fit results obtained under these models further evidence the absence of pooling in the mean model parameters.

The consistency of the inferred distributions corresponding to mean model parameters across the two datasets is vital. There are ten loci common to the Identifiler^{TM} and the NGM SElect^{TM} datasets, and except for these, the two datasets consisted of five and six more loci respectively. The posterior estimates of the parameters of normal inferred distributions of mean model parameters related to each hierarchical model are presented in Table 5.2 and Table 5.3 with the relevant p-vales of Kolmogorov-Smirnov goodness-of-fit test.

Table 5.2: Inferred distributions of intercept parameters

	NGM SElect^{TM}			Identifiler^{TM}		
Model	Mean	Stdev	p-value	Mean	Stdev	p-value
LN_2	-4.67	0.78	0.54	-4.94	0.87	0.66
MLN_2	-4.67	0.81	0.73	-4.87	0.88	0.58
G_2	-4.41	0.58	0.58	-4.67	0.72	0.86
N_2	-0.034	0.026	0.98	-0.034	0.021	0.78
MN_2	-0.034	0.026	0.99	-0.034	0.016	0.58
T_2	-0.034	0.027	0.94	-0.035	0.017	0.68
MT_2	-0.034	0.029	0.97	-0.035	0.020	0.71

Locus-specific intercept parameters of the mixture and non-mixture models based on normal and non-standardised Student's t distributions (N_2, T_2, MN_2, and, MT_2) fitted to both datasets have been derived from normal inferred distributions with almost similar location parameters. The dispersion of these intercept parameters is almost identical within the NGM SElect^{TM} dataset compared to the Identifiler^{TM} dataset. However, the intercept parameters of these four models fitted to the NGM SElect^{TM} dataset indicate more variability than that of the Identifiler^{TM} dataset. In addition, the inferred distributions of intercept parameters of these four models fitted for the NGM SElect^{TM} dataset reveal very high goodness-of-fit. The intercept parameters of the log-normal models (LN_2 and MLN_2) fitted to the NGM SElect^{TM} dataset have been modelled with almost identical normal hyper-prior distributions. The two inferred distributions related to the Identifiler^{TM} dataset reveal only a small change between the two location parameters.

The locus-specific slope parameters of normal models (N_2 and MN_2) and T_2 model fitted to both datasets followed normal distributions with similar location parameters. The

Table 5.3: Inferred distributions of slope parameters

	NGM SElect™			Identifiler™		
Model	Mean	Stdev	p-value	Mean	Stdev	p-value
LN_2	0.15	0.069	0.40	0.17	0.069	0.69
MLN_2	0.15	0.070	0.48	0.15	0.077	0.60
G_2	0.13	0.049	0.78	0.15	0.053	0.82
N_2	0.0075	0.0027	0.90	0.0074	0.0019	0.90
MN_2	0.0075	0.0026	0.86	0.0073	0.0015	0.95
T_2	0.0075	0.0027	0.88	0.0074	0.0016	1.00
MT_2	0.0075	0.012	0.03	0.0076	0.012	0.01

distributions of these three models fitted to the NGM SElect™ dataset are almost identical. However, the variances of locus-specific slope parameters of the models fitted to the NGM SElect™ dataset are noticeably larger than that of the Identifiler™ dataset. The larger p-values of the three normal inferred distributions reveal very high goodness-of-fit for both datasets. Even though the location and scale parameters of MT_2 are consistent across both datasets, each inferred distribution reveals very poor goodness-of-fit due to the variances which are larger than the variances of estimated locus-specific slope parameters. Both log-normal models also reveal approximately identical normal inferred distributions across both datasets.

5.3.2 Variance Parameters of Hierarchical Models

The intervals for the standard deviation parameters do not indicate any visible difference between hierarchical and non-hierarchical non-mixture models of log-normal, gamma, normal, and non-standardized Student's t models (Figures A.15 to A.18 in Appendix A). However, the credible interval plots of standard deviation parameters of mixture models, σ_0 and σ_1 (where σ_0^2 and $\sigma_0^2+\sigma_1^2$ are the variances of the two mixture components) reveal some discrepancies between hierarchical and non-hierarchical methods, for some models (Figures A.19 to A.24 in Appendix A). The pair of credible intervals that has been drawn for σ_0 parameter of non-hierarchical and hierarchical normal models for the locus D2S441 in the NGM SElect™ dataset do not overlap, whereas all the other pairs of credible intervals drawn for both σ_0 and σ_1 under all the mixture models overlap. The intervals of normal mixture models fitted to the NGM SElect™ dataset (Figure A.21) indicate

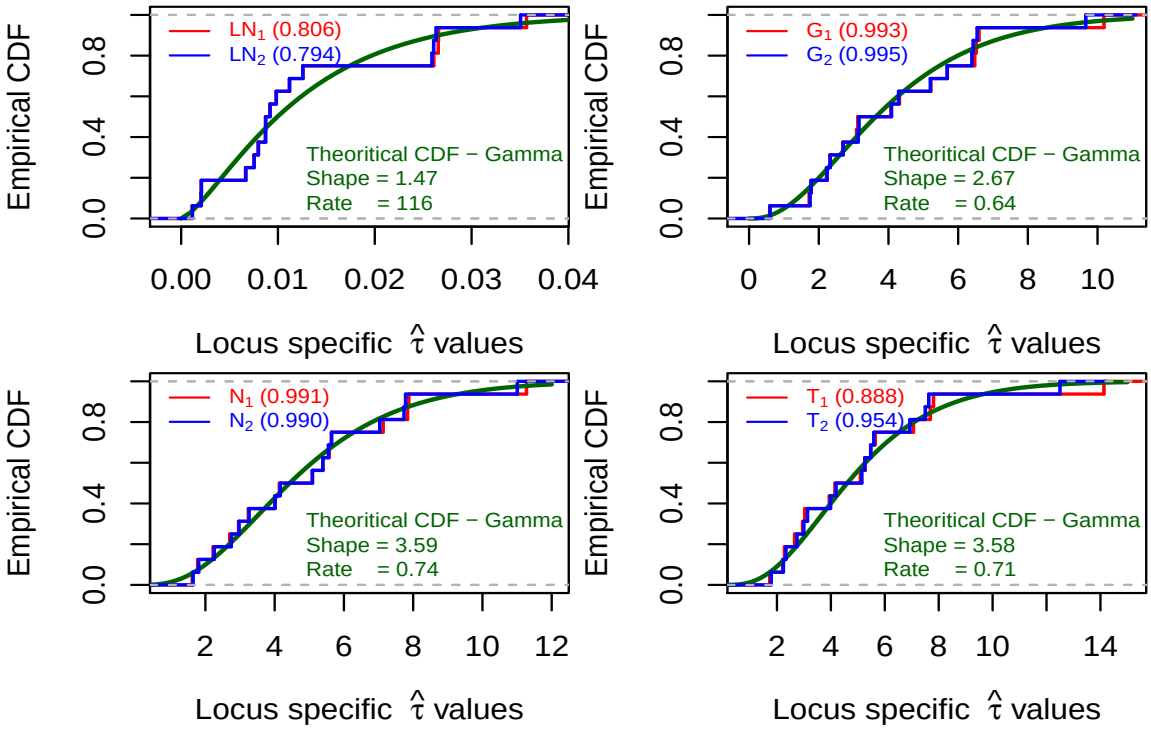

Figure 5.6: Inferred distributions of precision parameters (inverse variance τ) of the locus-specific variance non-mixture models for the NGM SElect^{TM} dataset

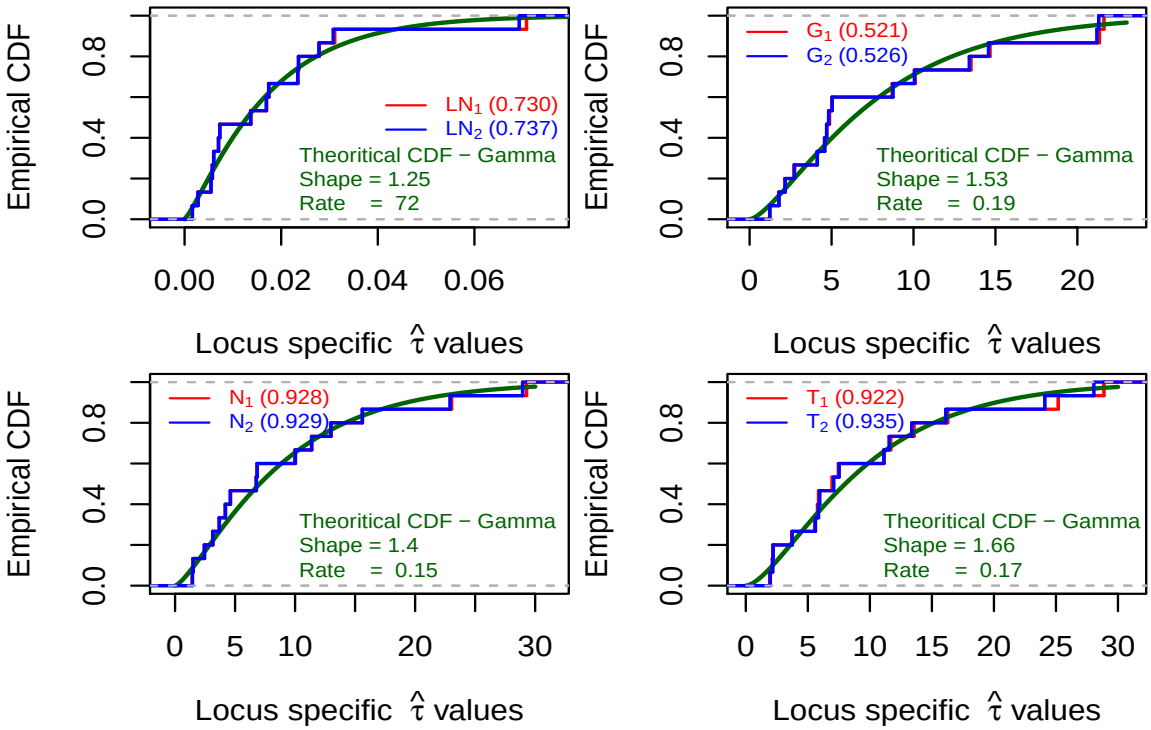

Figure 5.7: Inferred distributions of precision parameters (inverse variance τ) of the locus-specific variance non-mixture models for the Identifiler^{TM} dataset

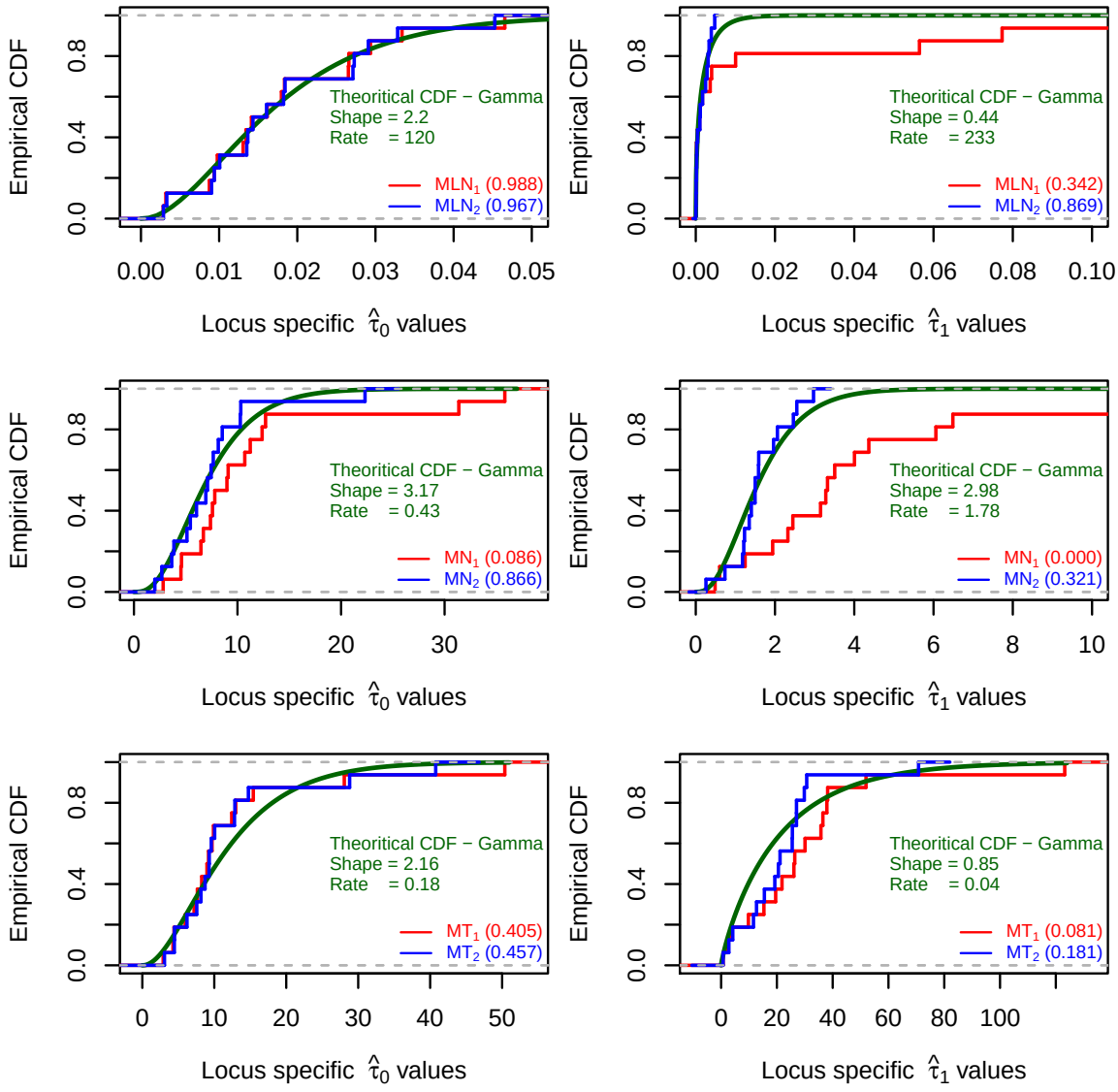

Figure 5.8: Inferred distributions of precision parameters (inverse variance τ) of the mixture models for the NGM SElect^{TM} dataset

larger gaps whereas the gaps for the Identifiler^{TM} dataset (Figure A.22) are negligible. The log-normal mixture models show (Figures A.19 and A.20) small but considerable differences only in the credible intervals of σ_1, for both datasets across the two methods. The differences between the intervals of standard deviation parameters of the hierarchical and non-hierarchical non-standardized Student's t mixture models are very small except for a few which show only slight deviations at some loci (Figures A.23 and A.24).

The empirical cumulative distributions of precision parameters are presented in Figures 5.6 and 5.8 for the NGM SElect^{TM} dataset and Figures 5.7 and 5.9 for the Identifiler^{TM}

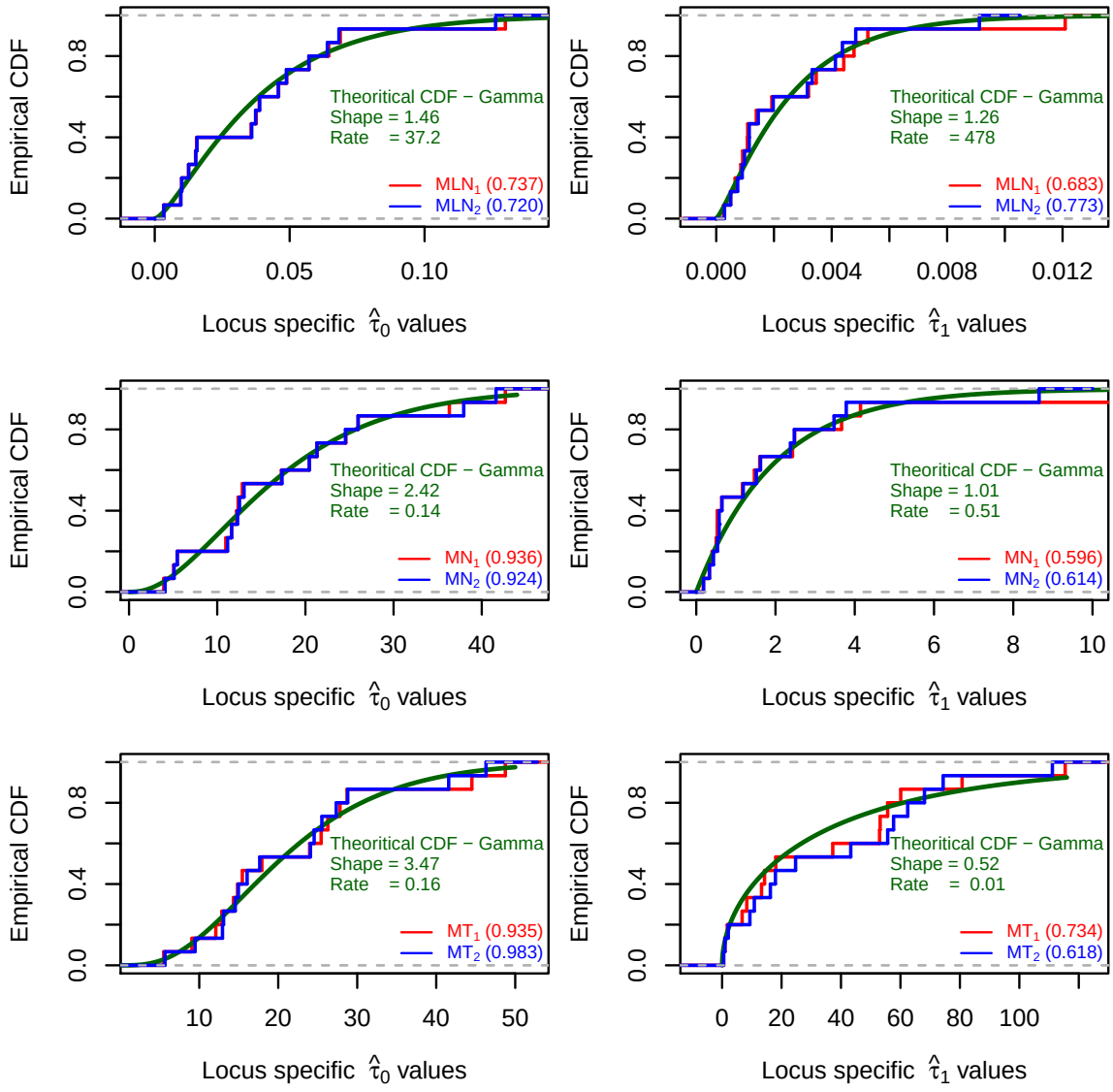

Figure 5.9: Inferred distributions of precision parameters (inverse variance τ) of the mixture models for the Identifiler^{TM} dataset

dataset. Locus-specific variance of the non-mixture models do not reveal any considerable difference between non-hierarchical and hierarchical methods in relation to the precision parameters of the respective models fitted to both datasets (Figures 5.6 and 5.7). Under each distribution, almost identical empirical cumulative distributions are observed for hierarchical and non-hierarchical models. Even though the larger p-values indicate high goodness-of-fit for each inferred distribution of precision parameter, they are not consistent across the two datasets under any model. The precision parameters of the three mixture models fitted to the Identifiler^{TM} dataset do not reveal considerable differences be-

tween two empirical cumulative distributions. In addition, Kolmogorov-Smirnov p-values are also equally large for each pair of empirical cumulative distributions. The precision parameters associated with the component with lower variance (τ_0) of the log-normal and non-standardised Student's t mixture models fitted to the NGM SElect^{TM} dataset also reveal approximately similar empirical cumulative distributions with larger goodness-of-fit p-values. The empirical cumulative distributions of the other precision parameter of these two models and both precision parameters of the normal mixture model reveal substantial differences. More interestingly, all the precision parameters of the mixture models reveal sufficient or extremely high goodness-of-fits for both datasets.

5.3.3 Changes in Log-likelihoods and Log Predictive Densities

Table 5.4: Log-likelihoods of the models

	NGM SElect^{TM}		Identifiler^{TM}	
Model	Hierarchical	Non-hierarchical	Hierarchical	Non-hierarchical
LN	14463	14464	22924	22924
G	14777	14778	23365	23365
N	15233	15234	23513	23515
T	15351	15354	24530	24531
MLN	15276	15278	24267	24268
MN	15472	15475	24617	24619
MT	15923	15938	25610	25639

It is intuitive to expect at least a small reduction in the log-likelihoods due to the effect of bias-variance trade-off in hierarchical models. The magnitude of the reduction in the log-likelihoods directly reflects the degree of bias or the level of pooling in the model parameters of the hierarchical model. However, many of the parameters associated with mean model and the variance model were nearly identical. Hence, considerable reductions in the log-likelihoods of hierarchical models in comparison with the respective non-hierarchical models cannot be expected. Almost identical log-likelihoods between each pair of hierarchical and non-hierarchical models presented in Table 5.4 clearly evidence the absence of pooling in the parameters of the hierarchical models.

Even though any substantial change cannot be expected in the log-likelihoods for any model, there are some small but considerable changes (increasing and decreasing) in the

estimated log point-wise predictive densities (clppd) of all the mixture models fitted to the NGM SElect^{TM} dataset (Figure 5.5) and non-standardised Student's t mixture model fitted to the Identifiler^{TM} dataset (Figure 5.6). In addition, the measures that are based on point-wise predictive densities: LOO-CV approximations (IS-LOO, TIS-LOO and PSIS-LOO), both versions of WAICs (WAIC and WAICalt), and their penalty terms (p_{WAIC} and $p_{\text{WAIC alt}}$) also revealed considerable changes in these models.

Table 5.5: Log predictive densities of the models fitted to the NGM SElect^{TM} dataset

Model	IS	TIS	PSIS	clppd	WAIC	WAICalt	p_{WAIC}	$p_{\text{WAIC alt}}$
LN_1	10207	11576	10855	15295	13680	12600	1616	2695
LN_2	10132	11563	10855	15292	13677	12596	1615	2696
G_1	11153	12259	11677	15533	14052	13132	1482	2401
G_2	11125	12251	11659	15527	14047	13122	1480	2405
N_1	11873	12784	12208	15926	14492	13596	1434	2330
N_2	11657	12759	12210	15927	14495	13596	1433	2331
T_1	8469	11105	10065	16488	14458	12889	2029	3598
T_2	8384	11065	10048	16487	14453	12878	2034	3609
MLN_1	9834	11871	11617	16475	14411	13636	2065	2839
MLN_2	10288	12125	11659	16354	13536	13536	1800	2817
MN_1	12053	13133	12609	16501	14903	13961	1598	2540
MN_2	12191	13254	12683	16838	15021	14061	1817	2777
MT_1	9751	11979	11147	16821	14734	13395	2088	3427
MT_2	9843	11930	11089	16544	14647	13320	1897	3224

Note: Penalty constants calculated in WAIC are given in the last two columns. WAIC and WAICalt are reported excluding the factor -2.

5.4 Discussion

The low degree of pooling in the group level parameters observed in this study can be explained as follows. Small samples require more information from the rest of the population than large samples, and hence the shrinkage is typically greater for the groups with relatively smaller number of observations [43]. According to Kruschke [108], a multilevel model produces very strong pooling or shrinkage for a dataset with many small groups. However, data-level errors are accurately approximated with their estimates for reasonably large groups [76]. The influence of the number of observations in group-levels of a multilevel model was examined in a simulation study, and it revealed that only

Table 5.6: Log predictive densities of the models fitted to the Identifiler^{TM} dataset

Mode l	IS	TIS	PSIS	clppd	WAIC	WAICalt	p_{WAIC}	$p_{\text{WAIC alt}}$
LN_1	16601	18627	17600	24000	21652	20101	2348	3900
LN_2	16650	18629	17606	24004	21659	20111	2345	3893
G_1	17807	19536	18687	24386	22211	20832	2175	3554
G_2	17828	19540	18700	24385	22211	20836	2173	3549
N_1	18283	19846	19022	24563	22425	21085	2138	3478
N_2	18483	19869	19011	24561	22422	21083	2138	3477
T_1	10187	15718	13675	26607	22880	19679	3727	6928
T_2	10019	15637	13632	26617	22885	19669	3732	6948
MLN_1	17974	20283	19363	26400	23668	21920	2733	4480
MLN_2	17983	20296	19378	26401	23666	21921	2735	4480
MN_1	19761	21378	20753	26725	24127	22623	2599	4102
MN_2	19698	21382	20749	26751	24137	22632	2614	4119
MT_1	14935	18876	17518	26984	23638	21369	3346	5615
MT_2	15244	18895	17610	26872	23584	21309	3288	5563

Note: Penalty constants calculated in WAIC are given in the last two columns. WAIC and WAICalt are reported excluding the factor -2.

the groups that are smaller than 50 observations lead to producing bias estimates [118]. Bias-variance trade-off is a consequence of pooling in the group-level parameters. The number of peak height information included in each locus (group) is very large (170 - 406 for the NGM SElect^{TM} dataset and 366 - 534 for the Identifiler^{TM} dataset) in this study. Therefore, the hierarchical models may result in virtually no pooling.

5.5 Summary

For both datasets, there is no visible pooling in the slope and intercept parameters of mean model under the hierarchical models. This can be due to the large samples of data utilised in the study. The locus-specific slopes and intercepts of non-hierarchical models (except the slope parameters of non-standardised Student's t mixture) are coming from normal distributions. The hierarchical models provided a way of estimating the parameters of these normal distributions. These inferred normal distributions are approximately consistent for both datasets for normal and non-standardised Student's t models, though slightly different for the rest of the models. Non-mixture models fitted to both datasets do not reveal any considerable pooling in standard deviation parameters. However, for mix-

ture models, there are some occasional minor changes in these parameters. The precision parameters of all the non-mixture models fitted to both datasets and the mixture models fitted to the Identifiler^{TM} dataset reveal high goodness-of-fit with their inferred distributions. However, these inferred distributions are not consistent across the two datasets.

For both datasets, the absence of bias-variance trade-off or lack of pooling in the group level parameters of hierarchical models, is clearly revealed by the almost identical log-likelihoods of each pair of hierarchical and non-hierarchical models. However, some minor changes are observed among the predictive measures estimated based on point-wise predictive densities. Although the hierarchical modelling approach has not been as effective as expected, obtaining inferred distributions of these parameters with high goodness-of-fit is an important outcome of this investigation.

Chapter 6

Bayesian Multiple Linear Regression with a Conjugate Prior Distribution

6.1 Introduction

Chapters 2, 4, and 5 already discussed the performance of two-component mixture models. Robust regression idea was the key motivation factor behind the two-component mixture models. A bulk of the observations that are modelled in real world datasets are reasonably well behaved, but in some cases there may be some unstable observations which can be modelled using another distribution. In 1960, Tukey [166] has discussed two-component heteroscedastic mixtures of normal densities under the family of contaminated normal distributions. In 2011, on his PhD thesis, Maheswaran [145] has also implemented and expanded the idea of robust statistical models using finite mixtures. The present study revealed a good performance in two-component normal mixture models in modelling stutter ratio. Mixture models in general, there is no reason to constraint only for two components. Practically, it can be used any number of components. However, in reality, no one knows how many components it should have. In traditional modelling, it assumes a finite number of components and estimates the parameters of these models. Treating the number of parameters of the model as another parameter is the key idea of infinite mixture modelling. An infinite mixture model is a data driven method of estimating the number of components in a model. Chapter 6 and 7 extend the idea of two-component

normal mixture models into infinite mixtures of linear regression models for PCR stutter.

This chapter explains the analytical process of the Bayesian version of multiple linear regression involving a fully conjugate prior distribution. However, it does not provide any novel results in relation to this sort of multiple linear regression models. It was unable to find a complete document related to this topic, and most of the existing literature provided important results along with partial theoretical derivations. This chapter attempts to fulfil the gaps in the theoretical derivations reviewing appropriate published literature, peer reviewed references and unpublished online resources.

Initially, the derivation of the likelihood function, the selection of conjugate prior distribution, and the combination of them to derive the posterior distribution of model parameters is discussed. Then, the long analytical process of deriving the prior predictive distribution of data from the posterior distribution of model parameters is presented and the posterior predictive distribution of new data is derived. The derived analytical relationships between the prior information, the observed data, and the parameters of posterior predictive distribution of the data are applied in the next chapter to develop an infinite mixture of linear regression models for predicting stutter ratio with improved accuracy.

The methods for studying how a dependent variable changes as a mathematical function of one or more independent variables (covariates/ predictors) are referred to as "regression", and have a history of more than 100 years. The concept of linear regression was first used by a British biologist Francis Galton in 1908 [178]. Since then, there has been a remarkable improvement in the regression methodology. It has already been applied in various scientific fields, and recognised as one of the most frequently-used analytical tools in statistical data analysis for both interpolation and extrapolation purposes.

6.1.1 Conditional Bayesian Regression Modelling

Let θ be the vector of regression parameters and ψ be the parameter vector that determines the distribution of X, $p(X|\psi)$. Gelman et al. [72] have shown that a Bayesian regression model ignores the information provided by X about the joint prior distribution of θ and ψ, $p(\theta, \psi)$. This can be explained as below.

In standard regression, it is assumed that the distribution of X, $p(X|\psi)$ does not pro-

vide any information about the conditional distribution of $\boldsymbol{y}$ given X, $p(\boldsymbol{y}|X)$. A full Bayesian model involves a joint likelihood of X and $\boldsymbol{y}$, $p(X,\boldsymbol{y}|\psi,\theta)$ with a prior distribution $p(\psi,\theta)$ and a conditional distribution $p(X|\psi)$. If ψ and θ are independent, then,

$$p(\psi,\theta) = p(\psi)p(\theta).$$

Consequently, the posterior distribution of (ψ,θ) can be factored out as

$$p(\psi,\theta|X,\boldsymbol{y}) = p(\psi|X)p(\theta|X,\boldsymbol{y}).$$

Without any loss of information, the second factor can be further expanded as a standard regression model:

$$p(\theta|X,\boldsymbol{y}) \propto p(\theta)p(\boldsymbol{y}|X,\theta).$$

Therefore, the distribution of X is no longer considered in fitting a regression model on $\boldsymbol{y}$.

6.1.2 Bayesian Multiple Linear Regression Modelling

A majority of Bayesian literature focuses on understanding theoretical and practical aspects of multiple linear regression models. Banerjee [12] provides a useful set of guidelines for developing Bayesian linear regression models including relevant theoretical derivations.

Let us consider a linear regression model with $p-1$ predictors. The random variable $\boldsymbol{Y}$ is an n-dimensional vector of data $\boldsymbol{y}$. X is an $n \times p$ design matrix which includes the values of covariates, and β is a $p \times 1$ vector of parameters specifying the effect of each predictor. ε is an $n \times 1$ vector of variables that represent the random errors corresponding to the dependent variable. Then the linear regression model can be presented in the form:

$$\mathbf{Y} = X\beta + \varepsilon.$$

When one assumes an independent and identically distributed (i.i.d.) normal distribution

with zero mean and unknown finite variance σ^2 for each error term $\big($i.e. $\varepsilon_i \sim N\big(0,\sigma^2\big)\big)$, the distribution of random error terms (ε) is specified as $\varepsilon \sim MVN\big(0,\sigma^2 I_n\big)$, where I_n is an $n \times n$ identity matrix.

When the multivariate normality is assumed for the random errors, the model is known as normal linear regression [91] and is written in the form:

$$\mathbf{Y}\big|\beta,\sigma^2 \sim MN_n\big(X\beta,\sigma^2 I_n\big). \tag{6.1}$$

There have been different approaches applied to linear regression modelling. During the last century, non-Bayesian methods were dominant in both statistical theory and practice. However, as a result of new computational techniques, a re-emergence of Bayesian approaches has been observed during the last few decades [72]. The Bayesian mechanism of model fitting consists of three essential steps [52]:

1. given the unknown parameters, derive the likelihood of the data;
2. propose suitable prior distributions to all unknown parameters;
3. given the data, determine the posterior distribution of the parameters using Bayes' theorem.

6.2 The Likelihood Function

The joint probability density function of the observed data, which is obtained as a function of the unknown parameters is called the likelihood. The likelihood function for the parameters including the regression coefficients and variance $\big(\beta,\sigma^2\big)$ is

$$\begin{aligned} p\big(\mathbf{y}\big|\beta,\sigma^2\big) &= \frac{1}{\big(2\pi\big)^{\frac{n}{2}}\big|\sigma^2 I_n\big|^{\frac{1}{2}}} exp\Big[\frac{-1}{2}\big(\mathbf{y}-X\beta\big)^T\big(\sigma^2 I_n\big)^{-1}\big(\mathbf{y}-X\beta\big)\Big] \\ &= \frac{1}{\big(2\pi\sigma^2\big)^{\frac{n}{2}}} exp\Big[\frac{1}{2\sigma^2}\big(\mathbf{y}-X\beta\big)^T\big(\mathbf{y}-X\beta\big)\Big]. \end{aligned} \tag{6.2}$$

6.3 Selection of Prior Distributions

In Bayesian methodology, the parameters of a model are treated as random quantities themselves. Therefore, in statistical perspective, it is essential to employ suitable probability distributions to describe them. The information that is related to the level of uncertainty in the model parameters is typically represented by the prior distributions. The probability distribution of a new dataset is combined with the prior distributions of the model parameters to derive the posterior distributions, which are customarily used to make statistical inferences involving model parameters [69]. In Bayesian hierarchical modeling, the parameters of the prior distributions are also assumed to be random, and modelled with hyper (multiple level) priors. Prior/hyper-prior distributions of the model parameters incorporate either the previous results of similar studies or the experience of the client/statistician or both. The level of uncertainty in the model parameters, reflected by the prior distributions is often very high. However, it can be overcome with the use of strong informative priors. These priors are usually defined based on the empirical results originated from similar studies with real world data.

Subjectivity is a common issue that can be seen in all the statistical models, because, each model is a mathematical realisation of the real world [72]. As a consequence of relying on prior distributions, Bayesian methods are sometimes extensively criticised regarding subjectivity. The reliance on the priors is the most controversial aspect of Bayesian statistics [127].

6.3.1 Conjugate Prior Distributions

Various types of priors are found in Bayesian literature. However, the advantages in the use of conjugate priors have been emphasised in many publications [14, 18, 37, 60, 72, 77, 100, 127, 130]. Let us assume that $p(\theta)$ is the prior density of parameter θ, and $p(\theta|y)$ is the posterior density of θ after observing data y. If the both prior and posterior densities, $p(\theta)$ and $p(\theta|y)$ belong to a class of parametric densities F, then the prior density $p(\theta)$ is said to be a conjugate prior with respect to likelihood $p(y|\theta)$ [100]. Therefore, the conjugacy of a prior distribution is a well-defined property of the prior density with re-

spect to the likelihood function; hence, conjugate priors are sometimes said to be "closed under sampling" [14, 100]. Intractability is a typical problem in many target posterior distributions, and can be avoided by the use of conjugate prior distributions [14, 127, 130]. When a conjugate prior distribution is used for the parameters of a given Bayesian model, the consequential posterior distribution certainly follows a known parametric form [72]. Therefore, the use of a conjugate prior distribution is always mathematically convenient. In addition to this advantage, the information containing in the conjugate prior distributions can be interpreted as additional data. Very often, the posterior parameters are expressed as weighted means of the conjugate prior parameters and maximum likelihood estimators [130].

6.3.2 The Joint Conjugate Prior

In the Bayesian approach, the variance of random error term σ^2 and the vector of unknown regression coefficients β of a linear regression model are treated as random quantities, and it is required to place appropriate prior distributions on them. The selection of conjugate prior distributions and the steps for deriving posterior distributions for linear regression models have been discussed by a number of authors [5, 60, 100, 127, 130]. Two independent conjugate prior distributions can be suggested for these two parameters (β, σ^2). A p-dimensional multivariate normal and inverse-gamma densities are the intuitive options for β and σ^2 respectively. A scaled inverse chi-square distribution with degrees of freedom ν_0 and scale parameter σ_0^2 is a convenient parametrisation of the inverse gamma distribution [72, 126] and is referred to as the natural conjugate prior for σ^2. According to this parametrisation, prior information is equivalent to an average squared deviation σ_0^2 of a sample of ν_0 observations. An inverse chi-square distribution with parameters ν_0 and σ_0^2 is equivalent to an inverse gamma distribution with parameters $\frac{\nu_0}{2}$ and $\frac{\nu_0\sigma_0^2}{2}$. Let us assume that μ_β and $\ddot{V}_\beta$ are the mean vector and variance-covariance matrix of the p-dimensional multivariate normal prior distribution of β respectively. The shape and scale parameters of the inverse gamma prior distribution for σ^2 are designated by a and b respectively. Then the joint prior distribution of (β, σ^2) is defined as the product of these

two independent conjugate prior distributions.

$$p(\beta,\sigma^2) = p(\beta)p(\sigma^2) = MN_p(\mu_\beta,\ddot{V}_\beta)IG(a,b)$$

Surprisingly, this joint prior is not conjugate with respect to the likelihood function defined in Equation 6.2. Conversely, the conditional densities $p(\beta|\sigma^2)$ and $p(\sigma^2|\beta)$ are individually conjugate with respect to the given likelihood. The priors with this property are sometimes labelled as "semi-conjugate" or "conditionally conjugate". Higher dimensionality of the parametric space and greater model complexity are the key factors affecting the level of difficulty of obtaining conjugate prior distributions [65]. Therefore, conditional conjugacy is a common problem in many complex models. When it is required to specify the joint probability distribution of a model with highly dimensional parametric space, it is intuitive to expect some harder circumstances. However, the semi-conjugacy problem of the joint prior distribution of (β,σ^2) is fixed with the use of following conditional distribution for β given σ^2:

$$p(\beta|\sigma^2) = MN_p(\mu_\beta,\sigma^2 V_\beta).$$

Here, the variance-covariance matrix of β has been defined as the product of unscaled variance-covariance matrix V_β and σ^2. Finally, the joint prior distribution for (β,σ^2) is defined in the following form, and it is a fully conjugate with respect to the likelihood function of the linear regression model.

$$p(\beta,\sigma^2) = p(\beta|\sigma^2)p(\sigma^2) = MN_p(\mu_\beta,\sigma^2 V_\beta)IG(a,b)$$

The density function of this joint conjugate prior,

$$\begin{aligned} p(\beta,\sigma^2) &= MN_p(\mu_\beta,\sigma^2 V_\beta) IG(a,b) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}\left|\sigma^2 V_\beta\right|^{\frac{1}{2}}} exp\Big[\frac{-1}{2}(\beta-\mu_\beta)^T(\sigma^2 V_\beta)^{-1}(\beta-\mu_\beta)\Big] \\ &\quad \frac{b^a}{\Gamma(a)}\Big(\frac{1}{\sigma^2}\Big)^{a+1} exp\Big(\frac{-b}{\sigma^2}\Big) \\ &= \frac{b^a}{(2\pi)^{\frac{p}{2}}\Gamma(a)\left|V_\beta\right|^{\frac{1}{2}}}\Big(\frac{1}{\sigma^2}\Big)^{a+\frac{p}{2}+1} exp\Big\{\frac{-1}{\sigma^2}\Big[b+\frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\Big]\Big\} \end{aligned} \tag{6.3}$$

is analogous to the density function of multivariate normal inverse gamma $(MNIG)$ distribution with the parameters μ_β, V_β, a, and b. Therefore, the joint conjugate prior distribution of (β,σ^2) is

$$p(\beta,\sigma^2) \sim MNIG(\mu_\beta, V_\beta, a, b). \tag{6.4}$$

6.4 The Posterior Distribution of Model Parameters

The Bayes' theorem is employed to combine the joint conjugate prior and the likelihood function to determine the posterior distribution of the parameters (β,σ^2). It is given by,

$$p(\beta,\sigma^2|\mathbf{y}) = \frac{p(\mathbf{y}|\beta,\sigma^2)p(\beta,\sigma^2)}{p(\mathbf{y})},$$

where; $p(\mathbf{y}) = \int p(\mathbf{y}|\beta,\sigma^2)p(\beta,\sigma^2)d\beta d\sigma^2$ is a constant for a given dataset. Therefore, the posterior distribution of (β,σ^2) is proportional to the product between the likelihood

and the prior distribution. Hence,

$$\begin{aligned}
p(\beta,\sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\beta,\sigma^2)p(\beta,\sigma^2) \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Big[\frac{-1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)\Big] \\
&\quad \frac{b^a}{(2\pi)^{\frac{p}{2}}\Gamma(a)|V_\beta|^{\frac{1}{2}}}\Big(\frac{1}{\sigma^2}\Big)^{a+\frac{p}{2}+1} exp\Big\{\frac{-1}{\sigma^2}\Big[b+\frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\Big]\Big\} \\
&\propto \Big(\frac{1}{\sigma^2}\Big)^{a+\frac{n+p}{2}+1} \\
&\quad exp\Big\{\frac{-1}{\sigma^2}\Big[b+\frac{1}{2}\Big((\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)+(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\Big)\Big]\Big\}.
\end{aligned} \tag{6.5}$$

Let us consider the expression $(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)+(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)$ in the exponential part of the above equation. Expanding both transposes, the succeeding multiplications are executed. Then the like terms of the simplified expression are combined as below.

$$\begin{aligned}
&(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)+(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta) \\
&\quad = (\mathbf{y}^T-\beta^T X^T)(\mathbf{y}-X\beta)+(\beta^T-\mu_\beta^T)V_\beta^{-1}(\beta-\mu_\beta) \\
&\quad = \mathbf{y}^T\mathbf{y}-\mathbf{y}^T X\beta-\beta^T X^T\mathbf{y}+\beta^T X^T X\beta \\
&\qquad +\beta^T V_\beta^{-1}\beta-\beta^T V_\beta^{-1}\mu_\beta-\mu_\beta^T V_\beta^{-1}\beta+\mu_\beta^T V_\beta^{-1}\mu_\beta \\
&\quad = (\beta^T V_\beta^{-1}\beta+\beta^T X^T X\beta)-(\mathbf{y}^T X\beta+\beta^T X^T\mathbf{y}) \\
&\qquad -(\beta^T V_\beta^{-1}\mu_\beta+\mu_\beta^T V_\beta^{-1}\beta)+(\mathbf{y}^T\mathbf{y}+\mu_\beta^T V_\beta^{-1}\mu_\beta) \\
&\quad = \beta^T(V_\beta^{-1}+X^T X)\beta-2\mathbf{y}^T X\beta-2\mu_\beta^T V_\beta^{-1}\beta+(\mathbf{y}^T\mathbf{y}+\mu_\beta^T V_\beta^{-1}\mu_\beta) \\
&\quad = \beta^T(V_\beta^{-1}+X^T X)\beta-2(\mu_\beta^T V_\beta^{-1}+\mathbf{y}^T X)\beta+(\mathbf{y}^T\mathbf{y}+\mu_\beta^T V_\beta^{-1}\mu_\beta) \\
&\quad = \beta^T(V_\beta^{-1}+X^T X)\beta-2(V_\beta^{-1}\mu_\beta+X^T\mathbf{y})^T\beta+(\mathbf{y}^T\mathbf{y}+\mu_\beta^T V_\beta^{-1}\mu_\beta).
\end{aligned} \tag{6.6}$$

Let us assume that U and α are $k\times 1$ matrices. When A is a $k\times k$ symmetric positive definite matrix, then $A=A^T$ and $A^{-1}=(A^T)^{-1}=(A^{-1})^T$. Let us consider the expression

$U^TAU - 2\alpha^TU$. It can be simplified as:

$$\begin{aligned}
U^TAU - 2\alpha^TU &= U^TAU - U^T\alpha - \alpha^TU \\
&= U^TAU - U^TAA^{-1}\alpha - \alpha^TA^{-1}AU \\
&= U^TA\big(U - A^{-1}\alpha\big) - \alpha^TA^{-1}A\big(U - A^{-1}\alpha\big) - \alpha^TA^{-1}AA^{-1}\alpha \\
&= \big(U^TA - \alpha^TA^{-1}A\big)\big(U - A^{-1}\alpha\big) - \alpha^TA^{-1}\alpha \\
&= \big(U^T - \alpha^TA^{-1}\big)A\big(U - A^{-1}\alpha\big) - \alpha^TA^{-1}\alpha \\
&= \big(U - A^{-1}\alpha\big)^TA\big(U - A^{-1}\alpha\big) - \alpha^TA^{-1}\alpha
\end{aligned}$$

The first two of three sub-expressions in the right hand side of Equation 6.6 can be further simplified by applying the result in the above equation with $U = \beta,\ A = \big(V_\beta^{-1} + X^TX\big)$, and $\alpha = \big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)$. This yields:

$$\begin{aligned}
&\big(\mathbf{y} - X\beta\big)^T\big(\mathbf{y} - X\beta\big) + \big(\beta - \mu_\beta\big)^TV_\beta^{-1}\big(\beta - \mu_\beta\big) \\
&= \beta^T\big(V_\beta^{-1} + X^TX\big)\beta - 2\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)^T\beta + \big(\mathbf{y}^T\mathbf{y} + \mu_\beta^TV_\beta^{-1}\mu_\beta\big) \\
&= \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big]^T\big(V_\beta^{-1} + X^TX\big) \\
&\quad \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big] \\
&\quad - \big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)^T\big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big) + \big(\mathbf{y}^T\mathbf{y} + \mu_\beta^TV_\beta^{-1}\mu_\beta\big) \\
&= \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big]^T\big(V_\beta^{-1} + X^TX\big) \\
&\quad \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big] \\
&\quad - \big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)^T\big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1} + X^TX\big)\big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big) \\
&\quad + \big(\mathbf{y}^T\mathbf{y} + \mu_\beta^TV_\beta^{-1}\mu_\beta\big) \\
&= \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big]^T\big(V_\beta^{-1} + X^TX\big) \\
&\quad \big[\beta - \big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\big] \\
&\quad - \Big[\big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\Big]^T\big(V_\beta^{-1} + X^TX\big)\Big[\big(V_\beta^{-1} + X^TX\big)^{-1}\big(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\big)\Big] \\
&\quad + \big(\mathbf{y}^T\mathbf{y} + \mu_\beta^TV_\beta^{-1}\mu_\beta\big).
\end{aligned}$$

Let us define $\widetilde{V}_\beta$ and $\widetilde{\mu}_\beta$ as:

$$\widetilde{V}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}$$
$$\widetilde{\mu}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}\left(V_\beta^{-1}\mu_\beta + X^T \mathbf{y}\right) = \widetilde{V}_\beta\left(V_\beta^{-1}\mu_\beta + X^T \mathbf{y}\right)$$

and substitute them in the above equation. Then,

$$\begin{aligned}
&\left(\mathbf{y} - X\beta\right)^T\left(\mathbf{y} - X\beta\right) + \left(\beta - \mu_\beta\right)^T V_\beta^{-1}\left(\beta - \mu_\beta\right) \\
&\qquad = \left(\beta - \widetilde{\mu}_\beta\right)^T \widetilde{V}_\beta^{-1}\left(\beta - \widetilde{\mu}_\beta\right) - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta + \left(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta\right) \\
&\qquad = \left(\beta - \widetilde{\mu}_\beta\right)^T \widetilde{V}_\beta^{-1}\left(\beta - \widetilde{\mu}_\beta\right) + \left(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta\right). \qquad (6.7)
\end{aligned}$$

Now the result in Equation 6.5 is modified in the light of the result obtained in Equation 6.7. Then,

$$\begin{aligned}
p\left(\beta, \sigma^2 | \mathbf{y}\right) &\propto \left(\frac{1}{\sigma^2}\right)^{a + \frac{n+p}{2} + 1} \\
&\quad exp\Big\{\frac{-1}{\sigma^2}\Big[b + \frac{1}{2}\Big(\left(\mathbf{y} - X\beta\right)^T\left(\mathbf{y} - X\beta\right) + \left(\beta - \mu_\beta\right)^T V_\beta^{-1}\left(\beta - \mu_\beta\right)\Big)\Big]\Big\} \\
&= \left(\frac{1}{\sigma^2}\right)^{a + \frac{n+p}{2} + 1} exp\Big\{\frac{-1}{\sigma^2}\Big[b + \frac{1}{2}\left(\beta - \widetilde{\mu}_\beta\right)^T \widetilde{V}_\beta^{-1}\left(\beta - \widetilde{\mu}_\beta\right) \\
&\qquad\qquad + \left(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta\right)\Big]\Big\}.
\end{aligned}$$

Let us define $\tilde{a}$ and $\tilde{b}$ as:

$$\tilde{a} = a + \frac{n}{2}$$
$$\tilde{b} = b + \frac{1}{2}\Big(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta\Big)$$

and substitute them in the expression of the posterior density of $\left(\beta, \sigma^2\right)$ given above. Consequently, it becomes

$$p\left(\beta, \sigma^2 | \mathbf{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\tilde{a} + \frac{p}{2} + 1} exp\Big\{\frac{-1}{\sigma^2}\Big[\tilde{b} + \frac{1}{2}\left(\beta - \widetilde{\mu}_\beta\right)^T \widetilde{V}_\beta^{-1}\left(\beta - \widetilde{\mu}_\beta\right)\Big]\Big\}. \qquad (6.8)$$

As discussed previously, the fully conjugate prior distribution of $\left(\beta, \sigma^2\right)$ is a multivariate normal inverse gamma distribution (i.e; $MNIG\left(\mu_\beta, V_\beta, a, b\right)$) whose density function was

in the form (see Equation 6.3):

$$p(\beta,\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1} exp\Big\{\frac{-1}{\sigma^2}\Big[b+\frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\Big]\Big\}. \tag{6.9}$$

As shown in Equations 6.8 and 6.9, the prior and posterior distributions of (β,σ^2) belong to one class of densities. Therefore; the posterior distribution of (β,σ^2) is also a multivariate normal inverse gamma distribution. Hence,

$$p(\beta,\sigma^2|\mathbf{y}) \sim MNIG(\widetilde{\mu}_\beta,\widetilde{V}_\beta,\tilde{a},\tilde{b}). \tag{6.10}$$

6.5 The Prior Predictive Distribution

The inferences that can be made on unknown observable data are very important in all fields of statistics. In the Bayesian point of view, it is often called predictive inference. The distribution of unknown observable data y is not conditional on prior observations. In addition, it can be used to predict them. Therefore, the distribution of observable quantity $\mathbf{y}$ is informatively labelled as prior predictive distribution [72]. It is also named as marginal distribution and can be obtained as an integral of the conditional density $p(\mathbf{y}|\sigma^2)$ with respect to σ^2. The evaluation of the conditional distribution $p(\mathbf{y}|\sigma^2)$ is given below.

$$\begin{aligned}
p(\mathbf{y}|\sigma^2) &= \int_\beta p(\mathbf{y}|\beta,\sigma^2)p(\beta|\sigma^2)d\beta \\
&= \int_\beta N_n(X\beta,\sigma^2 I_n)N_p(\mu_\beta,\sigma^2 V_\beta)d\beta \\
&= \int_\beta \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2 I_n|^{\frac{1}{2}}} exp\Big[\frac{-1}{2}(\mathbf{y}-X\beta)^T(\sigma^2 I_n)^{-1}(\mathbf{y}-X\beta)\Big] \\
&\quad \frac{1}{(2\pi)^{\frac{p}{2}}|\sigma^2 V_\beta|^{\frac{1}{2}}} exp\Big[\frac{-1}{2}(\beta-\mu_\beta)^T(\sigma^2 V_\beta)^{-1}(\beta-\mu_\beta)\Big]d\beta \\
&= \int_\beta \frac{1}{(2\pi)^{\frac{n+p}{2}}|V_\beta|^{\frac{1}{2}}(\sigma^2)^{\frac{n+p}{2}}} \\
&\quad exp\Big\{\frac{-1}{2\sigma^2}\Big[(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)+(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\Big]\Big\}d\beta
\end{aligned}$$

This is further simplified applying the result in the Equation 6.7.

$$
\begin{aligned}
p(\mathbf{y}|\sigma^2) = & \int_{\beta} \frac{1}{(2\pi)^{\frac{n+p}{2}} |V_\beta|^{\frac{1}{2}} (\sigma^2)^{\frac{n+p}{2}}} \\
& exp\Big\{ \frac{-1}{2\sigma^2} \Big[(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1} \mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1} \widetilde{\mu}_\beta) + (\beta - \widetilde{\mu}_\beta)^T \widetilde{V}_\beta^{-1} (\beta - \widetilde{\mu}_\beta) \Big] \Big\} d\beta \\
= & \frac{1}{(2\pi)^{\frac{n}{2}} |V_\beta|^{\frac{1}{2}} (\sigma^2)^{\frac{n}{2}}} exp\Big[\frac{-1}{2\sigma^2} \Big(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1} \mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1} \widetilde{\mu}_\beta \Big) \Big] \\
& |\widetilde{V}_\beta|^{\frac{1}{2}} \int_{\beta} \frac{1}{(2\pi)^{\frac{p}{2}} |\sigma^2 \widetilde{V}_\beta|^{\frac{1}{2}}} exp\Big\{ \frac{-1}{2} \Big[(\beta - \widetilde{\mu}_\beta)^T (\sigma^2 \widetilde{V}_\beta)^{-1} (\beta - \widetilde{\mu}_\beta) \Big] \Big\} d\beta .
\end{aligned}
$$

Now it considers the fact that

$$
\int_{\beta} \frac{1}{(2\pi)^{\frac{p}{2}} |\sigma^2 \widetilde{V}_\beta|^{\frac{1}{2}}} exp\Big\{ \frac{-1}{2} \Big[(\beta - \widetilde{\mu}_\beta)^T (\sigma^2 \widetilde{V}_\beta)^{-1} (\beta - \widetilde{\mu}_\beta) \Big] \Big\} d\beta = 1.
$$

Then the conditional distribution $p(\mathbf{y}|\sigma^2)$ becomes

$$
p(\mathbf{y}|\sigma^2) = \frac{|\widetilde{V}_\beta|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}} |V_\beta|^{\frac{1}{2}} (\sigma^2)^{\frac{n}{2}}} exp\Big[\frac{-1}{2\sigma^2} \Big(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1} \mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1} \widetilde{\mu}_\beta \Big) \Big]. \tag{6.11}
$$

The simplification of this result is highly dependent on the expression

$$
\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1} \mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1} \widetilde{\mu}_\beta .
$$

According to the definition of $\widetilde{\mu}_\beta$, it is known that

$$
\widetilde{\mu}_\beta = \widetilde{V}_\beta (V_\beta^{-1} \mu_\beta + X^T \mathbf{y}).
$$

Substituting $\widetilde{\mu}_\beta$ in the expression:

$$
\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1} \mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1} \widetilde{\mu}_\beta
$$

with $\widetilde{V}_\beta\left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right)$ it can be simplified as given below.

$$
\begin{aligned}
&\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T\widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta \\
&= \mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \left(\widetilde{V}_\beta\left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right)\right)^T\widetilde{V}_\beta^{-1}\left(\widetilde{V}_\beta\left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right)\right) \\
&= \mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right)^T\widetilde{V}_\beta\widetilde{V}_\beta^{-1}\left(\widetilde{V}_\beta V_\beta^{-1}\mu_\beta + \widetilde{V}_\beta X^T\mathbf{y}\right) \\
&= \mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \left(\mu_\beta^T V_\beta^{-1} + \mathbf{y}^T X\right)\left(\widetilde{V}_\beta V_\beta^{-1}\mu_\beta + \widetilde{V}_\beta X^T\mathbf{y}\right) \\
&= \mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \mu_\beta^T V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}\mu_\beta - \mu_\beta^T V_\beta^{-1}\widetilde{V}_\beta X^T\mathbf{y} - \mathbf{y}^T X\widetilde{V}_\beta V_\beta^{-1}\mu_\beta - \mathbf{y}^T X\widetilde{V}_\beta X^T\mathbf{y} \\
&= \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\widetilde{V}_\beta X^T\mathbf{y}\right) - \left(\mathbf{y}^T X\widetilde{V}_\beta V_\beta^{-1}\mu_\beta + \mu_\beta^T V_\beta^{-1}\widetilde{V}_\beta X^T\mathbf{y}\right) \\
&\quad + \left(\mu_\beta^T V_\beta^{-1}\mu_\beta - \mu_\beta^T V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}\mu_\beta\right) \\
&= \mathbf{y}^T\left(I_n - X\widetilde{V}_\beta X^T\right)\mathbf{y} - 2\mathbf{y}^T X\widetilde{V}_\beta V_\beta^{-1}\mu_\beta + \mu_\beta^T\left(V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}\right)\mu_\beta
\end{aligned}
\tag{6.12}
$$

A further simplification of this expression would be very advantageous in future calculations. According to Equation 6.12, the simplification is greatly dependent on two key expressions: $X\widetilde{V}_\beta V_\beta^{-1}$ and $V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}$. According to the definition of $\widetilde{V}_\beta$, it is known that $\widetilde{V}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}$, therefore, $\widetilde{V}_\beta^{-1} = V_\beta^{-1} + X^T X$. The simplification of $X\widetilde{V}_\beta V_\beta^{-1}$ can be progressed with the identity $\widetilde{V}_\beta\widetilde{V}_\beta^{-1} = I_p$.

$$
\begin{aligned}
\widetilde{V}_\beta\left(V_\beta^{-1} + X^T X\right) &= I_p \\
\widetilde{V}_\beta V_\beta^{-1} &= I_p - \widetilde{V}_\beta X^T X \\
X\widetilde{V}_\beta V_\beta^{-1} &= X - X\widetilde{V}_\beta X^T X \\
X\widetilde{V}_\beta V_\beta^{-1} &= \left(I_n - X\widetilde{V}_\beta X^T\right)X
\end{aligned}
\tag{6.13}
$$

The steps associated with the simplification of $V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}$ are substantially dependent on the **Sherman-Morrison-Woodbury** formula. Even though there are a number of advancements and/or generalisations related to this, the original formula has been used to derive the inverse of a matrix [51]. The formula is defined in the following way. Let us assume that, S and U^T represents $m \times n$ matrices. R and T are square matrices with $n \times n$

and $m \times m$ dimensions respectively. When R and T are non-singular [51, 93],

$$\big(R+STU\big)^{-1} = R^{-1} - R^{-1}S\big(T^{-1}+UR^{-1}S\big)^{-1}UR^{-1}.$$

The formula can be proved in several ways; however, succeeding to a few tricky steps in matrix operations it can be proved as follows.

$$\begin{aligned}
RHS &= R^{-1} - R^{-1}S\big(T^{-1}+UR^{-1}S\big)^{-1}UR^{-1} \\
&= \big(R+STU\big)^{-1}\big(R+STU\big)\big[R^{-1} - R^{-1}S\big(T^{-1}+UR^{-1}S\big)^{-1}UR^{-1}\big] \\
&= \big(R+STU\big)^{-1}\big[I_n + STUR^{-1} - \big(S+STUR^{-1}S\big)\big(T^{-1}+UR^{-1}S\big)^{-1}UR^{-1}\big] \\
&= \big(R+STU\big)^{-1}\big[I_n + STUR^{-1} - ST\big(T^{-1}+UR^{-1}S\big)\big(T^{-1}+UR^{-1}S\big)^{-1}UR^{-1}\big] \\
&= \big(R+STU\big)^{-1}\big[I_n + STUR^{-1} - STUR^{-1}\big] \\
&= \big(R+STU\big)^{-1} \\
&= LHS
\end{aligned}$$

The term $\widetilde{V}_\beta$ in the expression $V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}$ is replaced with its original expression: $\widetilde{V}_\beta = \big(V_\beta^{-1}+X^TX\big)^{-1}$. Subsequently, the result is modified into a form that is analogous to the Sherman-Morrison-Woodbury formula.

$$\begin{aligned}
V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1} &= V_\beta^{-1} - V_\beta^{-1}\big(V_\beta^{-1}+X^TX\big)^{-1}V_\beta^{-1} \\
&= V_\beta^{-1} - V_\beta^{-1}\big[\big((X^TX)^{-1}\big)^{-1} + V_\beta^{-1}\big]^{-1}V_\beta^{-1} \\
&= V_\beta^{-1} - V_\beta^{-1}I_p\big[\big((X^TX)^{-1}\big)^{-1} + I_pV_\beta^{-1}I_p\big]^{-1}I_pV_\beta^{-1}
\end{aligned}$$

This is further simplified by substituting $R = V_\beta$, $S = U = I_p$, and $T = (X^TX)^{-1}$ in the Sherman-Morrison-Woodbury formula. Then,

$$V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1} = \big(V_\beta + I_p(X^TX)^{-1}I_p\big)^{-1} = \big((X^TX)^{-1} + I_pV_\beta I_p\big)^{-1}.$$

Again the Sherman-Morrison-Woodbury formula is applied with a different setting to simplify this result. Here, it assumes $R = (X^T X)^{-1}$, $S = U = I_p$, and $T = V_\beta$. Subsequently,

$$\begin{aligned} V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1} &= X^T X - X^T X I_p \left(V_\beta^{-1} + I_p X^T X I_p\right)^{-1} I_p X^T X \\ &= X^T X - X^T X \left(V_\beta^{-1} + X^T X\right)^{-1} X^T X \\ &= X^T \left(I_n - X\left(V_\beta^{-1} + X^T X\right)^{-1} X^T\right) X. \end{aligned}$$

However, according to the definition of $\widetilde{V}_\beta$, $\left(V_\beta^{-1} + X^T X\right)^{-1} = \widetilde{V}_\beta$. Therefore,

$$V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1} = X^T \left(I_n - X\widetilde{V}_\beta X^T\right) X. \tag{6.14}$$

Then Equations 6.13 and 6.14 are employed in Equation 6.12 and get,

$$\begin{aligned} &\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T \widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta \\ &= \mathbf{y}^T \left(I_n - X\widetilde{V}_\beta X^T\right)\mathbf{y} - 2\mathbf{y}^T X\widetilde{V}_\beta V_\beta^{-1}\mu_\beta + \mu_\beta^T \left(V_\beta^{-1} - V_\beta^{-1}\widetilde{V}_\beta V_\beta^{-1}\right)\mu_\beta \\ &= \mathbf{y}^T \left(I_n - X\widetilde{V}_\beta X^T\right)\mathbf{y} - 2\mathbf{y}^T \left(I_n - X\widetilde{V}_\beta X^T\right) X\mu_\beta + \mu_\beta^T X^T \left(I_n - X\widetilde{V}_\beta X^T\right) X\mu_\beta \\ &= \left(\mathbf{y}^T - \mu_\beta^T X^T\right)\left(I_n - X\widetilde{V}_\beta X^T\right)\left(\mathbf{y} - X\mu_\beta\right) \\ &= \left(\mathbf{y} - X\mu_\beta\right)^T \left(I_n - X\widetilde{V}_\beta X^T\right)\left(\mathbf{y} - X\mu_\beta\right). \end{aligned} \tag{6.15}$$

Using the definition of $\widetilde{V}_\beta$; $\widetilde{V}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}$, the expression $I_n - X\widetilde{V}_\beta X^T$ is reorganised as

$$\begin{aligned} I_n - X\widetilde{V}_\beta X^T &= I_n - X\left(V_\beta^{-1} + X^T X\right)^{-1} X^T \\ &= (I_n)^{-1} - (I_n)^{-1} X \left(V_\beta^{-1} + X^T (I_n)^{-1} X\right)^{-1} X^T (I_n)^{-1}, \end{aligned}$$

and simplified in light of the Sherman-Morrison-Woodbury formula by replacing $R = I_n$, $S = X$, $U = X^T$, and $T = V_\beta$. Finally, the subsequent result,

$$I_n - X\widetilde{V}_\beta X^T = \left(I_n + XV_\beta X^T\right)^{-1}$$

is applied in the Equation 6.15, and then its form is reduced to:

$$\begin{aligned}\mathbf{y}^T\mathbf{y}+\mu_\beta^T V_\beta^{-1}\mu_\beta-\widetilde{\mu}_\beta^T\widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta &= \Big(\mathbf{y}-X\mu_\beta\Big)^T\Big(I_n-X\widetilde{V}_\beta X^T\Big)\Big(\mathbf{y}-X\mu_\beta\Big)\\ &= \Big(\mathbf{y}-X\mu_\beta\Big)^T\Big(I_n+XV_\beta X^T\Big)^{-1}\Big(\mathbf{y}-X\mu_\beta\Big). \qquad (6.16)\end{aligned}$$

In addition to this result, it is very important to simplify the ratio $\frac{\left|\widetilde{V}_\beta\right|^{\frac{1}{2}}}{\left|V_\beta\right|^{\frac{1}{2}}}$, to streamline the conditional distribution $p\big(\mathbf{y}\big|\sigma^2\big)$ (Equation 6.11) into a simple form. The simplification can be started with the following formula that is very useful in evaluating the determinant of a sum of matrices of the form of $R+STU$ [93]. Let R and T be non-singular matrices with dimensions $n\times n$ and $m\times m$ respectively. The dimensions of matrices S and U are $n\times m$ and $m\times n$ respectively. Then,

$$\big|R+STU\big| = \big|R\big|\big|T\big|\big|T^{-1}+UR^{-1}S\big|.$$

According to the definition of $\widetilde{V}_\beta$, $\widetilde{V}_\beta=\big(V_\beta^{-1}+X^TX\big)^{-1}$. This definition can be slightly modified as $\widetilde{V}_\beta=\big(V_\beta^{-1}+X^TI_n^{-1}X\big)^{-1}$. Hence, $\widetilde{V}_\beta^{-1}=V_\beta^{-1}+X^TI_n^{-1}X$. The determinant of $\widetilde{V}_\beta^{-1}$ is obtained considering the above formula of the determinants.

$$\begin{aligned}\left|\widetilde{V}_\beta^{-1}\right| &= \left|V_\beta^{-1}+X^TI_n^{-1}X\right|\\ &= \left|V_\beta^{-1}\right|\left|I_n^{-1}\right|\left|I_n+XV_\beta X^T\right|\\ \left|\widetilde{V}_\beta^{-1}\right| &= \left|V_\beta^{-1}\right|\left|I_n+XV_\beta X^T\right|\end{aligned}$$

This result has streamlined the ratio of the determinants of $\widetilde{V}_\beta$ and V_β as

$$\frac{\left|\widetilde{V}_\beta\right|}{\left|V_\beta\right|}=\frac{1}{\left|I_n+XV_\beta X^T\right|}.$$

Therefore;

$$\frac{\left|\widetilde{V}_\beta\right|^{\frac{1}{2}}}{\left|V_\beta\right|^{\frac{1}{2}}} = \frac{1}{\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}}.$$

The Equation 6.11 is simplified based on Equation 6.16 and the result of the above equation in order to obtain the final form of the conditional density $p(\mathbf{y}|\sigma^2)$.

$$\begin{aligned}
p(\mathbf{y}|\sigma^2) &= \frac{\left|\widetilde{V}_\beta\right|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\left|V_\beta\right|^{\frac{1}{2}}(\sigma^2)^{\frac{n}{2}}} exp\Big[\frac{-1}{2\sigma^2}\Big(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T\widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta\Big)\Big] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}(\sigma^2)^{\frac{n}{2}}} exp\Big[\frac{-1}{2\sigma^2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big(I_n + XV_\beta X^T\Big)^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\sigma^2\Big(I_n + XV_\beta X^T\Big)\right|^{\frac{1}{2}}} exp\Big\{\frac{-1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big[\sigma^2\Big(I_n + XV_\beta X^T\Big)\Big]^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big\}
\end{aligned}$$

According to this result it is obvious that

$$p(\mathbf{y}|\sigma^2) \sim N_n\Big(X\mu_\beta, \sigma^2\big(I_n + XV_\beta X^T\big)\Big).$$

The prior predictive (marginal) distribution of $\mathbf{y}$, $p(\mathbf{y})$ is calculated as an integral of the joint probability density of $\mathbf{y}$ and σ^2, with respect to σ^2. That is,

$$p(\mathbf{y}) = \int_0^\infty p(y, \sigma^2) p\sigma^2.$$

However, mathematically, it is very convenient to proceed with an integral of the conditional distribution of y. Hence, the computation of $p(\mathbf{y})$ progresses an integration of

$p(\mathbf{y}|\sigma^2)$ with respect to σ^2 as given below.

$$
\begin{aligned}
p(\mathbf{y}) &= \int_0^\infty p(\mathbf{y}|\sigma^2)p(\sigma^2)d\sigma^2 = \int_0^\infty N_n\Big(X\mu_\beta, \sigma^2\big(I_n + XV_\beta X^T\big)\Big) IG(a,b)d\sigma^2 \\
&= \int_0^\infty \frac{1}{(2\pi)^{\frac{n}{2}}\Big|\sigma^2\Big(I_n + XV_\beta X^T\Big)\Big|^{\frac{1}{2}}} exp\Big\{\frac{-1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big[\sigma^2\Big(I_n + XV_\beta X^T\Big)\Big]^{-1} \\
&\quad \Big(\mathbf{y} - X\mu_\beta\Big)\Big\}\frac{b^a}{\Gamma(a)}\Big(\frac{1}{\sigma^2}\Big)^{a+1} exp\Big(\frac{-b}{\sigma^2}\Big)d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\Big|I_n + XV_\beta X^T\Big|^{\frac{1}{2}}} \\
&\quad \int_0^\infty \Big(\frac{1}{\sigma^2}\Big)^{\left(\frac{2a+n}{2}\right)+1} exp\Big\{\frac{-1}{\sigma^2}\Big[b + \frac{1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big(I_n + XV_\beta X^T\Big)^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big]\Big\}d\sigma^2
\end{aligned}
\tag{6.17}
$$

Let us assume that S is a random variable such that, $S \sim IG(\alpha, \lambda)$. Then,

$$
p(S) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\Big(\frac{1}{S}\Big)^{\alpha+1} exp\Big(\frac{-\lambda}{S}\Big) \quad \text{and} \int_0^\infty p(S)dS = 1.
$$

Therefore,

$$
\int_0^\infty \Big(\frac{1}{S}\Big)^{\alpha+1} exp\Big(\frac{-\lambda}{S}\Big)dS = \frac{\Gamma(\alpha)}{\lambda^\alpha}.
$$

Replace $\lambda = \Big[b + \frac{1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big(I_n + XV_\beta X^T\Big)^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big]$, $S = \sigma^2$, and $\alpha = \Big(\frac{2a+n}{2}\Big)$ in above integration. Then,

$$
\begin{aligned}
&\int_0^\infty \Big(\frac{1}{\sigma^2}\Big)^{\left(\frac{2a+n}{2}\right)+1} exp\Big\{\frac{-1}{\sigma^2}\Big[b + \frac{1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big(I_n + XV_\beta X^T\Big)^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big]\Big\}d\sigma^2 \\
&= \frac{\Gamma\big(\frac{2a+n}{2}\big)}{\Big[b + \frac{1}{2}\Big(\mathbf{y} - X\mu_\beta\Big)^T\Big(I_n + XV_\beta X^T\Big)^{-1}\Big(\mathbf{y} - X\mu_\beta\Big)\Big]^{\left(\frac{2a+n}{2}\right)}}.
\end{aligned}
$$

In the light of this result, Equation 6.17 can be simplified as below.

$$
\begin{aligned}
p(\mathbf{y}) &= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}} \\
&\quad \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\left(\frac{2a+n}{2}\right)+1} exp\left\{\frac{-1}{\sigma^2}\left[b + \frac{1}{2}\left(\mathbf{y} - X\mu_\beta\right)^T\left(I_n + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\mu_\beta\right)\right]\right\} d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}} \\
&\quad \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\left[b + \frac{1}{2}\left(\mathbf{y} - X\mu_\beta\right)^T\left(I_n + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\mu_\beta\right)\right]^{\left(\frac{2a+n}{2}\right)}} \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}} \\
&\quad \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\left(b\right)^{\left(\frac{2a+n}{2}\right)}\left[1 + \frac{1}{2b}\left(\mathbf{y} - X\mu_\beta\right)^T\left(I_n + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\mu_\beta\right)\right]^{\left(\frac{2a+n}{2}\right)}} \\
&= \frac{\Gamma\left(\frac{2a+n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)\left(b\right)^{\frac{n}{2}}\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}} \\
&\quad \left[1 + \frac{1}{2b}\left(\mathbf{y} - X\mu_\beta\right)^T\left(I_n + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\mu_\beta\right)\right]^{-\left(\frac{2a+n}{2}\right)}
\end{aligned}
$$

It is important to reorganise required parts of this equation in the way of identifying the distribution of the prior predictive distribution of $\mathbf{y}$, $p(\mathbf{y})$.

$$\begin{aligned}
p(\mathbf{y}) &= \frac{\Gamma\left(\frac{2a+n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)(b)^{\frac{n}{2}}\left|I_n + XV_\beta X^T\right|^{\frac{1}{2}}} \\
&\quad \left[1+\frac{1}{2b}\left(\mathbf{y}-X\mu_\beta\right)^T\left(I_n+XV_\beta X^T\right)^{-1}\left(\mathbf{y}-X\mu_\beta\right)\right]^{-\left(\frac{2a+n}{2}\right)} \\
&= \frac{\Gamma\left(\frac{2a+n}{2}\right)}{(\pi)^{\frac{n}{2}}\Gamma(a)\left|(2b)\left(I_n+XV_\beta X^T\right)\right|^{\frac{1}{2}}} \\
&\quad \left[1+\frac{a}{2ab}\left(\mathbf{y}-X\mu_\beta\right)^T\left(I_n+XV_\beta X^T\right)^{-1}\left(\mathbf{y}-X\mu_\beta\right)\right]^{-\left(\frac{2a+n}{2}\right)} \\
&= \frac{\Gamma\left(\frac{2a+n}{2}\right)}{(\pi)^{\frac{n}{2}}\Gamma(a)\left|(2a)\left(\frac{b}{a}\right)\left(I_n+XV_\beta X^T\right)\right|^{\frac{1}{2}}} \\
&\quad \left[1+\frac{1}{2a}\left(\mathbf{y}-X\mu_\beta\right)^T\left[\left(\frac{b}{a}\right)\left(I_n+XV_\beta X^T\right)\right]^{-1}\left(\mathbf{y}-X\mu_\beta\right)\right]^{-\left(\frac{2a+n}{2}\right)}
\end{aligned}$$

Finally,

$$\begin{aligned}
p(\mathbf{y}) &= \frac{\Gamma\left(\frac{2a+n}{2}\right)}{(\pi)^{\frac{n}{2}}\Gamma\left(\frac{2a}{2}\right)(2a)^{\frac{n}{2}}\left|\left(\frac{b}{a}\right)\left(I_n+XV_\beta X^T\right)\right|^{\frac{1}{2}}} \\
&\quad \left[1+\frac{1}{2a}\left(\mathbf{y}-X\mu_\beta\right)^T\left[\left(\frac{b}{a}\right)\left(I_n+XV_\beta X^T\right)\right]^{-1}\left(\mathbf{y}-X\mu_\beta\right)\right]^{-\left(\frac{2a+n}{2}\right)}.
\end{aligned} \tag{6.18}$$

Suppose T is a d_T dimensional random variable that has a multivariate Student t distribution with mean μ_T, scale matrix Σ_T, and degrees of freedom ν_T $\left(\text{i.e. } T \sim MVSt_{(\nu_T)}(\mu_T,\Sigma_T)\right)$. Then the probability density function of T is in the form [127]:

$$p(t\left|\mu_T,\Sigma_T,\nu_T\right.) = \frac{\Gamma\left(\frac{\nu_T+d_T}{2}\right)}{(\pi)^{\frac{d_T}{2}}(\nu_T)^{\frac{d_T}{2}}\Gamma\left(\frac{\nu_T}{2}\right)\left|\Sigma_T\right|^{\frac{1}{2}}}\left|1+\frac{1}{\nu_T}(t-\mu_T)^T\Sigma_T^{-1}(t-\mu_T)\right|^{-\left(\frac{\nu_T+D_T}{2}\right)}. \tag{6.19}$$

The Equation 6.18 is analogous to the probability density function of a Multivariate Student's t distribution (Equation 6.19). Hence, the prior predictive of $\mathbf{y}$ is a multivariate Student's t distribution with mean $X\mu_\beta$, scale matrix $\left(\frac{b}{a}\right)\left(I_n + XV_\beta X^T\right)$, and degrees of freedom $2a$. It can be denoted as:

$$p(\mathbf{y}) \sim MVSt_{(2a)}\left(X\mu_\beta, \left(\frac{b}{a}\right)\left(I_n + XV_\beta X^T\right)\right).$$

In conclusion, the computation of prior predictive distribution of y can be summarised as:

$$\begin{aligned} p(\mathbf{y}) &= \int_0^\infty p(\mathbf{y}|\sigma^2)p(\sigma^2)d\sigma^2 \\ &= \int_0^\infty N_n\left(X\mu_\beta, \sigma^2\left(I_n + XV_\beta X^T\right)\right)IG(a,b)d\sigma^2 \\ &= \int_0^\infty MNIG\left(X\mu_\beta, \sigma^2\left(I_n + XV_\beta X^T\right), a, b\right)d\sigma^2 \\ &\sim MVSt_{(2a)}\left(X\mu_\beta, \left(\frac{b}{a}\right)\left(I_n + XV_\beta X^T\right)\right). \end{aligned} \tag{6.20}$$

Alternatively, the prior predictive distribution of y can be defined as:

$$p(\mathbf{y}) = \int_0^\infty \int_\beta p(\mathbf{y}|\beta, \sigma^2)p(\beta, \sigma^2)d\beta d\sigma^2.$$

Taking Equations 6.1 and 6.4 into account, the above result can be revised as:

$$p(\mathbf{y}) = \int_0^\infty \int_\beta MN_n(X\beta, \sigma^2 I_n)MNIG(\mu_\beta, V_\beta, a, b)d\beta d\sigma^2.$$

However, the prior predictive distribution of $\mathbf{y}$ has been already verified and shown in Equation 6.20. Therefore, it is obvious that,

$$\int_0^\infty \int_\beta MN_n(X\beta, \sigma^2 I_n)MNIG(\mu_\beta, V_\beta, a, b)d\beta d\sigma^2 \sim MVSt_{(2a)}\left(X\mu_\beta, \left(\frac{b}{a}\right)\left(I_n + XV_\beta X^T\right)\right). \tag{6.21}$$

6.6 The Posterior Predictive Distribution

Let us assume that, $\widetilde{X}_m$ is the matrix of known covariates that requires predicting the analogous outcome $\widetilde{\mathbf{y}}_m$. $\widetilde{X}_m$ and $\widetilde{\mathbf{y}}_m$ are $m \times p$ and $m \times 1$ matrices respectively. The distribution of $\widetilde{\mathbf{y}}_m$ is expected to predict conditional on the data that have been already observed. Therefore, the distribution of $\widetilde{\mathbf{y}}_m$, $p(\widetilde{\mathbf{y}}_m|\mathbf{y})$ is called the posterior predictive distribution. In general, the posterior predictive distribution is defined as an average that is calculated based on the conditional predictions over the posterior distribution of (β, σ^2). The process followed in calculating the prior predictive is extended to determine the posterior predictive distribution of unobserved data $\widetilde{\mathbf{y}}_m$ [72].

According to the definition of the posterior predictive distribution,

$$p(\widetilde{\mathbf{y}}_m, |\mathbf{y}) = \int_{\sigma^2}\int_{\beta} p(\widetilde{\mathbf{y}}_m, \beta, \sigma^2|\mathbf{y}) d\beta d\sigma^2 = \int_{\sigma^2}\int_{\beta} p(\widetilde{\mathbf{y}}_m|\beta, \sigma^2, \mathbf{y}) p(\beta, \sigma^2|\mathbf{y}) d\beta d\sigma^2,$$

and assuming the conditional independence of $\mathbf{y}$ and $\tilde{\mathbf{y}}_m$ given (β, σ^2) [72],

$$p(\widetilde{\mathbf{y}}_m, |\mathbf{y}) = \int_{\sigma^2}\int_{\beta} p(\widetilde{\mathbf{y}}_m|\beta, \sigma^2) p(\beta, \sigma^2|\mathbf{y}) d\beta d\sigma^2. \tag{6.22}$$

In case of known β and σ^2, the distribution of $\widetilde{\mathbf{y}}_m$ is multivariate normal with mean $\widetilde{X}_m\beta$ and variance-covariance matrix $\sigma^2 I_m$. That is,

$$p(\widetilde{\mathbf{y}}_m|\beta, \sigma^2) \sim MN_m\left(\widetilde{X}_m\beta, \sigma^2 I_m\right). \tag{6.23}$$

Even though, β and σ^2 are unknown, posterior samples of them can be used to overcome the problem of unknown parameters. The posterior predictive distribution given by Equation 6.22, is reformed as below, based on Equation 6.23 and the posterior distribution of

(β, σ^2), which is given by Equation 6.10,

$$p(\widetilde{\mathbf{y}}_m, |\mathbf{y}) = \int_{\sigma^2}\int_{\beta} p(\widetilde{\mathbf{y}}_m|\beta, \sigma^2) p(\beta, \sigma^2|\mathbf{y}) d\beta d\sigma^2$$
$$= \int_{\sigma^2}\int_{\beta} MN_m(\widetilde{X}_m\beta, \sigma^2 I_m) MNIG(\widetilde{\mu}_\beta, \widetilde{V}_\beta, \tilde{a}, \tilde{b}) d\beta d\sigma^2.$$

This result is directly analogous to the outcome of the Equation 6.21. Therefore, posterior predictive of $\tilde{\mathbf{y}}_m$ is a multivariate Student's t distribution with mean $\widetilde{X}_m\widetilde{\mu}_\beta$, scale matrix $\left(\frac{\tilde{b}}{\tilde{a}}\right)\left(I_m + \widetilde{X}_m\widetilde{V}_\beta\widetilde{X}_m^T\right)$, and degrees of freedom $2\tilde{a}$. It can be denoted as:

$$p(\widetilde{\mathbf{y}}_m, |\mathbf{y}) \sim MVSt_{(2\tilde{a})}\left(\widetilde{X}_m\widetilde{\mu}_\beta, \left(\frac{\tilde{b}}{\tilde{a}}\right)\left(I_m + \widetilde{X}_m\widetilde{V}_\beta\widetilde{X}_m^T\right)\right). \tag{6.24}$$

6.7 Summary

Stutter ratio can be modelled as a simple linear regression model of longest uninterrupted sequence (*LUS*). Therefore, the theoretical outcomes discussed in this chapter must be simplified to a simple linear regression model.

$\beta^T = (\beta_0, \beta_1)$ is the regression coefficient matrix and σ^2 is the unknown finite constant variance of the random errors of the regression model. μ_β, V_β, a, and b are respectively the mean vector, variance-covariance matrix, shape, and scale parameters of bi-variate normal inverse gamma joint prior distribution of β and σ^2. The respective parameters of the posterior distribution of β and σ^2 are $\widetilde{\mu}_\beta$, $\widetilde{V}_\beta$, $\widetilde{a}$, and $\widetilde{b}$. The joint prior distribution and the observed data are connected in the following way to calculate these parameters where n, $\mathbf{y}$, and X are the sample size, response variable, and the design matrix respectively.

$$\widetilde{V}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}$$
$$\widetilde{\mu}_\beta = \left(V_\beta^{-1} + X^T X\right)^{-1}\left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right) = \widetilde{V}_\beta\left(V_\beta^{-1}\mu_\beta + X^T\mathbf{y}\right)$$
$$\tilde{a} = a + \frac{n}{2}$$
$$\tilde{b} = b + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \widetilde{\mu}_\beta^T\widetilde{V}_\beta^{-1}\widetilde{\mu}_\beta\right)$$

The posterior predictive distribution of a new datum $(\widetilde{X},\widetilde{y})$ can be derived by replacing $(\widetilde{X}_m,\widetilde{\mathbf{y}}_m)$ by $(\widetilde{X},\widetilde{y})$ and the number of data points in the new dataset m by one, in Equation 6.24. Then the posterior predictive distribution of the new datum is a univariate non-standardised Student's t distribution with location, scale, and degrees of freedom parameters $\widetilde{X}\widetilde{\mu}_\beta$, $\left(\frac{\tilde{b}}{\tilde{a}}\right)\left(1+\widetilde{X}\widetilde{V}_\beta\widetilde{X}^T\right)$, and $2\tilde{a}$ respectively.

Chapter 7

Infinite Mixtures of Linear Regression Models

7.1 Introduction

Extending the idea of two-component mixture models presented in Chapter 2, this chapter introduces the infinite mixtures of linear regression modelling approach to develop more robust models for predicting stutter ratio. With the normality assumption of random error terms, a mixture of linear regression models can be interpreted as a mixture of Gaussian distributions. In addition, an infinite mixture is the infinite limit of a finite mixture model. Hence, a brief review of finite mixture models, especially the finite mixtures of normal densities is presented. Finally, this chapter briefly discusses some of the approaches including **Chinese Restaurant Process** (**CRP**) that can be used to deal with infinite mixtures and introduces some improvements to the collapsed Gibbs sampling algorithm in order to search better models in terms of log posterior density (log-likelihood).

In real-world phenomena, data that we are trying to model are often much more complex than would be expected under theoretical considerations. For instance, data that have originated from heterogeneous sources could indicate more than one peak in its probability distribution. The presence of bi-modal histograms with a sufficient gap between two peaks probably indicates a mixture of two or more sources for the data [123]. Occasionally, irregular combinations of high and low probability masses can also be seen in the

distribution of data. In an extreme case, a deviation or isolation of one or more parts of the distribution could also be demonstrated. Using a set of assumptions is an indispensable characteristic of any statistical model. In most of the model building methods, a single distribution is frequently assumed for the data. However, if the data originated from more than one distribution, then they may not be well-characterised by a single distribution [105]. It is not uncommon to encounter situations where the probability distribution of data is not compatible with any of the known statistical probability distributions. The modelling of datasets with above or similar characteristics is always challenging and requires much expertise in the state-of-the-art techniques. In circumstances where a standard single probability distribution does not help, a finite mixture of parametric probability distributions may be useful to model data.

7.2 Beyond Multiple Linear Regression Models

Generally, a multiple linear regression model is presented as

$$y_i = E(y_i) + \varepsilon_i,$$

where $E(y_i)$ and ε_i are the mathematical expectation and the random error term of the i^{th} observation of the response variable y_i respectively [178]. Let us assume that we have a multiple linear regression model with k explanatory variables $X_1, X_2, ..., X_k$. If there are n observations $(x_{11}, x_{21}, ..., x_{k1})$, $(x_{12}, x_{22}, ..., x_{k2})$, ..., $(x_{1n}, x_{2n}, ..., x_{kn})$ corresponding to a particular scenario, then the model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \quad for \quad i = 1, 2, ..., n.$$

Typically, we assume that the errors are independent and identically distributed (i.i.d) with zero mean (i.e. $\mathrm{E}(\varepsilon_i) = 0$) and that we have unknown finite constant variance (i.e. $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$). In addition, we usually assume that the errors are normally distributed. This assumption allows us to make inferential statements. These assumptions on the error lead to the following assumptions on the response variable y.

1. Independence
2. Normality
3. Constant variance

Any of these assumptions may be dropped by changing the model under specific conditions. For example, the independence assumption can be dropped and use a time series model where the data are time-based. Significant violations of the assumption of normality may lead to poor prediction performance, and misleading inferences. The presence of skewness (lack of symmetry), heavy-tailedness, and light-tailedness are some of the indications of violations of normality. When the error variance is fixed to a finite constant σ_ε^2, the data is said to be homoscedastic, otherwise it is referred to as heteroscedastic [178]. If the data have no constant variance or normality, generalised linear models (GLMs) may be possible under certain conditions (e.g. if the data are rightly-skewed, a log-normal or a gamma distribution may be appropriate). Incorrectly specified models and the skewed distribution of the explanatory variables have been frequently discussed as possible reasons for heteroscedasticity in linear regression models [91, 178], and is a problem in GLMs too. When a model is specified without one or more important variables, it is possible to arrive at a situation with heteroscedastic features. However,these features can often be removed with the introduction of appropriate variables to the model. Heteroscedasticity is also addressed by way of transformation of the response. For example, it is common to take the logarithm of the response when the variance increases with the mean.

If there is a lack of independence in data, then the estimators are inefficient and this leads to too small (underestimated) standard deviations. However, there is no problem related to this assumption in modelling stutter ratio. According to the central limit theorem, when we use large datasets, any violation of the assumptions does not have a considerable effect on model parameter estimates or confidence intervals for the mean. However, if a model is fitted for a predictive purpose, then a misspecified model or variance parameters with non-constant behaviour lead to poor performance in interpretations that depend on the predicted information. In addition, even with large datasets traditional modelling techniques may not be effective when the data has specific characteristics. For example,

when data exhibit a clustered behaviour, it is very difficult to model them using a known single distribution [179].

Transformation may not help in situations where the residuals of a linear regression model exhibit several modes. However, finite mixture models can be applied as an alternative, powerful, and flexible tool [105]. The flexibility that is provided by the mixture models to address the issues related to heavy-tailed, skewed (asymmetrical), multimodel, and leptokurtic or platykurtic data sets is widely appreciated in statistical modelling [105, 123, 179]. It is rather difficult to approximate these situations using most of the known distributions. Mixture models, in contrast, can be equally employed to assess such datasets that have originated from either known or unknown distributions. They are also capable in handling underdispersion, overdispersion, and heteroscedasticity issues related to traditional models [105].

Mixture models can be basically dichotomised as finite and infinite. In principle, a convex combination of two or more probability density functions is defined as a *finite mixture model*. More generally, it is a statistically weighted finite collection of probability density functions. The infinite limit of a finite mixture model is usually regarded as an *infinite mixture model*. The infinite mixture model is also known as a *Dirichlet process mixture model*, which is a distribution over distributions. It is widely recognised as a stochastic process that can be used in Bayesian non-parametric modelling of data. There are numerous applications of Dirichlet processes in recent literature [4, 133].

7.3 History of Finite Mixture Models

Finite mixtures of distributions are increasingly receiving attention from both theorists and practitioners as an extremely flexible method that provides a mathematical-based approach for modelling [123]. There are numerous applications in the fields of astronomy, biology, genetics, medicine, psychiatry, economics, and engineering which emphasise the effectiveness of finite mixture models in complex situations. When the populations of data are known or are suspected to be comprised of sub-populations, finite mixture models can be effectively used [57]. In cluster analysis, it is very convenient to address

the issue of heterogeneity, using finite mixtures of distributions [123]. In the univariate case, a finite mixture of normal distributions with common variance can be successfully employed to approximate any continuous distribution. Similarly, in multivariate cases, the common variance is replaced by a common variance-covariance matrix. Moreover, finite mixture models provide a convenient semi-parametric framework to model distributions with any unidentified shapes. A mixture model with an appropriate number of components is capable in representing reasonably complex models with a high degree of accuracy. Finite mixture models are attracting extensive interest in situations where a single parametric family reveals inability to produce a satisfactory model to represent the distributional characteristics of the observed data.

A paper entitled "*Contributions to the Mathematical Theory of Evolution*" [132] written in 1894 by reputed biometrician K. Pearson is regarded as one of the earliest to use finite mixture models. In his study on several body measures of crabs, the mixture model (two-component heterogeneous mixture of normal densities) provided a better description of the data because it models the sub-species differences. The methodology adopted in this analysis was based on the method of moments derived from the observed frequency distribution concerning the mid-points of ratios of the forehead to body length intervals. The moment-based method involved a massive amount of calculations in finding the five parameters of the mixture model. Rao [57] suggested the maximum likelihood estimation as an alternative technique to fit a two-component mixture model. In comparison with the moment-based estimation, maximum likelihood estimators have many desirable statistical properties. The maximum likelihood approach is appreciated as a method which is capable in producing highly efficient results [142]. However, the calculation of maximum likelihood estimates subjected to incomplete datasets is a common problem in many statistical applications. With the introduction of "*Expectation-Maximisation*" [EM] algorithm in 1977 by Dempster, Laird, and Rubin, the method of maximum likelihood estimation has been extensively recognised as the most commonly-used method of fitting finite mixture distributions [50, 123].

7.4 Finite Mixtures of Normal Densities

Finite mixture models, especially the Gaussian mixtures, are frequently employed as fundamental data analysis tools in clustering and classification of data. The interpretability of the results under these models is fully empowered by its comprehensive mathematical basis [62]. As a consequence of this, they are being progressively preferred by the modelling practitioners. Among the various clustering procedures, the ones that provide facilities to model overlapping clusters are highly sophisticated. The finite mixture model is regarded as a probabilistic approach combined with model based procedure, which has been capable of modelling overlapping clusters. In real-world problems where the unobserved heterogeneity in data cannot be disregarded, finite mixture models have been regarded as a powerful framework [53]. In the applications of these models, the observations are assumed to be drawn from more than one heterogeneous sources (populations). Each source is termed as a cluster or a sub-population. In model-based clustering, each source is reflected by a cluster and the clusters are modelled using parametric models. In Gaussian mixtures, the family of normal distributions are used to build parametric models for each cluster. Sometimes, these clusters are naturally defined, hence, the number of components in the model is known. However, in the problems where clusters cannot be naturally recognised, this has to be specified by the user. In such circumstances, the optimal number of components in the model can be assessed in terms of the likelihood of data under each competing model [62].

It is very convenient to illustrate the broadness of normal mixtures with the consideration of univariate case. A univariate Gaussian mixture density can be defined as a weighted average of individual Gaussian densities. Let us consider a K-component Gaussian mixture with mean, variance, and mixing proportion vectors $\mu = (\mu_1, \mu_2, \ldots, \mu_K)^T$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)^T$, and $\pi = (\pi_1, \pi_2, \ldots, \pi_K)^T$ respectively. Here; μ_k, σ_k^2, and π_k $(k = 1, 2, \ldots, K)$ denote the mean, variance and mixing proportion corresponding to the k^{th} component of the mixture respectively. Mixing proportions are non-negative and summed to one. Then the mixture density $f(y_j|\mu, \boldsymbol{\sigma^2}, \pi)$ of a random variable y_j can be defined in

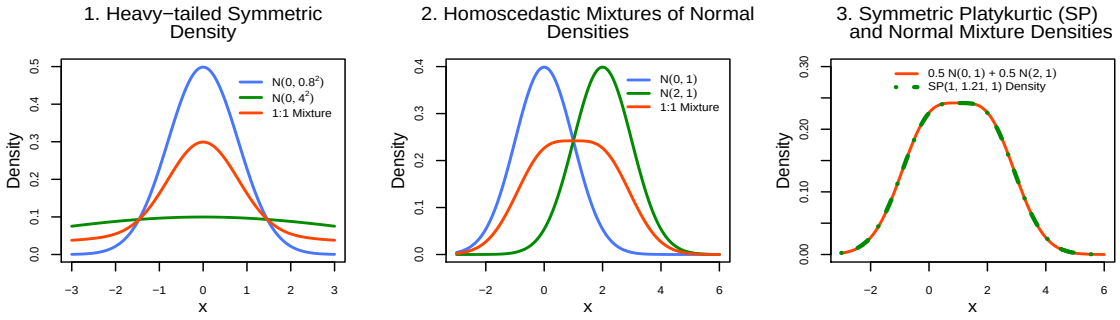

Figure 7.1: Density plots of two component 1:1 mixtures of univariate normal densities. $SP(\mu, \sigma^2, \gamma)$ denotes a symmetric platykurtic distribution with μ, σ^2, and γ as the location, scale, and kurtosis parameters respectively.

the form

$$f(y_j|\mu, \boldsymbol{\sigma^2}, \pi) = \sum_{k=1}^{K} \pi_k \phi_k(y_j|\mu_k, \sigma_k^2),$$

where

$$\phi_k(y_j|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\Big[-\frac{1}{2}\Big(\frac{y_j - \mu_k}{\sigma_k}\Big)^2\Big].$$

The flexibility of finite mixtures of Gaussians in approximating various types of distributions can be graphically illustrated. The first plot of Figure 7.1 illustrates the flexibility in approximating heavy-tailed symmetric distributions. In real-world applications, there may be some densities deviated substantially from elliptically symmetric behaviours and cannot be modelled even with the family of t-distributions. A mixture density as it illustrated in Figure 7.1, can be efficiently used to approximate such asymmetric densities. The family of *Symmetric Platykurtic* (SP) distributions is one of the best alternatives that can be used to overcome the poor fitting problems of normal densities especially with the presence of heavy tailed distributions [6]. The probability density function of a symmetric

platykurtic distribution with parameters μ, σ^2, and γ; $SP(\mu, \sigma^2, \gamma)$ is defined as

$$f(x|\mu,\sigma^2,\gamma) = \frac{\left[2\gamma + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{\gamma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sum_{k=0}^{\gamma} \binom{\gamma}{k} (2\gamma)^{\gamma-k} 2^{k+\frac{1}{2}} \Gamma(k+\frac{1}{2})\sigma},$$

where, μ and $\sigma(>0)$ are the location and scale parameters. The kurtosis parameter γ determines the shape of the distribution and takes only non-negative integers. The Gaussian distribution is a special case of symmetric platykuric distributions when $\gamma = 0$. The second plot in Figure 7.1 illustrates a symmetric platykurtic distribution as a mixture of two homoscedastic Gaussian densities with equal weights (1:1 mixture). The third plot visually evidences the goodness-of-fit of this Gaussian mixture in approximating symmetric platykurtic distribution with location, scale, and kurtosis parameters 1, 1.21, and 1 respectively.

The problems associated with heavy-tailed distributions are generally encountered with the family of Student's t-distributions. It is a rich member of the well-known bell-shaped symmetric distributions family and enables more flexibility and robustness in modelling. Degrees of freedom, the additional parameter empowers to accommodate heavier tails in the distribution than with the standard normal distribution. The non-standardised Student's t-distribution is a member of location-scale family and provides more facilities to accommodate various bell-shaped behaviours in statistical modelling. The mean and variance are defined when the degrees of freedom of the distribution is greater than 1 and 2 respectively. The mean of the distribution exactly equals the location parameter as it does in Gaussian distribution. In contrast, the scale parameter of the distribution does not exactly equal the variance, unlike the normal density. However, the variance is a function of both degrees of freedom ν and scale parameter σ^2 of the distribution, and is defined as $\frac{\nu}{\nu-2}\sigma^2$.

Representation of the Student-t distribution as a scale mixture of normal is a well-known fact in statistics. The Student's t-distribution with location, scale, and degrees of freedom parameters θ, σ^2, and ν respectively can be expressed as a mixture of normal distributions $N(\mu, \omega^2\sigma^2)$ assuming an inverse gamma distribution $IG(\frac{\nu}{2}, \frac{\nu}{2})$ for ω^2 [1, 38,

39]. Mathematically,

$$t_\nu(x|\theta,\sigma^2) = \int_0^\infty N(x|\theta,\omega^2\sigma^2) IG(\omega^2|\frac{\nu}{2},\frac{\nu}{2}) d\omega^2$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\omega^2\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\omega\sigma}\right)^2} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \omega^{-(\nu+2)} e^{\left(-\frac{\nu}{2\omega^2}\right)} d\omega^2.$$

Scale mixtures of normal densities are efficiently used in Bayesian estimation via Markov Chain Monte Carlo (MCMC) methods[55]. However, a lack of performance has been observed in the symmetric versions of scale mixtures with the presence of extreme outliers that are revealed only in one side of the data distribution. As an alternative, skewed scale mixtures that can accommodate both skewness and heavy-tailedness have been introduced to overcome the problems associated with their symmetric counterparts. For example, the symmetric Student-t distribution is replaced with the skewed Student-t density, which is a mixture of skewed normal distributions.

Finite mixtures of Gaussian densities are frequently found in various fields of modelling [57]. However, the lack of robustness with the presence of outliers is a key limitation of the finite mixtures of normal distributions [8]. The influence of outlying data on estimating the means and the variances (or variance-covariance matrices) is a well-known fact in statistics, and is regarded as one of the main reasons for the robustness problems in finite Gaussian mixtures. The behaviour of outliers in a dataset can be captured by increasing the number of components (clusters) in the model. However, increased complexity of the model as a result of these additional components needs to be considered. The problem of model complexity with the existence of outliers can be successfully overcome by using Student-t mixtures instead of Gaussian mixtures.

Among the literature on stochastic volatility models can be found some interesting applications of finite mixtures of normal densities that approximate log chi-square random variables with one degree of freedom (i.e.log- χ_1^2). Chi-square distributions are always positively skewed and the level of skewness decreases as the degrees of freedom increases. Hence, the distribution with one degree of freedom is the most right-skewed one among others. Log transformations are frequently used to obtain symmetric behaviours for pos-

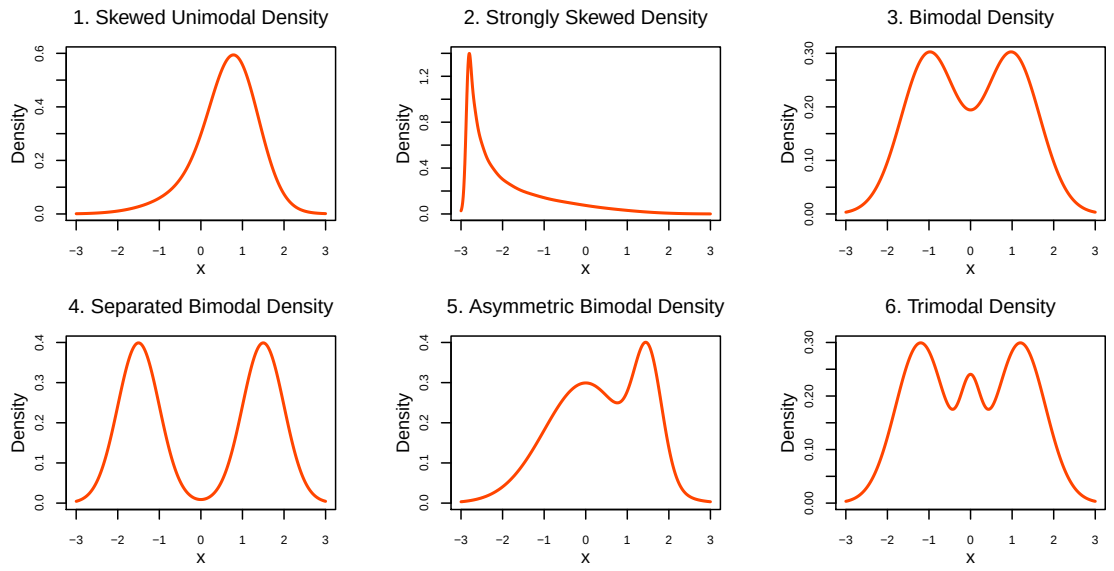

Figure 7.2: Various shapes of normal mixture densities

Table 7.1: Parameters for normal mixture densities shown in Figure 7.2

Mixture density	Weights and parameters of the component densities
1. Skewed unimodal	$\frac{1}{5}N(0,1)+\frac{1}{5}N\left(\frac{1}{2},(\frac{2}{3})^2\right)+\frac{3}{5}N\left(\frac{13}{15},(\frac{5}{9})^2\right)$
2. Strongly skewed	$\sum_{k=0}^{7}\frac{1}{8}N\left(3\left[(\frac{2}{3})^k-1\right],(\frac{2}{3})^{2k}\right)$
3. Bimodal	$\frac{1}{2}N\left(-1,(\frac{2}{3})^2\right)+\frac{1}{2}N\left(1,(\frac{2}{3})^2\right)$
4. Separated bimodal	$\frac{1}{2}N\left(-\frac{3}{2},(\frac{1}{2})^2\right)+\frac{1}{2}N\left(\frac{3}{2},(\frac{1}{2})^2\right)$
5. Asymmetric bimodal	$\frac{3}{4}N(0,1)+\frac{1}{4}N\left(\frac{3}{2},(\frac{1}{3})^2\right)$
6. Trimodal	$\frac{9}{20}N\left(-\frac{6}{5},(\frac{3}{5})^2\right)+\frac{9}{20}N\left(\frac{6}{5},(\frac{3}{5})^2\right)+\frac{1}{10}N\left(0,(\frac{1}{4})^2\right)$

Source: Adapted from Marron and Wand [121]

itively skewed random variables. However, the logarithm of chi-square distribution with one degree of freedom is moderately left-skewed. The log-square transformation of a standard normal (logarithm of a chi-square) random variable is a key problem in many stochastic volatility models for financial time series. In these models, log chi-square ran-

dom variables with one degree of freedom (log- χ_1^2) are very poorly approximated by normal distributions, especially with small samples [107]. As a consequences of this, Gaussian mixtures have been used as an approximation in many applications related to volatility models. For example, a seven-component mixture of normal densities was successfully used to approximate the log chi-square distribution by Kim et al.[107], in their study about stochastic volatility models. Following this study, Mahieu et al.[119], conducted an experiment on this approximation and settled for a three-component mixture (with mixing proportions 0.70, 0.25, and 0.05) instead of seven. In another application of stochastic volatility models with expectation-maximisation (EM) algorithm, a two component normal mixture has been used to replace the log chi-square distribution, and observed a robust fit [106]. Marron and Wand [121] presented a wide variety of finite normal mixtures in order to illustrate their flexibility in approximating various types of distributional features including different magnitudes of skewness and multimodality. Even though their article contained 15 such mixtures, only six of them are reproduced in Figure 7.2 to demonstrate the broadness of the normal mixtures in modelling. The weights and parameters corresponding to the individual components of these six mixture densities are presented in Table 7.1 for further information.

7.5 Finite Mixtures of Multiple Linear Regression Models

Finite mixture models for linear regression models can be regarded as a special case of Gaussian mixtures as it employs the family of normal distributions to model the errors within each component.

Let us assume a K-component finite mixture of multiple linear regression models with $p-1$ predictors. The conditional density h of an n-dimensional vector of data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of a dependent variable $\mathbf{Y}$ is in the form

$$h(\mathbf{y}|X, \beta_1, \beta_2, \ldots, \beta_K, \pi_1, \pi_2, \ldots, \pi_K) = \sum_{j=1}^{K} \pi_j \phi_j(\mathbf{y}_j|X_j, \beta_j, \sigma_j^2),$$

where π_j is the mixing probability of cluster j. π_j's are positive and summed to one (i.e. $0 < \pi_j \leq 1$, for all $j = 1,2,\ldots,K$ and $\sum_{j=1}^{K} \pi_j = 1$). X is an $n \times p$ design matrix which includes the values of the covariates, β_j and σ_j^2 are the $p \times 1$ vector of the regression coefficients and variance of random error terms for cluster j respectively. Assuming n_j number of observations for the j^{th} cluster, the density ϕ_j of an n_j-dimensional vector of dependent variable $\mathbf{y}_j^T = (y_{j1}, y_{j2}, \ldots, y_{jn_j})$ has been defined conditional on $n_j \times p$ design matrix X_j and β_j as follows.

$$\begin{aligned}\phi_j(\mathbf{y}_j|X_j,\beta_j,\sigma_j^2) &= \frac{1}{(2\pi)^{\frac{n_j}{2}}\left|\sigma_j^2 I_{n_j}\right|^{\frac{1}{2}}} exp\Big[\frac{-1}{2}\left(\mathbf{y}_j - X_j\beta_j\right)^T\left(\sigma_j^2 I_{n_j}\right)^{-1}\left(\mathbf{y}_j - X_j\beta_j\right)\Big] \\ &= \frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{n_j}{2}}} exp\Big[\frac{1}{2\sigma_j^2}\left(\mathbf{y}_j - X_j\beta_j\right)^T\left(\mathbf{y}_j - X_j\beta_j\right)\Big]\end{aligned}$$

Finite mixtures of regression models have been broadly discussed in existing literature. Various types of models including linear, generalised linear, and generalised linear mixed models have been reported in these literatures. Extending the usefulness of the concepts highlighted in the literature, software packages have been developed. Generally, finite mixtures of multiple linear regression models can be treated as the simplest among others. Various **R** packages developed to provide computational functionality of finite mixture models are found. For instance, the **R** package **flexmix** is a computational software solution that offers conveniences through flexible infrastructure for fitting mixtures of regression models [90]. Basically, the package empowers the functionality of the model-based clustering, and the Expectation-Maximization (EM) algorithm is used to estimate the model parameters. The **R** package **mclust** provides computational infrastructure for model-based clustering, classification, and density estimation using maximum likelihood technique via EM algorithm for parameter estimation [63]. The **R** package **mixtools** also offers a wide range of alternatives for fitting finite mixtures of regression models. Either the EM algorithm itself or the ideas developed based on it are used as the basis for many algorithms in the package [13].

Statistical Analysis Software (**SAS/STAT®**) is another leading computational software that provides a broad range of procedures in statistical modelling. Linear regression models are always treated as special cases of generalised linear models, and the **FMM**

procedure (**PROC FMM**) in SAS provides essential facilities for fitting mixtures of them [105]. Both maximum likelihood and Bayesian techniques are available in the **FMM** procedure. A majority of the **R** packages for fitting finite mixture models are heavily dependent on the EM algorithm in maximum likelihood estimation (e.g. **flexmix**, **mclust**, and **mixtools**). **PROC FMM** in **SAS**, in contrast, uses a dual quasi-Newton optimisation algorithm as the default method in parameter estimation in the context of mixture models. However, several other optimisation methods, namely, conjugate-gradient, double-dogleg, Nedler-Mead simplex, Newton-Raphson technique with ridging, and trust-region are also available to accomplish parameter estimation under finite mixture models. Alternatively, Gibbs sampling technique is used as the default option under the Bayesian version of **PROC FMM**. However, the Metropolis-Hasting algorithm that was originally proposed by Gamerman [64] is used especially for the situations where the Gibbs sampling is impossible [105].

7.5.1 Number of Components in a Finite Mixture Model

Finite mixture models provide a great flexibility in modelling data that are assumed to come from more than one source population. In fact, the number of sources that the data were generated under many practical phenomena is generally unknown. Hence, the selection of an appropriate number of sources, which is the number of components in the mixture model that reflects the optimal level of model complexity, is always problematic. A comprehensive review of finite mixture modelling, including the problem of selecting the number of components, has been provided by McLachlan et al. [123]. The issue of determining the number of components has been further reviewed with updated information by McLachlan et al. [124] in the context of Gaussian mixture models. The selection of number of components in a finite mixture model has been discussed in relation to several methods such as reversible jump sampler in relation to MCMC methods [89, 116, 125, 143], methods related to Bayes factors or BIC which approximates the logarithm of a Bayes factor [62, 103, 123, 124, 150], birth-and-death process [33, 125, 155, 158], and methods related to likelihood ratio tests [102, 117, 124]. In the context of density estimation, the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are

considered as adequate criteria to determine the appropriate number of mixture components [123]. However, the BIC has been highlighted as the most widely used measure to determine the number of components [124, 149].

7.6 Infinite Mixture Models

The model complexity is often expressed in terms of number of parameters in the model. A misfit between the model complexity and the amount of data at hand can lead to over-fitting or under-fitting problems in traditional parametric models that have fixed and finite number of parameters [163]. Bayesian non-parametric methods provide increased flexibility in applied statistical modelling as they do not restrict the number of parameters in the model. In fact, the models where the number of parameters can grow with the size of training dataset are more appropriate to be referred to as fully non-parametric. In Bayesian non-parametric modelling, Dirichlet processes, in particular Dirichlet process mixture models, are frequently used [163]. Specifying the number of components in finite mixture models however is practically difficult even though the calculations are relatively simple. Infinite mixture models, in contrast, do not require the user to specify the number of components. Instead, a Dirichlet process [61], which is an infinite-dimensional generalisation of the Dirichlet distribution [163] is used.

7.6.1 Dirichlet Process (DP)

Non-parametric models have been continuously influenced by Bayesian modelling techniques [163]. It is a general practice to assume that the data at hand have been drawn identically and independently from an unknown underlying distribution (say F). In the Bayesian approach, a prior distribution is assumed on the parameters of F, and the posterior distribution is derived based on the observed data and the prior. Non-parametric Bayesian approach, in contrast, places a prior distribution over a set of distributions [42]. A Dirichlet prior is a typical example for a non-parametric prior distribution. DP mixture models are also known as infinite mixtures. In these models, the number of components is countably infinite, which can technically be controlled by defining a prior distribution

on the mixing proportions [88, 128]. The Chinese restaurant process, the Stick-breaking construction, and the Pólya urn scheme are frequently used as Dirichlet prior distributions in Bayesian mixture models. These representations of DP offer different inference algorithms for DP-based mixture models [162]. However, there are some close relationships among them [3, 88, 164]. For example, Polya urn scheme is closely related to CRP, which is a distribution over partitions. This scheme can be generalised using exchangeability, and it leads to the DP.

A DP is a stochastic process [163] and is a distribution over distributions (a measure on measures) [42]. It is mathematically defined along with advanced principles of measure theory. Let Θ is measurable space, H is a probability measure defined on Θ, and α is a positive scaler [61, 163]. Then G is defined as a Dirichlet process with parameters H and α (i.e. $\mathrm{DP}(\alpha, H)$), if $G(B_1), \ldots, G(B_k)$ has a Dirichlet distribution for all finite measurable partitions $\big(B_1, \ldots, B_k\big)$ of Θ for every $k = 1, 2, \ldots$. In notational form

$$G \sim \mathrm{DP}(\alpha, H),$$
$$\text{if} \quad \big(G(B_1), \ldots, G(B_k)\big) \sim \mathrm{Dir}\big(\alpha H(B_1), \ldots, \alpha H(B_k)\big)$$

for every finite measurable partition $B_1, \ldots, B_k$ of Θ.

There are two parameters of Dirichlet process: the base distribution H and the concentration (strength) parameter α [48, 92, 128]. The base distribution is a joint prior distribution of the component parameters [88, 179], and the concentration parameter is a positive scaler. For any measurable set B of Θ (i.e. $B \subset \Theta$), the mean and the variance of $G(A)$ are formulated as below [163].

$$\mathrm{E}[G(A)] = H(A)$$
$$\mathrm{Var}[G(A)] = \frac{H(A)(1 - H(A))}{(\alpha + 1)}$$

Therefore, H and α can be understood as the mean and inverse variance of $\mathrm{DP}(\alpha, H)$. G is a random distribution for any finite measurable partition $B_1, \ldots, B_k$ of Θ, hence $G(B_1), \ldots, G(B_k)$ is also random. When H is continuous, the probability of any two samples of H being equal is theoretically zero. However, the samples drawn from a

DP are always discrete as it is made up of countably infinite collection of point masses [48, 154, 163]. Consequently, two samples of a DP can be collided with a non-zero probability. The discreteness of the samples drawn from a DP facilitates the clustering in DP mixture models.

Let $B_1,\ldots,B_k$ be a finite measurable partition of Θ, and $\theta_1,\ldots,\theta_n$ be a set of independent draws from G. The number of draws θ_is ($i=1,2,\ldots,n$) belonging to the r^{th} ($r=1,2,\ldots,k$) partition is denoted by n_r. Then the posterior distribution of G is also a DP such that [163]

$$\big(G(B_1),\ldots,G(B_k)\big)|\theta_1,\ldots,\theta_n \sim \text{Dir}\big(\alpha H(B_1)+n_1,\ldots,\alpha H(B_k)+n_k\big).$$

The result shown in the above equation is true for all finite measurable partitions of Θ. Consequently, the posterior distribution of G is also a DP. Hence, DP is very important as a non-parametric conjugate family of prior over distributions. The posterior distribution of G can be shown to be

$$G|\theta_1,\ldots,\theta_n \sim \text{DP}\Big(\alpha+n, \frac{\alpha}{\alpha+n}H+\frac{n}{\alpha+n}\frac{\sum_{i=1}^{n}\delta_{\theta_i}}{n}\Big),$$

where δ_{θ_i} is the distribution concentrated at θ_i [128], and $\frac{\sum_{i=1}^{n}\delta_{\theta_i}}{n}$ is the empirical distribution. It is very clear that the posterior base distribution is the weighted average of the base and empirical distributions where the weights are proportional to the concentration parameter α and the number of observations n respectively. The posterior concentration parameter is simply calculated as the sum of α and n. Hence, the concentration parameter of prior distribution α implies the prior information in terms of number of observations. This is one of the important properties of conjugate family of distributions.

Let $\theta_1,\ldots,\theta_n$ be independently and identically distributed draws from G. Suppose that it is expected to draw a new observation θ_{n+1} after observing n draws from G. This is equivalent to drawing the new observation from the posterior of G. Let B be a measurable subset of Θ (i.e. $B\subset\Theta$), then the probability that the new observation θ_{n+1} belongs to B

after observing n draws, can be calculated as [163]

$$\begin{aligned}\Pr\big(\theta_{n+1} \in B|\theta_1,\ldots,\theta_n\big) &= \mathrm{E}\big[G(B)|\theta_1,\ldots,\theta_n\big] \\ &= \frac{1}{\alpha+n}\Big(\alpha H(B)+\sum_{i=1}^{n}\delta_{\theta_i}(B)\Big).\end{aligned}$$

This is corresponding to the posterior base distribution G that is conditional on the initial sequence of n observations. Consequently, the predictive distribution of θ_{n+1} becomes the posterior base distribution given $\theta_1,\ldots,\theta_n$ i.e.

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{\alpha+n}\Big(\alpha H+\sum_{i=1}^{n}\delta_{\theta_i}\Big).$$

The discreteness of the draws from a DP generates repeated values among the realisations. These repeated values are directly involved in creating clusters and clustering property of DP. The relationship between discreteness and clustering properties of DP facilitates clustering via DP mixture models. Let $\tilde{\theta}_1,\ldots,\tilde{\theta}_K$ be the set of distinct (unique) draws among $\theta_1,\ldots,\theta_n$. Accordingly, the predictive distribution of θ_{n+1} can be written in the form

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{\alpha+n}\Big(\alpha H+\sum_{k=1}^{K} n_k\delta_{\tilde{\theta}_k}\Big),$$

where n_k be the number of repeats of $\tilde{\theta}_k$ among the n $(n=\sum_{k=1}^{K} n_k)$ draws. Hence, the probability that $\tilde{\theta}_k$ will be drawn as θ_{n+1} is proportional to the number of times it has already been observed. Consequently, the clusters that have more observations than others grow faster. This is called "**rich-gets-richer**" phenomenon. The mean and variance of the number of clusters K among n observations are approximated as follows.

$$\begin{aligned}\mathrm{E}\big(K|n\big) &\simeq \alpha\log\Big(1+\frac{n}{\alpha}\Big) \qquad \text{for } n,\alpha \gg 0 \\ \mathrm{Var}\big(K|n\big) &\simeq \alpha\log\Big(1+\frac{n}{\alpha}\Big) \qquad \text{for } n > \alpha \gg 0\end{aligned}$$

7.6.2 Stick-breaking Construction

In 1994, Sethuraman [153] introduced the stick-breaking construction as a simple constructive process. This is obviously more straightforward and simple than the other representations and proofs of DPs [163]. In this construction, a probability stick of length one is sequentially breaking into two pieces randomly according to a Beta distribution. The construction is represented in Figure 7.3 and can be summarised into the following steps.

1. Draw a random observation β_1 from Beta(1, α) distribution. Then break the stick into two pieces being proportional to $\beta_1 : (1-\beta_1)$. Select the length of the first piece β_1 as the first probability weight π_1. Then $\pi_1 = \beta_1$.

2. To derive the next probability weight π_k, draw another random observation β_k from Beta(1, α) distribution.

3. Break the remaining part of the stick of length $\prod_{b=1}^{k-1}(1-\beta_b)$ into two pieces being proportional to $\beta_k : (1-\beta_k)$. Calculate π_k such that $\pi_k = \beta_k \prod_{b=1}^{k-1}(1-\beta_b)$.

4. Repeat step 2 and 3.

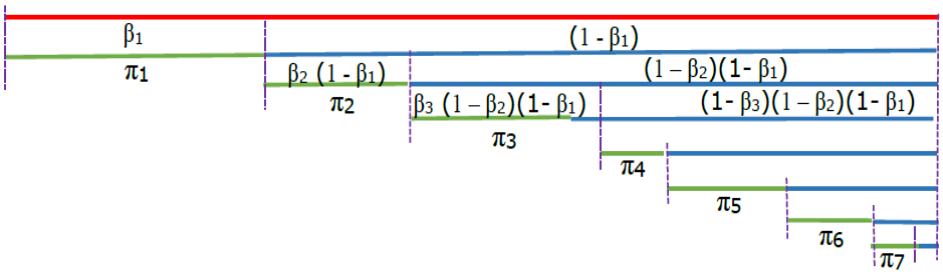

Figure 7.3: Stick-breaking construction

The vector of π_is needs not to be in the descending order of their magnitudes. Assuming a base distribution H, a concentration parameter α, and a sequence of probability weights $\pi = (\pi_1, \pi_2, \ldots)$, a Dirichlet process G can be expressed as below [154, 163].

$$\beta_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = \beta_k \prod_{b=1}^{k-1}(1-\beta_b)$$

$$\tilde{\theta}_k \sim H \qquad G \sim \sum_{k=1}^{\infty} \pi_k \delta_{\tilde{\theta}_k}$$

7.6.3 Pólya Urn Scheme

The Pólya urn construction is sometimes referred to as Blackwell-MacQueen scheme [163] to acknowledge the work of Blackwell and MacQueen [16] on this topic. The Pólya urn scheme represents draws from a Dirichlet process rather than the Dirichlet process itself [164]. The discreteness property of Pólya urn draws facilitates clustering property. In a Pólya urn scheme, the values of the parameter space Θ of the base distribution H are represented by a collection of balls with different colours. The parameter space Θ in relation to the urn problem is the set of all possible unique colours [154]. Suppose all the balls in the urn can be classified into K distinct colours $c_1, \ldots, c_K$ with frequencies $\alpha_1, \ldots, \alpha_K$ respectively. The balls are drawn randomly with equal probabilities and sampling with replacement. In addition, subsequently to each draw, a new ball with the same colour is added to the urn. Then the proportion of the balls in different colours in the urn G follows a Dirichlet distribution (i.e.$G \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$) in the limit of infinite number of draws from the urn.

In DP applications, the modelling of successive independent draws from G without a direct reference to it, is very important. In this approach, it uses an empty urn at the start and the subsequent steps are as below.

1. Draw the first colour θ_1 at random from the base distribution H and add a ball of that colour to the empty urn.

2. Suppose the number of balls in the urn is denoted by n. Draw a colour θ_{n+1} from the base distribution H with probability $\frac{\alpha}{\alpha+n}$ or draw a colour (ball) from the urn with probability $\frac{n}{\alpha+n}$.

3. Increase the number of balls in the urn by adding a new ball of the colour θ_{n+1}.

4. Repeat step 2 and 3.

According to the Pólya urn scheme, the predictive distribution of θ_{n+1} after observing $\theta_1, \ldots, \theta_n$ is derived as

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha+n}\Big(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}\Big).$$

The clustering property in the Pólya urn scheme can be represented by assuming $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ distinct colours in the urn before drawing θ_{n+1}. When n_k represents the number of balls of the colour $\tilde{\theta}_k$ in the urn, the predictive distribution of θ_{n+1} can be written in the form

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\Big(\alpha H + \sum_{k=1}^{K} n_k \delta_{\tilde{\theta}_k}\Big).$$

7.6.4 Chinese Restaurant Process

In 1978, Pitman [139] published an article named "An extension of de Finetti's theorem". Following the findings of this study with some additional contributions, in 1985, Aldous [3] named the Chinese restaurant process (CRP) as a distribution over partitions. The process is closely related to the Pólya Urn Scheme as a distribution over partitions.

CRP is described with a hypothetical restaurant with an infinite number of tables and unbounded number of seats for each table [154, 163, 164]. Each customer enters the restaurant one after the other and decides a table based on the number of customers already seated at each non-empty table. The steps associated with CRP can be summarised as below.

1. The first customer enters the restaurant and sits at Table 1.

2. Suppose the total number of customers already occupied are denoted by n and number of non-empty tables are labelled as Table 1, ..., Table K. The number of customers seated at Table k ($k = 1, \ldots, K$), is denoted by n_k.

3. When the customer $n+1$ arrives, he selects the Table k with probability $\frac{n_k}{\alpha+n}$ or a new table (say Table $K+1$) with probability $\frac{\alpha}{\alpha+n}$.

4. Repeat step 2 and 3 for all the customers.

CRP clearly illustrates the clustering property of the Dirichlet process. Each table in a CRP represents a cluster, and the customers at the table represent the observations belonging to that cluster. It assumes round tables to represent permutations of the observations within the cluster [163]. Conceptually, CRP assumes infinite number of clusters; however, use only a finite number of them to partition the observed data. The mixing

proportion related to any component in a mixture model tends to be zero as the number of components tends to infinity. However, combining all the empty partitions into a single component, the number of components can be fixed to a finite number. In a CRP, the total number of clusters cannot be greater than the number of observations at hand as n observations can be assigned at most n clusters. However, it can grow as the number of observation increases.

7.6.5 Pitman-Yor Process

The Pitman-Yor process [138] is a two parameter (α and β) generalisation of the CRP and has the rich-gets-richer property. In this process the $(n+1)^{\text{th}}$ customer selects the Table k (where $k = 1, \ldots, K$) with probability $\frac{n_k - \beta}{n+\alpha}$ and a new table (say Table $K+1$) with probability $\frac{\alpha+\beta K}{n+\alpha}$. As it noted in CRP, n_k and α represent the number of customers already occupied at Table k and the concentration parameter respectively. β is referred to as the discount parameter. Contrasting to the CRP, in a Pitman-Yor process, the probability of introducing a new table increases as the number of tables increases.

7.7 Collapsed Gibbs Sampling with CRP for Better Models

This section describes the procedure of fitting an infinite mixture model for stutter ratio (SR). Assuming Gaussian errors, a simple linear regression of longest uninterrupted sequence (LUS) is used to model SR. The theoretical aspects of Bayesian model fitting in the context of linear regression model, assuming a fully conjugate prior distribution has been discussed in Chapter 6. CRP was employed as a non-parametric (Dirichlet) prior to fit an infinite mixture of regression models for predicting SR. The CRP is selected as the representation of DP due to its simplicity in calculations and understanding. Its property of being the predictive probability proportional to the number of observations in the cluster is statistically attractive. However, CRP has been paid little attention in applications than other DP representations. Therefore, this study attempts to evaluate the behaviour of CRP in order to search the space of models, rather than drawing posterior samples. Mod-

Table 7.2: Notations used in the algorithm

Notation	Description
NN	Total number of observations
XX	Covariate matrix ($NN \times 2$)
YY	Column vector of SR
IGa	Shape parameter of the inverse gamma prior
IGb	Scale parameter of the inverse gamma prior
$IGATilda$	Posterior shape parameter of the inverse gamma
$IGBTilda$	Posterior scale parameter of the inverse gamma
$MuBeta$	Prior mean vector of β (2×1)
$ScaleBeta$	Prior scale matrix of β (2×2)
$MuBetaTilda$	Posterior mean vector of β (2×1)
$ScaleBetaTilda$	Posterior scale matrix of β (2×2)
KK	Number of active mixture components
$Alpha$	Concentration parameter of the Dirichlet prior
ZZ	A vector of length NN that keeps cluster indicators of all the observations
nn	A vector of length KK that keeps the number of observations in each cluster
ii	A vector of length NN that keeps an indicator ($1 : NN$) for each datum

ifications motivated by "data cloning" can focus the search on higher likelihood models. In this study, infinite mixture models are fitted only to the D2S1338 locus of the NGM SElect^{TM} dataset. The notation that has been used in the algorithm is given in Table 7.2. The study acknowledges the online materials and MATLAB code of Yee Whye Teh [161]. The steps related to the computation can be summarised as below.

Step 1: Read the observed SR into YY and corresponding LUS values into the second column of XX. All the elements of the first column are ones. Calculate NN as the length of YY vector. Set an initial value for the number of active mixture components KK and concentration parameter of the Dirichlet prior $Alpha$.

Step 2: Set the appropriate prior information for the normal-inverse gamma prior (IGa, IGb, $MuBeta$, $ScaleBeta$).

Step 3: Generate $KK+1$ number of empty clusters. Assign all the observations into the first KK clusters randomly. Keep a track of cluster indicators in ZZ. The component

$KK+1$ takes care of all the empty clusters.

Step 4: Remove each datum from the current model, then add it back in according to the conditional probability.

4a: Remove datum *ii* from the appropriate cluster.

– If there are any other observations in the cluster:

- reduce the number of observations in the particular cluster by one.
- update the posterior parameters: $IGATilda$, $IGBTilda$, $MuBetaTilda$, and $ScaleBetaTilda$ of the cluster.

– If there is no observation in the cluster, delete it and:

- reduce the number of active clusters by one.
- adjust the cluster indicators ZZ and the number of observations in each cluster *nn* appropriately.

4b: Add datum *ii* to an appropriate cluster according to the conditional probability.

- calculate the likelihoods (posterior predictive densities) of the datum *ii* under each of the active K clusters.
- calculate the likelihood (prior predictive density) of the datum *ii* under the empty $(K+1)$ cluster.
- calculate the conditional probability p_k $(k = 1, \ldots, K+1)$ of the datum *ii* fitting to cluster k, such that p_k is proportional to the number of observations nn_k and the likelihood (use $nn_{K+1} = \alpha$).

4c: Assign datum *ii* to cluster k with probability p_k.

– If there are any other observations in the selected cluster:

- increase the number of observations in the cluster by one.
- adjust the cluster indicator of the datum appropriately.
- update the posterior parameters: $IGATilda$, $IGBTilda$, $MuBetaTilda$, and $ScaleBetaTilda$ of the cluster.

– If it is a new cluster:

- set the previous empty cluster ($K+1$) into the new active cluster.
- set the cluster indicator and the number of observations of this cluster into $K+1$ and 1 respectively.
- update the posterior parameters: $IGATilda$, $IGBTilda$, $MuBetaTilda$, and $ScaleBetaTilda$ of the cluster.
- increase the number of active clusters by one introducing a new empty cluster .

4d: Repeat steps 4a, 4b, and 4c for all the observations in the dataset.

Step 5: Repeat step 4 until the required number of iterations (5000 for this study) is achieved.

7.7.1 Selection of Prior Parameters

The selection of the parameters of normal inverse gamma prior distribution is a key problem in this application. The calculation of these parameters based on a small random sample of observation from the dataset is a convenient option, and it was labelled as **prior A**. However, the number of observations (n_0) in this sample is directly related with the two parameters (a, b) of the inverse gamma prior distribution. When a multiple linear regression model with $k-1$ predictors is considered, the two parameters are estimated in the following way, where MSE is the mean squared error of the regression model fitted to the selected sample.

$$a = \frac{n_0 - k}{2}$$
$$b = \frac{n_0 - k}{2} MSE$$

This study selected five different random samples of size five ($n_0 = 5$) from the dataset of size 406 to evaluate the behaviour of CRP across different samples. As LUS is the only predictor used in this application, $k = 2$, hence the quantity $\frac{n_0-k}{2}$ equals to 1.5. This factor was reduced to 0.1 to minimise the effect of prior sample on the performance of collapsed Gibbs sampling with CRP, and used as the second option (**prior B**) of deriving the prior parameters. The performance of this modification was also examined based

on the same five samples. In the third option (**prior C**), in addition to the factor 0.1, the sample of five observations was replaced with the full dataset. The mean vector μ_β and the unscaled covariance matrix V_β were also calculated based on the selected sample under each method.

7.7.2 Initial Allocation of Clusters

The Gibbs sampling can be initialised with any number of clusters that is not more than the size of the dataset. This study used two settings, one with a single starting cluster and the other with ten initial clusters. The one that starts with a single cluster was labelled as **forward** method and the other as **backward** method. These two methods were separately applied with each of the three methods proposed for prior parameter estimation. The observations in the dataset were randomly allocated to these clusters prior to the Gibbs sampling. As there are a massive number of deletions of existing clusters and creations of new clusters, the Gibbs sampler like this can be regarded as a **birth-and-death** process of clusters.

7.7.3 Concentration Parameter (α)

The effect of CRP depends on the concentration parameter α, which controls the number of active clusters in the mixture . As the effect of CRP in collapsed Gibbs sampling cannot be predicted, this study repeated each combination of prior selection (prior A, prior B, and prior C) and the initial cluster allocation (forward and backward) for 20 different values of α ($10^{-15}, 10^{-14}, \ldots, 10^{4}$).

7.7.4 Improvements made to Collapsed Gibbs Sampling Algorithm for better Models

In Gibbs sampling, each datum is allocated to either an existing cluster or a new cluster based on the respective conditional probabilities. The conditional probability of a given observation corresponding to a given cluster is proportional to the number of observations and the likelihood (posterior predictive density) of the observation. When the number

of observations and the likelihood of a new observation belonged to the cluster k of K clusters are denoted by n_k and p_k, then the conditional probability of the observation fitting to cluster k, P_k, is calculated as below.

$$P_k \propto n_k p_k \quad ;k = 1, \ldots, K \tag{7.1}$$

For a given observation, a new cluster is created based on the probability P_{K+1}, which is proportional to the product of the concentration parameter α and the likelihood (prior predictive density) of the observation p_{K+1}. The new cluster is represented by the cluster indicator $(K+1)$. Therefore,

$$P_{K+1} \propto \alpha p_{K+1}$$

Hence the concentration parameter α can be considered as the number of observations in the $(K+1)^{th}$ cluster, which is actually empty.

Since the conditional probabilities in the original collapsed Gibbs sampling are proportional to the size of the cluster (n_k), large clusters get larger in many occasions. This is called the 'rich-gets-richer' phenomenon. When there are two or more clusters with approximately similar likelihoods in relation to a given observation, being selected the one with more observations is a good statistical property in clustering. However, a large cluster even with a small likelihood can be favourable than the other small clusters with relatively high likelihoods. This characteristic can desperately affect the performance of clustering. Hence, this study proposes to increase the relative priority of the likelihoods in calculating conditional probabilities.

When the clusters are overlapping or not far apart from each other, there may be at least few clusters with relatively large likelihoods for a given observation. Under these circumstances, the cluster with the highest likelihood for the observation may have a smaller conditional probability of being selected than not being selected. The same situation can be expected with relatively large number of clusters. If our goal is finding high likelihood models, this is a shortfall of the original version of the collapsed Gibbs sampling algorithm with CRP. Therefore, this study proposes to use either the second or third power of

these likelihoods instead of the original value. This is not a completely new idea as it has already been adopted in data cloning which utilizes both Bayesian framework and MCMC computational methods. In this method, it assumes the likelihood of k independent copies of data instead of the likelihood of the observed data themselves [113]. When the conditional probabilities are calculated under collapsed Gibbs sampling, the use of second power of the likelihood corresponds to an additional independent copy of the observation. Similarly, the third power corresponds to two additional copies of the observation. This modification may be really helpful in isolating the cluster with the highest likelihood for a given observation, from the rest of the clusters. Considering this suggestion, Equation 7.1 is revised as

$$P_k \propto n_k p_k^d \quad ; k = 1, \ldots, K+1,\ n_{K+1} = \alpha, \text{ and } d = 1,2,3. \tag{7.2}$$

7.7.5 Computational Limitations

Prior A and prior B were evaluated on five different samples, hence there were 11 possibilities (levels) of prior. When each possibility was considered with two settings of cluster allocations and 20 different values of α, it made 440 ($11 \times 2 \times 20$) testing conditions. This study evaluated the performance of collapsed Gibbs sampling with CRP for better models at each level of d, hence there were 1320 (440×3) testing conditions altogether. Since this required a massive amount of computational resources to test the performance under the above conditions, this study used the stutter information of only one locus (D2S1338) of the NGM SElect^{TM} dataset that has the largest number (406) of stutter peaks.

7.8 Results and Discussion

Posterior samples were drawn using collapsed Gibbs sampling, considering each combination of prior selection (11 levels), cluster initialisation (2 levels), and concentration parameter α (20 levels). After 3000 burn-in steps, each sampler was run for another 2000 iterations. Using a thinning interval of four, 500 posterior draws were recorded under each testing condition. Even though the performance of collapsed Gibbs sampling with

Table 7.3: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-15}$

	$P_k \propto n_k p_k$			$P_k \propto n_k p_k^2$			$P_k \propto n_k p_k^3$		
Condition	NC	N	LL	NC	N	LL	NC	N	LL
AF1	1	500	1230	1	500	1230	4	500	1629
AF2	1	500	1229	1	500	1229	4	500	1633
AF3	1	500	1228	1	500	1228	4	500	1623
AF4	1	500	1233	1	500	1233	3	500	1487
AF5	1	500	1231	1	500	1231	4	500	1622
AB1	2	500	1362	4	500	1584	5	500	1669
AB2	2	500	1367	4	500	1596	5	500	1724
AB3	2	500	1366	5	500	1682	5	500	1706
AB4	2	500	1351	3	500	1474	4	500	1584
AB5	2	500	1358	4	500	1569	4	500	1622
BF1	1	500	1245	1	500	1245	4	500	1658
BF2	1	500	1245	1	500	1245	4	500	1657
BF3	1	500	1245	1	500	1245	4	500	1678
BF4	1	500	1245	1	500	1245	4	500	1656
BF5	1	500	1245	1	500	1245	4	500	1656
BB1	2	500	1382	4	500	1621	6	500	1807
BB2	2	500	1383	5	500	1712	6	500	1785
BB3	2	500	1384	6	500	1766	7	500	1848
BB4	2	500	1378	4	500	1631	6	500	1783
BB5	2	500	1381	3	500	1632	6	500	1797
CF	1	500	1245	1	500	1245	2	500	1384
CB	1	500	1245	2	500	1364	3	500	1435

CRP for better models was tested for 20 different values of the concentration parameter α, the results corresponding to a set of selected α values are presented.

The following notations have been used in summarising the results. A testing condition is labelled as **Combination** which is represented by two letters and a number. The first letter (A, B, or C) is used to denote the prior method (prior A, prior B, or prior C respectively). The initial cluster allocation: "forward" (starting with a single cluster) or "backward" (starting with ten clusters) is represented by the second letter (F or B respectively). A number between 1 and 5 is used (only with prior A and prior B) as the third symbol to indicate the sample that has been used to derive the prior parameters. The notation **NC** is used to denote the number of active (non-empty) clusters in the infinite mixture model. This can be varied even within a single testing condition. Hence, the number of MCMC samples out of 500 posterior draws that is compatible with the given NC value is represented by **N**. The average log-likelihood calculated over the posterior draws is represented by **LL**.

Table 7.3 to 7.7 present the variations in the performance of collapsed Gibbs sam-

Table 7.4: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-11}$

	$P_k \propto n_k p_k$			$P_k \propto n_k p_k^2$			$P_k \propto n_k p_k^3$		
Condition	NC	N	LL	NC	N	LL	NC	N	LL
AF1	1	500	1230	3	500	1485	4	500	1629
AF2	1	500	1229	1	500	1229	4	500	1633
AF3	1	500	1228	3	500	1488	4	500	1644
AF4	1	500	1233	3	500	1474	3	500	1487
AF5	1	500	1231	3	500	1480	4	500	1622
AB1	2	500	1362	4	500	1584	5	500	1669
AB2	2	499	1367	4	500	1596	5	129	1724
	3	1	1441				6	371	1743
AB3	2	500	1366	5	500	1682	5	500	1706
AB4	2	499	1351	3	500	1474	4	500	1584
	3	1	1384						
AB5	2	499	1358	4	500	1569	4	500	1622
	3	1	1390						
BF1	1	500	1245	4	500	1637	4	500	1658
BF2	1	500	1245	4	500	1632	4	500	1657
BF3	1	500	1245	3	500	1506	4	500	1685
BF4	1	500	1245	3	500	1497	4	500	1656
BF5	1	500	1245	3	500	1503	4	500	1656
BB1	2	499	1382	5	500	1693	6	500	1807
	3	1	1438						
BB2	2	498	1383	5	435	1708	6	500	1785
	3	2	1431	6	65	1768			
BB3	2	500	1384	6	500	1766	7	500	1848
BB4	2	493	1378	4	500	1631	6	500	1783
	3	7	1437						
BB5	2	499	1382	5	500	1690	6	500	1797
	3	1	1419						
CF	1	500	1245	2	500	1365	3	500	1435
CB	1	500	1245	2	500	1364	3	500	1435

pling with CRP, under different combinations of prior selection methods, initial cluster allocations, and concentration parameter values. In addition, the consistency of prior A and prior B across various samples was examined. In each table, the effects of the second and third powers of the likelihoods, which have been used in calculating the conditional probabilities, on the performance are compared. The additional information (cluster sizes and the standard deviations of log-likelihoods) relevant to the results given in Table 7.3 to 7.7 are presented in Appendix B.

The overall picture of the results reveals an increasing trend in the variation of the number of clusters as the number of clusters increases. In addition, the study frequently observed small clusters (containing few observations) as the number of clusters increases (Appendix B). The effect of prior sample used to estimate the parameters of prior A and B was evaluated for five different random samples. The same five samples were used in all

Table 7.5: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-8}$

	$P_k \propto n_k p_k$			$P_k \propto n_k p_k^2$			$P_k \propto n_k p_k^3$		
Condition	NC	N	LL	NC	N	LL	NC	N	LL
AF1	1	500	1230	3	500	1485	4	500	1629
AF2	1	500	1229	3	139	1488	4	500	1633
				4	361	1580			
AF3	1	500	1228	3	500	1488	4	500	1644
AF4	1	495	1233	3	500	1474	3	500	1487
	2	5	1249						
AF5	1	500	1231	3	500	1480	4	500	1604
AB1	2	161	1361	4	500	1584	5	500	1669
	3	339	1418						
AB2	2	189	1366	4	500	1611	6	500	1743
	3	311	1422						
AB3	2	322	1366	5	500	1682	5	500	1706
	3	178	1423						
AB4	2	435	1351	3	500	1474	4	500	1584
	3	65	1405						
AB5	2	323	1358	4	500	1569	4	500	1622
	3	177	1411						
BF1	1	500	1245	4	500	1637	4	500	1658
BF2	1	500	1245	4	500	1632	4	500	1657
BF3	1	500	1245	3	500	1506	4	87	1684
							5	413	1761
BF4	1	500	1245	4	500	1630	5	500	1714
BF5	1	500	1245	4	500	1632	4	500	1656
BB1	2	56	1386	5	500	1693	6	500	1807
	3	444	1441						
BB2	2	82	1380	5	1	1738	6	500	1785
	3	418	1439	6	499	1768			
BB3	2	351	1384	6	500	1766	7	500	1848
	3	149	1439						
BB4	2	278	1376	5	500	1680	6	500	1783
	3	222	1433						
BB5	2	144	1380	5	500	1690	6	500	1797
	3	356	1437						
CF	1	500	1245	2	497	1364	3	500	1435
				3	3	1372			
CB	1	500	1245	2	489	1365	3	500	1435
				3	11	1380			

the occasions. The results reveal a moderate variation in the number of active clusters due to the differences of these samples. The variation is clearly visible with backward method of initial cluster allocation, especially with smaller values of the concentration parameter.

A clear difference between the forward and backward methods of initial cluster allocation in relation to the performance of collapse Gibbs sampling can be observed for both prior A and B irrespective of the value of the concentration parameter. When the Gibbs sampler starts with ten initial clusters, the sampler converges into more clusters than when it starts with one cluster. In contrast, prior C gives almost similar convergence

Table 7.6: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-6}$

	$P_k \propto n_k p_k$			$P_k \propto n_k p_k^2$			$P_k \propto n_k p_k^3$		
Condition	NC	N	LL	NC	N	LL	NC	N	LL
AF1	1	355	1230	4	500	1603	4	500	1629
	2	9	1279						
	3	136	1419						
AF2	1	500	1229	4	141	1582	4	500	1633
				5	359	1635			
AF3	1	500	1228	3	249	1488	5	500	1718
				4	251	1622			
AF4	1	491	1233	3	500	1474	3	500	1487
	2	9	1249						
AF5	2	13	1356	3	296	1480	4	500	1604
	3	487	1411	4	204	1589			
AB1	3	500	1417	4	500	1584	5	500	1669
AB2	2	1	1381	4	500	1611	6	500	1743
	3	499	1422						
AB3	2	4	1365	5	500	1668	5	500	1706
	3	496	1422						
AB4	2	28	1348	3	500	1474	4	500	1584
	3	472	1403						
AB5	2	13	1356	4	500	1569	4	500	1622
	3	487	1411						
BF1	1	499	1245	4	500	1637	4	500	1658
	2	1	1259						
BF2	1	500	1245	4	500	1632	4	500	1657
BF3	1	497	1245	3	500	1506	5	500	1764
	2	3	1259						
BF4	1	500	1245	4	500	1630	5	500	1714
BF5	1	499	1245	4	500	1632	5	500	1718
	2	1	1259						
BB1	3	500	1440	5	500	1693	6	500	1773
BB2	2	10	1383	6	500	1768	6	500	1785
	3	499	1439						
BB3	2	13	1384	6	500	1766	7	500	1848
	3	487	1441						
BB4	2	4	1380	5	500	1680	6	500	1783
	3	496	1433						
BB5	2	3	1370	5	500	1689	6	500	1797
	3	497	1436						
CF	1	499	1245	2	321	1365	3	497	1435
		1	1259	3	179	1374	4	3	1442
CB	1	500	1245	2	241	1364	3	499	1435
				3	259	1384	4	1	1447

for both methods of initial cluster allocation except with the lowest value of concentration parameter considered in this study (Table 7.3).

The number of observations in the prior sample is directly represented in the prior distribution under prior A. However as previously mentioned, prior B and C were defined by minimising this effect. As expected, prior B produced more stable results in terms of the number of clusters than prior A. In addition, the number of active clusters that resulted

Table 7.7: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-2}$

	$P_k \propto n_k p_k$			$P_k \propto n_k p_k^2$			$P_k \propto n_k p_k^3$		
Condition	NC	N	LL	NC	N	LL	NC	N	LL
AF1	3	481	1417	4	337	1571	5	215	1657
	4	19	1422	5	158	1585	6	283	1691
				6	5	1581	7	2	1686
AF2	3	480	1422	4	293	1611	5	497	1681
	4	19	1425	5	205	1627	6	3	1680
	5	1	1411	6	2	1620			
AF3	3	484	1422	5	491	1683	6	445	1767
	4	16	1426	6	8	1684	7	54	1787
				7	1	1684	8	1	1786
AF4	3	495	1401	3	479	1474	4	442	1584
	4	5	1401	4	21	1477	5	58	1586
AF5	3	492	1411	4	475	1568	4	177	1622
	4	8	1414	5	24	1573	5	321	1633
				6	1	1583	6	2	1638
AB1	3	481	1417	4	17	1584	5	44	1673
	4	19	1422	5	474	1608	6	453	1692
				6	9	1610	7	3	1686
AB2	3	476	1420	5	495	1637	6	500	1743
	4	24	1421	6	5	1644			
AB3	3	484	1422	5	475	1681	6	484	1734
	4	16	1426	6	24	1684	7	16	1740
				7	1	1694			
AB4	3	495	1403	3	477	1474	4	436	1584
	4	5	1408	4	23	1476	5	64	1588
AB5	3	492	1411	4	475	1568	4	158	1622
	4	8	1414	5	24	1573	5	340	1633
				6	1	1583	6	2	1630
BF1	3	496	1440	6	500	1768	6	499	1780
	4	4	1446				7	1	1773
BF2	3	497	1438	5	497	1681	6	500	1807
	4	3	1443	6	3	1675			
BF3	3	496	1441	6	496	1795	7	498	1843
	4	4	1440	7	4	1798	8	2	1845
BF4	3	496	1433	4	498	1630	6	499	1783
	4	4	1436	5	2	1630	7	1	1788
BF5	3	497	1436	5	498	1690	5	499	1718
	4	3	1447	6	2	1694	6	1	1725
BB1	3	493	1440	5	499	1692	6	495	1780
	4	7	1435	6	1	1716	7	5	1793
BB2	3	497	1439	6	499	1769	7	500	1827
	4	3	1442	7	1	1776			
BB3	3	496	1441	6	497	1795	7	500	1848
	4	4	1440	7	3	1799			
BB4	3	496	1433	5	498	1680	6	500	1783
	4	4	1436	6	2	1675			
BB5	3	496	1436	5	498	1689	6	499	1797
	4	4	1435	6	2	1691	7	1	1810
CF	1	18	1245	3	372	1383	3	23	1434
	2	478	1259	4	128	1399	4	477	1438
	3	4	1262						
CB	1	22	1245	3	458	1380	3	27	1435
	2	476	1259	4	42	1402	4	473	1438
	3	2	1262						

under prior B is slightly higher than that of under prior A, on average. Prior C, unlike prior B, is defined based on the overall estimates of the dataset. When prior C is compared with prior A and B, it always converged to a relatively low number of clusters.

The use of the second and third powers of likelihoods in calculating the conditional probabilities makes a significant change in the models explored by collapsed Gibbs sampling with CRP. The number of active clusters increases with the increase in the order of likelihoods. The variation in the number of active clusters even for a fixed set of values of: the concentration parameter, initial cluster allocation, prior, and the power of likelihood, is a key feature of the results. This variation increases with the increase of concentration parameter as they produce more clusters than with smaller values. Even though the use of higher orders of likelihoods results in a large number of active clusters, the variation of them decreases.

The behaviour of the number of active clusters along with their variations according to the sample selection for estimating the prior parameters, forward and backward methods of initial cluster allocation, use of different prior distributions, and employing different powers of likelihoods have been discussed so far. A performance evaluation of the collapsed Gibbs sampling with CRP in the context of log-likelihood with respect to these factors is vital. The log-likelihoods presented in Tables 7.3 to 7.7 are combined over different values of concentration parameter and five samples used under prior A and B, and are displayed in Table 7.8. The number of MCMC samples (**N**) used as the weights in calculating the weighted averages of log-likelihoods presented in this table.

Even though there are some differences in the number of active clusters due to the cluster allocation method, they both produce approximately similar log-likelihoods for a given combination of: prior, the power of likelihood, and the number of active clusters. However, there are three cases (highlighted in the table) that produced considerably large differences in the log-likelihoods. Prior C gives almost identical log-likelihoods under both methods of initial cluster allocation than prior A and B.

The results presented in Table 7.8 clearly indicate an increasing trend in the magnitude of log-likelihoods as the power of likelihood increases for all the three prior distributions. Furthermore, a gradual decrease in the standard deviations (SD) of log-likelihoods is evi-

Table 7.8: The variations in the performance of collapsed Gibbs sampling with CRP in terms of log-likelihoods

Prior	Power	Cluster Allocation	Number of Active Clusters 1	2	3	4	5	6	7	8
A	1	Forward	1230	1295	1414	1421	1411			
		Backward	1231	1360	1416	1421				
	2	Forward	1230		1480	1589	1645	1637	1684	
		Backward			1474	1584	1663	1659	1694	
	3	Forward			1487	1625	1676	1737	1783	1786
		Backward				1602	1687	1733	1731	
B	1	Forward	1245	1259	1438	1442				
		Backward		1381	1438	1437				
	2	Forward	1245		1504	1633	1686	1781	1798	
		Backward			1632	1627	1691	1770	1793	
	3	Forward				1660	1731	1790	1843	1845
		Backward						1790	1884	
C	1	Forward	1245	1259	1262					
		Backward	1245	1259	1262					
	2	Forward	1245	1365	1380	1399				
		Backward		1364	1381	1402				
	3	Forward		1384	1435	1438				
		Backward			1435	1438				

dent as the power of likelihood increases (Appendix B). Even though the study did not observe a large number of active clusters under prior C, a lack of gain in the log-likelihoods is clearly visible even with fewer numbers of active clusters. Therefore, the estimation of prior parameters based on the whole dataset can be clearly ruled out by comparing with the performance of prior A and B. It is obvious to expect an approximate difference of 20 in the log-likelihoods under prior B over A as prior B considers 406 observations against 401 under prior A. However, the results reveal a larger improvement than expected in the log-likelihoods. Hence, the use of prior B with the third power of likelihood can be recommended as the best method in terms of the gain in log-likelihoods. The study observed a substantial reduction in the overall machine time with the backward method of initial cluster allocation than the forward method. Hence, this study further recommends initialising Gibbs sampler with more clusters than with a single cluster.

In a finite mixture model, the number of components is fixed. An infinite mixture model, in contrast, does not assume a fixed number, and rather changes during the Gibbs sampler as it consists of a birth-and-death mechanism of components. However, the number of components is always less than or equal to the size of the dataset. Moreover, it

converges to a finite number or a set of numbers. In an infinite mixture model, the number of components is not controlled directly. However, the number or numbers are greatly dependent on the value of the concentration parameter. Hence, the selection of a suitable value for the concentration parameter is the ultimate technique for controlling the number of active clusters in an infinite mixture model.

As discussed in section 7.5.1, BIC is the most widely used method in determining the number of components in a finite mixture model. Therefore, BIC is an adequate criterion even for an infinite mixture model to decide the number of components in the final model. In some cases, this study revealed a set of consecutive numbers (e.g. 5, 6, and 7) as the active number of components (clusters). In these situations, the number of parameters in the model is not certain as the number of clusters is not fixed. According to the results obtained in the study (see Appendix B), in majority of the situations, there are only a few observations in the additional clusters (e.g. 6th and 7th clusters), hence they can be simply treated as tiny additional clusters. Therefore, practically, larger differences in the models cannot be observed due to varying number of clusters. In such situations, the cluster that produced the largest fraction of MCMC samples could be a practically convenient and reasonably accurate criterion to be used in deciding the best number of clusters among them. When there are considerably larger differences among the models due to their number of clusters, and if there is a model that has produced a fraction of MCMC samples that is close to unity, then it can be selected as the best number of clusters for the model. However, when there are larger differences among the models due to number of clusters, and in the presence of two or more clusters with relatively larger fractions of MCMC samples, the best model can be selected based on the BIC (likelihood).

The optimal number of active clusters is decided by minimising the value of BIC, which is defined as $-2\text{log-likelihood} + k\ln(n)$, where k and n are the number of parameters in the final model and the size of the dataset. Assume that the collapsed Gibbs sampler does not have multiple convergence of active number clusters for any value of the concentration parameter. The minimum log-likelihood gain required for every additional cluster

(Δ_{ll}) can be calculated as

$$\Delta_{ll} = \frac{1}{2}k\ln(n).$$

Every cluster consists of seven more parameters of bivariate normal inverse gamma joint distribution of the slope and intercept parameters of the mean model $\beta^T = (\beta_0, \beta_1)$ and error variance σ^2: mean vector (μ_β) (2 parameters), variance-covariance matrix (V_β) (3 independent parameters), shape and scale parameters (a, b) (2 parameters). Therefore, the introduction of every additional cluster increases the model complexity by an additional eight parameters including the mixing proportion. As the locus D2S1338 of the NGM SElect^{TM} dataset consists of 406 stutter peaks, $n = 406$ and $k = 8$. Therefore, Δ_{ll} becomes

$$\Delta_{ll} = \frac{1}{2}8\ln(406) \simeq 24.$$

Considering the results presented in Table 7.8 and the value of Δ_{ll}, it is recommended to use a seven-component mixture model for predicting stutter at this locus. However, an evaluation of the performance of collapsed Gibbs sampling with CRP beyond seven clusters and extending the model investigation into other loci will be required before making a final decision on the concentration parameter (number of components).

7.8.1 Performance of Infinite Mixture Models compared to Previously selected Non-hierarchical Models

The log-likelihoods of the locus-specific variance normal ($N1$) and the two-component normal mixture models for the D2S1338 locus in NGM SElect^{TM} dataset are 1245 and 1247 respectively. The infinite mixture models corresponding to these two models are the ones with one and two clusters respectively. The infinite mixture model with two clusters yielded relatively larger log-likelihoods under prior B than the two-component mixture model (Table 7.8). Therefore better performance of the infinite mixture model against two-component mixture model is obvious. The other models with more than two clusters also yielded larger log-likelihood gains even after adjusting for model complexity.

Therefore, based on the findings in terms of log-likelihoods (or BIC), infinite mixture models up to seven clusters clearly demonstrate a potential to generate better predictions.

7.9 Summary

Initially, this chapter explains the importance of investigating infinite mixture models for predicting stutter. These models have several advantages in terms of robustness, flexibility and increased capability of dealing with non-normality and heteroscedasticity issues in traditional statistical models. A Dirichlet process (DP) is a stochastic process that enables placing a distribution over distributions, and different representations of DPs are: the Stick- breaking construction, the Pólya (Blackwell-MacQueen) urn scheme, the Pitman-Yor process, and the Chinese restaurant process (CRP). This study uses an algorithm based on the collapsed Gibbs sampling that uses CRP as a non-parametric DP prior, for fitting an infinite mixture of simple linear regression models for SR using LUS as the predictor. In addition, the study proposes and illustrates the use of second and third powers of the likelihoods in calculating the conditional probabilities of Gibbs sampling. In this study, the performance of collapsed Gibbs sampling with CRP has been varied based on several factors including the prior samples (five fixed random samples) used to estimate prior parameters, different priors (A, B, and C that used three methods in estimating prior parameters), initial cluster allocation (forward and backward - staring with a single cluster and ten clusters respectively), and the power of likelihood in calculating conditional probabilities of Gibbs sampling (1 - the original version of collapsed Gibbs sampling with CRP, 2 and 3 - the proposed versions of of collapsed Gibbs sampling with CRP). In addition, the effect of the concentration parameter (α) was tested at 20 different values. The performance of collapsed Gibbs sampler with CRP was tested in terms of the variation in the number of active clusters and the log-likelihoods of the data.

As expected, an increasing trend in the number of clusters was observed in CRP as the value of α increases. In addition, there was an overall increasing trend in the variation of the number of active clusters as the number of clusters increases. A moderate variation in the number of active clusters due to different initial samples of prior parameter esti-

mation was evident for the backward method of initial cluster allocation, especially with smaller α. When the Gibbs sampler initialised with ten clusters, it converged to more clusters than initialising with a single cluster, under prior A and B but not under C. In terms of the number of clusters, prior B yielded more stable results than prior A. Prior C resulted in smaller number of clusters than A and B. When the third power of likelihoods was used in calculating the conditional probabilities of Gibbs sampling, the results revealed more stability in the number of active clusters even with large number of active clusters. The log-likelihoods were clearly increased with the order of the likelihood used in calculating the conditional probabilities. Prior B revealed the biggest improvement in log-likelihoods with the third power of likelihood, and the Gibbs sampler used a fairly longer machine time with an initialisation of a single cluster. Therefore, the study recommends using: prior B, the third power of likelihood, and initialisation of the Gibbs sampler with more clusters. Based on the results obtained, a seven-component mixture can be selected (in terms of log-likelihoods) as the best option for improving stutter prediction at the D2S1338 locus of the NGM SElect^{TM} dataset. Some important suggestions for future research based on possible extensions of this study, are explained in Chapter 8.

Chapter 8

Conclusions and Future Work

8.1 Introduction

The objective of this chapter is to summarise the findings of the previous chapters and introduce the possible directions for future work related to the study. This research re-examines existing works on modelling stutter ratio (*SR*) and develops new advanced Bayesian models to improve the accuracy of stutter prediction. In addition, this study contributes to the field of statistics by exploring important theoretical aspects related to the fitting and evaluation of Bayesian models of various types (hierarchical, non-hierarchical, and mixture). It also makes significant advances in the field of infinite mixture models that use collapsed Gibbs sampling with Chinese restaurant process, a representation of the Dirichlet process, as a non-parametric prior.

Forensic DNA analysis is an extremely valuable human identification technique used in criminal investigations. Since the amount of template DNA extracted from biological samples collected in crime scenes is often very small (approximately 10^{-12}g), template DNA is amplified using the polymerase chain reaction (PCR) process. This amplification allows length variants in the DNA, known as short tandem repeats (STRs). An electropherogram (epg) is the graphical display of the signal detected in STR when a sample is exposed to laser light. The presence of a peak in an epg corresponds to the alleles (the variants or polymorphism of a gene) in the DNA sample, which can be used to describe differences between individuals. Due to its intrinsic probabilistic nature, statistics is es-

sentially employed in DNA evidence interpretation. There are four main types of models facilitating the interpretation, namely: classical, binary, semi-continuous, and continuous. Continuous models ensure relatively greater reliability than other types of models in the evaluation of DNA evidence. However, statistical models for PCR phenomena are required to implement these models.

Minor peaks in an epg at positions other than the parental allelic positions are known as stutters. The presence of these stutters in an epg has been a key problem in DNA mixture evidence interpretation. Sophisticated methodologies to distinguish between stutters and real alleles are essential for accurate interpretation. Thus, practitioners use various approaches to understand the behaviour of stutters and work hard to make the interpretations more precise. Production of PCR stutter is usually studied in terms of stutter ratio (*SR*), which is defined as the observed stutter peak height as a ratio of the height of the parent allelic peak.

This research reviews existing models for PCR stutter ratio and develops new advanced Bayesian models to increase the efficiency of stutter prediction. The two sets of data: NGM SElect^{TM} and Identifiler^{TM} that include stutter peak information related to 4646 and 6949 heterozygous loci respectively were used throughout the study. All the improvements made for modelling stutter ratio can be summarised under three categories, namely:

1. non-hierarchical (non-mixture and two-component mixture) models
2. hierarchical (non-mixture and two-component mixture) models
3. infinite mixture models

In contrast to previous research, this study performs rigorous performance evaluations for all types of models fitted except infinite mixture models, which have been evaluated based on likelihoods only.

8.2 Non-hierarchical Models

The longest uninterrupted sequence (LUS) has been previously used as the key covariate in explaining the behaviour of SR. Bright et al. [21] investigated the performance of five models (two gamma, two log-normal, and one log-normal mixture) for predicting stutter ratios. In Chapter 2, this work has been extended by introducing six new alternative models including two normal, two non-standardised Student's t, and two two-component mixture models based on normal and non-standardised Student's t distributions. All these 11 models assumed a locus-specific model for the mean of SR. In variance modelling of SR, the models were classified into three categories, namely, profile-wide variance, locus-specific variance, and two-component mixture models. The performance of both the existing and proposed models was evaluated in Chapter 2 and 4.

In relation to the slope and intercept parameters, the normal and non-standardised Student's t models fitted to both datasets have indicated a high degree of concordance. The concordance in the parameters of log-normal and gamma models is moderate. The 11 models indicated significant differences in locus-specific slopes and intercepts. In the mean models of normal and log-normal distributions, the above differences were observed only at the TPOX locus of the Identifiler^{TM} dataset. Similarly, the parameters were different for gamma models at the TH01 locus of the NGM SElect^{TM} dataset. The locus-specific and profile-wide standard deviations were significantly different for all the non-mixture models fitted to both datasets. In the log-normal mixture model fitted to the Identifiler^{TM} dataset and the normal and non-standardised Student's mixture models fitted to the NGM SElect^{TM} dataset, the component with larger variance captured a relatively larger percentage of points. The non-standardised Student's t mixture model has shown the best ability to capture the stutter ratios, with approximately a 20% increase compared to the normal mixture model. In comparison with the locus-specific variance models, the profile-wide variance non-standardised Student's t models revealed more heavy-tailed behaviours as their degrees of freedom parameters are small. Regardless of the revealed heavy-tailed behaviour of the Student's t-mixture model, large upper bounds of both degrees of freedom parameters enables approximating normal-like tail behaviours.

In order to identify appropriate measures to be used in the performance evaluation of

the old and new non-hierarchical models presented in Chapter 2, Chapter 3 reviewed the measures available for the Bayesian model assessment. This assessment comprised detailed reviews of usefulness, shortfalls, and required conditions for using model evaluation methods: information criteria, cross-validation measures, and Bayesian p-values. Based on this evaluation, two information criteria (BIC and WAIC), LOO-CV approximations, and Bayesian p-values were used in Chapter 4 for evaluating the 11 models.

The graphical evaluation with Q-Q and P-P plots was carried out for normal and log-normal models since such evaluation is not applicable for the other models. This assessment has identified several lack-of-fit issues in normal and log-normal models. These problems are critical in the tails of the distributions, and the models fitted to the Identifiler^{TM} dataset have mostly evidenced this. Hence, a better performance in the tail behaviour can be expected from two-component mixture models; however, this cannot be graphically tested.

The BIC results have implied that, non-standardised Student's t mixture and the two-component normal mixture are the first and second best models regardless of the dataset. However, considering the relative complexity in mixture models, the non-standardised Student's t model is selected as the best option. Although WAIC, which uses posterior predictive distribution rather than point estimates of parameters in calculating likelihoods, is considered as the best information criterion, it has a certain condition for its validity. Since the posterior log-predictive densities of more than 95% of the observations in each dataset exceed 0.4, the validity requirement of WAIC is not satisfied in this study. Therefore, LOO-CV was selected as the best measure for evaluating the models, based on their predictive accuracy. Due to the high computational cost associated with exact LOO-CV, it was approximated using three measures, namely: IS (importance sampling), TIS (truncated importance sampling), and PSIS (Pareto smoothed importance sampling). Despite the limitations in WAIC and LOO-CV measures, they all have confirmed that the models based on normal distribution outperform in all the modelling categories (profile-wide variance, locus-specific variance, and two-component mixture), for both datasets. This finding is of high importance, as normal distribution is the most widely used, well-known, simple continuous probability distribution among practitioners. Therefore, the implemen-

tation of a normal model for stutter ratio, in a continuous Bayesian model of DNA mixture interpretation would be relatively easy.

Evaluation of model performance based on Bayesian p-values and L-measures offers mixed findings while not contradicting the above conclusions related to the best models. The p-values representing marginal predictive distribution have not indicated any problem in any model fitted to the datasets as they all were close to the desired value 0.5. However, the p-values based on chi-square discrepancy measure have revealed issues in predictions of some models. The deviations of predictions in comparison with actual observations, from the estimated mean of the respective distribution were quite large in gamma models fitted to the NGM SElect^{TM} dataset. These values were larger for both profile-wide and locus-specific non-standardised Student's t models fitted to the Identifiler^{TM} dataset. Large L-measures were observed for the log-normal mixture model, and this indicates larger variations in the predictions. Producing a few unbelievably large predicted values by the log-normal mixture model has been identified as the possible reason for this. Table 8.1 summarises the comparative performance of all the non-hierarchical models developed and evaluated in this study.

Table 8.1: Comparative performance of the fitted models

		NGM SElect^{TM}		Identifiler^{TM}	
Modelling Category	Model Rank	BIC	LOO-CV	BIC	LOO-CV
Profile-wide variance					
	First	T_0	N_0	T_0	N_0
	Second	N_0	G_0	G_0	G_0
	Third	G_0	LN_0	N_0	LN_0
	Fourth	LN_0	T_0	LN_0	T_0
Locus-specific variance					
	First	T_1	N_1	T_1	N_1
	Second	N_1	G_1	N_1	G_1
	Third	G_1	LN_1	G_1	LN_1
	Fourth	LN_1	T_1	LN_1	T_1
Two-component mixture					
	First	MT_1	MN_1	MT_1	MN_1
	Second	MN_1	MLN_1	MN_1	MLN_1
	Third	MLN_1	MT_1	MLN_1	MT_1

8.3 Hierarchical Models

Complete pooling, or analysis of a whole set of data ignoring its inherent hierarchical structure suppresses variations in the data and may impact on the overall objective of a study. Performing a separate analysis for each source of the data, or no pooling, also tends to provide misleading inferences. Therefore, hierarchical models for seven non-hierarchical models including four locus-specific variance models (gamma, normal, log-normal, and non-standardised Student's t) and three two-component mixture models (normal, log-normal, and non-standardised Student's t) were investigated in Chapter 5. It was expected to find an increased accuracy through a shrinkage in locus-specific model parameters in stutter prediction. The reason for selecting only these models was that in Chapter 2 and 4, they revealed relatively better performance in comparison with the profile-wide variance models.

The almost identical log-likelihoods of each pair of hierarchical and non-hierarchical models indicated the absence of bias-variance trade-off or lack of pooling in the group level parameters of hierarchical models fitted to both datasets. However, some minor changes have been observed among the predictive measures that were estimated based on point-wise predictive densities. In particular, there was no visible pooling in the slope and intercept parameters of mean model under the hierarchical models. The hierarchical models have provided a way to estimate the parameters of the normal distributions from which the locus-specific slopes and intercepts of non-hierarchical models (except the slope parameters of non-standardised Student's t mixture), are coming. For normal and non-standardised Student's t models, a consistency was observed in these inferred normal distributions across the two datasets while revealing slight differences for the other models. Even though the non-mixture models fitted to the datasets did not reveal any considerable pooling in standard deviation parameters, some occasional minor changes in these parameters have been observed in mixture hierarchical models.

The precision parameters of all the non-mixture models fitted to both datasets and the mixture models fitted to the Identifiler^{TM} dataset revealed high goodness-of-fit with their inferred distributions which were not consistent across the two datasets. In re-examining the results, the study has found that hierarchical models are more effective with smaller

sample sizes. Hence, including a large number of peak height information in each locus may be the key reason for the lack of pooling revealed in the group level parameters.

8.4 Infinite Mixture Models

The analytical relationships between the prior information, the observed data, and the parameters of posterior predictive distribution of data, derived in Chapter 6 were applied in Chapter 7 to develop an infinite mixture of simple linear regression models for predicting stutter. Fitting an infinite mixture model for *SR* is important in DNA mixture evidence interpretation as it provides a more robust, flexible, and accurate prediction for PCR stutter. When data are coming from different sub-populations, effective modelling requires taking this existing clustering into consideration. Finite mixture models provide a great flexibility in modelling data that are assumed to come from more than one source population. Since in many practical contexts, the number of sources that the data were generated is unknown, the selection of an appropriate number of sources or the number of components in the mixture model is problematic. In Bayesian non-parametric methods that provide increased flexibility by not restricting the number of parameters, Dirichlet processes (DP)-based mixture models are frequently used.

A DP is a stochastic process that enables placing a distribution over distributions. The Stick-breaking construction, Pólya (Blackwell-MacQueen) urn scheme, Pitman-Yor process, and the Chinese restaurant process (CRP) are different representations of DPs. Considering its relative simplicity and having predictive probabilities proportional to the number of observations in clusters, this study selected CRP as a non-parametric DP prior within a collapsed Gibbs sampling algorithm when fitting an infinite mixture of simple linear regression models for *SR* using *LUS* as the predictor. Due to the associated computational cost and the time constraints, this study has developed infinite mixture models only to the D2S1338 locus of the NGM SElect^{TM} dataset. The performance of collapsed Gibbs sampling with CRP was tested in terms of the variation in the number of active clusters and the log-likelihoods of data. In addition, the effect of the concentration parameter (α) was tested at 20 different values.

The prior sample used, the method of estimating prior parameters (A, B, and C), the initial cluster allocation method, and the power of likelihood in calculating conditional probabilities of Gibbs sampling, were the criteria that defined the testing conditions in the study. The study has found some important associations between the above parameters, in relation to the performance of collapsed Gibbs sampling with CRP. The number of clusters was increased as the value of α increases, and there was an increasing trend in the variation of the number of active clusters as the number of clusters increases. As this implies, higher values of α lead to increased variation in the active number of clusters in collapsed Gibbs sampling with CRP. Although no variation in the number of active clusters is generally expected due to different initial samples of prior parameter estimation, a moderate variation has been observed across the initial samples under the backward method of initial cluster allocation (starting with ten clusters), particularly with smaller α. Prior A and B lead to producing more clusters under the backward method than the forward method (starting with a single cluster) while prior C does not reveal such change based on the initial cluster allocation. Prior B has exhibited more stability in the results in terms of number of clusters. In addition, the number of clusters produced by prior C is noticeably less than that of prior A and B.

The study proposed the second and third powers of the likelihoods (in addition to the likelihood itself, which is originally used in the collapsed Gibbs sampling with CRP) for calculating the conditional probabilities of Gibbs sampling. With the use of the third power of the likelihood, the results have revealed more stability in the number of active clusters, even with large number of active clusters. In addition, the log-likelihoods indicated a clear increase with the order of the likelihood. The biggest improvement in the log-likelihoods was evident with prior B and the third power of likelihood. In addition, the Gibbs sampler used a rather longer machine time in the initialisation with a single cluster in comparison with many (10) clusters. Based on these findings, the study recommends using: prior B, the third power of likelihood, and initialisation of the Gibbs sampler with more clusters, when developing infinite mixture models for *SR*. According to the log-likelihood results obtained in this study, a seven-component mixture is selected as the best option for improving stutter prediction at the D2S1338 locus of the

NGM SElect^{TM} dataset. Implementation of this model or a similar infinite mixture model across all the loci, in DNA evidence interpretation software, will essentially improve the quality of DNA mixture interpretation. Furthermore, this study extends infinite mixture models to the simple linear regression case. The improvements on the collapsed Gibbs sampling with CRP that are recommended in this study significantly contribute to theory as the existing amount of applications is rather low.

8.5 Directions for Future Research

Future work related to this study can be broadly classified as: the work related to model comparison criteria, infinite mixture models for stutter prediction, and the problems associated with CRP as a representation of DP. The specific details about the possible research directions are as follows:

1. After discussing various model comparison criteria, this study selected WAIC as the best for any set of models including hierarchical and mixture models. However, WAIC starts to fail when posterior variances of log predictive densities exceed 0.4. The study also used importance sampling approximations to the exact cross-validation, as it is computationally expensive when evaluating models fitted to large datasets. The study observed low performance in WAIC and these approximations when posterior variances exceed 0.4. Hence, it is important to improve these methods for comparing models fitted to large datasets and produce large posterior variances.

2. The study has recommended an improved version of collapsed Gibbs sampling CRP to fit infinite mixture models for predicting *SR* using *LUS* as a predictor. The study selected a seven-component mixture as the best since the study was conducted only for a selected set of values of concentration parameter. Therefore, the study may extend to larger values (greater than 10^{-2}) with the recommended version and examine infinite mixture models for all the loci of both datasets considered in the overall study.

3. The inverse proportionality of the variance of stutter height to the amount of template DNA is a useful relationship discussed under the modelling for *SR*. However, infinite mixture models fitted in this study have not considered this relationship since the conventional Bayesian linear regression model adopted does not facilitate incorporating it. Therefore, developing an infinite mixture model taking this relationship into account would be better.

4. Once an infinite mixture model is finalised, a comparison between the performance of that model and the corresponding finite mixture model, for predicting *SR* would be an interesting work.

5. Having multi-valued active number of clusters increases the model complexity in infinite mixture modelling without a substantial improvement in the fit. The suggested use of third power of the likelihood in calculating conditional probabilities under Gibbs sampling reduces the risk of having a multi-valued active number of clusters. However, when this problem cannot be avoided, it is important to investigate a suitable technique to minimise the model complexity.

Appendix A

Locus-specific Variation of Hierarchical Vs Non-hierarchical Model Parameters

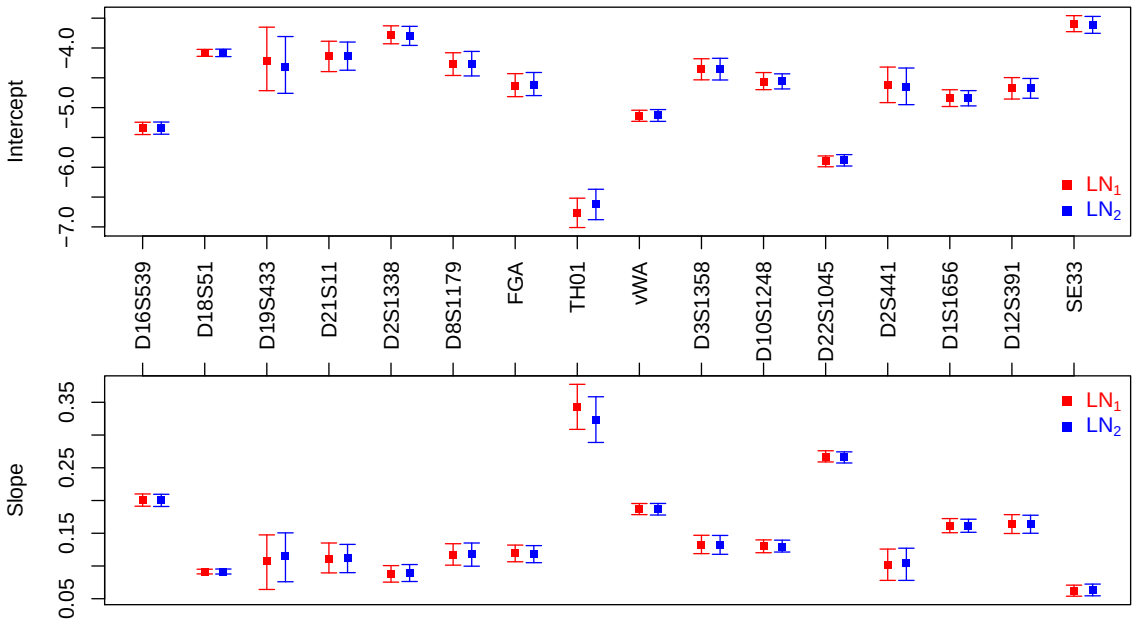

Figure A.1: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical log-normal models for the NGM SElect^{TM} dataset

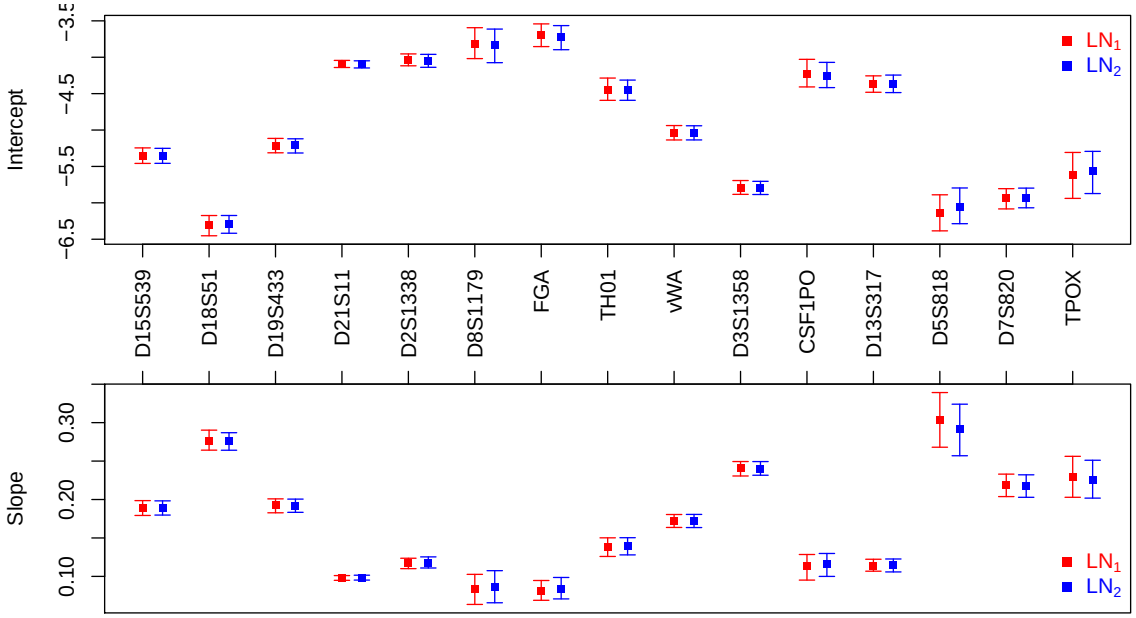

Figure A.2: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical log-normal models for the Identifiler^{TM} dataset

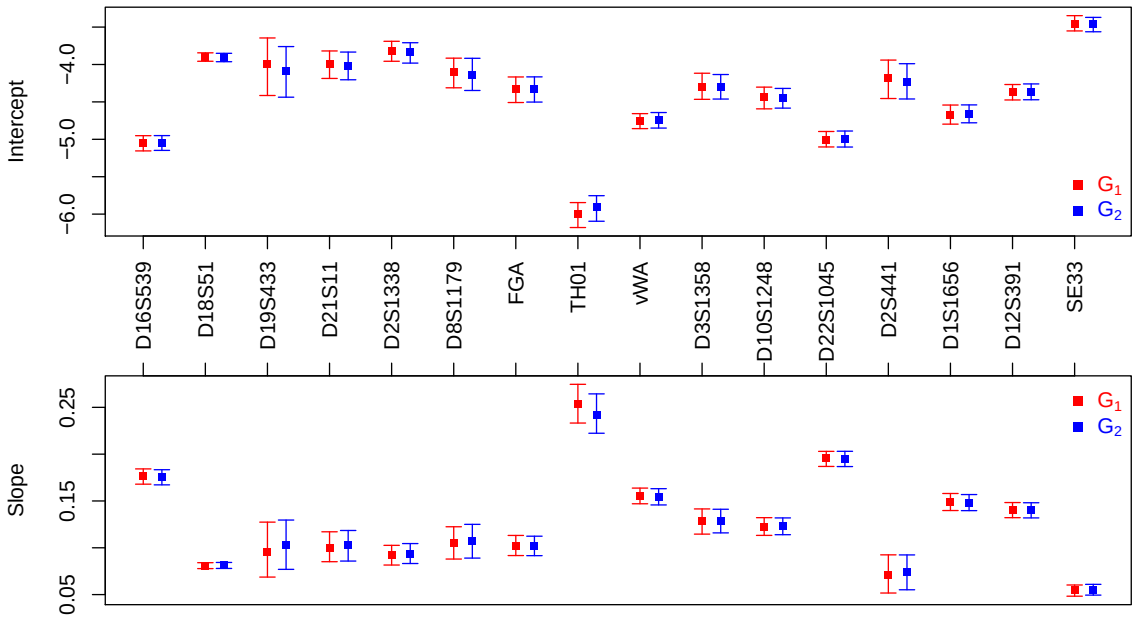

Figure A.3: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical gamma models for the NGM SElect^{TM} dataset

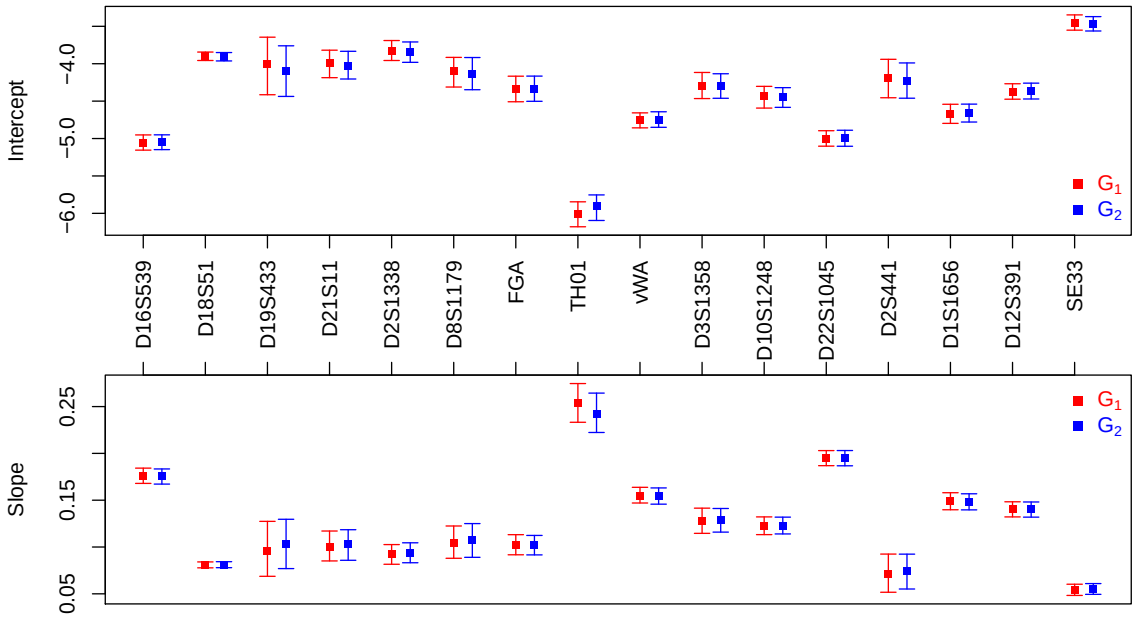

Figure A.4: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical gamma models for the Identifiler^{TM} dataset

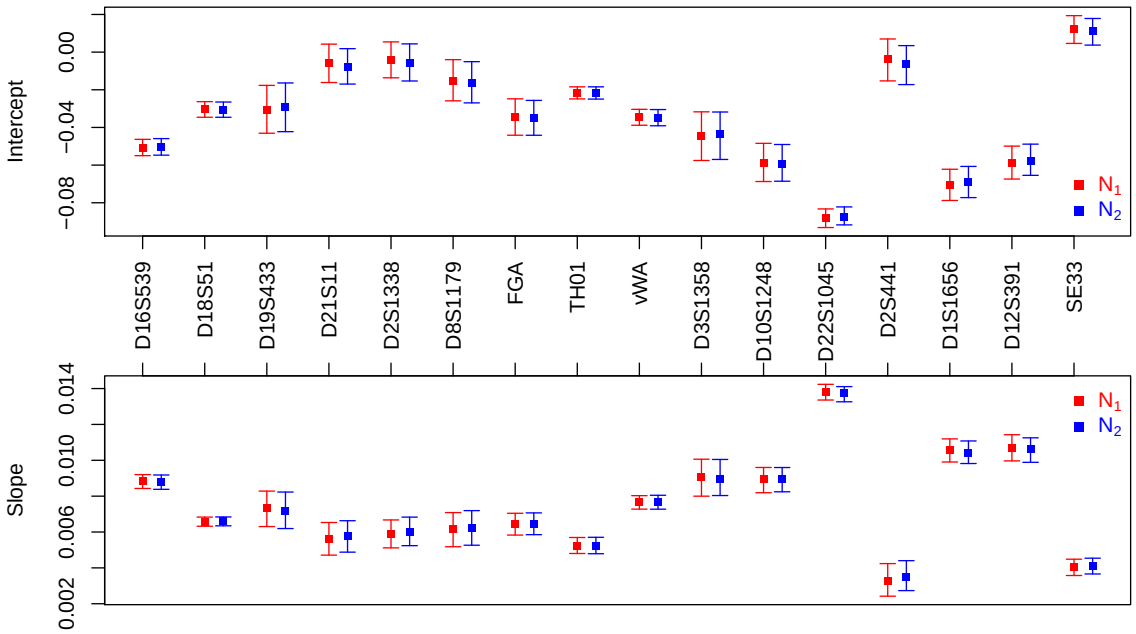

Figure A.5: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical normal models for the NGM SElect^{TM} dataset

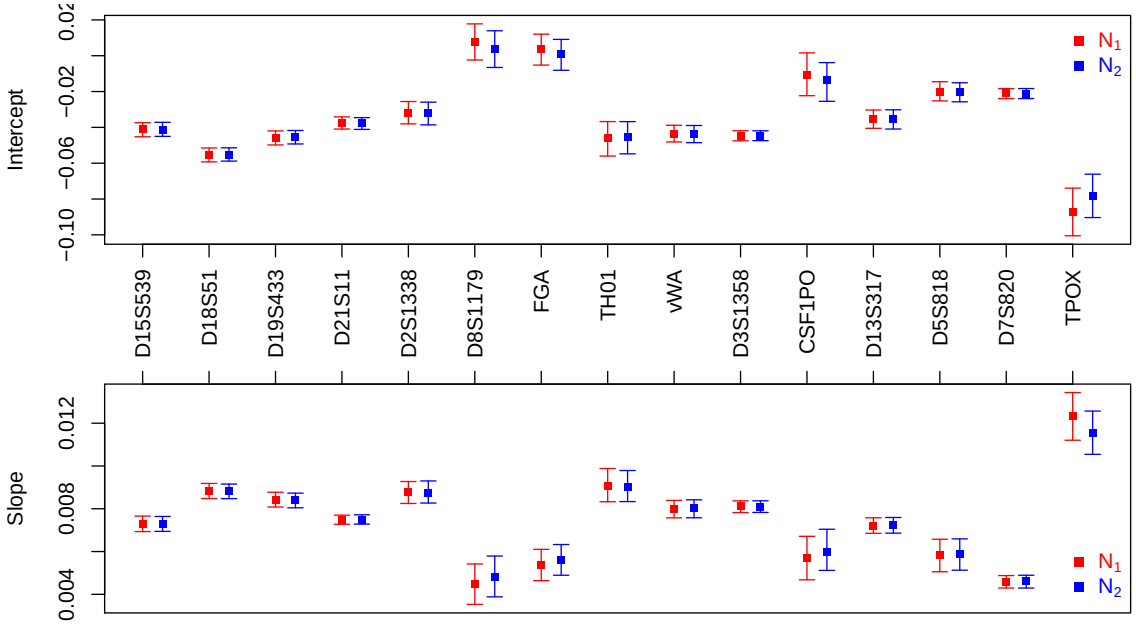

Figure A.6: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical normal models for the Identifiler^{TM} dataset

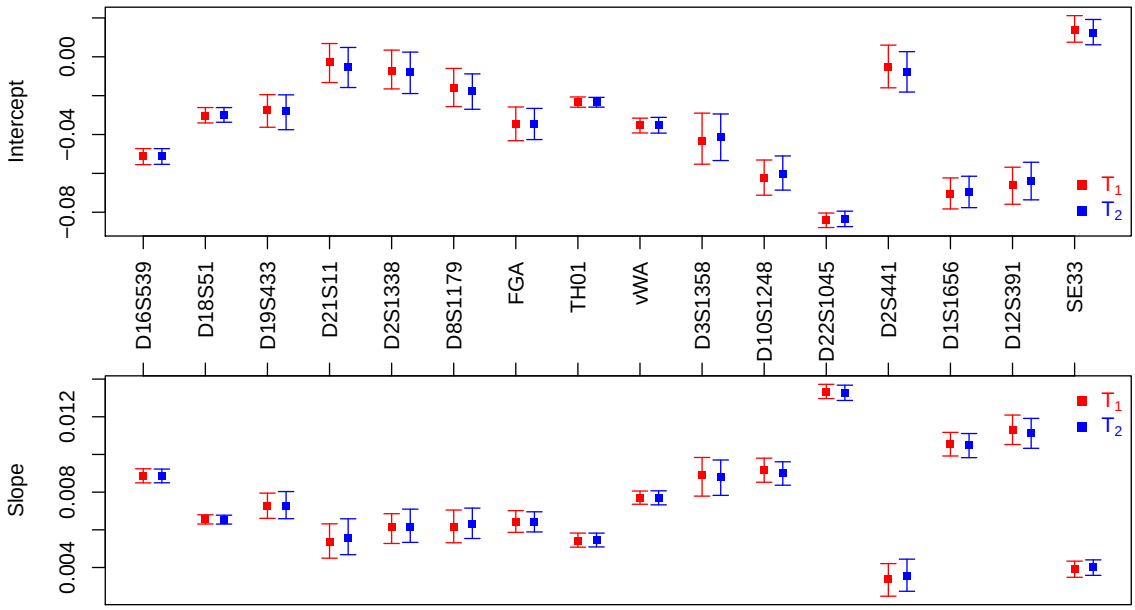

Figure A.7: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical non-standardised Student's t models for the NGM SElect^{TM} dataset

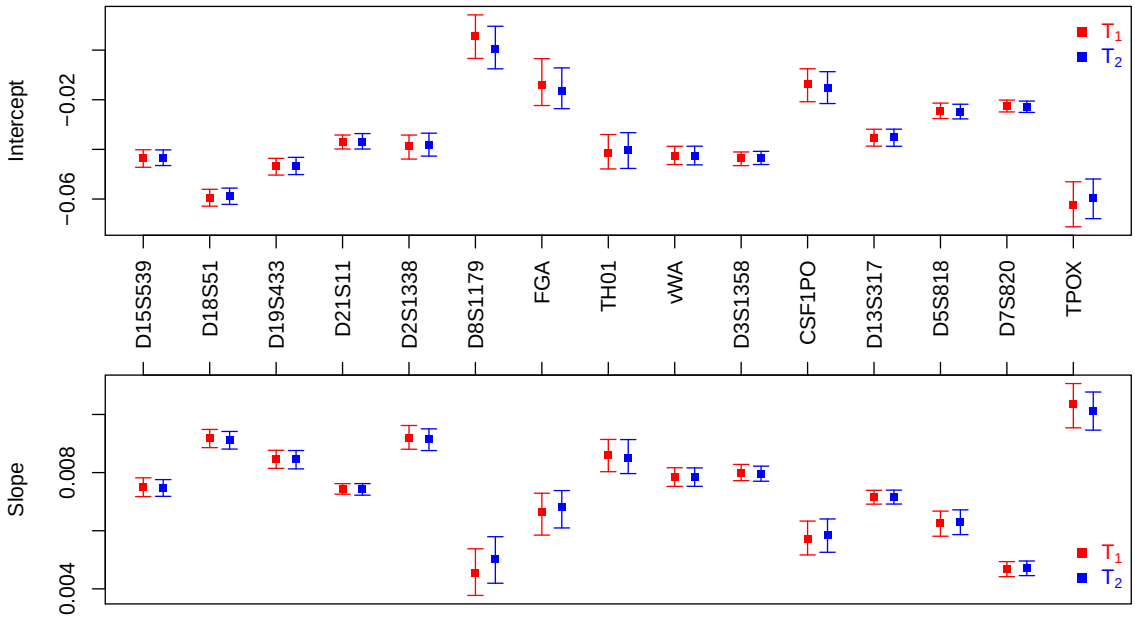

Figure A.8: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical non-standardised Student's t models for the Identifiler^{TM} dataset

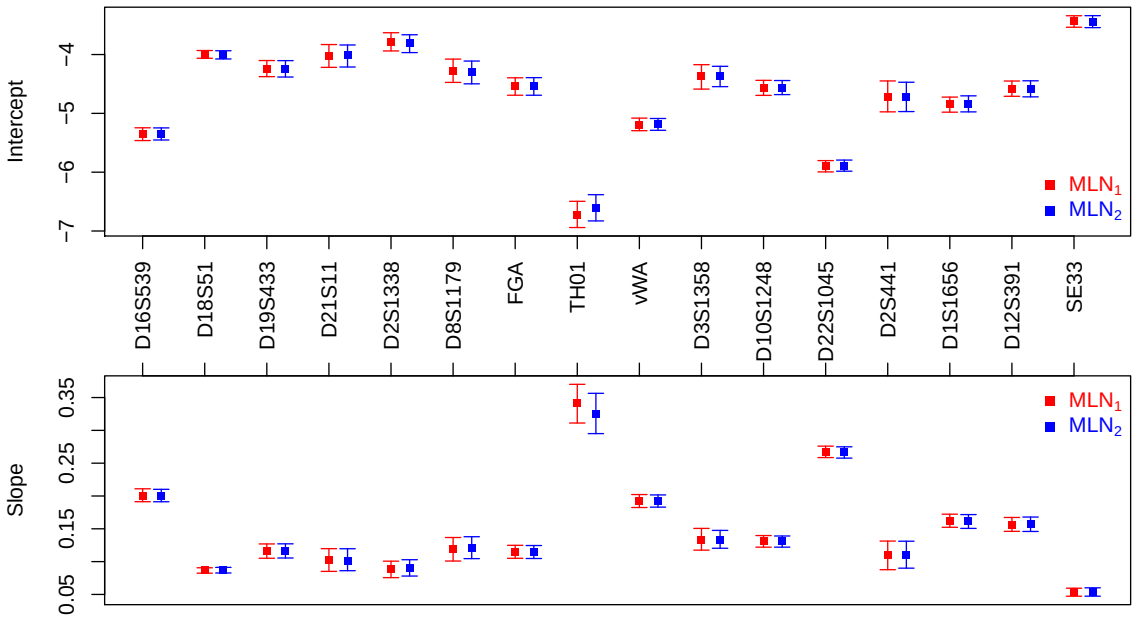

Figure A.9: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical log-normal mixture models for the NGM SElect^{TM} dataset

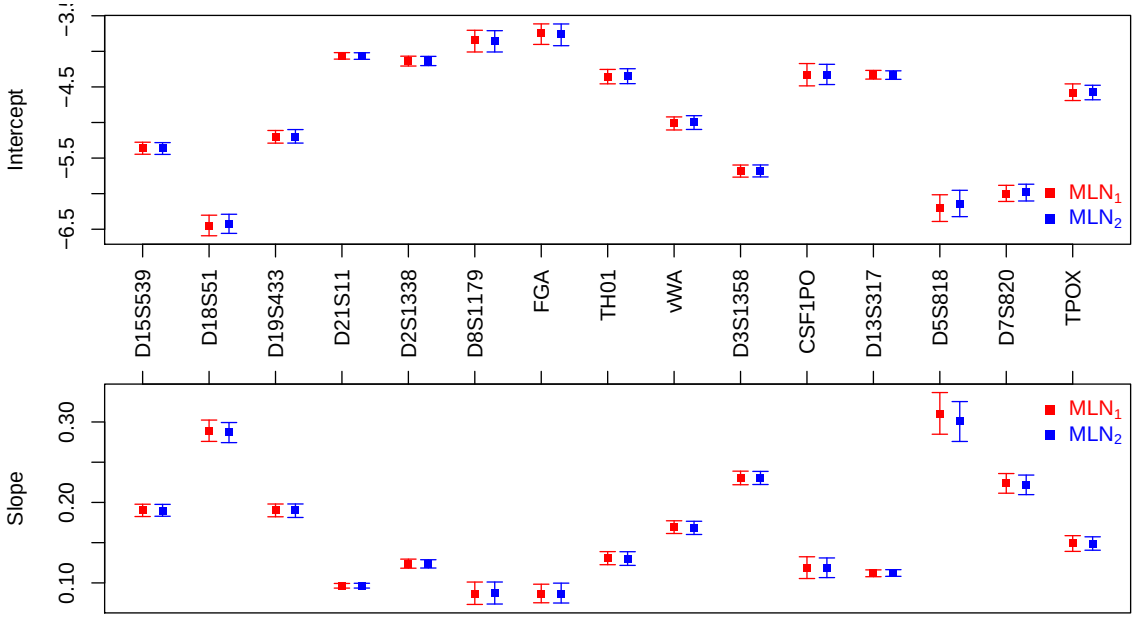

Figure A.10: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical log-normal mixture models for the Identifiler^{TM} dataset

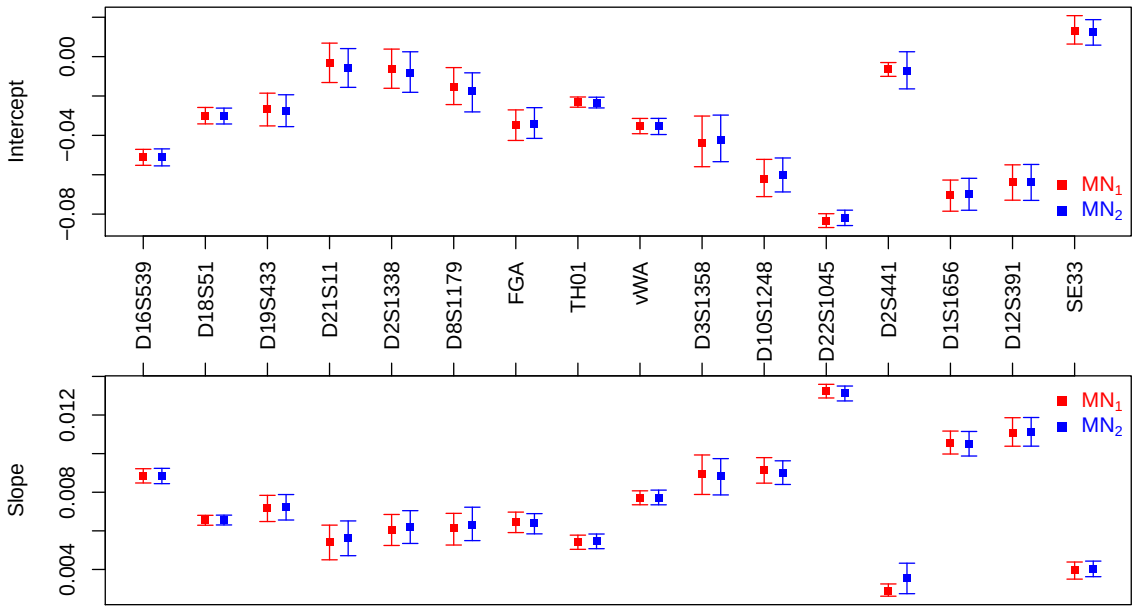

Figure A.11: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical normal mixture models for the NGM SElect^{TM} dataset

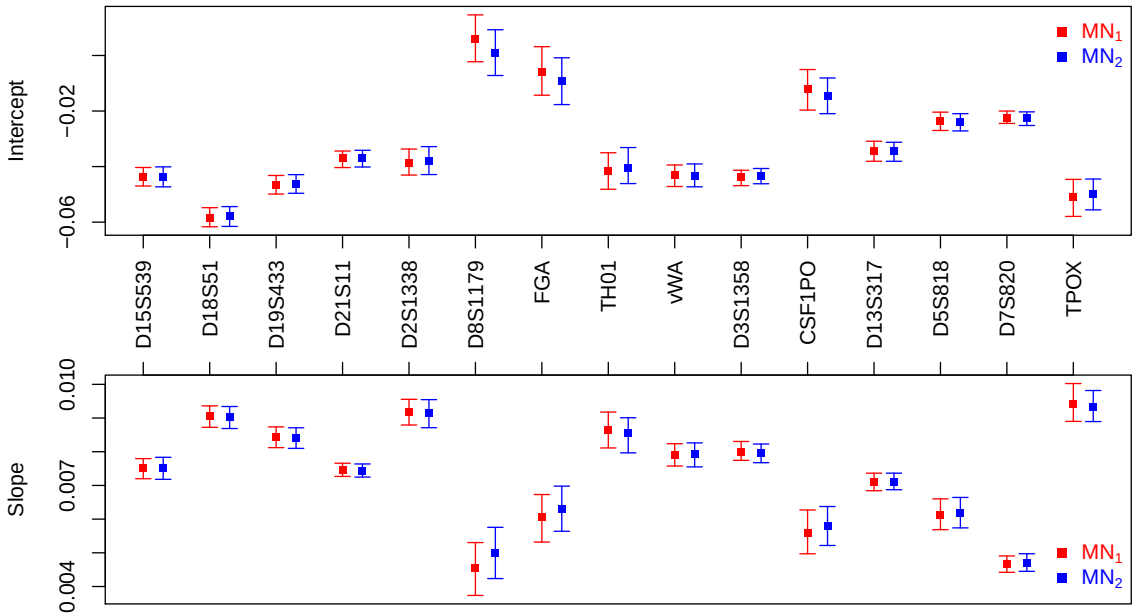

Figure A.12: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical normal mixture models for the Identifiler^{TM} dataset

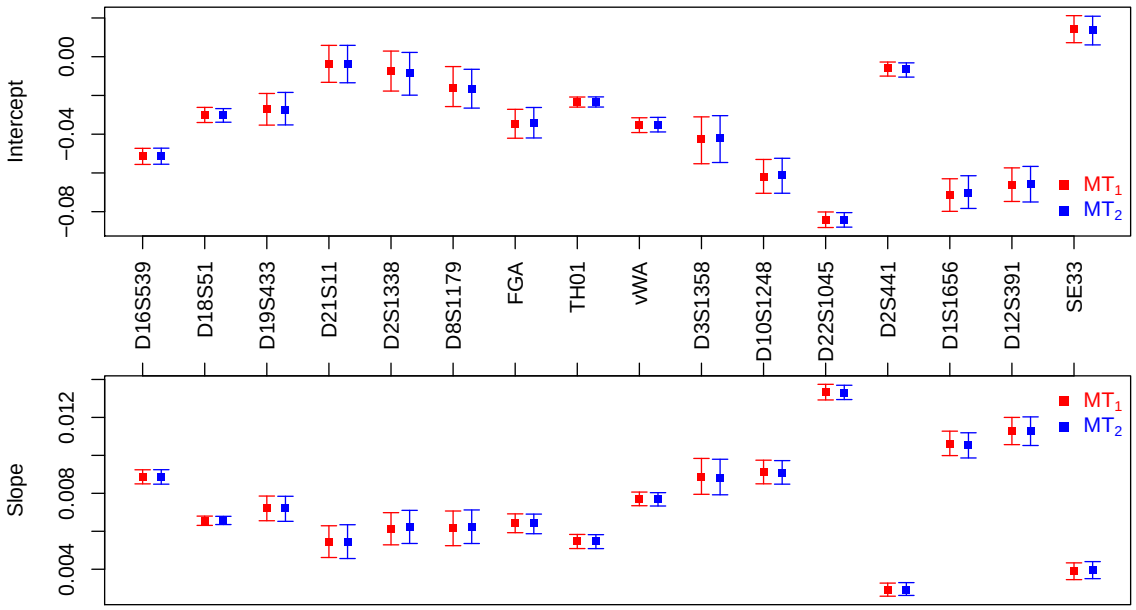

Figure A.13: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical non-standardised Student's t mixture models for the NGM SElect^{TM} dataset

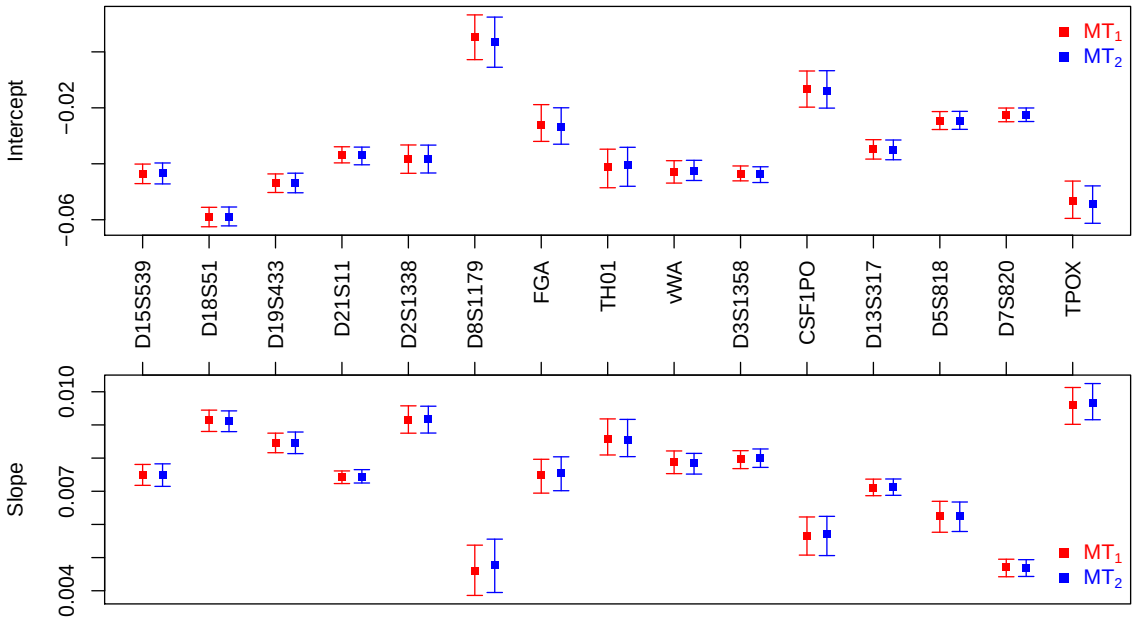

Figure A.14: Locus-specific variation (95% credible interval with posterior median) of the mean model parameters (slope β_0 and intercept β_1) of hierarchical and non-hierarchical non-standardised Student's t mixture models for the Identifiler^{TM} dataset

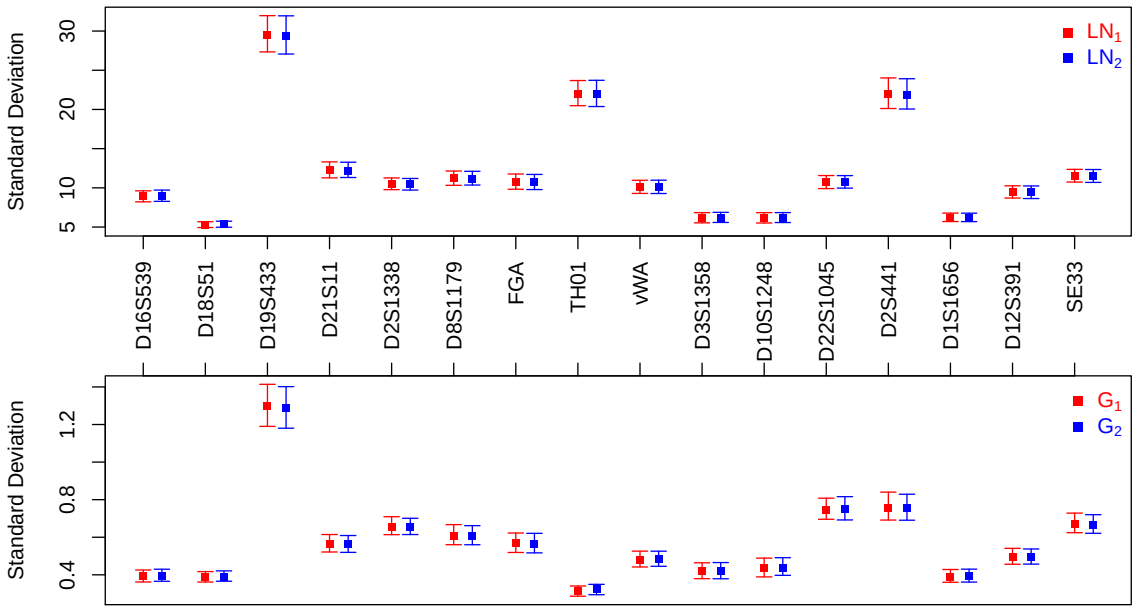

Figure A.15: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of log-normal and gamma models (hierarchical and non-hierarchical) for the NGM SElect^{TM} dataset

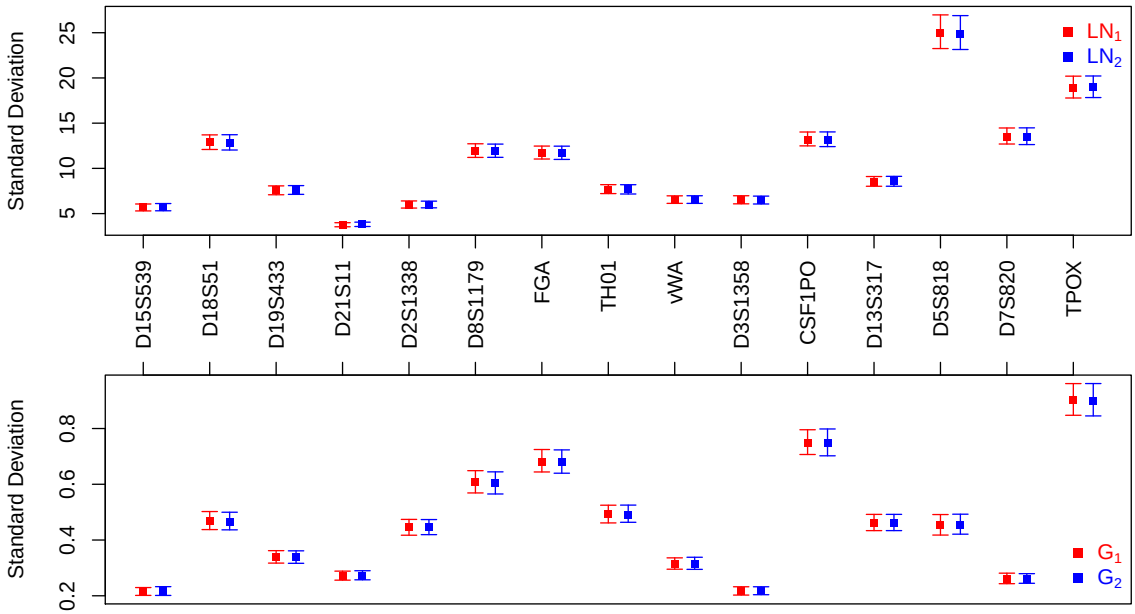

Figure A.16: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of log-normal and gamma models (hierarchical and non-hierarchical) for the Identifiler^{TM} dataset

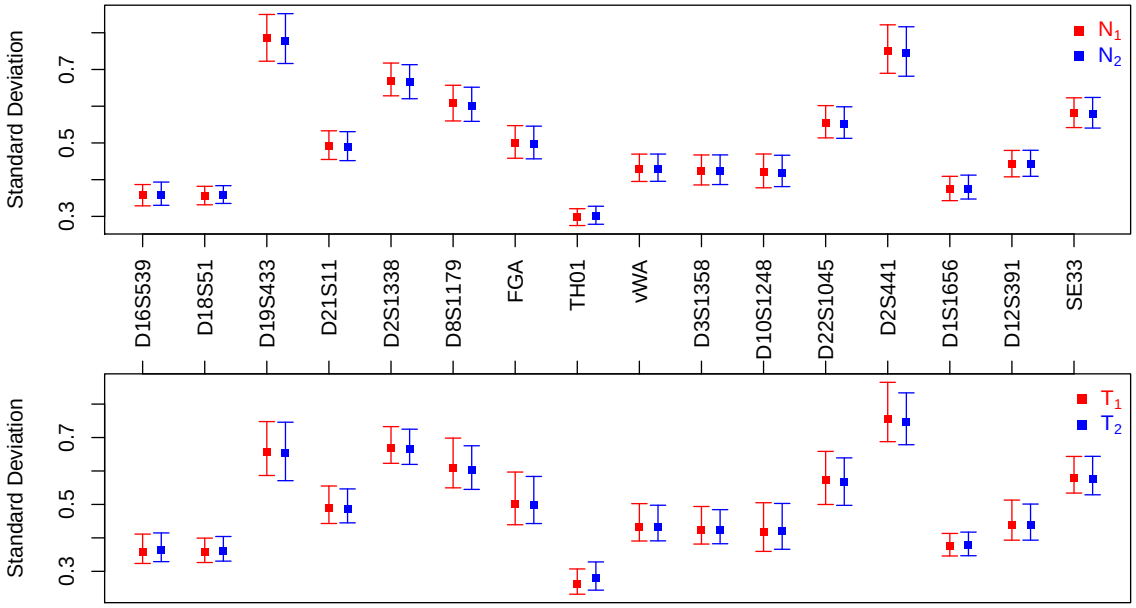

Figure A.17: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of normal and non-standardised Student's t models (hierarchical and non-hierarchical) for the NGM SElect^{TM} dataset

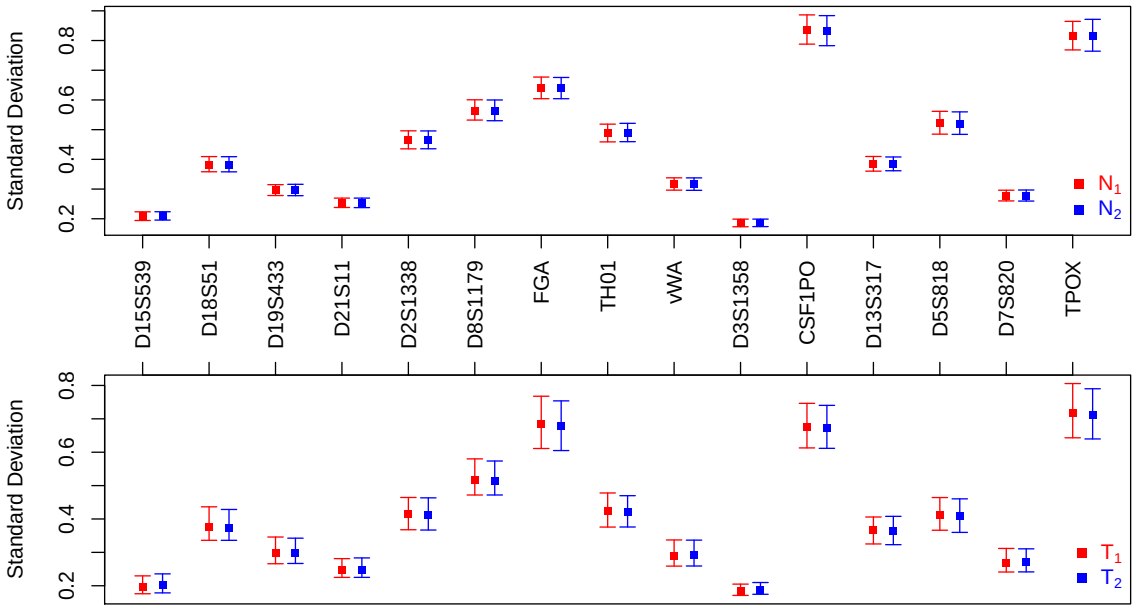

Figure A.18: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of log-normal and non-standardised Student's t models (hierarchical and non-hierarchical) for the Identifiler^{TM} dataset

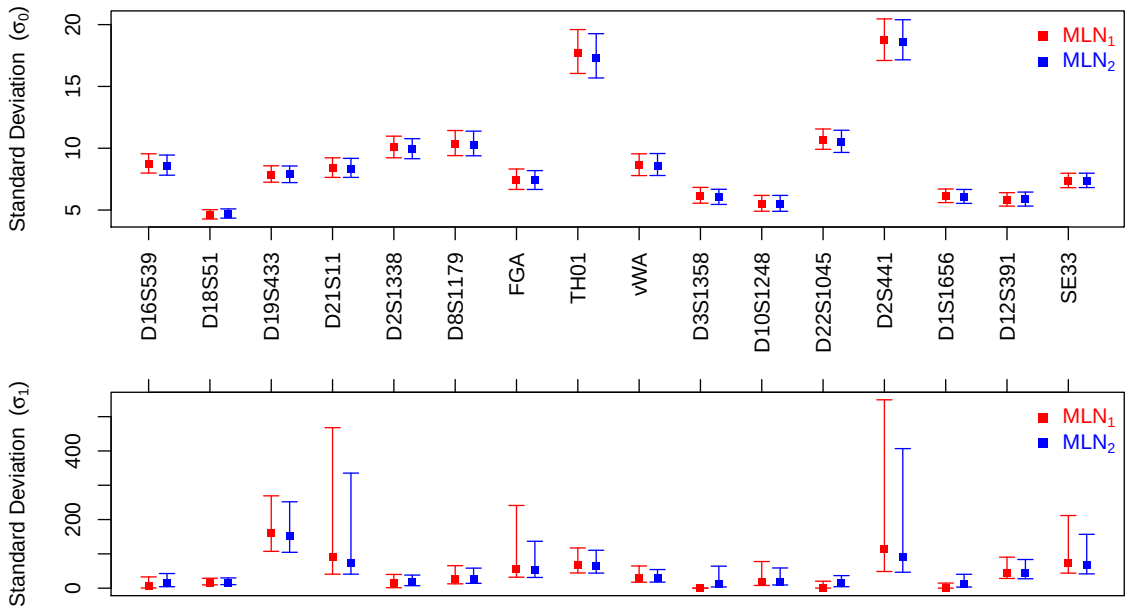

Figure A.19: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical log-normal mixture models for the NGM SElect^{TM} dataset

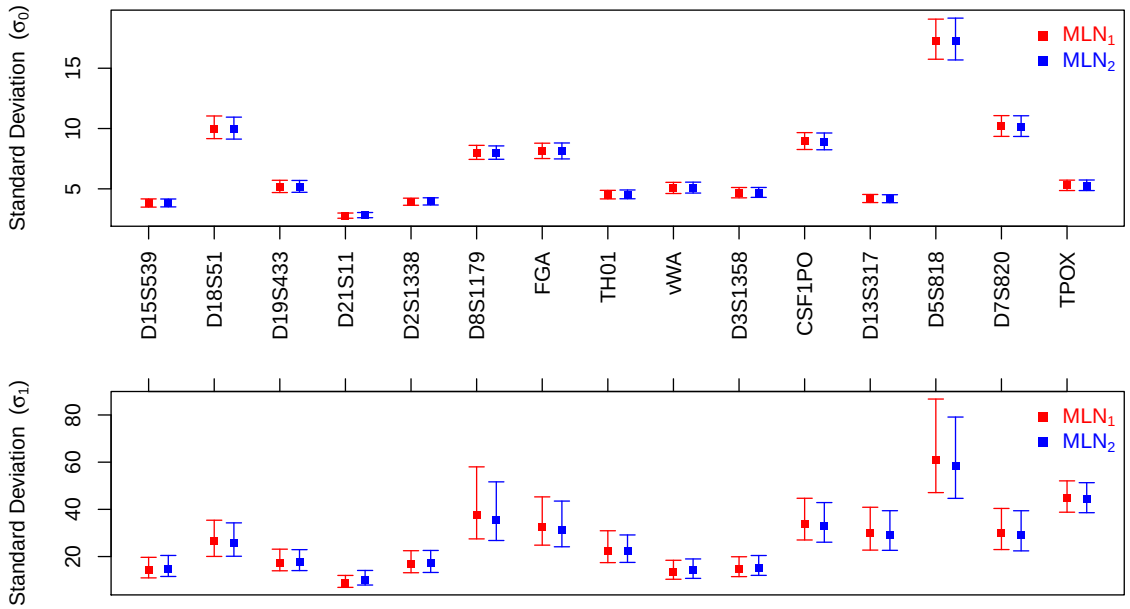

Figure A.20: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical log-normal mixture models for the Identifiler^{TM} dataset

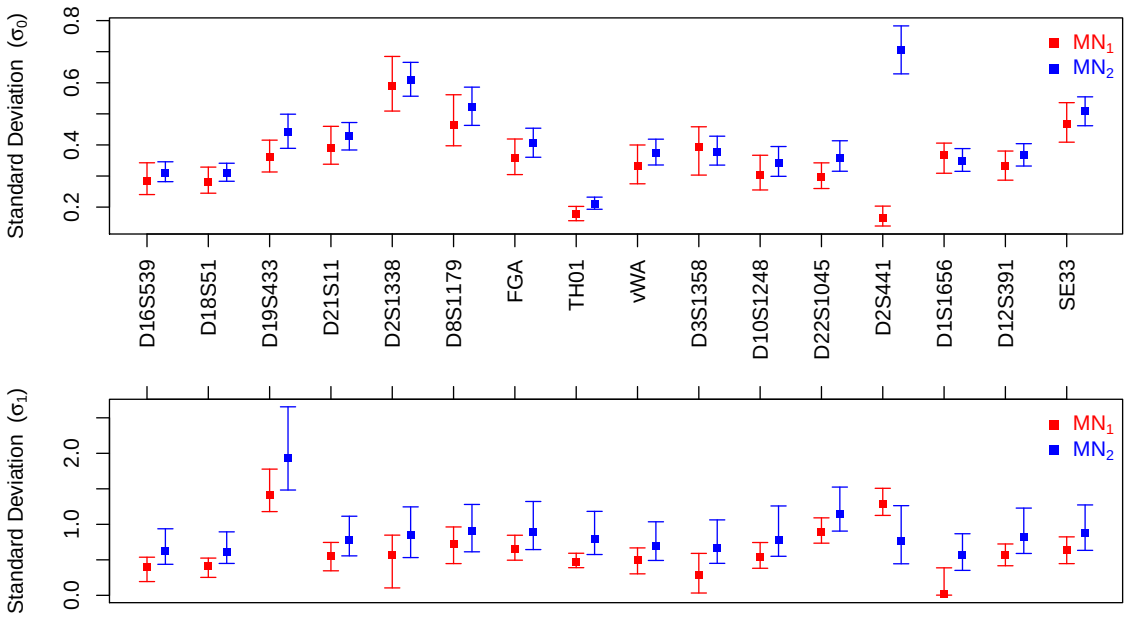

Figure A.21: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical normal mixture models for the NGM SElect^{TM} dataset

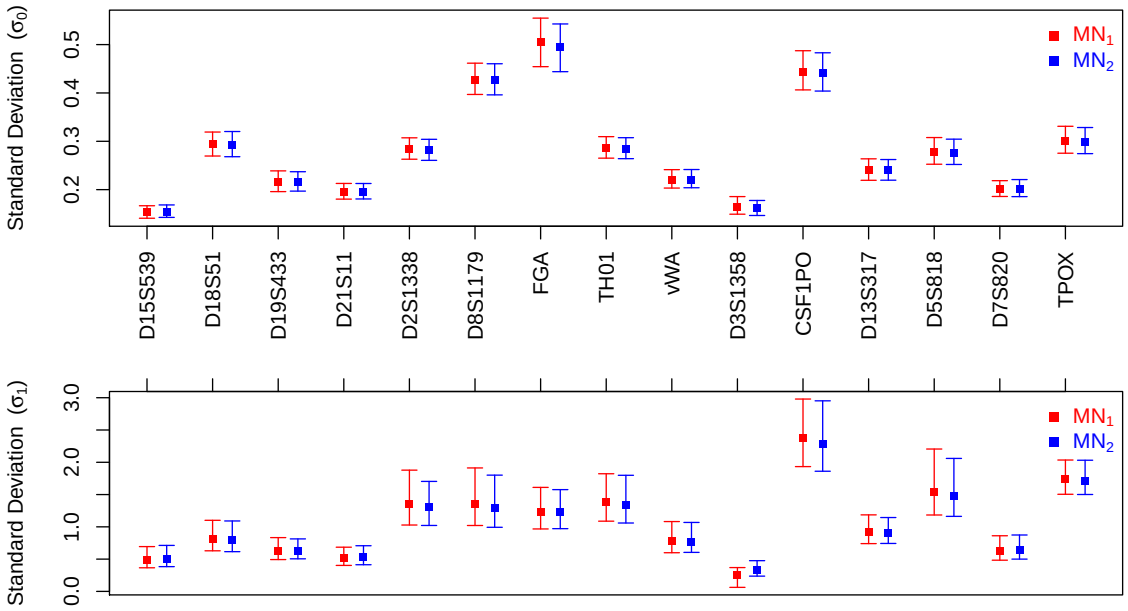

Figure A.22: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical normal mixture models for the Identifiler^{TM} dataset

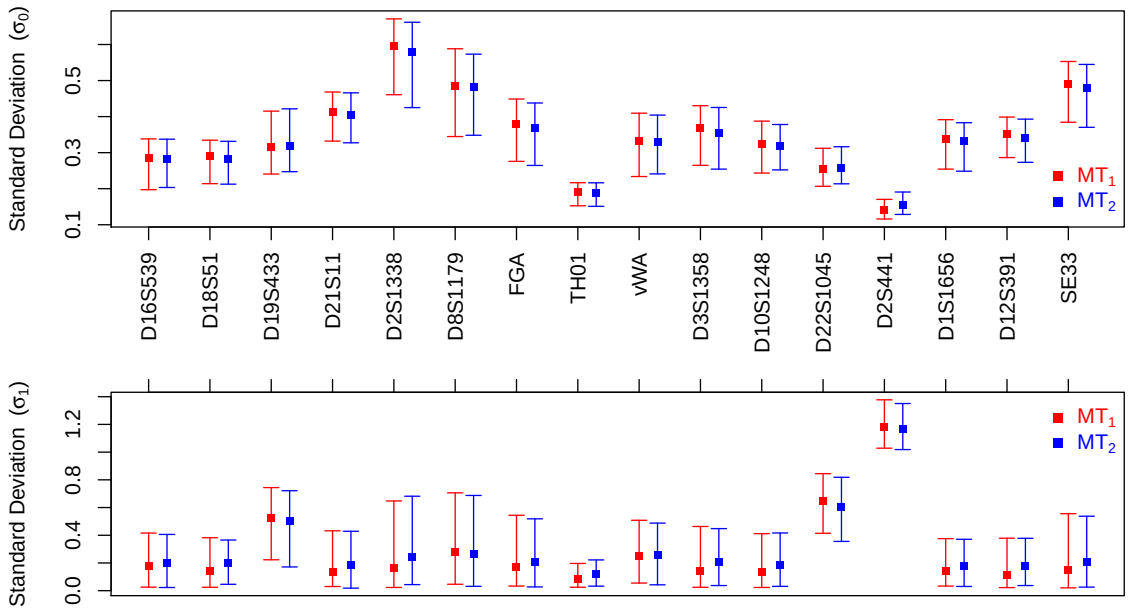

Figure A.23: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical non-standardised Student's t mixture models for the NGM SElect^{TM} dataset

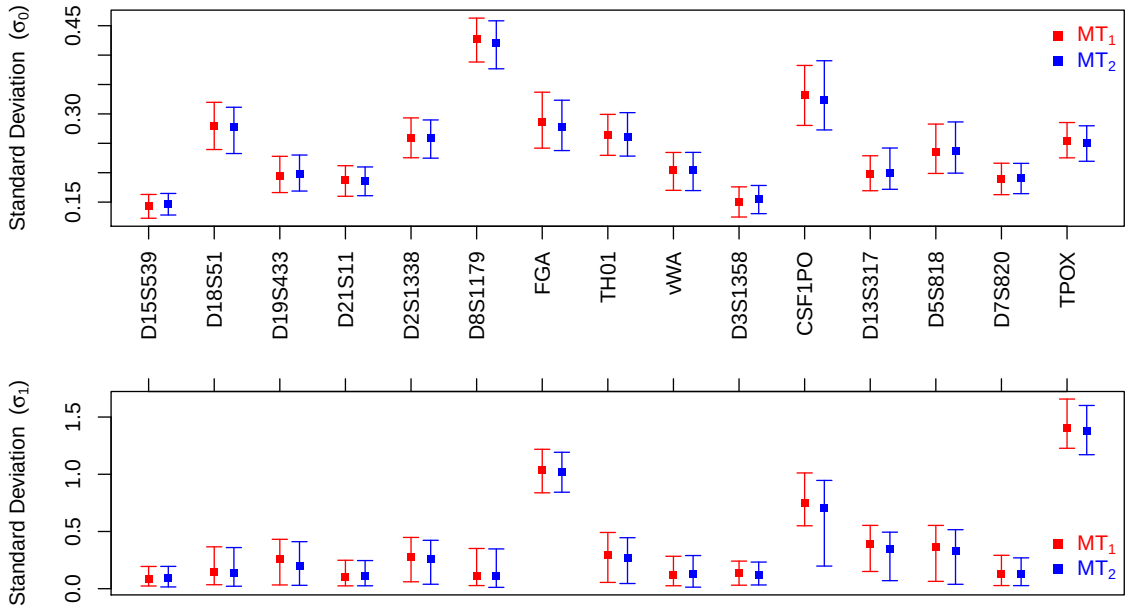

Figure A.24: Locus-specific variation (95% credible interval with posterior median) of the standard deviation parameters of hierarchical and non-hierarchical non-standardised Student's t mixture models for the Identifiler^{TM} dataset

Appendix B

The Additional Information Relevant to the Performance of Collapsed Gibbs Sampling with CRP

Table B.1: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-15}$ and $P_k \propto n_k p_k$

Condition	NC	N	LL	SD	C1	C2
AF1	1	500	1230		401	
AF2	1	500	1229		401	
AF3	1	500	1228		401	
AF4	1	500	1233		401	
AF5	1	500	1231		401	
AB1	2	500	1362	15	253	148
AB2	2	500	1367	14	253	148
AB3	2	500	1366	13	231	170
AB4	2	500	1351	17	258	143
AB5	2	500	1358	15	255	146
BF1	1	500	1245		406	
BF2	1	500	1245		406	
BF3	1	500	1245		406	
BF4	1	500	1245		406	
BF5	1	500	1245		406	
BB1	2	500	1382	15	251	155
BB2	2	500	1383	14	250	156
BB3	2	500	1384	14	234	172
BB4	2	500	1378	15	254	152
BB5	2	500	1381	14	255	151
CF	1	500	1245		406	
CB	1	500	1245		406	

Table B.2: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-11}$ and $P_k \propto n_k p_k$

Condition	NC	N	LL	SD	C1	C2	C3
AF1	1	500	1230		401		
AF2	1	500	1229		401		
AF3	1	500	1228		401		
AF4	1	500	1233		401		
AF5	1	500	1231		401		
AB1	2	500	1362	15	253	148	
AB2	2	499	1367	14	253	148	
	3	1	1441		266	134	1
AB3	2	500	1366	13	231	170	
AB4	2	499	1351	17	258	143	
	3	1	1384		268	130	3
AB5	2	499	1358	15	255	146	
	3	1	1390		289	101	11
BF1	1	500	1245		401		
BF2	1	500	1245		401		
BF3	1	500	1245		401		
BF4	1	500	1245		401		
BF5	1	500	1245		401		
BB1	2	499	1382	15	250	156	
	3	1	1438		266	138	2
BB2	2	498	1383	14	250	156	
	3	2	1431	16	285	120	2
BB3	2	500	1384	14	234	172	
BB4	2	493	1378	15	254	152	
	3	7	1437	16	270	133	3
BB5	2	499	1382	14	255	151	
	3	1	1419		289	116	1
CF	1	500	1245		406		
CB	1	500	1245		406		

Table B.3: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-8}$ and $P_k \propto n_k p_k$

Condition	NC	N	LL	SD	C1	C2	C3
AF1	1	500	1230		401		
AF2	1	500	1229		401		
AF3	1	500	1228		401		
AF4	1	495	1233		401		
	2	5	1249	3	395	6	
AF5	1	500	1231		401		
AB1	2	161	1361	15	253	148	
	3	339	1418	12	270	127	4
AB2	2	189	1366	13	253	148	
	3	311	1422	13	271	126	5
AB3	2	322	1366	13	232	169	
	3	178	1423	11	256	142	3
AB4	2	435	1351	16	257	144	
	3	65	1405	13	265	133	3
AB5	2	323	1358	15	253	148	
	3	177	1411	14	271	127	4
BF1	1	500	1245		406		
BF2	1	500	1245		406		
BF3	1	500	1245		406		
BF4	1	500	1245		406		
BF5	1	500	1245		406		
BB1	2	56	1386	15	255	151	
	3	444	1441	12	273	131	3
BB2	2	82	1380	16	251	155	
	3	418	1439	12	273	131	3
BB3	2	351	1384	14	234	172	
	3	149	1439	14	260	143	3
BB4	2	278	1376	15	253	153	
	3	222	1433	13	270	133	3
BB5	2	144	1380	14	256	150	
	3	356	1437	13	273	131	3
CF	1	500	1245		406		
CB	1	500	1245		406		

Table B.4: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-6}$ and $P_k \propto n_k p_k$

Condition	NC	N	LL	SD	C1	C2	C3
AF1	1	355	1230		401		
	2	9	1279	48	369	32	
	3	136	1419	12	269	129	5
AF2	1	500	1229		401		
AF3	1	500	1228		401		
AF4	1	491	1233		401		
	2	9	1249	3	397	4	
AF5	2	13	1356	16	249	152	
	3	487	1411	14	270	128	3
AB1	3	500	1417	12	270	127	4
AB2	2	1	1381		277	124	
	3	499	1422	12	271	125	5
AB3	2	4	1365	20	226	175	
	3	496	1422	12	156	141	4
AB4	2	28	1348	18	255	146	
	3	472	1403	15	266	133	2
AB5	2	13	1356	16	249	152	
	3	487	1411	14	270	128	3
BF1	1	499	1245		406		
	2	1	1259		405	1	
BF2	1	500	1245		406		
BF3	1	497	1245		406		
	2	3	1259	1	405	1	
BF4	1	500	1245		406		
BF5	1	499	1245		406		
	2	1	1259		405	1	
BB1	3	500	1440	12	272	131	3
BB2	2	10	1383	13	249	163	
	3	499	1439	12	273	130	3
BB3	2	13	1384	12	227	179	
	3	487	1441	13	260	143	3
BB4	2	4	1380	19	274	132	
	3	496	1433	13	269	134	3
BB5	2	3	1370	18	254	152	
	3	497	1436	13	273	131	3
CF	1	499	1245		406		
		1	1259		405	1	
CB	1	500	1245		406		

Table B.5: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-2}$ and $P_k \propto n_k p_k$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	3	481	1417	12	270	127	4				
	4	19	1422	11	269	121	9	2			
AF2	3	480	1422	12	270	126	5				
	4	19	1425	8	279	110	10	2			
	5	1	1411		283	115	1	1	1		
AF3	3	484	1422	12	256	141	4				
	4	16	1426	14	250	144	5	2			
AF4	3	495	1401	14	266	133	3				
	4	5	1401	8	273	125	2	1			
AF5	3	492	1411	14	269	129	3				
	4	8	1414	12	272	124	4	1			
AB1	3	481	1417	12	270	127	4				
	4	19	1422	11	269	121	9	2			
AB2	3	476	1420	12	273	123	4				
	4	24	1421	16	270	120	9	2			
AB3	3	484	1422	12	256	141	4				
	4	16	1426	14	250	144	5	2			
AB4	3	495	1403	15	266	133	3				
	4	5	1408	15	266	132	2	1			
AB5	3	492	1411	14	269	129	3				
	4	8	1414	12	272	124	4	1			
BF1	3	496	1440	12	273	130	3				
	4	4	1446	5	271	132	2				
BF2	3	497	1438	13	272	130	3				
	4	3	1443	15	271	130	3	1			
BF3	3	496	1441	13	260	143	3				
	4	4	1440	8	263	137	5	1			
BF4	3	496	1433	13	270	134	3				
	4	4	1436	13	265	137	3	1			
BF5	3	497	1436	13	273	131	3				
	4	3	1447	17	270	126	7	3			
BB1	3	493	1440	12	272	131	3				
	4	7	1435	9	277	121	7	1			
BB2	3	497	1439	12	273	130	3				
	4	3	1442	12	273	127	5	1			
BB3	3	496	1441	13	260	143	3				
	4	4	1440	8	263	137	5	1			
BB4	3	496	1433	13	270	134	3				
	4	4	1436	13	265	137	3	1			
BB5	3	496	1436	13	271	133	3				
	4	4	1435	6	266	130	8	3			
CF	1	18	1245	0	406						
	2	478	1259	2	402	4					
	3	4	1262	4	367	38	2				
CB	1	22	1245	0	406						
	2	476	1259	2	403	3					
	3	2	1262	1	402	2	2				

Table B.6: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-15}$ and $P_k \propto n_k p_k^2$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6
AF1	1	500	1230		401					
AF2	1	500	1229		401					
AF3	1	500	1228		401					
AF4	1	500	1233		401					
AF5	1	500	1231		401					
AB1	4	500	1584	9	185	111	97	8		
AB2	4	500	1596	8	179	119	92	10		
AB3	5	500	1682	10	144	108	91	53	5	
AB4	3	500	1474	7	257	137	7			
AB5	4	500	1569	11	195	107	93	7		
BF1	1	500	1245		406					
BF2	1	500	1245		406					
BF3	1	500	1245		406					
BF4	1	500	1245		406					
BF5	1	500	1245		406					
BB1	4	500	1621	8	189	111	96	11		
BB2	5	500	1712	27	149	100	78	56	23	
BB3	6	500	1766	10	136	96	87	51	29	7
BB4	4	500	1631	8	194	116	89	7		
BB5	3	500	1632	8	197	113	89	7		
CF	1	500	1245		406					
CB	2	500	1364	6	302	104				

Table B.7: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-11}$ and $P_k \propto n_k p_k^2$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6
AF1	3	500	1485	6	256	134	12			
AF2	1	500	1229		401					
AF3	3	500	1488	6	250	140	12			
AF4	3	500	1474	6	257	137	7			
AF5	3	500	1480	7	258	134	10			
AB1	4	500	1584	9	185	111	97	8		
AB2	4	500	1596	8	179	119	92	10		
AB3	5	500	1682	10	144	108	91	53	5	
AB4	3	500	1474	7	257	137	7			
AB5	4	500	1569	11	195	107	93	7		
BF1	4	500	1637	7	192	113	94	7		
BF2	4	500	1632	12	194	112	94	7		
BF3	3	500	1506	6	254	142	11			
BF4	3	500	1497	6	281	116	8			
BF5	3	500	1503	6	261	134	11			
BB1	5	500	1693	13	165	106	76	54	5	
BB2	5	435	1708	26	151	100	78	55	22	
	6	65	1768	11	137	96	63	58	50	2
BB3	6	500	1766	10	136	96	87	51	30	7
BB4	4	500	1631	8	194	116	89	7		
BB5	5	500	1690	9	189	107	56	49	5	
CF	2	500	1365	6	300	106				
CB	2	500	1364	6	301	105				

Table B.8: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-8}$ and $P_k \propto n_k p_k^2$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6
AF1	3	500	1485	6	256	134	12			
AF2	3	139	1488	6	259	130	12			
	4	361	1580	10	203	103	83	12		
AF3	3	500	1488	6	250	140	12			
AF4	3	500	1474	6	257	137	7			
AF5	3	500	1480	7	258	134	10			
AB1	4	500	1584	9	185	111	97	8		
AB2	4	500	1611	9	178	124	92	7		
AB3	5	500	1682	10	144	108	91	53	5	
AB4	3	500	1474	7	257	137	7			
AB5	4	500	1569	11	195	107	93	7		
BF1	4	500	1637	7	192	113	94	7		
BF2	4	500	1632	12	194	112	94	7		
BF3	3	500	1506	6	254	142	11			
BF4	4	500	1630	8	194	116	89	7		
BF5	4	500	1632	8	197	112	89	7		
BB1	5	500	1693	13	165	106	76	54	5	
BB2	5	1	1738		141	94	78	56	37	
	6	499	1768	12	135	98	63	58	50	2
BB3	6	500	1766	10	136	96	87	51	30	7
BB4	5	500	1680	11	180	113	59	50	4	
BB5	5	500	1690	9	189	107	56	49	5	
CF	2	497	1364	6	300	106				
	3	3	1372	5	315	89	2			
CB	2	489	1365	6	300	106				
	3	11	1380	7	315	80	11			

Table B.9: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-6}$ and $P_k \propto n_k p_k^2$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6
AF1	4	500	1603	10	182	123	91	6		
	4	141	1582	10	200	106	84	12		
AF2	5	359	1635	15	167	115	70	46	4	
AF3	3	249	1488	5	250	139	12			
	4	251	1622	11	167	125	104	5		
AF4	3	500	1474	6	257	131	7			
AF5	3	296	1480	6	258	134	10			
	4	204	1589	14	185	125	86	6		
AB1	4	500	1584	9	185	111	97	8		
AB2	4	500	1611	9	178	124	92	7		
AB3	5	500	1668	10	139	106	89	61	6	
AB4	3	500	1474	7	257	137	7			
AB5	4	500	1569	11	195	107	93	7		
BF1	4	500	1637	7	192	113	94	7		
BF2	4	500	1632	12	194	112	94	7		
BF3	3	500	1506	6	254	142	11			
BF4	4	500	1630	8	194	116	89	7		
BF5	4	500	1632	8	197	112	89	7		
BB1	5	500	1693	13	165	106	76	54	5	
BB2	6	500	1768	12	135	98	63	58	50	2
BB3	6	500	1766	10	136	96	87	51	30	7
BB4	5	500	1680	11	180	113	59	50	4	
BB5	5	500	1689	9	189	107	56	49	5	
CF	2	321	1365	6	303	104				
	3	179	1374	6	313	90	3			
CB	2	241	1364	6	303	103				
	3	259	1384	13	280	97	29			

Table B.10: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-2}$ and $P_k \propto n_k p_k^2$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	4	337	1571	9	204	104	84	10			
	5	158	1585	12	199	103	80	17	3		
	6	5	1581	11	207	100	74	16	3	2	
AF2	4	293	1611	8	178	125	92	7			
	5	205	1627	13	172	121	74	31	4		
	6	2	1620	8	179	126	80	14	2	1	
AF3	5	491	1683	10	144	108	91	53	5		
	6	8	1684	6	151	106	85	52	5	2	
	7	1	1684		148	91	89	66	4	2	1
AF4	3	479	1474	6	257	137	7				
	4	21	1477	6	256	138	5	2			
AF5	4	475	1568	11	195	107	93	7			
	5	24	1573	8	197	104	91	7	2		
	6	1	1583		195	108	82	14	1	1	
AB1	4	17	1584	9	187	109	98	8			
	5	474	1608	11	179	108	83	29	3		
	6	9	1610	13	177	105	79	36	2	2	
AB2	5	495	1637	14	173	113	70	43	3		
	6	5	1644	13	164	117	66	50	3	1	
AB3	5	475	1681	11	143	107	91	54	5		
	6	24	1684	10	144	103	83	56	13	3	
	7	1	1694		131	101	89	72	5	2	1
AB4	3	477	1474	7	257	137	7				
	4	23	1476	7	259	135	5	2			
AB5	4	475	1568	11	195	107	93	7			
	5	24	1573	8	197	104	91	7	2		
	6	1	1583		195	108	82	14	1	1	
BF1	6	500	1768	11	136	97	63	58	51	2	
BF2	5	497	1681	11	192	98	61	52	3		
	6	3	1675	4	191	103	57	52	1	1	
BF3	6	496	1795	10	116	97	90	59	39	6	
	7	4	1798	9	114	98	89	59	38	7	1
BF4	4	498	1630	8.1	194	116	89	7			
	5	2	1630	5	191	115	92	8	1		
BF5	5	498	1690	9	189	107	56	49	5		
	6	2	1694	2	189	105	54	51	7	1	
BB1	5	499	1692	13	165	106	76	54	5		
	6	1	1716		161	99	86	54	5	1	
BB2	6	499	1769	10	133	99	63	58	51	2	
	7	1	1776		131	100	62	57	54	1	1
BB3	6	497	1795	10	116	97	89	58	40	6	
	7	3	1799	7	118	95	89	58	41	5	1
BB4	5	498	1680	10	179	114	60	49	5		
	6	2	1675	0	180	111	59	50	6	1	
BB5	5	498	1689	10	189	107	57	49	5		
	6	2	1691	10	196	101	55	48	7	1	
CF	3	372	1383	12	282	95	29				
	4	128	1399	9	207	132	65	2			
CB	3	458	1380	11	292	92	22				
	4	42	1402	11	202	135	68	1			

Table B.11: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-15}$ and $P_k \propto n_k p_k^3$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	4	500	1629	4	170	134	91	7			
AF2	4	500	1633	4	170	134	91	7			
AF3	4	500	1623	6	155	122	111	13			
AF4	3	500	1487	3	270	123	8				
AF5	4	500	1622	5	170	137	87	7			
AB1	5	500	1669	8	140	112	77	66	7		
AB2	5	500	1724	5	116	106	77	72	31		
AB3	5	500	1706	6	130	108	85	69	8		
AB4	4	500	1584	5	179	121	94	6			
AB5	4	500	1622	4	170	138	87	7			
BF1	4	500	1658	4	180	125	94	8			
BF2	4	500	1657	4	178	126	95	7			
BF3	4	500	1678	4	165	122	108	11			
BF4	4	500	1656	5	180	128	87	10			
BF5	4	500	1656	4	183	126	89	8			
BB1	6	500	1807	5	118	106	68	59	52	3	
BB2	6	500	1785	5	130	107	62	54	46	8	
BB3	7	500	1848	6	109	85	80	55	50	19	7
BB4	6	500	1783	6	117	107	74	54	50	4	
BB5	6	500	1797	5	119	108	71	53	50	5	
CF	2	500	1384	3	277	129					
CB	3	500	1435	5	196	138	72				

Table B.12: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-11}$ and $P_k \propto n_k p_k^3$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	4	500	1629	4	170	134	91	7			
AF2	4	500	1633	4	170	134	91	7			
AF3	4	500	1644	4	163	128	102	8			
AF4	3	500	1487	3	270	123	8				
AF5	4	500	1622	5	170	137	87	7			
AB1	5	500	1669	8	140	112	77	66	7		
AB2	5	129	1724	5	116	106	77	71	31		
	6	371	1743	6	115	106	74	68	37	2	
AB3	5	500	1706	6	130	108	85	69	8		
AB4	4	500	1584	5	179	121	94	6			
AB5	4	500	1622	4	170	138	87	7			
BF1	4	500	1658	4	180	125	94	8			
BF2	4	500	1657	4	178	126	95	7			
BF3	4	500	1685	7	171	113	93	29			
BF4	4	500	1656	5	180	128	87	10			
BF5	4	500	1656	4	183	126	89	8			
BF1	6	500	1807	5	118	106	68	59	52	3	
BF2	6	500	1785	5	130	107	62	54	46	8	
BF3	7	500	1848	6	109	85	80	55	50	19	7
BF4	6	500	1783	6	117	107	74	54	50	4	
BF5	6	500	1797	5	119	108	71	53	50	5	
CF	3	500	1435	5	196	137	73				
CB	3	500	1435	5	198	135	73				

Table B.13: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-8}$ and $P_k \propto n_k p_k^3$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	4	500	1629	4	170	134	91	7			
AF2	4	500	1633	4	170	134	91	7			
AF3	4	500	1644	4	163	128	102	8			
AF4	3	500	1487	3	270	123	8				
AF5	4	500	1604	5	179	120	93	9			
AB1	5	500	1669	8	140	112	77	66	7		
AB2	6	500	1743	6	115	106	74	68	37	2	
AB3	5	500	1706	6	130	108	85	69	8		
AB4	4	500	1584	5	179	121	94	6			
AB5	4	500	1622	4	170	138	87	7			
BF1	4	500	1658	4	180	125	94	8			
BF2	4	500	1657	4	178	126	95	7			
BF3	4	87	1684	6	171	113	93	28			
	5	413	1761	7	159	91	83	64	9		
BF4	5	500	1714	4	165	120	61	55	6		
BF5	4	500	1656	4	183	126	89	8			
BB1	6	500	1807	5	118	106	68	59	52	3	
BB2	6	500	1785	5	130	107	62	54	46	8	
BB3	7	500	1848	6	109	85	80	55	50	19	7
BB4	6	500	1783	6	117	107	74	54	50	4	
BB5	6	500	1797	5	119	108	71	53	50	5	
CF	3	500	1435	5	199	135	73				
CB	3	500	1435	5	197	136	73				

Table B.14: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-6}$ and $P_k \propto n_k p_k^3$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7
AF1	4	500	1629	4	170	134	91	7			
AF2	4	500	1633	4	170	134	91	7			
AF3	5	500	1718	5	145	108	80	61	7		
AF4	3	500	1487	3	270	123	8				
AF5	4	500	1604	5	179	120	93	9			
AB1	5	500	1669	8	140	112	77	66	7		
AB2	6	500	1743	6	115	106	74	68	37	2	
AB3	5	500	1706	6	130	108	85	69	8		
AB4	4	500	1584	5	179	121	94	6			
AB5	4	500	1622	4	170	138	87	7			
BF1	4	500	1658	4	180	125	94	8			
BF2	4	500	1657	4	178	126	95	7			
BF3	5	500	1764	4	158	93	85	63	8		
BF4	5	500	1714	4	165	120	61	55	6		
BF5	5	500	1718	4	174	116	57	53	6		
BB1	6	500	1773	12	135	107	69	59	32	4	
BB2	6	500	1785	5	130	107	62	54	46	8	
BB3	7	500	1848	6	109	85	80	55	50	19	7
BB4	6	500	1783	6	117	107	74	54	50	4	
BB5	6	500	1797	5	119	108	71	53	50	5	
CF	3	497	1435	5	200	132	73				
	4	3	1442	2	179	158	68	1			
CB	3	499	1435	5	194	140	73				
	4	1	1447		178	159	68	1			

Table B.15: The results of collapsed Gibbs sampling with CRP at $\alpha = 10^{-2}$ and $P_k \propto n_k p_k^3$

Condition	NC	N	LL	SD	C1	C2	C3	C4	C5	C6	C7	c8
AF1	5	215	1657	9	161	129	71	37	3			
	6	283	1691	8	140	120	66	56	17	3		
	7	2	1686	18	144	116	65	59	15	2	1	
AF2	5	497	1681	5	155	123	70	50	3			
	6	3	1680	9	158	123	62	54	3	1		
AF3	6	445	1767	8	113	89	77	64	52	6		
	7	54	1787	7	102	93	75	70	41	17	2	
	8	1	1786		101	99	75	66	40	17	2	1
AF4	4	442	1584	6	179	122	94	6				
	5	58	1586	6	180	121	93	5	2			
AF5	4	177	1622	5	170	138	87	7				
	5	321	1633	6	169	137	81	12	2			
	6	2	1638	5	159	150	79	12	1	1		
AB1	5	44	1673	7	143	124	68	59	7			
	6	453	1692	8	139	120	66	56	17	3		
	7	3	1686	10	139	123	65	57	14	2	1	
AB2	6	500	1743	6	115	105	73	68	37	2		
AB3	6	484	1734	7	129	105	76	60	26	5		
	7	16	1740	6	128	105	75	60	26	6	2	
AB4	4	436	1584	5	179	121	94	6				
	5	64	1588	7	178	122	93	6	2			
AB5	4	158	1622	5	169	138	87	7				
	5	340	1633	6	169	137	81	12	2			
	6	2	1630	8	178	129	82	12	1	1		
BF1	6	499	1780	5	143	101	60	54	44	4		
	7	1	1773		141	107	58	53	45	1	1	
BF2	6	500	1807	5	118	106	67	59	52	3		
BF3	7	498	1843	6	110	83	78	53	48	32	2	
	8	2	1845	10	108	82	80	54	52	31	1	1
BF4	6	499	1783	6	117	108	73	54	50	4		
	7	1	1788		116	111	72	53	49	4	1	
BF5	5	499	1718	4	174	116	57	53	6			
	6	1	1725		159	126	58	57	4	2		
BB1	6	495	1780	20	135	107	68	57	36	3		
	7	5	1793	16	130	101	66	62	34	12	2	
BB2	7	500	1827	6	117	105	61	54	37	30	2	
BB3	7	500	1848	6	109	85	80	55	50	19	7	
BB4	6	500	1783	6	117	107	74	54	50	4		
BB5	6	499	1797	5	119	108	71	53	50	5		
	7	1	1810		116	107	73	52	50	6	2	
CF	3	23	1434	4	203	129	73					
	4	477	1438	5	185	149	70	1				
CB	3	27	1435	4	195	137	74					
	4	473	1438	5	185	149	70	1				

Appendix C

JAGS Model Specifications

G_0 model

```
model{
  for(i in 1:N){
    sr[i] ~ dgamma(shape[i], rate[i])
    log.mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    mu[i] <- exp(log.mu[i])
    tau[i] <- height[i]*invsigmasq
    s[i] <- 1/sqrt(tau[i])
    shape[i] <- mu[i]*mu[i]/(s[i]*s[i])
    rate[i] <- mu[i]/(s[i]*s[i])
    sr.new[i] ~ dgamma(shape[i], rate[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0.1, 0.001)
  }
    invsigmasq ~ dgamma(0.001, 0.001)
    sigma <- sqrt(1/invsigmasq)
}
```

G_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dgamma(shape[i], rate[i])
    log.mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    mu[i] <- exp(log.mu[i])
    tau[i] <- height[i]*invsigmasq[marker[i]]
    s[i] <- 1/sqrt(tau[i])
    shape[i] <- mu[i]*mu[i]/(s[i]*s[i])
    rate[i] <- mu[i]/(s[i]*s[i])
    sr.new[i] ~ dgamma(shape[i], rate[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
    invsigmasq[i] ~ dgamma(0.001, 0.001)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
}
```

G_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dgamma(shape[i], rate[i])
    log.mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    mu[i] <- exp(log.mu[i])
    tau[i] <- height[i]*invsigmasq[marker[i]]
    s[i] <- 1/sqrt(tau[i])
    shape[i] <- mu[i]*mu[i]/(s[i]*s[i])
    rate[i] <- mu[i]/(s[i]*s[i])
    sr.new[i] ~ dgamma(shape[i], rate[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(mb0, pb0)
    b1[i] ~ dnorm(mb1, pb1)
    invsigmasq[i] ~ dgamma(alpha, beta)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.000001, 0.001)
  alpha ~ dgamma(0.001, 0.001)
  beta ~ dgamma(0.001, 0.001)
}
```

LN_0 model

```
model{
  for(i in 1:N){
    sr[i] ~ dlnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq
    sr.new[i] ~ dlnorm(mu[i], tau[i])
  }
  ##priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0.1, 0.001)
    }
  invsigmasq ~ dgamma(0.001, 0.001)
  sigma <- sqrt(1/invsigmasq)
}
```

LN_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dlnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq[marker[i]]
    sr.new[i] ~ dlnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.000001)
    invsigmasq[i] ~ dgamma(0.001, 0.001)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
}
```

LN_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dlnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq[marker[i]]
    sr.new[i] ~ dlnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(mb0, pb0)
    b1[i] ~ dnorm(mb1, pb1)
    invsigmasq[i] ~ dgamma(alpha, beta)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.001, 0.001)
  alpha ~ dgamma(0.001, 0.001)
  beta ~ dgamma(0.001, 0.001)
}
```

MLN_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dlnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
    d[i] ~ dbern(pp)
    sr.new[i] ~ dlnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
    t0[i] ~ dgamma(0.001, 0.001)
    t1[i] ~ dgamma(0.001, 0.001)
    sigma0[i] <- 1/sqrt(t0[i])
    sigma1[i] <- 1/sqrt(t1[i])
  }
  pp ~  dunif(0,1)
}
```

MLN_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dlnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
    d[i] ~ dbern(pp)
    sr.new[i] ~ dlnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(mb0, pb0)
    b1[i] ~ dnorm(mb1, pb1)
    t0[i] ~ dgamma(alpha.t0, beta.t0)
    t1[i] ~ dgamma(alpha.t1, beta.t1)
    sigma0[i] <- 1/sqrt(t0[i])
    sigma1[i] <- 1/sqrt(t1[i])
```

```
  }
  pp ~  dunif(0,1)
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.001, 0.001)
  alpha.t0 ~ dgamma(0.001, 0.001)
  beta.t0 ~ dgamma(0.001, 0.001)
  alpha.t1 ~ dgamma(0.001, 0.001)
  beta.t1 ~ dgamma(0.001, 0.001)
}
```

N_0 model

```
model{
  for(i in 1:N){
    sr[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq
    sr.new[i] ~ dnorm(mu[i], tau[i])
  }
  ##priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
  }
  invsigmasq ~ dgamma(0.001, 0.001)
  sigma <- sqrt(1/invsigmasq)
}
```

N_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq[marker[i]]
    sr.new[i] ~ dnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
    invsigmasq[i] ~ dgamma(0.001, 0.001)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
}
```

N_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]*invsigmasq[marker[i]]
    sr.new[i] ~ dnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(mb0, pb0)
    b1[i] ~ dnorm(mb1, pb1)
    invsigmasq[i] ~ dgamma(alpha, beta)
  sigma[i] <- sqrt(1/invsigmasq[i])
```

```
  }
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.001, 0.001)
  alpha ~ dgamma(0.001, 0.001)
  beta ~ dgamma(0.001, 0.001)
}
```

MN_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
    d[i] ~ dbern(pp)
    sr.new[i] ~ dnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
    t0[i] ~ dgamma(0.001, 0.001)
    t1[i] ~ dgamma(0.001, 0.001)
    sigma0[i] <- 1/sqrt(t0[i])
    sigma1[i] <- 1/sqrt(t1[i])
  }
  pp ~  dunif(0,1)
}
```

MN_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
    d[i] ~ dbern(pp)
    sr.new[i] ~ dnorm(mu[i], tau[i])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(mb0, pb0)
    b1[i] ~ dnorm(mb1, pb1)
    t0[i] ~ dgamma(alpha.t0, beta.t0)
    t1[i] ~ dgamma(alpha.t1, beta.t1)
    sigma0[i] <- 1/sqrt(t0[i])
    sigma1[i] <- 1/sqrt(t1[i])
  }
  pp ~  dunif(0,1)
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.001, 0.001)
  alpha.t0 ~ dgamma(0.001, 0.001)
  beta.t0 ~ dgamma(0.001, 0.001)
  alpha.t1 ~ dgamma(0.001, 0.001)
  beta.t1 ~ dgamma(0.001, 0.001)
}
```

T_0 model

```
model{
  for(i in 1:N){
    sr[i] ~ dt(mu[i], tau[i], nu[marker[i]])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    invssq[i] <- height[i]*invsigmasq
    tau[i] <- invssq[i]*nu[marker[i]]/(nu[marker[i]]-2)
    sr.new[i] ~ dt(mu[i], tau[i], nu[marker[i]])
  }
  ##priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.001)
    b1[i] ~ dnorm(0, 0.001)
    log_nu[i] ~ dunif(1.1, 5)
    nu[i] <- exp(log_nu[i])
  }
  invsigmasq ~ dgamma(0.001, 0.001)
  sigma <- sqrt(1/invsigmasq)
}
```

T_1 model

```
model{
  for(i in 1:N){
    sr[i] ~ dt(mu[i], tau[i], nu[marker[i]])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
    invssq[i] <- height[i]*invsigmasq[marker[i]]
    tau[i] <- invssq[i]*nu[marker[i]]/(nu[marker[i]]-2)
    sr.new[i] ~ dt(mu[i], tau[i], nu[marker[i]])
  }
  #priors
  for(i in 1:nMarkers){
    b0[i] ~ dnorm(0, 0.000001)
    b1[i] ~ dnorm(0, 0.000001)
    log_nu[i] ~ dunif(1.1, 5)
    nu[i] <- exp(log_nu[i])
    invsigmasq[i] ~ dgamma(0.000001, 0.000001)
    sigma[i] <- sqrt(1/invsigmasq[i])
  }
}
```

T_2 model

```
model{
  for(i in 1:N){
    sr[i] ~ dt(mu[i], tau[i], nu[marker[i]])
    mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
        invssq[i] <- height[i]*invsigmasq[marker[i]]
        tau[i] <- invssq[i]*nu[marker[i]]/(nu[marker[i]]-2)
        sr.new[i] ~ dt(mu[i], tau[i], nu[marker[i]])
  }
  #priors
  for(i in 1:nMarkers){
        b0[i] ~ dnorm(mb0, pb0)
        b1[i] ~ dnorm(mb1, pb1)
        log_nu[i] ~ dunif(1.1, 5)
        nu[i] <- exp(log_nu[i])
        invsigmasq[i] ~ dgamma(alpha, beta)
        sigma[i] <- sqrt(1/invsigmasq[i])
  }
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
```

```
  pb1 ~ dgamma(0.001, 0.001)
  alpha ~ dgamma(0.001, 0.001)
  beta ~ dgamma(0.001, 0.001)
}
```

MT_1 model

```
model{
  for(i in 1:N){
        sr[i] ~ dt(mu[i], tau[i], nu[i])
        mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
        tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
        nu[i] <- (1 - d[i])*pnu1[marker[i]] + d[i]*pnu2[marker[i]]
        d[i] ~ dbern(p)
  }
  #priors
  for(i in 1:nMarkers){
        b0[i] ~ dnorm(0, 0.001)
        b1[i] ~ dnorm(0, 0.001)
        t0[i] ~ dgamma(0.001, 0.001)
        t1[i] ~ dgamma(0.001, 0.001)
        sigma0[i] <- 1/sqrt(t0[i])
        sigma1[i] <- 1/sqrt(t1[i])
        log_nu1[i] ~ dunif(0, 5)
        pnu1[i] <- exp(log_nu1[i])
        log_nu2[i] ~ dunif(0, 5)
        pnu2[i] <- exp(log_nu2[i])
  }
  p ~  dunif(0,1)
}
```

MT_2 model

```
model{
  for(i in 1:N){
        sr[i] ~ dt(mu[i], tau[i], nu[i])
        mu[i] <- b0[marker[i]] + b1[marker[i]]*LUS[i]
        tau[i] <- height[i]/(1/t0[marker[i]] + d[i]/t1[marker[i]])
        nu[i] <- (1 - d[i])*pnu1[marker[i]] + d[i]*pnu2[marker[i]]
        d[i] ~ dbern(pp)
  }
  #priors
  for(i in 1:nMarkers){
        b0[i] ~ dnorm(mb0, pb0)
        b1[i] ~ dnorm(mb1, pb1)
        t0[i] ~ dgamma(alpha.t0, beta.t0)
        t1[i] ~ dgamma(alpha.t1, beta.t1)
        sigma0[i] <- 1/sqrt(t0[i])
        sigma1[i] <- 1/sqrt(t1[i])
        log_nu1[i] ~ dunif(1.1, 5)
        pnu1[i] <- exp(log_nu1[i])
        log_nu2[i] ~ dunif(1.1, 5)
        pnu2[i] <- exp(log_nu2[i])
  }
  pp ~  dunif(0,1)
  mb0 ~ dnorm(0, 0.001)
  mb1 ~ dnorm(0, 0.001)
  pb0 ~ dgamma(0.001, 0.001)
  pb1 ~ dgamma(0.001, 0.001)
  alpha.t0 ~ dgamma(0.001, 0.001)
  beta.t0 ~ dgamma(0.001, 0.001)
  alpha.t1 ~ dgamma(0.001, 0.001)
  beta.t1 ~ dgamma(0.001, 0.001)
}
```

Appendix D

R codes for Infinite mixture models

The same R codes are publicly available in github.com (URL: https://github.com/sampathf73/InfMixtures).

```
Alpha <- 1e-15
locus <- "D2S1338"
locnum <- 4
Power <- 3
mypath <- "./"

progsim <- function (fn, Alpha=Alpha, locus=locus, locnum=locnum,
                     Power=Power, mypath=mypath) {
  Alpha <- Alpha
  locus <- locus
  locnum <- locnum
  Power <- Power
  KKvalue <- 1   # Number of active components KK
  filenumber <- fn
  niter <- 5000

  Resultfile1 <- paste(mypath, "/P", Power, "CS-Loc-",
                 locnum, "-Alp-", Alpha, "-fn", filenumber,
                 "A.csv", sep = "")
  Resultfile2 <- paste(mypath, "/P", Power, "LL-Loc-",
                 locnum, "-Alp-", Alpha, "-fn", filenumber,
                 "A.csv", sep = "")
  Resultfile3 <- paste(mypath, "/P", Power, "SumLL-Loc-",
                 locnum, "-Alp-", Alpha, "-fn", filenumber,
                 "A.csv", sep = "")
  library(Matrix)
  library(mnormt)

  ###################################################################

  setClass("GaussianLR", representation(hh = "list", GLR = "list"),
  contains="list")

GaussianLR <- function(hh) {
  new("GaussianLR",
  # Total number of observations N
    hh$NN <- as.integer(hh$nn),
  # Cavariates matrix (Nx2)
    hh$XX <- as.matrix(hh$xx),
  # A column matrix of responses
    hh$YY <- as.matrix(hh$yy),
  # Shape parameter (a) of Inverse Gamma prior
```

```
    hh$IGA <- as.numeric(hh$IGa),
  # Scale parameter (b) of Inverse Gamma prior
    hh$IGB <- as.numeric(hh$IGb),
  # Mean vector of betas (2X1)
    hh$MuBeta <- as.matrix(hh$mubeta),
  # Scale matrix of betas (2X2)
    hh$ScaleBeta <- as.matrix(hh$ScaleBeta),

  # Total number of observations N
    GLR.NN <- as.integer(hh$NN),
  # Cavariates matrix (Nx2)
    GLR.XX <- as.matrix(hh$XX),
  # A column matrix of responses (Nx1)
    GLR.YY <- as.matrix(hh$YY),
  # Shape parameter (a) of Inverse Gamma prior
    GLR.IGA <- as.numeric(hh$IGA),
  # Scale parameter (b) of Inverse Gamma prior
    GLR.IGB <- as.numeric(hh$IGB),
  # Prior mean vector of betas (2X1)
    GLR.MuBeta <- as.matrix(hh$MuBeta),
  # Prior scale matrix of betas (2X2)
    GLR.ScaleBeta <- as.matrix(hh$ScaleBeta),
    GLR.LL <- as.numeric(0),

  if (hh$NN != 0) {
    stop("Number of observations: nn must be zero")
  } else {
     # Posterior shape parameter (a) of Inverse Gamma
       GLR.IGATilda <- hh$IGA
     # Posterior scale parameter (b) of Inverse Gamma
       GLR.IGBTilda <- hh$IGB
     # Posterior mean vector of betas (2X1)
       GLR.MuBetaTilda <- hh$MuBeta
     # Posterior scale matrix of betas (2X2)
       GLR.ScaleBetaTilda <- hh$ScaleBeta
     },


GLR <- as.list(list(NN = GLR.NN, XX = GLR.XX, YY = GLR.YY,
  IGA = GLR.IGA, IGB = GLR.IGB,
  MuBeta = GLR.MuBeta,
  ScaleBeta = GLR.ScaleBeta,
  IGATilda = GLR.IGATilda,
  IGBTilda = GLR.IGBTilda,
  MuBetaTilda = GLR.MuBetaTilda,
  ScaleBetaTilda = GLR.ScaleBetaTilda,
  LL = GLR.LL)))
  GLR
} # End of GaussianLR


logpredictive <- function(GLR, XTilda, YTilda) {
  # Calculates the log of the predictive probability of a datum
  # ll = logpredictive(X, A, B, Mu, Beta, Xstar)
  # log predictive probability of YTilda given XTilda
  #   and other data items in the component
  # log p(YTilda|XTilda, (x_1, y_1),...,(x_n, y_n))

  # Degrees of freedom of MVSt distribution
    df <- 2 * GLR$IGATilda
  # Mean of MVSt distribution
    Mean <- XTilda %*% GLR$MuBetaTilda
  # Scale parameter of MVSt distribution
  # Var_Cov = Scale * df / (df - 2)
    ScaleTilda <- (GLR$IGBTilda / GLR$IGATilda)*
                  (1 + XTilda %*% GLR$ScaleBetaTilda %*% t(XTilda))

  ll <- dmt(YTilda, Mean, ScaleTilda, df, log=TRUE)
  ll
} # End of logpredictive
```

```
CLL <- function(GLR) {
  # Calculates the log likelihood of a cluster
  CLL <- 0
  for (item in 1:nrow(GLR$YY))
     CLL <- CLL + logpredictive(GLR, matrix(c(GLR$XX[item,]),1,2),
     matrix(c(GLR$YY[item,]),1,1))
  CLL
} # End of CLL


additem2D <- function(GLR, xx, yy) {
  # Adds data item to a component
  # GLR = additem2D(GLR, xx, yy)
  # Ddds datum (xx, yy) into component GLR
  # xx is a 1x2 matrix in the form (1 x)

  if (GLR$NN == 0) { # Add data to an empty component
   # Number of observations in the component
     GLR$NN <- nrow(yy)
   # Observed covariates of the component
     GLR$XX <- matrix(rbind(xx), ncol=2)
   # Observed responses of the component
     GLR$YY <- matrix(rbind(yy), ncol=1)
   } else { # Add data to non-empty components
     # Number of observations
       GLR$NN <- GLR$NN + nrow(yy)
     # Observed covariates of the component
       GLR$XX <- matrix(rbind(GLR$XX, xx), ncol=2)
     # Observed responses of the component
       GLR$YY <- matrix(rbind(GLR$YY, yy), ncol=1)
     }

  # Posterior shape parameter (a) of Inverse Gamma
    GLR$IGATilda <- GLR$IGA + 0.5 * GLR$NN
  # Posterior scale parameter (b) of Inverse Gamma
    GLR$IGBTilda <- GLR$IGB + 0.5 * as.numeric(t(GLR$YY -
                    GLR$XX %*% GLR$MuBeta)%*%
                    solve(Diagonal(GLR$NN)
                    + GLR$XX %*% GLR$ScaleBeta %*% t(GLR$XX))%*%
                    (GLR$YY - GLR$XX %*% GLR$MuBeta))
  # Posterior scale matrix of betas (2X2)
    GLR$ScaleBetaTilda <- solve(solve(GLR$ScaleBeta)
                          + t(GLR$XX) %*% GLR$XX)
  # Posterior mean vector of betas (2X1)
    GLR$MuBetaTilda <- GLR$ScaleBetaTilda %*%
                       (solve(GLR$ScaleBeta) %*%
                       GLR$MuBeta + t(GLR$XX) %*% GLR$YY)
  GLR$LL <- CLL(GLR)
  GLR
} # End of additem2D


delitem2D <- function(GLR, xx, yy) {
  # Deletes a data item from a component
  # GLR = delitem2D(GLR, xx, yy)
  # Deletes datum (xx, yy) from component GLR
  # xx is a 1x2 matrix in the form (1 x)

  if (GLR$NN == 1) {
  # Coverts to an empty component after detetion
  # Number of observations in the component
    GLR$NN <- 0
  # NO observed covariates in the component
    GLR$XX <- NA
  # No observed responses in the component
    GLR$YY <- NA
  # Posterior shape parameter (a) of Inverse Gamma
    GLR$IGATilda <- GLR$IGA
  # Posterior scale parameter (b) of Inverse Gamma
```

```
    GLR$IGBTilda <- GLR$IGB
  # Posterior mean vector of betas (2X1)
    GLR$MuBetaTilda <- GLR$MuBeta
  # Posterior scale matrix of betas (2X2)
    GLR$ScaleBetaTilda <- GLR$ScaleBeta
    GLR$LL <- 0
  } else { # Deletes a data item from a component
    index = -999
    j = 1

    while (index < 0) { # Identify the datum to delete
      if ((GLR$XX[j,2] == xx[1,2]) && (GLR$YY[j] == yy[1,1])) {
       index = j # Row number of the datum to be deleted
      }
    j = j + 1
    } # End of while

   # Number of observations in the component
     GLR$NN <- GLR$NN - 1
   # Observed covariates of the component
     GLR$XX <- matrix(GLR$XX[-index,], ncol=2)
   # Observed responses of the component
     GLR$YY <- matrix(GLR$YY[-index], ncol=1)

   # Posterior shape parameter (a) of Inverse Gamma
     GLR$IGATilda <- GLR$IGA + 0.5 * GLR$NN
   # Posterior scale parameter (b) of Inverse Gamma
     GLR$IGBTilda <- GLR$IGB
                     + 0.5 * as.numeric(t(GLR$YY - GLR$XX %*%
                     GLR$MuBeta)%*% solve(Diagonal(GLR$NN)
                     + GLR$XX %*% GLR$ScaleBeta %*%
                     t(GLR$XX) )%*% (GLR$YY
                     - GLR$XX %*% GLR$MuBeta))
   # Posterior scale matrix of betas (2X2)
     GLR$ScaleBetaTilda <- solve(solve(GLR$ScaleBeta)
                           + t(GLR$XX) %*% GLR$XX)
   # Posterior mean vector of betas (2X1)
     GLR$MuBetaTilda <- GLR$ScaleBetaTilda %*%
                        (solve(GLR$ScaleBeta) %*%
                        GLR$MuBeta + t(GLR$XX) %*% GLR$YY)
   GLR$LL <- CLL(GLR)
 } # End of else
GLR
} # End of delitem2D


DPMLR_Init <- function(KK,Alpha,GLR0,xx,yy,zz) {
  # Initialize DP mixture model
  # GLR0 empty GLR component with hh prior,
  # Active mixture components
    DPM.KK <- KK
  # Total number of observations
    DPM.NN <- nrow(yy)
  # Concentration parameter of DP prior
    DPM.Alpha <- Alpha
  # Mixture Components
    DPM.GLR <- vector(mode = "list", length = KK+1)
    DPM.XX <- xx # Covarites
    DPM.YY <- yy # Responses
    DPM.ZZ <- zz # Initial cluster assignments (between 1 and KK).
    DPM.nn <- matrix(0,1,KK) # KK number of mpty clusters
    DPM <- list(KK=DPM.KK, NN=DPM.NN, Alpha=DPM.Alpha, GLR=DPM.GLR,
    XX=DPM.XX, YY=DPM.YY, ZZ=DPM.ZZ, nn=DPM.nn)

  # Initialize mixture components
  # Component KK+1 takes care of all inactive components
    for (kk in 1:(KK+1)) {
      # Generating KK+1 number of empty clusters
        DPM$GLR[[kk]] <- GLR0
    }
```

```
  # Add data items into mixture components
    for (ii in 1:DPM$NN) {
      kk = zz[ii] # Identify the cluster index of the datum ii
      DPM$GLR[[kk]] <- additem2D(DPM$GLR[[kk]],matrix(xx[ii,],1,2),
                        matrix(yy[ii],1,1))
    # Add the datum to the cluster
      DPM$nn[[kk]] <- DPM$nn[[kk]] + 1# Cluster size increase by 1
    }
  DPM
} # End of DPMLR_Init

  temp4 <- matrix(NA,nrow=500,ncol=103)
  Resultfile4 <- paste(mypath, "/P", Power, "Gibs-Loc-",
                 locnum, "-Alp-", Alpha, "-fn", filenumber,
                 "A.csv", sep = "")

DPMLR_Gibbs <- function(DPMLR, niter, temp4) {
  # Gibbs sampler for DPMLR
  KK <- DPMLR$KK # Number of active clusters
  NN <- DPMLR$NN # Total number of data items
  Alpha <- DPMLR$Alpha # Dispersion parameter of DP prior
  GLR <- DPMLR$GLR # A vector of mixture components
  XX <- DPMLR$XX # A 2-column matrix of covariates
  YY <- DPMLR$YY # A column matrix
  ZZ <- DPMLR$ZZ # Cluster indicators
  nn <- DPMLR$nn # Number of data items in each cluster

  for (i in 1:niter) {
    # In each iteration, remove each data item from the model
    # Then add it back according to the conditional probabilities

    for (ii in 1:NN) { # iterate over data items ii
     # Remove data item xx[ii] from component GLR[kk]
       kk <- ZZ[ii] # Current component data item ii belongs to
     # Number of data items in component kk is reduced by 1
       nn[kk] <- nn[kk] - 1
       GLRtemp1 <- GLR[kk][[1]] # Component kk
     # Remove data item from component kk
       GLRtemp2 <- delitem2D(GLRtemp1, matrix(XX[ii,],1,2),
                             matrix(YY[ii],1,1))
     # Component kk after removing the data item
       GLR[kk][[1]] <- GLRtemp2

  # Delete the component if it has become empty
    if (nn[kk] == 0) {
      KK <- KK - 1 # Number of components is reduced by 1
      GLR <- GLR[-kk] # Delete the empty component
      # nn related to empty component is removed
        nn <- nn[-kk]
        idx <- which(ZZ>kk)
      #  Adjust all the indicators of the components
      #    after component kk
         ZZ[idx] <- ZZ[idx] - 1
    }

  # compute conditional probabilities pp(kk)
  #    of data item ii belonging to each component kk
  # compute probabilities in log domain, then exponentiate
  # logpredictive(N, X, Y, A, B, MuBeta, ScaleBeta,
  #    XTilda, YTilda)
  pp <- log(c((nn), Alpha))
  for (kkk in 1:(KK+1)) {
    GLRtemp3 <- GLR[kkk][[1]]
    pp[kkk] <- pp[kkk] + Power * logpredictive(GLRtemp3,
               matrix(XX[ii,],1,2), matrix(YY[ii],1,1))
  }
  pp <- exp(pp - max(pp)) # -max(p) for numerical stability
  pp <- pp / sum(pp)

  # Select component kk by sampling from conditional
```

```
  #    probabitilies
  uu <- runif(1)
  kk <- 1+sum(uu>cumsum(pp))

  # When a new active component is required
  if (kk == KK+1) {
   # Increse number of components by 1
     KK <- KK + 1
   # Number of observations in the new component
     nn[kk] <- 0
   # Increse the indicator of previous empty
   #     component by 1
     GLR[kk+1] <- GLR[kk]
  }

  # Add data item xx[ii] back into model (component GLR[kk])
    ZZ[ii] <- kk
  # Number of data items in component kk is reduced by 1
    nn[kk] <- nn[kk] + 1
    GLRtemp3 <- GLR[kk][[1]] # Component kk
  # Add data item to component kk
    GLRtemp4 <- additem2D(GLRtemp3, matrix(XX[ii,],1,2),
                matrix(YY[ii],1,1))
  # Component kk after adding the data item
    GLR[kk][[1]] <- GLRtemp4
} # End of iteration over data items


 if ((i > 3000)&(i%%4 == 0)) {
   r <- (i - 3000)/4
   temp4[r,1] = r
   temp4[r,2:(KK+1)]=nn
   sumt <- 0
   for (k in 1:KK) {
     temp4[r,(51+k)]=GLR[[k]]$LL
     sumt <- sumt + GLR[[k]]$LL
   }
   temp4[r,102]= KK
   temp4[r,103]= sumt
 }
} # End of iteration

 write.csv(temp4, Resultfile4)

  # Update DPMLR object
    DPMLR$GLR <- GLR
    DPMLR$ZZ <- ZZ
    DPMLR$nn <- nn
    DPMLR$KK <- KK
    DPMLR

} # End of Gibbs sampler

#################### APPLICATION ############################

##### NGM DATA #####

df0 <- read.csv("NGMdata.csv",header=TRUE)
df1 <- df0[,-1]
df2 <- df1
library(plyr)
df2 <- rename(df2, c("Marker" = "locus",
                     "Marker.code" = "marker" ,
                     "Stutter.Height" = "stheight",
                     "Allele" = "allele",
                     "LUS" = "LUS",
                     "Allele.Height" = "height",
                     "SR" = "sr"))
locname1 <- as.character(df2$locus[df2$marker==locnum][1])
df <- df2[df2$locus==locname1,]
sampsize <- nrow(df)
```

```
set.seed(as.integer(filenumber))
priordataIndex <- sort(sample(1:sampsize,5,replace=F))
sr1 <- df$sr
LUS1 <- df$LUS
y0 <- sr1[priordataIndex]
x0 <- matrix(c(rep(1,5), LUS1[priordataIndex]),
             ncol=2, byrow=FALSE)
lm0 <- lm(y0 ~ x0[,2])

Beta0 <- summary(lm0)[['coefficients']][1,1] # Beta0
Beta1 <- summary(lm0)[['coefficients']][2,1] # Beta1
MuBeta0 <- matrix(c(Beta0, Beta1),2,1) # (Beta0, Beta1)^T
sigma0 <- summary(lm0)[['sigma']]
cov.unscaled <- matrix(summary(lm0)[['cov.unscaled']],2,2)
ScaleBeta0 <- cov.unscaled # VBeta
# sigma0 = Residual standard error of the linear model
#        = summary(lm0)[['sigma']] = sqrt(MSE)
# Mean(Beta) = MuBeta
# Var-Cov(Beta) = VBeta * Sigmasquared

################ With n0, x0, y0 historical data: ############
# Mean(Beta)0 = MuBeta0
#  = Beta coefficients of linear model
#       based on the sample of size n0
# Mean(Beta)0 = (X0^TX0)^(-1)X0^TY0
# VBeta0 = cov.unscaled = (X0^TX0)^(-1)
# Var-Cov(Beta)0 = (sigma0)^2 * ScaleBeta0
#                = (sigma0)^2 * VBeta0
#                = sigma0^2 (X0^TX0)^(-1)
# sigmasquare ~ IG(A,B)
# A = (n0 - k)/2 and B = (n0 - k)* MSE /2

##### Posterior Predictive Distribution #############

# yTilda|y ~ MVSt(2ATilda)[XTilda*MuBetaTilda,
#             (BTilda/ATilda)*(I+XTilda*VBetaTilda*XTilda^T)]

A <- 1.5 # since k=2, n0 = 5
B <- 1.5*sigma0^2

hh0 <- list(nn=0, xx=NA, yy=NA, IGa=NA, IGb=NA,
mubeta=NA, ScaleBeta=NA)
hh0$IGa <- A
hh0$IGb <- B
hh0$mubeta <- MuBeta0
hh0$ScaleBeta <- ScaleBeta0
GLR0 <- GaussianLR(hh0)

sr2 <- sr1[-(priordataIndex)]
LUS2 <- LUS1[-(priordataIndex)]

xxall <- matrix(c(rep(1,sampsize-5), LUS2), ncol=2,
                byrow=FALSE)
yyall <- matrix(c(sr2), ncol=1)

# Initial parameters of Gibbs sampler
  KK <- KKvalue # Number of active components
#Alpha  # Dispersion parameter of Dirichlet prior
  NN <- nrow(yyall) # Total number of observations N
# Initial component indicaters
  zz <-  ceiling(runif(NN)*KK)

# Initialize DPMLR object
  DPMLR = DPMLR_Init(KK,Alpha,GLR0,xxall,yyall,zz)

temp1 <- matrix(NA,nrow=1,ncol=50)
temp2 <- matrix(NA,nrow=1,ncol=50)
temp3 <- matrix(NA,nrow=1,ncol=2)


RDatafile <- paste(mypath, "/P", Power, "Loc-",
```

```
            locnum, "-Alp-", Alpha, "-fn", filenumber,
            "A.RData", sep = "")
DPMLRtemp <- DPMLR_Gibbs(DPMLR, niter, temp4)
save(DPMLRtemp, file = RDatafile)
temp1[1,1:length(DPMLRtemp$nn)]=DPMLRtemp$nn
sum <- 0
for (j in 1:length(DPMLRtemp$nn)) {
  temp2[1,j] <- DPMLRtemp$GLR[[j]]$LL
  sum <- sum + DPMLRtemp$GLR[[j]]$LL
}
temp3[1,1] <- length(DPMLRtemp$nn)
temp3[1,2] <- sum

write.csv(temp1, Resultfile1)
write.csv(temp2, Resultfile2)
write.csv(temp3, Resultfile3)

} # End of progsim

############################################################

RunProg <- function(funnum) {
  if (funnum == 1) progsim (fn=1, Alpha=Alpha, locus=locus,
                            locnum=locnum, Power=Power,
                            mypath=mypath)
  if (funnum == 2) progsim (fn=2, Alpha=Alpha, locus=locus,
                            locnum=locnum, Power=Power,
                            mypath=mypath)
  if (funnum == 3) progsim (fn=3, Alpha=Alpha, locus=locus,
                            locnum=locnum, Power=Power,
                            mypath=mypath)
  if (funnum == 4) progsim (fn=4, Alpha=Alpha, locus=locus,
                            locnum=locnum, Power=Power,
                            mypath=mypath)
  if (funnum == 5) progsim (fn=5, Alpha=Alpha, locus=locus,
                            locnum=locnum, Power=Power,
                            mypath=mypath)
} # End of RunProg

library("doParallel")
library("foreach")
cl <- makeCluster(5)
registerDoParallel(cl)
prognum <- 1:5
foreach(i=1:length(prognum)) %dopar% RunProg(prognum[i])
stopCluster(cl)
csvfile1 <- paste("P", Power, "CS-Loc-", locnum,
            "-Alp-", Alpha, "-fn", 1, "A.csv", sep = "")
csvfile2 <- paste("P", Power, "LL-Loc-", locnum,
            "-Alp-", Alpha, "-fn", 1, "A.csv", sep = "")
csvfile3 <- paste("P", Power, "SumLL-Loc-", locnum,
            "-Alp-", Alpha, "-fn", 1, "A.csv", sep = "")
mydata1 = read.csv(csvfile1, header=T)
mydata2 = read.csv(csvfile2, header=T)
mydata3 = read.csv(csvfile3, header=T)

for (ff in 2:5) {
  csvfiletemp1 <- paste("P", Power, "CS-Loc-",
                  locnum, "-Alp-", Alpha, "-fn", ff,
                  "A.csv", sep = "")
  csvfiletemp2 <- paste("P", Power, "LL-Loc-",
                  locnum, "-Alp-", Alpha, "-fn", ff,
                  "A.csv", sep = "")
  csvfiletemp3 <- paste("P", Power, "SumLL-Loc-",
                  locnum, "-Alp-",
                  Alpha, "-fn", ff, "A.csv", sep = "")

mydatatemp1 = read.csv(csvfiletemp1, header=T)
mydatatemp2 = read.csv(csvfiletemp2, header=T)
mydatatemp3 = read.csv(csvfiletemp3, header=T)
```

```
mydata1 = rbind(mydata1, mydatatemp1)
mydata2 = rbind(mydata2, mydatatemp2)
mydata3 = rbind(mydata3, mydatatemp3)

}

Fullcsvfile1 <- paste(mypath, "/P", Power,
                "CS-Loc-", locnum, "-Alp-", Alpha,
                "-Complete-", "A.csv", sep = "")
Fullcsvfile2 <- paste(mypath, "/P", Power,
                "LL-Loc-", locnum, "-Alp-", Alpha,
                "-Complete-", "A.csv", sep = "")
Fullcsvfile3 <- paste(mypath, "/P", Power,
                "SumLL-Loc-", locnum, "-Alp-", Alpha,
                "-Complete-", "A.csv", sep = "")

write.csv(mydata1, Fullcsvfile1)
write.csv(mydata2, Fullcsvfile2)
write.csv(mydata3, Fullcsvfile3)

########################
```

Bibliography

[1] Ahsanullah, M., Kibria, B.M.G., and Shakil, M. *Normal and Student's t distributions and their applications*. Springer, Paris, 2014.

[2] Akaike, H. Statistical inference and measurement of entropy. In *Scientific inference, data analysis, and robustness*, pages 165–189, Tokyo, 1983. The Institute of Statistical Mathematics.

[3] Aldous, D., Ibragimov, I.A., and Jacod, J. Exchangeability and related topics. In Hennequin, P.L., editor, *École d'Été de Probabilités de Saint-Flour XIII-1983*, pages 1–198. Springer, Berlin, 1985.

[4] Almomani, R.and Dong, M. and Zhu, D. Object tracking via Dirichlet process-based appearance models. *Neural Computing and Applications*, 2016.

[5] Alston, C.L., Mengersen, K.L., and Pettitt, A.N. *Case studies in Bayesian statistical modelling and analysis*. John Wiley & Sons, West Sussex, 2012.

[6] Andargie, A.A. and Rao, K.S. Estimation of a linear model with two-parameter symmetric platykurtic distributed errors. *Journal of Uncertainty Analysis and Applications*, 1(1):1–19, 2013.

[7] Ando, T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2).

[8] Archambeau, C. and Verleysen, M. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.

[9] Atukorala, R. and Srianantha kumar, S. A comparison of the accuracy of asymptotic approximations in the dynamic regression model using Kullback-Leibler information. *Economic Modelling*, 45:169–174, 2015.

[10] Bakker, S.C., Sinke, R.J., and Pearson, P.L. Differences in stutter intensities between microsatellites are related to length and sequence of the repeat. *Unravelling the Genetics of Schizophrenia and ADHD*, pages 81–94, 2005.

[11] Balding,D.J. and Buckleton, J.S. Interpreting low template DNA profiles. *Forensic Science International: Genetics*, 4(1):1 – 10, 2009.

[12] Banerjee, S. The bayesian linear model - gory details (advanced optional reading). Available at `http://www.biostat.umn.edu/~brad/ph7440.html`. [Accessed 30 Sep. 2015].

[13] Benaglia, T., Chauveau, D., Hunter, D.R., and Young, D.S. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[14] Bernardo, J.M. and Smith, A.F.M. *Bayesian theory*. John Wiley & Sons, New York, 2000.

[15] Bhattacharya, S. and Haslett, J. Importance re-sampling MCMC for cross-validation in inverse problems. *Bayesian Analysis*, 2(2):385–407, 2007.

[16] Blackwell, D. and MacQueen, J.B. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[17] Bleka, Ø. and Gill, P. Interpretation of a complex STR DNA profile using EuroForMix. *Forensic Science International: Genetics Supplement Series*, 5:e405–e406, 2015.

[18] Bolstad, W.M. *Understanding computational Bayesian statistics*. John Wiley & Sons, New Jersey, 2010.

[19] Box, G.E.P. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[20] Bright, J.A., Curran, J.M., and Buckleton, J.S. Investigation into stutter ratio variance. *Australian Journal of Forensic Sciences*, 46(3):313–316.

[21] Bright, J.A., Curran, J.M., and Buckleton, J.S. Investigation into the performance of different models for predicting stutter. *Forensic Science International: Genetics*, 7(4):422–427, 2013.

[22] Bright, J.A. et al. Developmental validation of STRmix^{TM}, expert software for the interpretation of forensic DNA profiles. *Forensic Science International: Genetics*, 23:226–239, 2016.

[23] Bright, J.A., Taylor, D., Curran, J.M., and Buckleton, J.S. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2):296–304, 2013.

[24] Brookes, C., Bright, J.A., Harbison, S.A., and Buckleton, J.S. Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1):58–63, 2012.

[25] Buckleton, J.S., Curran, J.M., and Gill, P. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics*, 1(1):20–28, 2007.

[26] Buckleton, J.S. and Gill, P.D. Preferential degradation, 2003. P17961 International Patent.

[27] Buckleton, J.S., Triggs, C.M., and Walsh, S.J. *Forensic DNA evidence interpretation*. CRC press, Florida, 2005.

[28] Budowle, B. et al. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework*. *Journal of Forensic Sciences*, 54(4):810–821, 2009.

[29] Burnham, K.P. and Anderson, D. *Model selection and multi-model inference*. Springer, New York, 2nd edition, 1998.

[30] Butler, J.M. *Advanced topics in forensic DNA typing: Methodology*. Elsevier, London, 2011.

[31] Butler, J.M. *Advanced topics in forensic DNA typing: Interpretation*. Elsevier, London, 2014.

[32] Butler, J.M. and Reeder, D.J. Short Tandem Repeat DNA Internet DataBase (STRBase). Available at `http://www.cstl.nist.gov/biotech/strbase/`. [Accessed 08 Sep. 2016].

[33] Cappé, O., Robert, C.P., and Rydén, T. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700, 2003.

[34] Carlin, B.P. and Louis, T.A. *Bayesian methods for data analysis*. Taylor & Francis, Florida, 3rd edition, 2008.

[35] Castillo, E. and Hadi, A.S. Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620, 1997.

[36] Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674, 2006.

[37] Chen, M.H. and Ibrahim, J.G. Conjugate priors for generalized linear models. *Statistica Sinica*, 13(2):461–476, 2003.

[38] Choy, S.T.B. and Chan, C.M. Scale mixtures distributions in insurance applications. *Astin Bulletin*, 33(01):93–104, 2003.

[39] Choy, S.T.B. and Smith, A.F.M. Hierarchical models with scale mixtures of normal distributions. *Test*, 6(1):205–221, 1997.

[40] Christensen, R., Johnson, W., Branscum, A., and Hanson, T.E. *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Taylor & Francis Group, Florida, 2011.

[41] Clark, J.S. and Gelfand, A.E. *Hierarchical modelling for the environmental sciences: Statistical methods and applications*. Oxford University Press Inc., New York, 2006.

[42] Cohen, S. Bayesian analysis in natural language processing. *Synthesis Lectures on Human Language Technologies*, 9(2):1–274, 2016.

[43] Congdon, P.D. *Applied Bayesian hierarchical methods*. Taylor & Francis Group, Florida, 2010.

[44] Cowell, R.G., Graversen, T., Lauritzen, S.L., and Mortera, J. Analysis of forensic *dna* mixtures with artefacts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):1–48, 2015.

[45] Cowell, R.G., Lauritzen, S.L., and Mortera, J. A gamma model for DNA mixture analyses. *Bayesian Analysis*, 2(2):333–348, 2007.

[46] Cowell, R.G., Lauritzen, S.L, and Mortera, J. Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*, 5(3):202–209, 2011.

[47] Curran, J.M. A MCMC method for resolving two person mixtures. *Science & Justice*, 48(4):168–177, 2008.

[48] Daumé III, H. Fast search for Dirichlet process mixture models. In *Proceedings of the 11th International conference on Artificial Intelligence and Statistics*, pages 83–90, 2007.

[49] de Zea Bermudez, P. and Kotz, S. Parameter estimation of the generalized Pareto distribution–Part I. *Journal of Statistical Planning and Inference*, 140(6):1353–1373, 2010.

[50] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[51] Deng, C.Y. A generalization of the Sherman–Morrison–Woodbury formula. *Applied Mathematics Letters*, 24(9):1561–1564, 2011.

[52] Denison, D.G.T., Homes, C.C., Mallick, B.K., and Smith, A.F.M. *Bayesian methods for nonlinear classification and regression*. John Wiley & Sons, West Sussex, 2002.

[53] Dias, J.G., Vermunt, J.K., and Ramos, S. Mixture hidden Markov models in finance research. In *Advances in data analysis, data handling and business intelligence*, pages 451–459. Springer, Berlin, 2010.

[54] Dror, I.E. and Hampikian, G. Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice*, 51(4):204–208, 2011.

[55] Ehlers, R.S. A study of skewed heavy-tailed distributions as scale mixtures. *American Journal of Mathematical and Management Sciences*, 34(1):40–66, 2015.

[56] Epifani, I., MacEachern, S.N., and Peruggia, M. Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2:774–806, 2008.

[57] Everitt, B.S. An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, 5(2):107–127, 1996.

[58] Evett, I.W., Gill, P.D., and J.A. Lambert. Taking account of peak areas when interpreting mixed DNA profiles.

[59] Evett, I.W. and Weir, I.W. *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sinauer Associates, Massachusetts, 1998.

[60] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. *Regression: Models, methods and applications*. Springer, Berlin, 2007.

[61] Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[62] Fink, A., Lausen, B., Seidel, W., and Ultsch, A., editors. *Advances in data analysis, data handling and business intelligence*. Springer, Berlin, 2008.

[63] Fraley, C. and Raftery, A.E. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

[64] Gamerman, D. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68, 1997.

[65] Gamerman, D. and Lopes, H.F. *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. Taylor & Francis Group, Florida, 2nd edition, 2006.

[66] Gan, F.F. and Koehler, K.J. Goodness-of-fit tests based on p-p probability plots. *Technometrics*, 32(3):289–303, 1990.

[67] Gelfand, A.E. Model determination using sampling-based methods. In Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., editors, *Markov chain Monte Carlo in practice*, pages 145–161. Chapman and Hall, London, 1996.

[68] Gelfand, A.E., Dey, D.K., and Chang, H. Model determination using predictive distributions with implementation via sampling-based methods. Technical Report 462, Department of Statistics, Stanford University, California, 1992.

[69] Gelman, A. Prior distribution. In El-Shaarawi, A.H. and Piegorsch, W.W., editors, *Encyclopedia of environmetrics*, pages 1634–1637. John Wiley & Sons, Chichester, 2002.

[70] Gelman, A. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

[71] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. *Bayesian data analysis*. Chapman & Hall/CRC, Florida, 2nd edition, 2004.

[72] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. *Bayesian data analysis*. Taylor & Francis Group, Florida, 3rd edition, 2014.

[73] Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York, 2006.

[74] Gelman, A., Hwang, J., and Vehtari, A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[75] Gelman, A., Meng, X.L., and Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

[76] Gelman, A. and Pardoe, I. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

[77] Ghosh, J.K., Delampady, M., and Samanta T. *An introduction to Bayesian analysis: Theory and methods*. Springer, New York, 2006.

[78] Gibb, A.J., Huell, A.L., Simmons, M.C., and Brown,R.M. Characterisation of forward stutter in the AmpF*l*STR® SGM Plus® PCR. *Science & Justice*, 49(1):24–31, 2009.

[79] Gilks, W.R., Richardson, S., and Spiegelhalter, D. *Markov chain Monte Carlo in practice*. Chapman & Hall, Cambridge, 1996.

[80] Gill, J. *Bayesian methods: A social and behavioral sciences approach*. Chapman & Hall/CRC, Florida, 3rd edition, 2002.

[81] Gill, P., Curran, J.M., and Elliot, K. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, 2005.

[82] Gill, P. et al. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91(1):41–53, 1998.

[83] Gill, P. et al. DNA commission of the international society of forensic genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2-3):90–101, 2006.

[84] Gill. P. et al. Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics*, 2(2):91–103, 2008.

[85] Gill, P., Puch-Solis, R., and Curran, J. The low-template-DNA (stochastic) threshold-Its determination relative to risk analysis for national DNA databases. *Forensic Science International: Genetics*, 3(2):104–111, 2009.

[86] Gill, P., Sparkes, B., and Buckleton, J.S. Interpretation of simple mixtures of when artefacts such as stutters are present - with special reference to multiplex STRs used by the forensic science service. *Forensic Science International*, 95(3):213 – 224, 1998.

[87] Good, P.I. *Resampling methods: A practical guide to data analysis*. Birkhäuser, Boston, 2006.

[88] Görür, D. and Rasmussen, C.E. Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):615–626, 2010.

[89] Green, P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[90] Grün, B. and Leisch, F. Applications of finite mixtures of regression models. Available at `http://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf`. [Accessed 20 Feb. 2017].

[91] Gujarati, D.N. *Basic econometrics*. McGraw Hill, 4th edition, 2003.

[92] Hannah, L.A., Blei, D.M., and Powell, W.B. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

[93] Harville, D.A. *Matrix algebra from a statistician's perspective*. Springer, New York, 1997.

[94] Hauge, X.Y. and Litt, M. A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics*, 2(4):411–415, 1993.

[95] Hoff, P.D. *A first course in Bayesian statistical methods*. Springer Science & Business Media, New York, 2009.

[96] Hurvich, C.M. and Tsai, C.L. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[97] Ibrahim, J.G., Chen, M.H., and Sinha, D. Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 11(2):419–443, 2001.

[98] Ibrahim, J.G. and Laud, P.W. A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, 89(425):309–319, 1994.

[99] Ionides, E.L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[100] Jackman, S. *Bayesian analysis for the social sciences*. John Wiley & Sons, New Jersey, 2009.

[101] Johnson, D. and Sinanovic, S. Symmetrizing the Kullback-Leibler distance. *IEEE Transactions on Information Theory*, 2001.

[102] Kasahara, H. and Shimotsu, K. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645, 2015.

[103] Kass, R.E. and Raftery, A.E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[104] Kelly, H., Bright, J.A., Buckleton, J.S., and Curran, J.M. A comparison of statistical models for the analysis of complex forensic DNA profiles. *Science & Justice*, 54(1):66–70, 2014.

[105] Kessler, D. and McDowell, A. Introducing the FMM procedure for finite mixture models. In *Proceedings of the SAS Global Forum*, 2012.

[106] Kim, J. and Stoffer, D.S. Fitting stochastic volatility models in the presence of irregular sampling via particle methods and the EM algorithm. *Journal of Time Series Analysis*, 29(5):811–833, 2008.

[107] Kim, S., Shephard, N., and Chib, S. Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393, 1998.

[108] Kruschke, J.K. *Doing Bayesian data analysis: A tutorial with R and BUGS*. Elsevier, Massachusetts, 2011.

[109] Kruschke, J.K. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Elsevier, London, 2nd edition, 2014.

[110] Kullback, S. *Information theory and statistics*. Dover Publications, Inc., New York, 1997.

[111] Kullback, S. and Leibler, R.A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[112] Laud, P.W. and Ibrahim, J.G. Predictive model selection. *Journal of the Royal Statistical Society: Series B*, 57(1):247–262, 1995.

[113] Lele, S.R., Dennis, B., and Lutscher, F. Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology letters*, 10(7):551–563, 2007.

[114] Link, W.A. and Barker, R.J. *Bayesian inference: With ecological applications*. Elsevier, London, 2010.

[115] Link, W.A. and Eaton, M.J. On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115, 2012.

[116] Liu, W., Zhang, B., Zhang, Z., Tao, J., and Branscum, A.J. Model selection in finite mixture of regression models: A Bayesian approach with innovative weighted g priors and reversible jump Markov chain Monte Carlo implementation. *Journal of Statistical Computation and Simulation*, 85(12):2456–2478, 2015.

[117] Lo, Y. Likelihood ratio tests of the number of components in a normal mixture with unequal variances. *Statistics & Probability Letters*, 71(3):225–235, 2005.

[118] Maas, C.J.M. and Hox, J.J. Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3):86–92, 2005.

[119] Mahieu, R.J. and Schotman, P.C. An empirical application of stochastic volatility models. *Journal of Applied Econometrics*, 13(4):333–360, 1998.

[120] Marchetti, S., Dolci, C., Riccadonna, S., and Furlanello, C. Bayesian hierarchical model for small area disease mapping: A breast cancer study. SIS2010 Scientific Meeting, Italy, 2010.

[121] Marron, J.S. and Wand, M.P. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.

[122] Mazerolle, M.J. Appendix 1: Making sense out of information criterion (AIC): Its use and interpretation in model selection and inference from ecological data. *Mouvements et Reproduction des Amphibiens en Tourbiéres Perturbées*, pages 174–190, 2004.

[123] McLachlan, G. and Peel, D. *Finite mixture models*. John Wiley & Sons, New Jersey, 2004.

[124] McLachlan, G.J. and Rathnayake, S. On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, 2014.

[125] Meligkotsidou, L. Bayesian multivariate Poisson mixtures with an unknown number of components. *Statistics and Computing*, 17(2):93–107, 2007.

[126] Murphy, K.P. Conjugate Bayesian analysis of the Gaussian distribution, 2007.

[127] Murphy, K.P. *Machine learning: A probabilistic perspective*. MIT press, London, 2012.

[128] Neal, R.M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[129] Nelder, J.A. and Wedderburn, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972.

[130] Ntzoufras, I. *Bayesian modeling using WinBUGS*. John Wiley & Sons, New Jersey, 2009.

[131] Parsons, L. and Bright, J.A. A manual and automated method for the forensic analysis of DNA from buccal samples on Whatman Indicating FTA Elute Cards. *Australian Journal of Forensic Sciences*, 44(4):393–402, 2012.

[132] Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[133] Peralta, B., Caro, A., and Soto, A. A proposal for supervised clustering with Dirichlet Process using labels. *Pattern Recognition Letters*, 80:52–57, 2016.

[134] Perlin, M.W. Explaining the likelihood ratio in DNA mixture interpretation. In *Proceedings of Promega's Twenty First International Symposium on Human Identification*, 2010.

[135] Perlin, M.W. et al. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430–1447, 2011.

[136] Perlin, M.W., Lancia, G., and Ng, S.K. Toward fully automated genotyping: Genotyping microsatellite markers by deconvolution. *American Journal of Human Genetics*, 57(5):1199–1210, 1995.

[137] Pickands III, J. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[138] Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

[139] Pitman, J.W. An extension of de Finetti's theorem. *Advances in Applied Probability*, 10(2):268–270, 1978.

[140] Press, S.J. *Subjective and objective Bayesian statistics: Principles, models, and applications*. John Wiley & Sons, New Jersey, 2nd edition, 2003.

[141] Puch-Solis, R. and Clayton, T. Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software. *Forensic Science International: Genetics*, 11:220–228, 2014.

[142] Rao, C.R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B*, 10(2):159–203, 1948.

[143] Richardson, S. and Green, P.J. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B*, 59(4):731–792, 1997.

[144] Riggan, W.B., Manton, K.G., Creason, J.P., Woodbury, M.A., and Stallard, E. Assessment of spatial variation of risks in small populations. *Environmental Health Perspectives*, 96:223–238, 1991.

[145] Rohan, M. *Using finite mixtures to robustify statistical models*. PhD thesis, University of Waikato, 2011.

[146] Norah Rudin and Keith Inman. *An introduction to forensic DNA analysis*, volume 3. CRC press, Florida, 2nd edition, 2002.

[147] Ruitberg, C.M., Reeder, D.J., and Butler, J.M. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research*, 29(1):320–322, 2001.

[148] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[149] Scrucca, L. Identifying connected components in Gaussian finite mixture models for clustering. *Computational Statistics and Data Analysis*, 93:5–17, 2016.

[150] Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, 8(1):289–317, 2016.

[151] Seghouane, A.K. and Amari, S.I. The AIC criterion and symmetrizing the Kullback–Leibler divergence. *IEEE Transactions on Neural Networks*, 18(1):97–106, 2007.

[152] Seo, S.B., Ge, J., King, J.L., and Budowle, B. Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples. *Forensic Science International: Genetics*, 8(1):213 – 218, 2014.

[153] Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[154] Sharif-Razavian, N. and Zollmann, A. An overview of nonparametric Bayesian models and applications to natural language processing. *Science*, pages 71–93, 2008.

[155] Shi, J.Q., Murray-Smith, R., and Titterington, D.M. Birth-death MCMC methods for mixtures with an unknown number of components. Technical Report 117, Department of Computing Science, University of Glasgow, Scotland, 2002.

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

[157] Steele, C.D. and Balding, D.J. Statistical evaluation of forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1:361–384, 2014.

[158] Stephens, M. Bayesian analysis of mixture models with an unknown number of components–an alternative to reversible jump methods. *Annals of statistics*, pages 40–74, 2000.

[159] Sugiura, N. Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods*, 7(1):13–26, 1978.

[160] SWGDAM. (2012). Validation guidelines for DNA analysis methods. Available at `http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf`. [Accessed 7 May 2015].

[161] Teh, Y.W. Adaptive modelling of complex data. Available at `http://www.gatsby.ucl.ac.uk/~ywteh/teaching/`. [Accessed 08 Oct. 2015].

[162] Teh, Y.W. Dirichlet processes: Tutorial and practical course. *Gatsby Computational Neuroscience Unit, University College London*, 2007.

[163] Teh, Y.W. Dirichlet process. In Sammut, C. and Webb, G.I., editors, *Encyclopedia of machine learning*, pages 280–287. Springer, New York, 2010.

[164] Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association: Theory and Methods*, 101(476):1566–1581, 2006.

[165] Thode, H.C. *Testing for normality*. Marcel Dekker, Inc., New York, 2002.

[166] Tukey, J.W. A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., and Mann, H.B., editors, *Contributions to probability and statistics*, pages 448–485. Stanford University Press, Califonia, 1960.

[167] Tvedebrink, T. Allelic drop-out in forensic genetics: Importance and estimation. University speech, Available at `http://people.math.aau.dk/~tvede/publications/auckland.pdf`, 2013. [Accessed 7 May 2015].

[168] Tvedebrink, T., Eriksen, P.S., Mogensen, H.S., and Morling, N. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, 3(4):222–226, 2009.

[169] Tvedebrink, T., Eriksen, P.S., Mogensen, H.S., and Morling, N. Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Science International: Genetics*, 6(1):97–101, 2012.

[170] Vaida, F. and Blanchard, S. Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

[171] Vehtari, A. and Gelman, A. WAIC and cross-validation in Stan. 2014.

[172] Vehtari, A. and Gelman, A. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[173] Vehtari, A., Gelman, A., and Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 2016.

[174] Vehtari, A. and Ojanen, J. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[175] Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[176] Westen, A.A. et al. Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. *Forensic Science International: Genetics*, 6(6):708–715, 2012.

[177] Weusten, J. and Herbergs, J. A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications. *Forensic Science International: Genetics*, 6(1):17–25, 2012.

[178] Yan, X. and Su, X.G. *Linear regression analysis: Theory and computing*. World Scientific, Singapore, 2009.

[179] Yerebakan, H.Z., Rajwa, B., and Dundar, M. The infinite mixture of infinite Gaussian mixtures. In *Advances in Neural Information Processing Systems 27*, pages 28–36, 2014.

[180] Zhang, J. and Stephens, M.A. A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, 51(3):316–325, 2009.