Data Selection Strategies for Bayesian Analysis with Filtering of Genetic Data

Josh Bromell

Supervisors: Prof David Bryant
Dr Peter Dillingham

A thesis submitted in fulfilment of the requirements of the degree of Master of Science in Mathematics

Department of Mathematics and Statistics
University of Otago

October, 2020

Abstract

The aim of this thesis is to look into data selection strategies for selecting data to be used for Bayesian analysis of genotyping by sequencing (GBS) data. Each selection of data leads to a different distribution on the model parameters. Methods for analysing the different resulting posterior distributions will be discussed and compared. The most applicable method will be applied to a set of simulated genetic markers.

Traditionally, GBS data sets are constructed so that each marker is a polymorphic (non-constant) site, for example a single nucleotide polymorphism (SNP). However, there is evidence to show that this might not be the optimal method. The best method may in fact be to include a certain proportion of sites which are not filtered on being polymorphic sites and are allowed to be constant sites.

To understand whether there is truth in this, we begin by analysing simplified problems with simpler distributions. These simpler problems will be studied analytically and using Monte Carlo samples. This decision making process is to decide the optimal proportion of which class of data points to include in the marker data set. The chosen method will then be first applied to a simulated marker data set and then the results analysed in order to show there appears to be an optimal mixture of data which should be used in any future phylogenetic analysis.

Acknowledgements

To start with, I would like to thank my two supervisors David Bryant and Peter Dillingham. All of your combined knowledge and support throughout this project has allowed me to learn so much, while also giving me the ability to finish this project to the best of my ability. Thank you, I could not have done it with out you two.

To my family and in particular my parents, I would like to thank you for all the support you have given me throughout my year of work.

I would like to thank my office mates for spending so much of my life with me over the past year, you all have been great for both work advice and help and also some much needed distractions and breaks.

Finally, I would like to thank all of my friends who outside of research have been able to make 2020 as good as possible for me (considering world events) and encouraging me to keep working hard and finish.

Contents

Abstract **2**

Acknowledgements **3**

1 Introduction and Background **6**

1.1 Bayesian Statistics in Evolutionary Biology 6

1.2 Challenges with Bayesian Analysis 9

1.3 Selecting SNP data 11

1.4 Thesis Outline 16

2 Data selection for a Gaussian Example **18**

2.1 Analytic expressions for the Posterior Densities: Mean unknown 18

2.2 Quantifying informativeness of posterior distributions 21

2.3 Analytical formula for scoring rules in the single data selection problem 22

2.4 Optimisation of Scoring Rules 29

2.5 Analytic expressions for the Posterior Densities: Variance unknown 35

3 Determining Mixing Proportions from Monte Carlo samples 37

3.1 Markov Chain Monte Carlo 38

3.2 Estimating optimal mixture proportions from sampled data - Mean unknown 44

3.2.1 Optimal value of p for the given scoring rule 50

3.3 Estimating optimal mixture proportions from sampled data - Variance unknown 55

3.3.1 Variance Matrix Restriction 61

3.3.2 Optimal value of p for the given scoring rule 63

3.4 Comparing Scoring Rules 67

3.4.1 Sensitivity of the Scoring Rules 67

3.4.2 Accuracy over Repeat Simulations 70

4 Analysis of Phylogentic Tree Posterior Distribution 74

4.1 First Analysis: highly divergent sequences 76

4.2 Second Analysis: less divergent sequences 81

4.2.1 Alternative Scoring Rule Check 90

4.3 Conclusions 92

5 Conclusions 94

Chapter 1

Introduction and Background

1.1 Bayesian Statistics in Evolutionary Biology

To begin, we must define what Bayesian inference and Bayesian analysis is. Bayesian inference is a process of fitting a probability model using Bayes' Theorem. This model is observed and parameters studied as a probability distribution. This 'posterior' distribution is obtained by a combination of prior knowledge of the parameters being inferred and a data set which has been observed in relation to the same parameters [10], [8]. The result from obtaining this distribution is being able to expand our knowledge of the parameters of interest.

The key theoretical tool within Bayesian inference is Bayes' Theorem, which is used in order to define and draw from a posterior distribution. This poste-

rior distribution is proportional to the likelihood function multiplied by the prior distribution. The prior distribution represents knowledge about the parameters before any data is observed and the likelihood function represents the probability of the observed data conditional on given parameters.

When designing a study, we have multiple ways to collect and observe data. How do these choices effect the posterior distribution? How would this effect the results and summaries obtained from this distribution? Is there an optimal way of collecting this data in order to obtain the most information?

With Bayes' Theorem, as the form of a posterior distribution can be a proportionality, a computational approach is frequently used in order to define and analyse the distribution. The computational method which is commonly used is Markov chain Monte Carlo (MCMC).

In the context of evolutionary biology, the Bayesian inference conducted will be analysing the posterior distribution of phylogenetic trees using a given model of evolution and some prior probabilities. The resulting posterior distribution relates to the posterior probability of the ith phylogenetic tree given the DNA or SNP data.

As these methods of phylogenetic analysis have become more popular, so has the availability of programs which utilise these methods. These programs include but are not limited to BEAST [2], SNAPPER [18] and Mr Bayes [14].

Mr Bayes is an open source program which was designed to perform Bayesian inference of phylogeny using MCMC to estimate and obtain draws from the posterior distribution of choice [12]. It has very similar methods for obtaining these posterior distribution as found in other programs like Beast2, however, there are quite different model assumptions, different operators and possible prior models etc between the two programs.

BEAST2 is another program which was designed for Bayesian phylogenetic analysis of molecular sequences [2]. This program is used to estimate time-measured phylogenies using different forms of molecular clocks. BEAST2 uses MCMC to average over the tree space, so that each tree is weighted proportional to its posterior probability.

SNAPPER [18] is a specific package which is run on the BEAST2 platform [18]. It was created to infer species trees from independent biallelic data at multiple loci for multiple individuals. What this package is designed to return are samples drawn from the joint posterior distribution of species phylogenies, species divergence times and effective population sizes. The MCMC method is very similar to that which is described in the BEAST2 program above.

For the following work we will be only using BEAST2 and SNAPPER, but Mr Bayes would have been a viable option as well.

1.2 Challenges with Bayesian Analysis

With the positives of Bayesian analysis, we also have practical challenges with this form of analysis. One of the main and most obvious challenges is the computational time required to obtain the draws from the posterior distribution. The main factors which effect computational speed are the complexity of the model and the size of the data sets.

With phylogenetic analysis, for the calculations, an evolutionary model is needed to be able to simulate from the posterior distribution of the phylogenetic trees. These models are substitution models and they model the different genetic events which can occur like a base substitution occurring which is how a SNP will appear. Examples of substitution models are the Hasegawa-Kishino-Yano (HKY) substitution model or the General time-reversible (GTR) substitution model. These models are computationally expensive and complex and hence take a longer time to run in comparison with some simple model.

The other factor for computational speed is the size of the data set and we are working with phylogenetic data sets. This means the data sets consist of DNA information representing specific SNP data or potentially the whole genome. These therefore will be very large data sets consisting of a lot of information and data points. Combining these facts of a large data set being used in a complex model a lot of times throughout a MCMC analysis, this will increase the computational time at a large rate.

Another main challenge which exists with Bayesian Analysis is lack of convergence of the chain and hence problems obtaining draws from the desired posterior distribution. Lack or partial lack of convergence with the MCMC run can be observed by looking at the trace plot of each of the draws and visually checking if the chain appears to be converging around one point tightly or not. If there is a partial lack of convergence, then the sample will be a less reliable indicator of the posterior distribution. This aim for better convergence is a normal and frequent problem within a MCMC analysis. Because of this we wish to find a method for improving convergence and obtaining a "smaller" posterior variance from our draws of the posterior distribution of phylogenetic trees.

In this thesis we will be examining a strategy for addressing these challenges by carefully choosing parts of the data in an alternative method. In general, we have correlated multivariate data X_i and Y_i which we wish to analyse, and for each i we can choose one of X_i or Y_i to be present in the data set. We study and analyse the posterior distributions and find some mixture of the data sets that returns the most accurate results. This means the data that will be observed will be chosen at a certain proportion from each set which is used, for example 0.4 of X_i and 0.6 of Y_i. This selection of proportion will effect the results which are obtained from the analysis and so this choice of proportion will be chosen purposefully in order to improve the signal and increase the accuracy of the results.

1.3 Selecting SNP data

A single nucleotide polymorphism (SNP) is a DNA sequence variation in which a single DNA base (adenine, thymine, cytosine or guanine) differs across the group of populations or species being studied. SNPs have now become a common source of genetic information and are one of the most abundant sources of genetic variation [19]. They can provide information on both evolutionary relationships and ancestral demographics [4].

The big question is whether a judicious choice at each locus of a SNP or of any biallelic marker could improve the inference. We focus specifically on a recent Bayesian method, SNAPPER [4], for inferring species relationships and demographics from markers. At present, SNAPPER only uses polymorphisms. However there is evidence to suggest that a mixture of markers of which only some are constrained to be polymorphic could be advantageous.

We will consider two primary sources of SNP data; genotyping by sequencing (GBS) and whole genome comparisons. GBS is a marker assisted selection tool. It is considered a form of Next Generation Sequencing (NGS), which is a new and frequently used method for genome sequencing [16]. The idea of NGS is the genome is broken up into small fragments and then random fragments are sampled and sequenced. As these fragments are created at random, there will exist some overlap between the fragments and therefore repeat sequencing of the same point, which increases the accuracy.

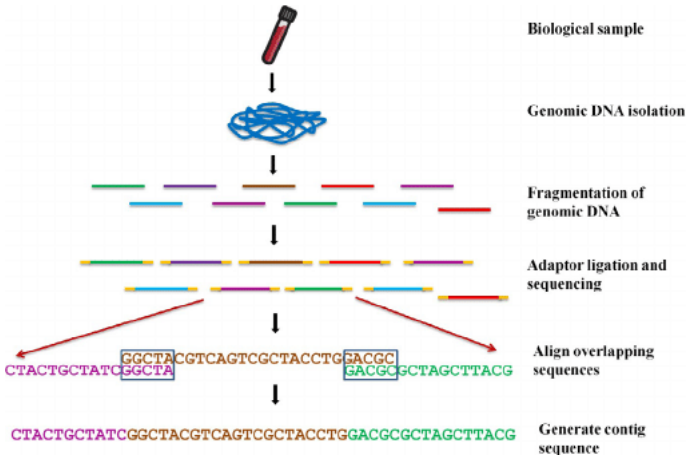


Figure 1.1: NGS Cartoon Representation [13]

For GBS, there is a specific change to this process. The adaptors that are ligated onto the end of the fragments are targeted to attach to the SNPs of interest. After the amplification steps, the differences which are found in the overlapping fragments are the SNPs which we were searching for.

GBS data is a commonly used technique for genotyping. Some reasons for the popularity of this method include the lower cost per sample, the ability to sequence predetermined areas of genetic variation and the fact it permits comparative analyses across different samples even in the lack of a reference genome.

The species history information obtained through Bayesian analysis is represented in the form of a phylogenetic tree. In fact, in any form of biological analysis, there needs to be some sort of method in order to be able to organise and group what we are trying to research and understand. The commonly used system is to organise based on evolutionary relationships between or-

ganisms and species, hence to organise based on the organisms or species phylogeny. The phylogeny can refer to physical similarities/differences or it can refer to genetic similarities/differences which is the case we will be considering.

The idea that everything is related at some level and therefore can be connected at when their most recent common ancestor was dates back to the 19th Century with Charles Darwin and "The Origin of Species" [6]. As every species is assumed to be related, we are able to connect all species into a directed graph with how far back they are connected being decided based on the similarity of their phylogeny of choice, for us that phylogeny is their genetic similarities.

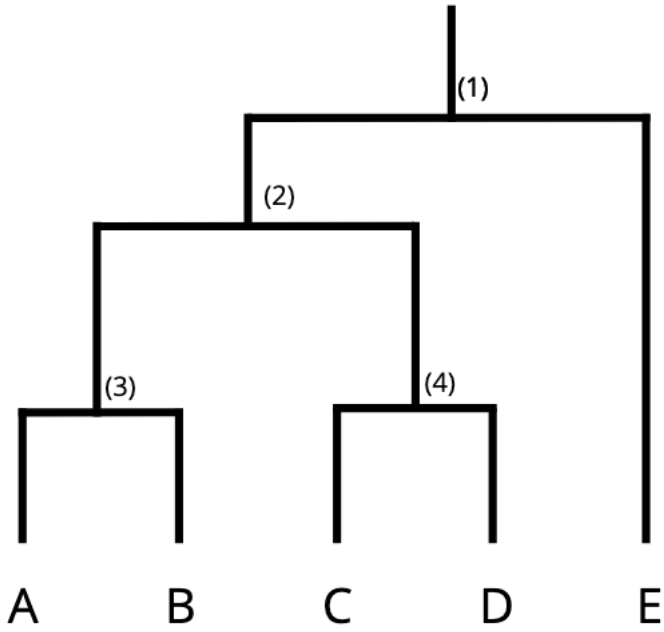


Figure 1.2: Phylogenetic Tree Example

What we can see in Figure 1.2 is an example of a phylogenetic tree. As mentioned, a phylogenetic tree is a directed graph which consists of branches and nodes. Branches are defined to connect the nodes on the graph and they go in the direction of time, traditionally this is going from top to bottom. Nodes are the points on the graph and are connected by branches. Examples of nodes consists of but are not limited to A, B, (2) and (3). Nodes like (2) and (3) which are connected by multiple different branches are defined as internal nodes. Nodes with only one branch connected to them like A and B are called leaves. Leaves are species at the current time and hence represent

the species which we are able to sample and obtain the phylogenetic information from.

Previously, in order to obtain species history information from some form of a biallelic marker, like a SNP, the method has included implementing very large simulation-based sampling of both species trees and the the actual gene trees at each locus. This was until this process was improved in [4] which as stated, computes the tree likelihoods directly from the biallelic markers.

This new improved modelling process consists of two main components; the model for the gene trees with a coalescent model and the model for the biallelic markers evolving down the gene tree with a mutation model. One issue that appears with SNP data and attempting to model mutation is effective non-identifiability of some parameters. Under certain mutation models, branch lengths of the species tree and the population size (θ) are confounded parameters and hence changing one or the other will have the same effect and hence can not be differentiated between the two. When interpretating the θ values, there is the chance of non-identifiability, especially if the amount of mutation is low.

The method of GBS for creating a SNP data set is to search for SNPs in short sequences of DNA originating at a restriction site. An alternative approach is to not restrict attention for SNPs but instead select any short sequence following the restriction site. This means the initial base in the read could actually be a constant site. The question now is to figure out a way to

decide what is the optimal proportion of these "unfiltered" points to include in order to decrease the bias and variance seen in the results, without loosing any of the species history information which is provided from the SNP's.

1.4 Thesis Outline

Chapter 2 starts with the analytical work when there is a multivariate Gaussian posterior distribution. We derive the different aspects of the different posterior distributions resulting from different mixtures of data and then attempt to define possible scoring rules and their analytical form for this specific posterior distribution. Formulas for optimal mixtures are also derived.

Chapter 3 deals with the same defined posterior distributions as in Chapter 2, however we now attempt to analyse these distributions only through Monte Carlo samples. The differing scoring rules are used to analyse these posterior distributions and these values are used to find an optimal posterior distribution. Every result is compared with the analytical results previously found if available.

Chapter 4 attempts to take the knowledge and understanding obtained from the multivariate Gaussian posterior distribution work and apply it to a less understood posterior distribution of parameters on phylogenetic trees. We take the chosen optimal scoring rule and obtain an optimal percentage of unfiltered GBS data to include in the analysis.

Chapter 5 discusses and summarises the main conclusions obtained.

Chapter 2

Data selection for a Gaussian Example

In this chapter we examine a simplified problem where the observed data are selected from one of two transformations of underlying multivariate normal data. Our goal will be to determine the proportion of data from each transformation that provides the most information about the model parameters. For each data selection strategy we can determine posterior densities analytically, thereby providing measures to determine which strategy provides the most information. Our investigations with these simple distributions inform the more complicated analyses in later chapters.

2.1 Analytic expressions for the Posterior Densities: Mean unknown

Initially we will assume Σ is known and μ is being inferred.

Theorem 1 ([10, p. 71])**.** *If $x_i \sim N(\mu, \Sigma)$ for all i and μ has a $N(\mu_0, \Sigma_0)$ prior distribution, then $\mu | x_1, \cdots, x_n \sim N(\mu_n, \Sigma_n)$ where*

$$\mu_n = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}(n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0) \tag{2.1}$$

$$\Sigma_n = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}. \tag{2.2}$$

Now we let $a, b \in \mathbb{R}^d$ and we suppose that for each i we can choose between $a^T x_i$ and $b^T x_i$. We consider the posterior density of the mean given a mixed data sample of

$$a^T x_1, a^T x_2, \cdots, a^T x_m, b^T x_{m+1}, \cdots, b^T x_n \qquad a, b \in \mathbb{R}^d. \tag{2.3}$$

Theorem 2. *If $x_i \sim N(\mu, \Sigma)$, μ has a $N(\mu_0, \Sigma_0)$ prior distribution and Σ is known, then*

$$\mu | a^T x_1, \cdots, a^T x_m, b^T x_{m+1}, \cdots, b^T x_n \sim N(\mu_n^{(m)}, \Sigma_n^{(m)})$$

where

$$\begin{aligned}
\Sigma_n^{(m)} &= \left(m\frac{aa^T}{a^T\Sigma a} + (n-m)\frac{bb^T}{b^T\Sigma b} + \Sigma_0^{-1} \right)^{-1}, \\
\mu_n^{(m)} &= \left(m\frac{aa^T}{a^T\Sigma a} + (n-m)\frac{bb^T}{b^T\Sigma b} + \Sigma_0^{-1} \right)^{-1} \left(\sum_{i=1}^{m} \frac{aa^T}{a^T\Sigma a} x_i + \sum_{i=m+1}^{n} \frac{bb^T}{b^T\Sigma b} x_i + \Sigma_0^{-1}\mu_0 \right).
\end{aligned} \tag{2.4}$$

Proof. From Bayes' Rule.

$$
\begin{aligned}
&2\log\pi(\mu|a^Tx_1,\cdots,a^Tx_m,b^Tx_{m+1},\cdots,b^Tx_n)\\
&=\sum_{i=1}^{m}\frac{(a^Tx_i-a^T\mu)^2}{a^T\Sigma a}+\sum_{i=m+1}^{n}\frac{(b^Tx_i-b^T\mu)^2}{b^T\Sigma b}+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)+\text{const}\\
&=\sum_{i=1}^{m}\frac{(\mu^Ta-x_i^Ta)(a^T\mu-a^Tx_i)}{a^T\Sigma a}+\sum_{i=m+1}^{n}\frac{(\mu^Tb-x_i^Tb)(b^T\mu-b^Tx_i)}{b^T\Sigma b}\\
&+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)+\text{const}\\
&=\sum_{i=1}^{m}\frac{\mu^Taa^T\mu-2\mu^Taa^Tx_i+(a^Tx_i)^2}{a^T\Sigma a}+\sum_{i=m+1}^{n}\frac{\mu^Tbb^T\mu-2\mu^Tbb^Tx_i+(b^Tx_i)^2}{b^T\Sigma b}\\
&+\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0+\mu_0^T\Sigma_0^{-1}\mu_0+\text{const}. \qquad (2.5)
\end{aligned}
$$

From this log-posterior, we simplify by absorbing terms not dependent on μ into the constant and we pull the terms out of the sum which are not dependent on i.

$$
\begin{aligned}
(2.5)&=m\frac{\mu^Taa^T\mu}{a^T\Sigma a}+\sum_{i=1}^{m}\frac{-2\mu^Taa^Tx_i}{a^T\Sigma a}+(n-m)\frac{\mu^Tbb^T\mu}{b^T\Sigma b}+\sum_{i=m+1}^{n}\frac{-2\mu^Tbb^Tx_i}{b^T\Sigma b}\\
&+\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0+\text{const}\\
&=\mu^T(m\frac{aa^T}{a^T\Sigma a}+(n-m)\frac{bb^T}{b^T\Sigma b}+\Sigma_0^{-1})\mu\\
&-2\mu^T(\sum_{i=1}^{m}\frac{aa^T}{a^T\Sigma a}x_i+\sum_{i=m+1}^{n}\frac{bb^T}{b^T\Sigma b}x_i+\Sigma_0^{-1}\mu_0)+\text{const}\\
&=(\mu-\mu_n^{(m)})^T(\Sigma_n^{(m)})(\mu-\mu_n^{(m)})+\text{const}. \qquad (2.6)
\end{aligned}
$$

□

2.2 Quantifying informativeness of posterior distributions

The goal is to select a value for m so that the resulting posterior distribution provides the the most information about the mean. There are number of ways to assess the information provided by a distribution, known as scoring rules [11]. We consider several options here.

When μ is from a univariate distribution, selecting m that minimises the posterior variance is an obvious metric to use. However, when μ is multivariate, there are multiple measures that can be used. We consider rules based on functions of the posterior variance matrix, with larger values corresponding to greater variability, or less information.

The trace of the covariance matrix is a standard scoring rule for the variability of a distribution [15]. It is the sum of the variances for each of the d dimensions present in the data:

$$f(\Sigma_n^{(m)}) = \sum_i (\Sigma_n^{(m)})_{ii}. \tag{2.7}$$

The downside of using this function is that it ignores the covariance terms of the variance matrix. If the covariance terms are small, the trace is a useful overall metric; if these terms are large, optimising for the trace may be a sub-optimal choice.

Another widely used scoring rule is the determinant of the variance matrix, also known as the generalised variance [9], [20]. This function uses both the main diagonal values and the off covariance values.

A third and closely related measure is the entropy, which is given by

$$H(X) = -\int_X f(x)\log(f(x))dx. \tag{2.8}$$

If we are analysing normal distributions with $X \sim N(\mu, \Sigma)$, it is given by [17]

$$H(X) = \frac{1}{2}\ln\det(2\pi e\Sigma). \tag{2.9}$$

Unfortunately, computing entropy for other distributions can be very complicated.

2.3 Analytical formula for scoring rules in the single data selection problem

Here we derive expressions for the different scoring rules evaluated on the posterior density for the mean μ, given a fixed variance Σ and selection parameter m. Throughout these calculations, we will be utilising two different formulas. These are the Sherman-Morrison formula and the Woodbury Matrix Identity.

Lemma 1 (Sherman-Morrison)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u, v \in \mathbb{R}^n$ are column vectors. Then $A + uv^T$ is invertible if and only if $1 + v^T A^{-1} u \neq 0$, in which case*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u} \tag{2.10}$$

$$\det(A + uv^T) = (1 + v^T A^{-1} u)\det(A). \tag{2.11}$$

The result in Equation 2.10 generalises to the

Lemma 2 (Woodbury Identity)**.** *Suppose A, U, C and V are all matrices of the correct dimensions. Specifically $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times n}$ and $C \in \mathbb{R}^{k\times k}$*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{2.12}$$

Recall that $\Sigma_n^{(m)}$ is the posterior variance of the mean, given that we have a known covariance matrix and reduced data as described in Theorem 2. The first scoring rule we determine analytically is the trace of the posterior variance.

Theorem 3. *The trace of the posterior distribution's variance matrix is*

$$tr(\Sigma_n^{(m)}) = tr(\Sigma_0^{-1}) - \frac{\alpha_1 m^2 + \beta_1 m + \gamma_1}{\alpha_2 m^2 + \beta_2 m + \gamma_2} \tag{2.13}$$

With the following definitions

$$
\begin{aligned}
\alpha_1 &= -b^T\Sigma_0 b a^T\Sigma_0^2 a - a^T\Sigma_0 a b^T\Sigma_0^2 b + 2b^T\Sigma_0 a a^T\Sigma_0^2 b \\
\beta_1 &= n b^T\Sigma_0 b a^T\Sigma_0^2 a + a^T\Sigma_0^2 a b^T - 2n b^T\Sigma_0 a a^T\Sigma_0^2 b + n a^T\Sigma_0 a b^T\Sigma_0^2 b - b^T\Sigma_0^2 b a^T\Sigma a \\
\gamma_1 &= n b^T\Sigma_0^2 b a^T\Sigma a \\
\alpha_2 &= -a^T\Sigma_0 a b^T\Sigma_0 b + (a^T\Sigma_0 b)^2 \\
\beta_2 &= n a^T\Sigma_0 a b^T\Sigma_0 b + a^T\Sigma_0 a b^T\Sigma b - b^T\Sigma_0^2 b a^T\Sigma a - n(a^T\Sigma_0 b)^2 \\
\gamma_2 &= n b^T\Sigma_0^2 b a^T\Sigma a
\end{aligned}
\tag{2.14}
$$

Proof. We rearrange the variance defined in Theorem 2 by first defining new variables and then applying the Woodbury Identity (2.12).

$$
\begin{aligned}
&(m\frac{aa^T}{a^T\Sigma a} + (n-m)\frac{bb^T}{b^T\Sigma b} + \Sigma_0^{-1})^{-1} = (A + uu^T + vv^T)^{-1} \\
&\text{with } A := \Sigma_0^{-1} \quad u := \sqrt{\frac{m}{a^T\Sigma a}}a \quad v := \sqrt{\frac{n-m}{b^T\Sigma b}}b
\end{aligned}
\tag{2.15}
$$

With the vectors u, v we also now define a matrix $U \in \mathbb{R}^{n\times 2}$ as

$$
U := [u \;\; v] \tag{2.16}
$$

where the first column is made up of the vector u and the second column is made up of the vector v.

We see that $UU^T = uu^T + vv^T$ and hence we define the variance as follows.

$$\left(m\frac{aa^T}{a^T\Sigma a} + (n-m)\frac{bb^T}{b^T\Sigma b} + \Sigma_0^{-1}\right)^{-1} = (A + UU^T)^{-1} \tag{2.17}$$

We apply this to (2.12) with $A = \Sigma_0$, $U = U$ and $C = I$.

$$\begin{aligned}(A + UU^T)^{-1} = (A + UCU)^{-1} &= A^{-1} - A^{-1}U(C^{-1} + U^TA^{-1}U)^{-1}U^TA^{-1} \\ &= \Sigma_0 - \Sigma_0 U(I + U^T\Sigma_0 U)^{-1}U^T\Sigma_0 \end{aligned} \tag{2.18}$$

We then calculate the 3 main parts of this matrix multiplication in order to simplify the expression.

$$\Sigma_0 U = \begin{bmatrix}\Sigma_0 u & \Sigma_0 v\end{bmatrix}, \qquad U^T\Sigma_0 = \begin{bmatrix} u^T\Sigma_0 \\ v^T\Sigma_0 \end{bmatrix} \tag{2.19}$$

The final requirement is the inverse, which in this case is a 2x2 matrix and hence the inverse can be solved for simply.

$$\begin{aligned}(I + U^T\Sigma_0 U)^{-1} &= \frac{1}{\Delta}\begin{bmatrix} (v^T\Sigma_0 v + 1) & -v^T\Sigma_0 u \\ -u^T\Sigma_0 v & (u^T\Sigma_0 u + 1)\end{bmatrix} \\ \text{with} \quad \Delta &:= (u^T\Sigma_0 u + 1)(v^T\Sigma_0 v + 1) - (u^T\Sigma_0 v)^2\end{aligned} \tag{2.20}$$

These can be multiplied together to give us the form of posterior variance

which we desire.

$$\begin{aligned}
\Sigma_n^{(m)} &= \Sigma_0 - \frac{1}{\Delta}\begin{bmatrix}\Sigma_0 u & \Sigma_0 v\end{bmatrix}\begin{bmatrix}(v^T\Sigma_0 v+1) & -v^T\Sigma_0 u \\ -u^T\Sigma_0 v & (u^T\Sigma_0 u+1)\end{bmatrix}\begin{bmatrix}u^T\Sigma_0 \\ v^T\Sigma_0\end{bmatrix} \\
&= \Sigma_0 - \frac{1}{\Delta}\Big[\Sigma_0 u(v^T\Sigma_0 v+1)u^T\Sigma_0 - \Sigma_0 u v^T\Sigma_0 u v^T\Sigma_0 - \Sigma_0 v u^T\Sigma_0 v u^T\Sigma_0 \\
&+ \Sigma_0 v(u^T\Sigma_0 u+1)v^T\Sigma_0\Big] \qquad (2.21)
\end{aligned}$$

With this form of the posterior variance, we define the trace of this matrix and then derive the simpler analytic form. In this derivation we use the fact that $\mathrm{tr}(aa^T) = a^T a$.

$$\mathrm{tr}(\Sigma_n^{(m)}) = \mathrm{tr}(\Sigma_0) - \frac{(v^T\Sigma_0 v+1)u^T\Sigma_0^2 u - 2v^T\Sigma_0 u u^T\Sigma_0^2 v + (u^T\Sigma_0 u+1)v^T\Sigma_0^2 v}{(u^T\Sigma_0 u+1)(v^T\Sigma_0 v+1) - (u^T\Sigma_0 v)^2} \qquad (2.22)$$

Using the definition for u and v, then rearranging and solving for m gives us the rational function.

$$\mathrm{tr}(\Sigma_n^{(m)}) = \mathrm{tr}(\Sigma_0^{-1}) - \frac{\alpha_1 m^2 + \beta_1 m + \gamma_1}{\alpha_2 m^2 + \beta_2 m + \gamma_2} \qquad (2.23)$$

With the following defined constants:

$$
\begin{aligned}
\alpha_1 &= -b^T\Sigma_0 b a^T\Sigma_0^2 a - a^T\Sigma_0 a b^T\Sigma_0^2 b + 2b^T\Sigma_0 a a^T\Sigma_0^2 b \\
\beta_1 &= n b^T\Sigma_0 b a^T\Sigma_0^2 a + a^T\Sigma_0^2 a b^T - 2n b^T\Sigma_0 a a^T\Sigma_0^2 b + n a^T\Sigma_0 a b^T\Sigma_0^2 b - b^T\Sigma_0^2 b a^T\Sigma a \\
\gamma_1 &= n b^T\Sigma_0^2 b a^T\Sigma a \\
\alpha_2 &= -a^T\Sigma_0 a b^T\Sigma_0 b + (a^T\Sigma_0 b)^2 \\
\beta_2 &= n a^T\Sigma_0 a b^T\Sigma_0 b + a^T\Sigma_0 a b^T\Sigma b - b^T\Sigma_0^2 b a^T\Sigma a - n(a^T\Sigma_0 b)^2 \\
\gamma_2 &= n b^T\Sigma_0^2 b a^T\Sigma a.
\end{aligned}
\tag{2.24}
$$

□

Next, we derive an analytical formula for the determinant of the posterior covariance matrix, the second of the scoring rules we considered.

Theorem 4. *The determinant of the posterior distribution's covariance matrix is as follows*

$$
\begin{aligned}
&det(\Sigma_n^{(m)}) \\
&= \det(\Sigma_0)\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)^{-1}.
\end{aligned}
\tag{2.25}
$$

Proof. From Theorem 2, the precision matrix P for the posterior distribution of μ is

$$
P = m\frac{aa^T}{a^T\Sigma a} + (n-m)\frac{bb^T}{b^T\Sigma b} + \Sigma_0^{-1}. \tag{2.26}
$$

This matrix can be written as

$$P = A + uu^T + vv^T \tag{2.27}$$

where

$$A := \Sigma_0^{-1} \quad u := \sqrt{\frac{m}{a^T \Sigma a}} a \quad v := \sqrt{\frac{n-m}{b^T \Sigma b}} b \tag{2.28}$$

allowing use of (2.10).

$$\begin{aligned}
&\det(A + uu^T + vv^T) \\
&\qquad = det(A + uu^T)(1 + v^T (A + uu^T)^{-1} v) \\
&\qquad = \det(A)(1 + u^T A^{-1} u)(1 + v^T (A + uu^T)^{-1} v) \\
&\qquad = \det(A)(1 + u^T A^{-1} u)(1 + v^T (A^{-1} - \frac{A^{-1} uu^T A^{-1}}{1 + u^T A^{-1} u}) v) \\
&\qquad = \det(A)((1 + u^T A^{-1} u) + (1 + u^T A^{-1} u) v^T A^{-1} v - v^T A^{-1} uu^T A^{-1} v) \\
&\qquad = \det(A)(1 + u^T A^{-1} u + v^T A^{-1} v + u^T A^{-1} uv^T A^{-1} v - (v^T A^{-1} u)^2) \\
&\qquad = \det(A)((1 + u^T A^{-1} u)(1 + v^T A^{-1} v) - (v^T A^{-1} u)^2). \qquad (2.29)
\end{aligned}$$

As this is the determinant value for the precision matrix, in order to get the

determinant for the variance matrix, we can invert it.

$$\det(\Sigma_n^{(m)}) = \frac{1}{\det(A)((1+u^TA^{-1}u)(1+v^TA^{-1}v)-(v^TA^{-1}u)^2)}$$
$$= \det(\Sigma_0)\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)^{-1} \tag{2.30}$$

□

The third scoring rule we derive analytically is that for the entropy function. This formula follows directly from the previous scoring rule derivation, the determinant.

Corollary 1. *The entropy of the posterior distribution's covariance matrix is as follows*

$$H(\Sigma_n^{(m)})$$
$$= \log\det(\Sigma_0) - \log\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right). \tag{2.31}$$

2.4 Optimisation of Scoring Rules

If we know that each of the scoring rules are convex, this would allow us to use the known results in convex optimisation, including the local minimum is equal to the global minimum [3]. With this, we look at trying to minimise each of these scoring rules and show that the minimum is unique.

Theorem 5. *The value for m value giving the minimum trace of the posterior distributions variance matrix is*

$$m_1^* = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \tag{2.32}$$

with the following definitions

$$\begin{aligned} A &= \alpha_1\beta_2 - 2\alpha_2\beta_1 \\ B &= 2\alpha_1\gamma_2 - 2\alpha_2\gamma_1 \\ C &= \beta_1\gamma_2 - \beta_2\gamma_1. \end{aligned} \tag{2.33}$$

This optimal value is unique.

The definitions for each of the α_i, β_i and γ_i are the same as defined in (2.14).

Proof.

$$\begin{aligned} &\frac{\partial}{\partial m}(\mathrm{tr}(\mathrm{var}(\mu|a^T x_i, b^T x_i, \Sigma))) \\ &= \frac{\partial}{\partial m}(\mathrm{tr}(\Sigma_0) - \frac{\alpha_1 m^2 + \beta_1 m + \gamma_1}{\alpha_2 m^2 + \beta_2 m + \gamma_2}) \\ &= \frac{(2\alpha_1 m + \beta_1)(\alpha_2 m^2 + \beta_2 m + \gamma_2) - (2\alpha_2 m + \beta_2)(\alpha_1 m^2 + \beta_1 m + \gamma_1)}{(\alpha_2 m^2 + \beta_2 m + \gamma_2)^2}. \end{aligned} \tag{2.34}$$

With this derivative, we let $\frac{\partial}{\partial m}(\mathrm{tr}(\mathrm{var}(\mu|a^T x_i, b^T x_i, \Sigma))) = 0$.

$$0 = (2\alpha_1 m + \beta_1)(\alpha_2 m^2 + \beta_2 m + \gamma_2) - (2\alpha_2 m + \beta_2)(\alpha_1 m^2 + \beta_1 m + \gamma_1) \tag{2.35}$$

Expanding, we get

$$0 = 2\alpha_1\alpha_2 m^3 + 2\alpha_1\beta_2 m^2 + 2\alpha_1\gamma_2 m + \beta_1\alpha_2 m^2 + \beta_1\beta_2 m + \beta_1\gamma_2$$
$$- (2\alpha_1\alpha_2 m^3 + 2\alpha_2\beta_1 m^2 + 2\alpha_2\gamma_1 m + \alpha_1\beta_2 m^2 + \beta_1\beta_2 m + \beta_2\gamma_1) \tag{2.36}$$

Simplifying the terms and factoring yields

$$(\alpha_1\beta_2 - \alpha_2\beta_1)m^2 + (2\alpha_1\gamma_2 - 2\alpha_2\gamma_1)m + (\beta_1\gamma_2 - \beta_2\gamma_1) = 0 \tag{2.37}$$

As this is a quadratic which we want the roots of, we therefore use the quadratic equation with the given constants, giving m_1^*.

To prove uniqueness, we note that $f_1(X) = \mathrm{tr}(X^{-1})$ is convex on the space of positive definite matrices [3] and $\Sigma_n^{(m)}$ is a linear function of m. □

Theorem 6. *The optimal m in order to minimise the determinant of the posterior variance matrix is*

$$m_2^* = \frac{n}{2} + \frac{a^T\Sigma_0 a b^T \Sigma b - a^T \Sigma a b^T \Sigma_0 b}{2a^T\Sigma_0 a b^T \Sigma_0 b - 2(a^T\Sigma_0 b)^2} \tag{2.38}$$

and this optimal value is unique.

Proof. We will be using the derived formula in Theorem 4 in order to optimise

it.

$$
\begin{aligned}
&\frac{\partial}{\partial m}\det(\mathrm{var}(\mu|a^Tx, b^Tx, \Sigma)) \\
&= \frac{\partial}{\partial m}\frac{1}{\det(\Sigma_0^{-1})\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)} \\
&= \frac{1}{\det(\Sigma_0^{-1})}\frac{\partial}{\partial m}\frac{1}{\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)} \\
&= \frac{-1}{\det(\Sigma_0^{-1})}\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)^{-2} \\
&\times \frac{\partial}{\partial m}\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right).
\end{aligned}
\tag{2.39}
$$

We expand the final factor of this expression.

$$
\begin{aligned}
&\frac{\partial}{\partial m}\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right) \\
&= \frac{\partial}{\partial m}\Bigg(1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m + \frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m) + \frac{a^T\Sigma_0 a b^T\Sigma_0 b}{a^T\Sigma a b^T\Sigma b}(mn-m^2) \\
&- \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(mn-m^2)\Bigg) \\
&= \left(\frac{a^T\Sigma_0 a}{a^T\Sigma a} - \frac{b^T\Sigma_0 b}{b^T\Sigma b} + \frac{a^T\Sigma_0 a b^T\Sigma_0 b}{a^T\Sigma a b^T\Sigma b}(n-2m) + \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(2m-n)\right).
\end{aligned}
\tag{2.40}
$$

We substitute this into (2.39) and equate this to zero.

$$
\begin{aligned}
&\frac{\partial}{\partial m}\det(\mathrm{var}(\mu|a^Tx, b^Tx, \Sigma)) \\
&= \frac{-1}{\det(\Sigma_0^{-1})}\left((1+\frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1+\frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(m(n-m))\right)^{-2} \\
&\times\left(\frac{a^T\Sigma_0 a}{a^T\Sigma a} - \frac{b^T\Sigma_0 b}{b^T\Sigma b} + \frac{a^T\Sigma_0 a b^T\Sigma_0 b}{a^T\Sigma a b^T\Sigma b}(n-2m) + \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(2m-n)\right) = 0
\end{aligned}
\tag{2.41}
$$

When solving this equation, we can multiply both sides by the two different denominators and we are left with a simpler linear equation to solve for.

$$
\left(\frac{a^T\Sigma_0 a}{a^T\Sigma a} - \frac{b^T\Sigma_0 b}{b^T\Sigma b} + \frac{a^T\Sigma_0 a b^T\Sigma_0 b}{a^T\Sigma a b^T\Sigma b}(n-2m) + \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma a b^T\Sigma b}(2m-n)\right) = 0
$$

$$
a^T\Sigma_0 a b^T\Sigma b - a^T\Sigma a b^T\Sigma_0 b + a^T\Sigma_0 a b^T\Sigma_0 b(n-2m) + (a^T\Sigma_0 b)^2(2m-n) = 0
$$

$$
a^T\Sigma_0 a b^T\Sigma b - a^T\Sigma a b^T\Sigma_0 b + a^T\Sigma_0 a b^T\Sigma_0 bn - (a^T\Sigma_0 b)^2 n = 2a^T\Sigma_0 a b^T\Sigma_0 bm - 2(a^T\Sigma_0 b)^2 m
$$

$$
\frac{a^T\Sigma_0 a b^T\Sigma b - a^T\Sigma a b^T\Sigma_0 b + (a^T\Sigma_0 a b^T\Sigma_0 b - (a^T\Sigma_0 b)^2)n}{2a^T\Sigma_0 a b^T\Sigma_0 b - 2(a^T\Sigma_0 b)^2} = m
$$

$$
m = \frac{n}{2} + \frac{a^T\Sigma_0 a b^T\Sigma b - a^T\Sigma a b^T\Sigma_0 b}{2a^T\Sigma_0 a b^T\Sigma_0 b - 2(a^T\Sigma_0 b)^2} \tag{2.42}
$$

To show uniqueness for this solution, we observe that $f_2(X) = (\det(X))^{1/n}$ is convex [3] and so $\det(\Sigma_n^{(m)})$ is an increasing function composed with a convex function. $\square$

Corollary 2. *The optimal m in order to minimise the entropy of the posterior variance matrix is*

$$m_3^* = \frac{n}{2} + \frac{a^T\Sigma_0 ab^T\Sigma b - a^T\Sigma ab^T\Sigma_0 b}{2a^T\Sigma_0 ab^T\Sigma_0 b - 2(a^T\Sigma_0 b)^2} \tag{2.43}$$

Proof. We will be using the derived formula for the entropy as stated in Corollary 1.

$$\begin{aligned} H = &\frac{p}{2}\ln(2\pi) + \frac{p}{2} - \frac{1}{2}\ln(\det(\Sigma_0^{-1})) \\ &- \frac{1}{2}\ln\left((1 + \frac{a^T\Sigma_0 a}{a^T\Sigma a}m)(1 + \frac{b^T\Sigma_0 b}{b^T\Sigma b}(n-m)) - \frac{(a^T\Sigma_0 b)^2}{a^T\Sigma ab^T\Sigma b}(m(n-m))\right) \end{aligned} \tag{2.44}$$

We observe that the entropy of the posterior variance is proportional to the log-det part of the function when considering it as a function of the variable m. This then gives the fact that it will be optimised at the same value $m_2^* = m_3^*$ as the determinant function. This is because the natural logarithm is a monotonically increasing function, hence doesn't effect the optimisation. □

So far with this analytical work, we now have a lot of different analytical forms of the desired distributions we wished to describe. The analytical forms of the posterior distributions are known with our transformed data and as the distributions are able to be defined analytically, then we have also been able to define the closed forms for the different scoring rules. These scoring rules were all defined and shown to be convex functions, hence we were able to solve for the global minimum. These minimum values give us

the information of how to transform our data in order to ensure that we obtain the most information from the posterior distribution.

2.5 Analytic expressions for the Posterior Densities: Variance unknown

We now consider a more complex scenario in which the mean is known and the variance matrix is being inferred. This implies that we will still be using data which has a Gaussian distribution but now the prior distribution is for the variance matrix and the posterior distribution will be over variance matrices.

We again assume the data is normally distributed:

$$x_1, \cdots, x_n \sim N(\mu, \Sigma) \quad x_i \in \mathbb{R}^d. \tag{2.45}$$

We use an inverse-Wishart prior distribution for Σ.

Theorem 7 ([10, p. 73]). *For fixed mean μ and prior $\Sigma \sim W^{-1}(\Psi, \nu)$, $[\Sigma | x_1, \cdots, x_n \sim W^{-1}(\Sigma_n, \nu_n)]$ with*

$$\begin{aligned} \Sigma_n &= \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T + \Psi \\ \nu_n &= \nu + n. \end{aligned} \tag{2.46}$$

Now suppose $a, b \in \mathbb{R}^d$ and $m \in [n]$. We attempted to extend the earlier

analysis to the posterior distribution $\Sigma|a^T x_1, \cdots, a^T x_m, b^T x_{m+1}, \cdots, b^T x_n$

Attempting to derive the analytical form of the posterior definition is important if we wish to be able to analyse different aspects of the posterior distribution, in particular the different scoring rules applied to the posterior variance matrix. The previous method for deriving these analytical forms was to use the observed conjugacy between the likelihood function and the prior distribution in order to have a posterior distribution of a known form, for example a normal distribution or a Wishart distribution. With this specific case, the likelihood function is now a product of univariate normal distributions, either being of the form of a transformation with the vector a or b and the prior distribution on the variance matrix is still a Wishart distribution. These are not conjugate and hence for the posterior distribution, when we apply Bayes' Theorem we are not left with a known form for a certain distribution.

Not having conjugacy occurring is actually the common outcome when applying Bayesian analysis to any form of real life data. Because of this, to continue on with analysis of this more complex scenario and infact any of the future scenarios we wish to analyse, we require another form of analysing and attempting to define the posterior distribution. This will be done numerically using Markov Chain Monte Carlo (MCMC) in order to sample from the posterior distribution and then use these samples to approximate the different aspects of the posterior distribution, for example the variance matrix approximated from the variance of the posterior draws.

Chapter 3

Determining Mixing Proportions from Monte Carlo samples

The previous chapter looked at the effect of different mixture proportions for a simple family of distributions for which we had analytical expressions for posterior densities. To step closer to the application of GBS analysis, we now consider the case when our knowledge of the posterior distributions is limited to MCMC samples. The reason for wanting to use MCMC to analyse the posterior distribution is due to the complexity of the posterior distributions which we wish to draw from. The target posterior distributions for the final application will not be a general known distribution and therefore will not have a simple closed form which could be analysed analytically.

As an interim step, in this chapter we will use numerical methods to analyse

the multivariate Gaussian distribution and then compare it with the analytic results from Chapter 2.

3.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a process in which we iteratively sample from a posterior distribution without knowing the analytical form. Here, we use the Metropolis-Hastings algorithm [5] to create a Markov chain which has a limiting distribution equal to the posterior distribution we wish to sample from. We then extract a sample from this chain. In comparison to direct sampling from a conjugate prior, MCMC is more computationally intensive. However, it does also allow us to draw samples from complex posteriors which have no conjugate prior.

We use the Metropolis-Hastings algorithm in order to implement MCMC. In this algorithm, a proposal distribution q is used to propose a new value y from a given value x. This value y is accepted with probability α with

$$\alpha = \min\left(1, \frac{\pi(y)q(y|x)}{\pi(x)q(x|y)}\right), \tag{3.1}$$

where $\pi(X)$ is a known distribution which is proportional to the desired target posterior distribution. See Algorithm 1 for an overview of MCMC.

Algorithm 1: Draw from the Posterior Distribution with MCMC

Result: Obtain independent posterior draws

```
Define the parameters: Σ, N, x(1);
for n = 2:N do
    Choose random dimension to alter;
    Draw Δx uniform on (−ϵ, ϵ);
    Add Δx to x = x(n − 1) in the chosen dimension to obtain the
      candidate draw y;
    Calculate the probability α with (3.1);
    Accept candidate draw with probability α;
    if y Accepted then
        x(n) = y;
    else
        x(n) = x(n − 1);
    end
end
Remove the first percentage of the posterior draws x as burn in;
return Every 100th x(n) value
```

Using this base algorithm we are able to obtain approximately independent draws from each of the posterior distributions.

In the proposal steps of this algorithm, Δx is calculated by drawing a uniform random value on $(-\epsilon, \epsilon)$. The choice of ϵ was made using visual inspection of the resulting trace plot of the chain for each dimension. If the chain was moving very slowly and not appearing to converge around a single point then

ϵ was increased. If the chain stayed at the same values for long periods of times and have a "skyscraper" look to it then ϵ was decreased. The optimal value for ϵ that we ended on was $\epsilon = 0.25$ and this gave us a trace plot like a "fuzzy caterpillar", as desired.

The required burn in percentage was also found using visual check of the trace of the chain. Once the chain was no longer in a likelihood climb or drop and it had reached a constant point it was fluctuating around, we knew the chain had reach convergence and each value in the chain was a draw from the posterior distribution. From visual inspection we saw that for each of these cases it was at the 10% point in the chain where convergence was reached. Hence we removed 10% of the chain values and the remaining 90% were now accepted as posterior draws.

The analytical results from Chapter 2 allow us to validate both the MCMC implementation and inference of the posterior variance.

For validation, the data will be simulated using the following parameters, that will be staying constant through all of the following analysis in this chapter.

$$n = 1000, \;\; m = 500, \;\; \Sigma = \begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix}, \;\; \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$
$$\mu = \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \;\; \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.2}$$

For the first test, we simulated

$$x_1, \cdots, x_n \sim N(\mu, \Sigma)$$

and then used MCMC to sample from the posterior density

$$\pi(\mu | x_1, \cdots, x_n)$$

with a $N(\mu_0, \Sigma_0)$ prior.

The estimated posterior mean and analytical posterior mean were

$$\mu_1 = \begin{pmatrix} 1.8969 \\ 4.9284 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 1.8979 \\ 4.9292 \end{pmatrix} \tag{3.3}$$

respectively. Where μ_1 is the numerical estimate and μ_2 is the analytical

solution.

The estimated posterior variance matrix and analytical posterior variance matrix were

$$\Sigma_1 = \begin{pmatrix} 0.003044 & 0.001951 \\ 0.001951 & 0.004773 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 0.002987 & 0.001984 \\ 0.001984 & 0.004971 \end{pmatrix} \tag{3.4}$$

respectively. Where Σ_1 is the numerical estimate and Σ_2 is the analytical solution.

These means and variances are equal up to numerical error and thus help to confirm that the analytical form for the posterior distribution is matching the distribution being sampled by the MCMC.

For the second test, we defined vectors

$$a = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad b = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{3.5}$$

We again simulated

$$x_1, \cdots, x_n \sim N(\mu, \Sigma)$$

and then used MCMC to sample from the posterior density

$$\pi(\mu | a^T x_1, \cdots, a^T x_m, b^T x_{m+1}, \cdots, b^T x_n)$$

with a $N(\mu_0, \Sigma_0)$ prior.

The estimated posterior mean and analytical posterior mean were

$$\mu_1 = \begin{pmatrix} 2.0440 \\ 4.8602 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 2.0452 \\ 4.878 \end{pmatrix} \tag{3.6}$$

respectively. Where μ_1 is the numerical estimate and μ_2 is the analytical solution.

The estimated posterior variance matrix and analytical variance matrix were

$$\Sigma_1 = \begin{pmatrix} 0.006041 & -0.000188 \\ -0.000188 & 0.010380 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 0.005964 & 0 \\ 0 & 0.009901 \end{pmatrix} \tag{3.7}$$

respectively. Where Σ_1 is the numerical estimate and Σ_2 is the analytical solution.

These variance matrices can again be assumed to be equal up to numerical error, once again confirming the analytical form for the posterior distribution is matching the distribution being sampled by the MCMC.

3.2 Estimating optimal mixture proportions from sampled data - Mean unknown

With knowing our posterior distribution that we are sampling from is correct, we begin analysing the best way to transform the data, by optimising the m value. In order to keep this more general and not dependent on n, we define the proportion value $p = \frac{m}{n}$ to optimise over instead. We will use the proposed score functions from 2.2 and use them to analyse the posterior distribution's variance matrix.

The protocol for this experiment is as follows.

Algorithm 2: Calculating scoring rule values over differing data mixture proportions with the mean unknown

Result: Obtain scoring rule values

```
for i = 1:N do
    Simulate x_1, ..., x_n from N(μ, Σ);
    for p=0.1,0.2,...,0.9 do
        Use MCMC to sample from
        π(μ|a^T x_1, ··· , a^T x_m, b^T x_{m+1}, ··· , b^T x_n)
        with a N(μ_0, Σ_0) prior and defining m = ⌈np⌉;
        Estimate the variance matrix and evaluate the scoring rules
    end
end
```

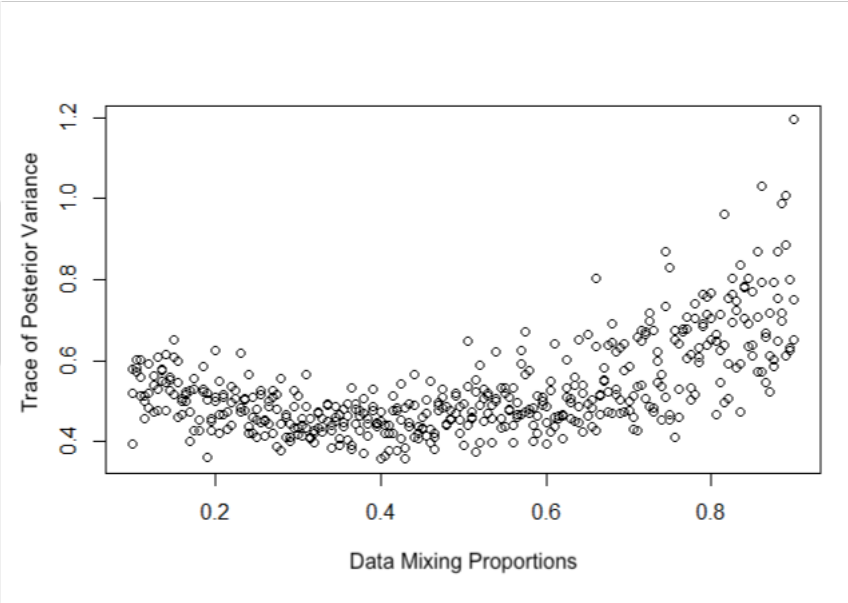


Figure 3.1: Trace Values of Posterior Variance Matrices shows an optimal value near $p = 0.4$

Figure 3.1 shows a plot of estimates for the trace of the posterior variance matrix for each proportion p, for three replicated samples at each of 160 values of p. We can see there is a clear convex curve occurring with the minimum occurring at a p value around 0.4. The data appears to have a medium to strong relationship with the parabolic curve, allowing us to make a more accurate prediction for the minimum. These data points also appear to follow the trend of a convex function which we expected.

The next scoring rule we look at and analyse is the determinant of the the

posterior distribution variance matrix.

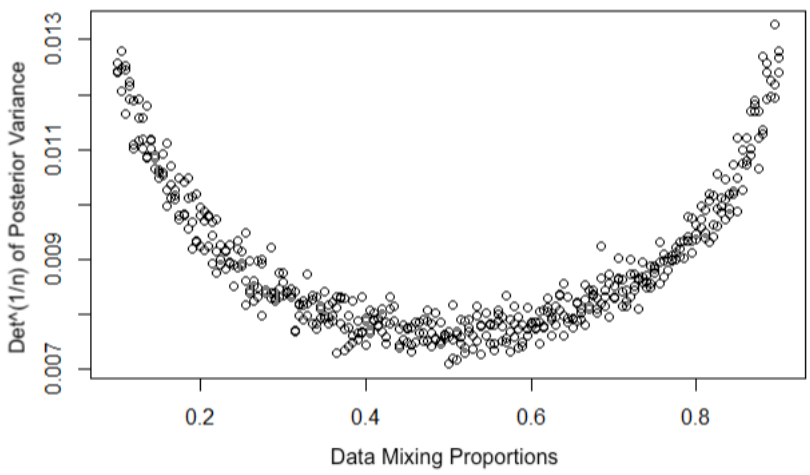


Figure 3.2: $\text{Det}^{1/n}$ Values of Posterior Variance Matrices shows an optimal value near $p = 0.5$

Figure 3.2 shows a plot of estimates for the nth root of the determinant of the posterior variance matrix for each proportion p, for three replicated samples at each of 160 values of p. We see there is a clear convex curve occurring with the minimum occurring at a p value around 0.5. This again shows that the optimal posterior distribution occurs with a mixture of data rather than at a boundary value.

The final scoring rule we will analyse is the entropy of the posterior vari-

ance. As this particular posterior distribution is still a normal distribution, then that means we can use the known closed form for the entropy of it [7].

$$H = \frac{1}{2} \ln \det 2\pi e \Sigma \tag{3.8}$$

Using this definition, we look at the entropy of all of the different distributions defined with different data mixture proportions.

As mentioned, for the multivariate case which we are analysing, there exits a closed form for the entropy of the posterior distribution. For a more general case, this will not exist and it definitely will not exist for the GBS application which we hoped to use it for. This means if we wish to move ahead further with entropy as the scoring rule then we will have to find a method to solve for the entropy, potentially numerical integration.

We expect to see the entropy values having the same relationship as the nth root of the determinant scoring rule. The reason for this is similar to the comparison we see in the optimisation of the two rules. The entropy for this particular distribution is a function of the determinant function as seen in Equation 3.8.

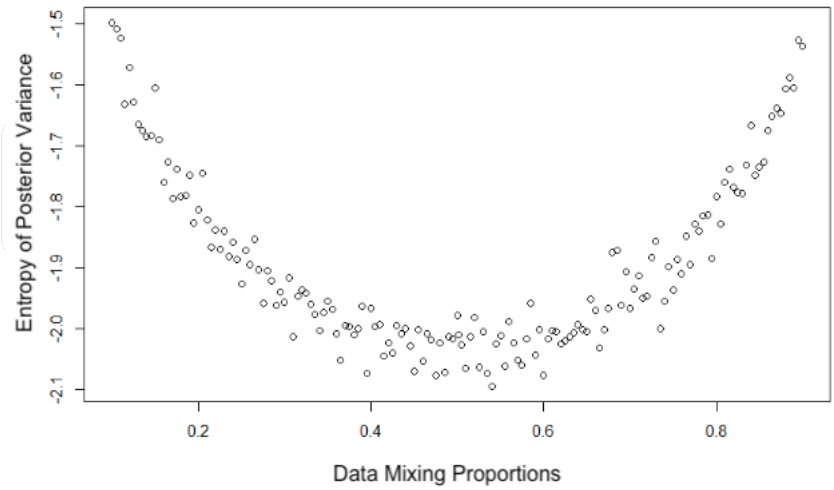


Figure 3.3: Entropy Values of Posterior Variance Matrices shows an optimal value near $p = 0.5$

Unsurprisingly, we do see that the relationship looks very similar to the determinant score function and is following the same relationship.

The problem of finding the best value for p is one of optimising a noisy function using as few data points as possible, since every data point is obtained using a computationally expensive MCMC run. A simple method is to fit a quadratic function to the points and then optimise that. In the next section we explore the suitability of this strategy.

3.2.1 Optimal value of p for the given scoring rule

We use the same data plotted as in Figure 3.1 to obtain a fitted curve for the data using both an empirical smoother (LOESS) and a quadratic fit. The reason for choosing a quadratic to attempt to match is because of the earlier work where we were able to show that the scoring rules of the posterior variance are a convex function. The quadratic is an obvious convex function with similar shape to the fitted LOESS curve.

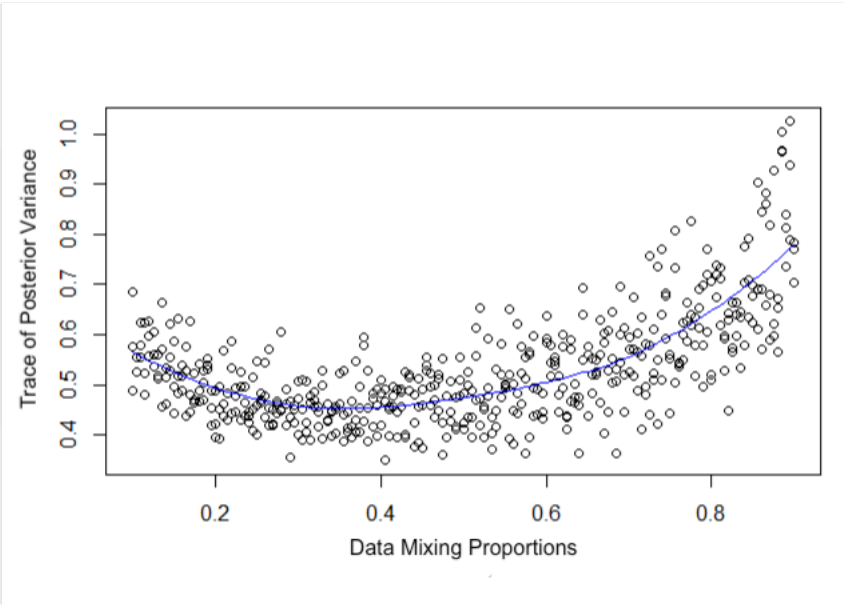


Figure 3.4: LOESS Fitted Curve - Posterior Variance Trace Values

Using the linear model function we define an approximate quadratic function

for the fitted curve.

$$y = 1.218x^2 - 0.9606x + 0.6410 \tag{3.9}$$

This is a simple quadratic which can be used to analyse the trace values of the posterior distribution's variance with different data reductions. Before we go analyse this function further, we will compare it with the original LOESS curve to ensure that it is a good approximation.

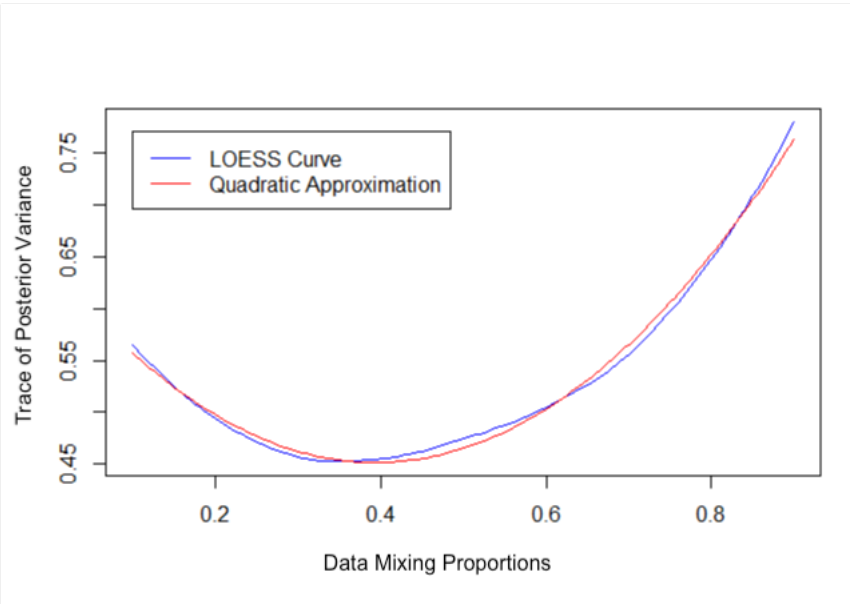


Figure 3.5: LOESS Fitted Curve and Quadratic Function Comparison

This confirms the quadratic function is a good approximation for these specific trace values and hence can be used to make inferences about the posterior

distribution and its corresponding variance matrix.

Optimising this quadratic function, we obtain the following optimal p value,

$$\hat{p} = \frac{0.9606}{2.436} = \frac{1601}{4060} \approx 0.394335. \tag{3.10}$$

Based on the trace score of the posterior variance matrix and these particular defined numerical values, the optimal mixing proportion for the sampled data is 0.394.

Viewing the fitted quadratic, we observe that a broad range of p values actually are close to minimising the trace of the posterior distribution variance matrix. This optimal range is approximately $(0.3, 0.5)$ and hence any posterior distribution defined with this p will have a resulting near minimal trace value.

We follow the same procedure that we did for the trace but now for the determinant and using the LOESS method in order to fit a curve to Figure 3.2.

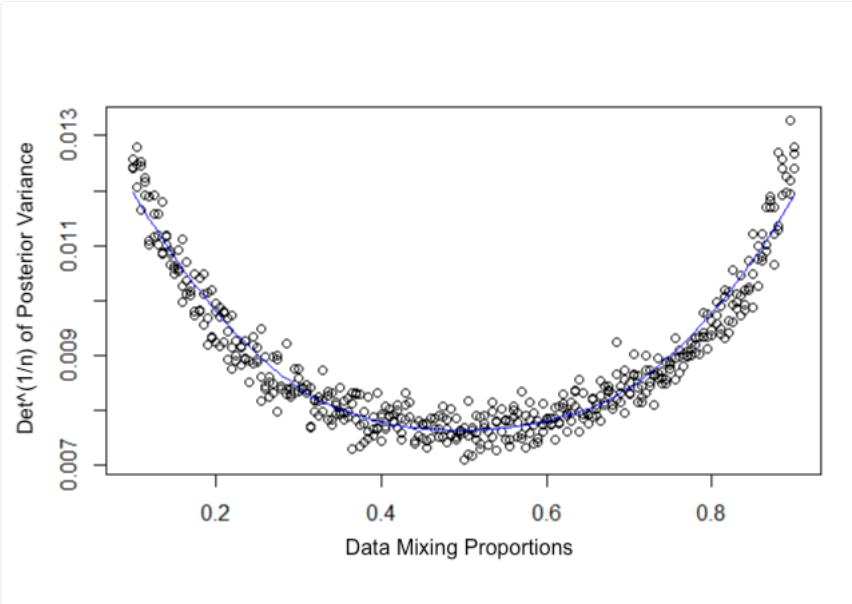


Figure 3.6: LOESS Fitted Curve - Posterior Variance $\text{Det}^{1/n}$ Values

We obtain from this LOESS fit, the approximate formula for the fitted curve as

$$y = 0.02645x^2 - 0.0265x + 0.01413. \tag{3.11}$$

Before we do the analysis of this curve, we ensure that it is a good approximation for the LOESS curve.

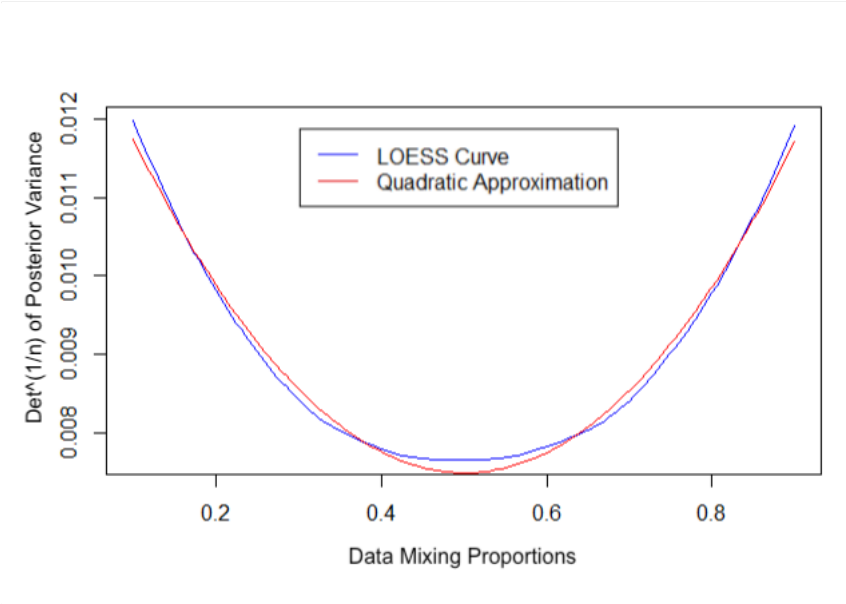


Figure 3.7: LOESS Fitted Curve and Quadratic Function Comparison

This confirms the quadratic function is a good approximation for these specific determinant values and hence can be used to make assumptions about the posterior distribution and its corresponding variance matrix.

Optimising this quadratic function, we obtain the following optimal p value,

$$\hat{p} = \frac{0.0265}{0.0529} = \frac{265}{529} \approx 0.500945 \tag{3.12}$$

What we conclude from this optimum, is that based on the determinant score of the posterior variance and these particular defined numerical values, the

optimal percentage of data reduction is at roughly 0.501.

Again, we see that the quadratic does have a broad range of values which will nearly optimise the determinant. This range is approximately (0.3,0.7) and hence any distribution defined with any p value in this range will nearly minimise the determinant.

When we apply the constants which we defined in Equation 3.2 to the optimum formula for the determinant in Equation 2.42 we get an analytical minimum of 0.501. This matches perfectly with the numerical result we got from the MCMC approach above. This in turn reinforces the accuracy of both the analytic approach and the numerical approach for optimising the determinant of the variance matrix.

The fitted curve will minimise at the same point if fitted to the entropy plot as well. This is due to the similarity between the closed forms for the determinant scoring rule and the entropy scoring rule of the multivariate normal posterior distribution.

3.3 Estimating optimal mixture proportions from sampled data - Variance unknown

In order to obtain different values of the scoring rules, we need to numerically define the different posterior distributions through MCMC and more

specifically following Algorithm 1. As these distributions are assuming that the mean is known and the variance matrix is being sampled, there will be different distributions required. In order to look at the variance matrix using multivariate distributions, we will reshape it to be in vector form.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{(1,2)} \\ \sigma_{(1,2)} & \sigma_2^2 \end{pmatrix} \mapsto \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_{(1,2)} \end{pmatrix} \tag{3.13}$$

In a similar approach to Algorithm 2, we check the different scoring rules of the posterior in the case of the mean being known and the variance matrix being inferred.

Algorithm 3: Calculating scoring rule values over differing data mixture proportions with the variance unknown

Result: Obtain scoring rule values

```
for i = 1:N do
    Simulate x_1, ..., x_n from N(μ, Σ);
    for p=0.1,0.2,...,0.9 do
        Use MCMC to sample from
        π(Σ|a^T x_1, ··· , a^T x_m, b^T x_{m+1}, ··· , b^T x_n)
        with a W^{-1}(Ψ, ν) prior and defining m = ⌈np⌉;
        Estimate the posterior variance matrix and evaluate the scoring
        rules
    end
end
```

For Algorithm 3, we require two more parameters to those that are defined in Equation 3.2. They are required for the inverse-Wishart prior and they are

$$\Psi = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \quad \nu = 3. \tag{3.14}$$

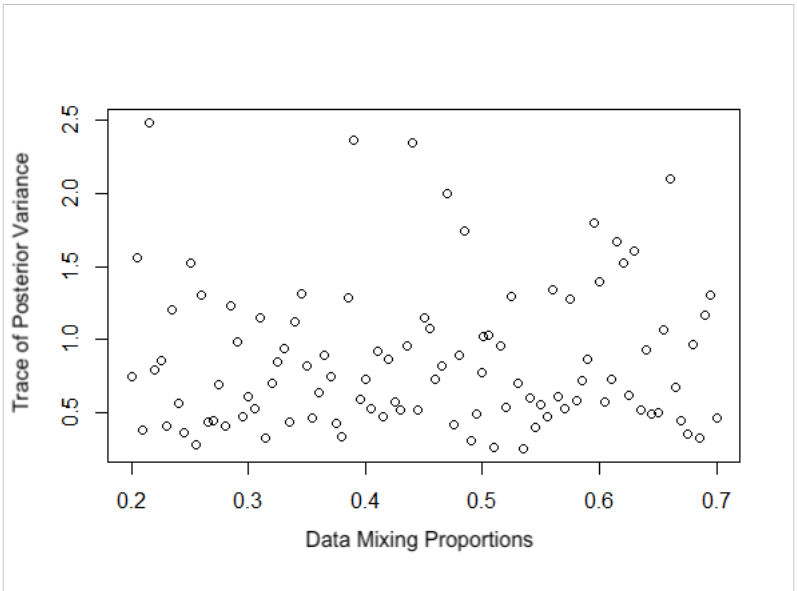


Figure 3.8: Trace Values of Posterior Variance Matrices

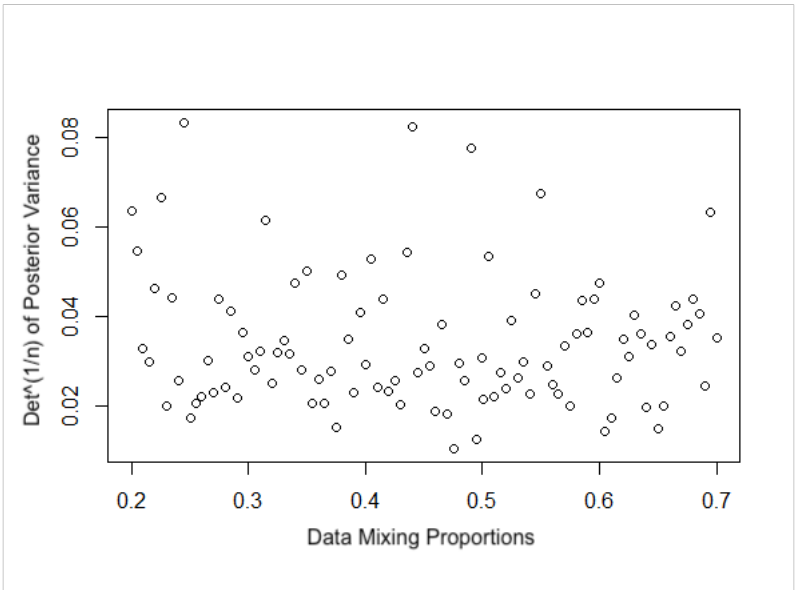


Figure 3.9: $\text{Det}^{1/n}$ Values of Posterior Variance Matrices

The first thing we notice, is that there are no entropy values calculated. The reason for this is because when the mean is known with the variance being inferred with data reduction, it is no longer a basic known distribution and hence there is no nice closed form for the entropy.

As we see with Figure 3.8 and Figure 3.9 , there is no obvious pattern visible with the different proportions for the data transformations. The reason for this is because of the covariance term in the variance matrix. The problem is for this example, the bivariate data is transformed to either look at the first coordinate or to look at the second. So at no point are the coordinates considered together. This means there is no chance for a comparison to occur and hence there is no way for the covariance term of the variance matrix to converge. This means the covariance term does a random walk throughout the entire MCMC chain. This will create a random covariance term for each of the different distributions and thus make it is of no use to use any scoring rule of the variance matrix of the distribution as it will be stochastic and therefore unusable.

In order to resolve this problem, we fix the covariance term and only allow the two variance terms on the main diagonal of the variance matrix to change. This solution to the problem has both positives and negatives with the analysis of the variance matrix of the different posterior distributions. Firstly, the main positive is that there is now none of the sampling error and randomness occurring through each MCMC run and hence that will make the resulting evaluations more accurate and informative. The negative with

doing this and keeping one of the terms constant, is that now the variance in the third coordinate of the posterior variance is always going to be zero. The problem with doing this is now any scoring rule which includes multiplication, e.g. the determinant function will return a zero value and not give any information.

3.3.1 Variance Matrix Restriction

The restriction which we have mentioned was assuming the covariance term of the variance matrix is constant and is equal to zero. Also, as previously mentioned, we will only be analysing the trace score function as it is additive, hence the zero value will not effect the other information obtained from the other variance terms.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \mapsto \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ 0 \end{pmatrix} \tag{3.15}$$

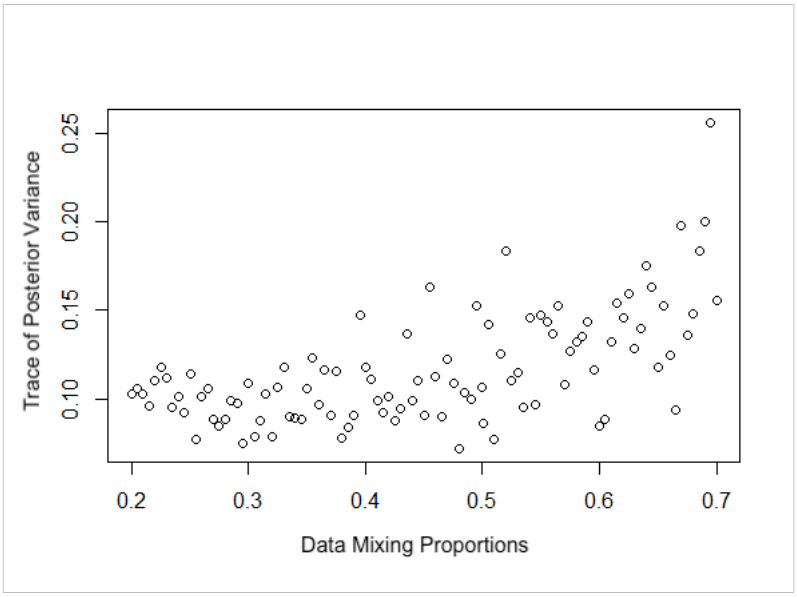


Figure 3.10: Trace Values of Restricted Posterior Variance Matrices

Figure 3.10 is now an improvement on Figure 3.8 as there is no longer the stochastic covariance term and hence there is now a pattern over the different data reductions occurring. In order to help confirm this pattern, we will simulate this again with a larger number of simulations by decreasing Δp and doing three replicates for each p.

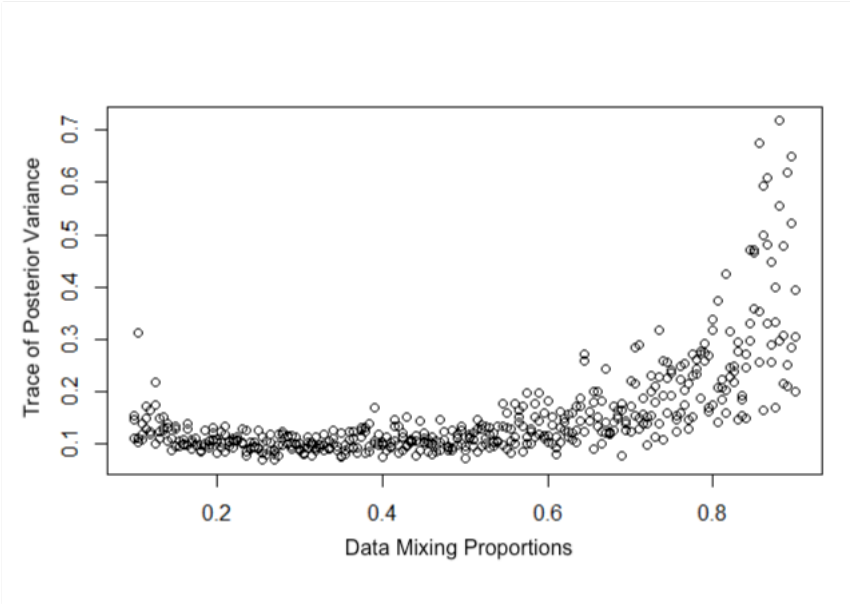


Figure 3.11: Trace Values of Restricted Posterior Variance Matrices, with Increased Trials

From Figure 3.11, it appears that the variance increases as p increases, for this situation that means as there is less of the second coordinate being considered. This matches with what we expect as in order to decrease variance, we would need more of the more variable data point.

3.3.2 Optimal value of p for the given scoring rule

We again use LOESS in order to obtain a fitted curve to the data plotted in Figure 3.11 and then derive an approximate function to represent the fitted curve.

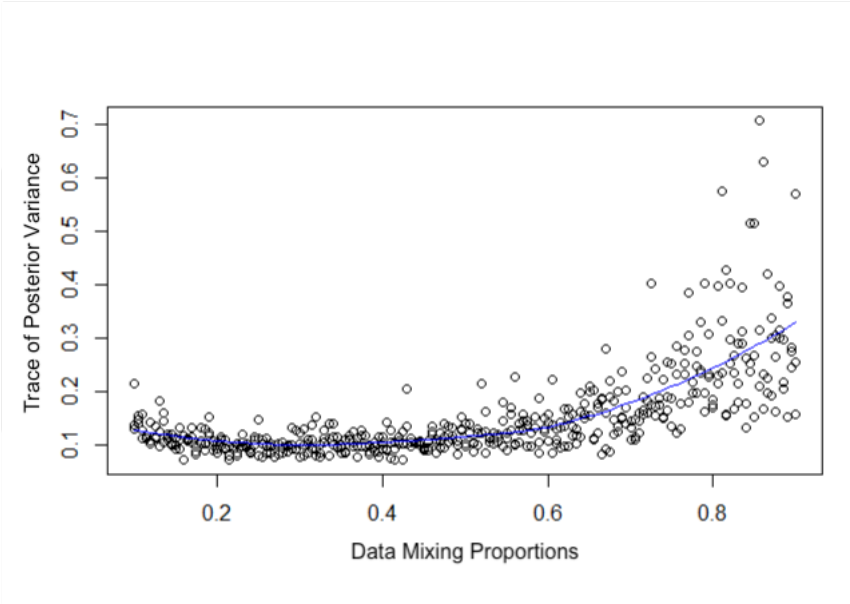


Figure 3.12: LOESS Fitted Curve - Posterior Variance Trace Values

From this we obtain the approximate formula

$$y = 0.7153x^2 - 0.4987x + 0.1776 \tag{3.16}$$

To ensure that this is an accurate approximation for the LOESS curve, we will compare it solely with the actual graph of the fitted curve.

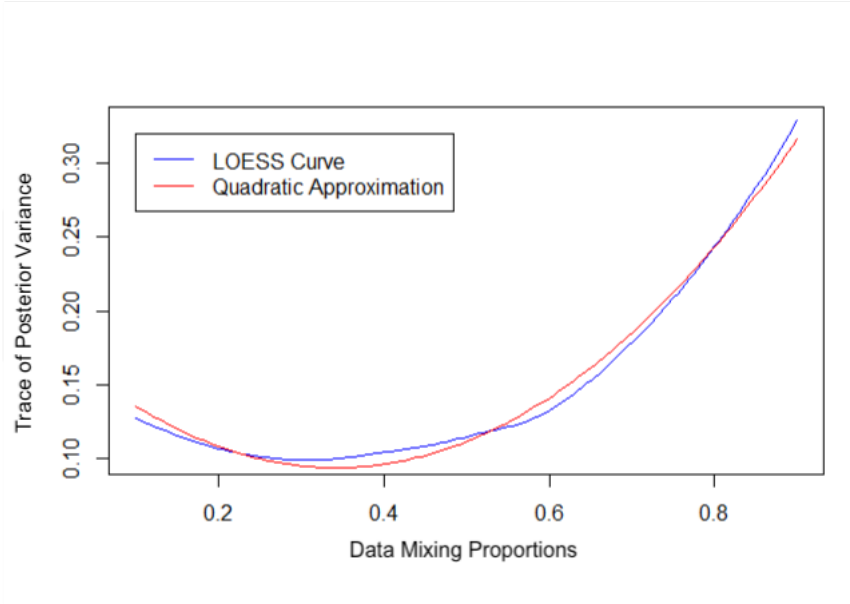


Figure 3.13: LOESS Fitted Curve and Quadratic Function Comparison

This confirms the quadratic function is a good approximation for these specific trace values and hence can be used to make assumptions about the posterior distribution and its corresponding variance matrix.

Optimising this quadratic function, we obtain the following optimal p value,

$$\hat{p} = \frac{0.4987}{1.4306} = \frac{4987}{14306} \approx 0.348595 \tag{3.17}$$

What we conclude from this is that based on the trace function, in order to minimise the variance of the posterior distribution with mean value unknown,

we should reduce the data with 0.349 of the vector a and the rest times the vector b. However, any proportion in the range of $(0.3, 0.5)$ will approximately minimise the trace value.

3.4 Comparing Scoring Rules

In order to help determine which scoring rule should be used for the optimisation, we need to find methods and tests which can be used in order to select the better of the possible scoring rules analysed.

For the following analysis' we will use the variance known with reduced data case and still use the values defined in Equation 3.2.

3.4.1 Sensitivity of the Scoring Rules

In order to compare the sensitivity of both the determinant and the trace over all of the posterior distributions covariance matrices, we will plot their own respective values against each other. From this plot we will then be able to see the sensitivity of the scoring rule by how they change over the possible range of posterior distributions.

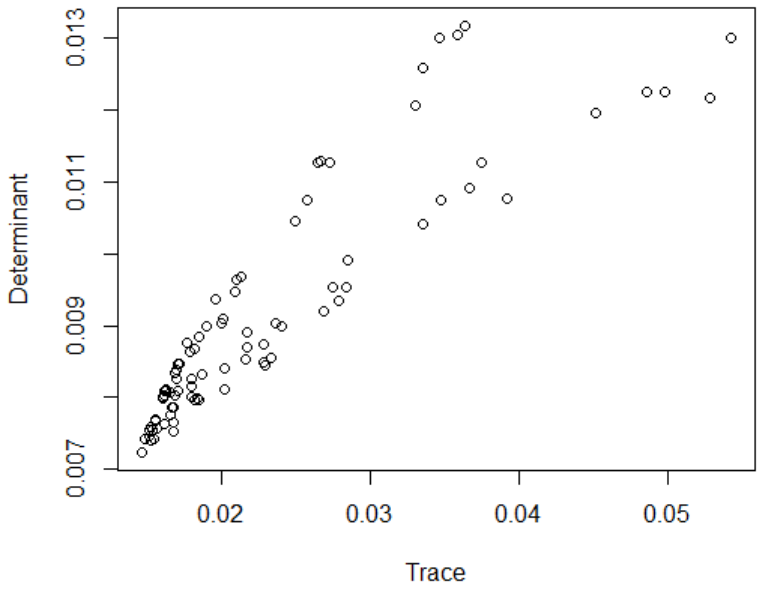


Figure 3.14: Trace Vs Determinant Comparison, Different Scale

From Figure 3.14, the first observation we can make is that there are points in which the determinant scoring rule has returned the same value and which the trace scoring rule has returned different values. This is a positive for the trace scoring rule as we are expecting to obtain different results and are wanting differentiation between the different posterior distribution's.

However, it is important to note that the two scoring rules are on very different scales and hence we will also re-plot it with equal axis in order to see the true comparison of the two rules.

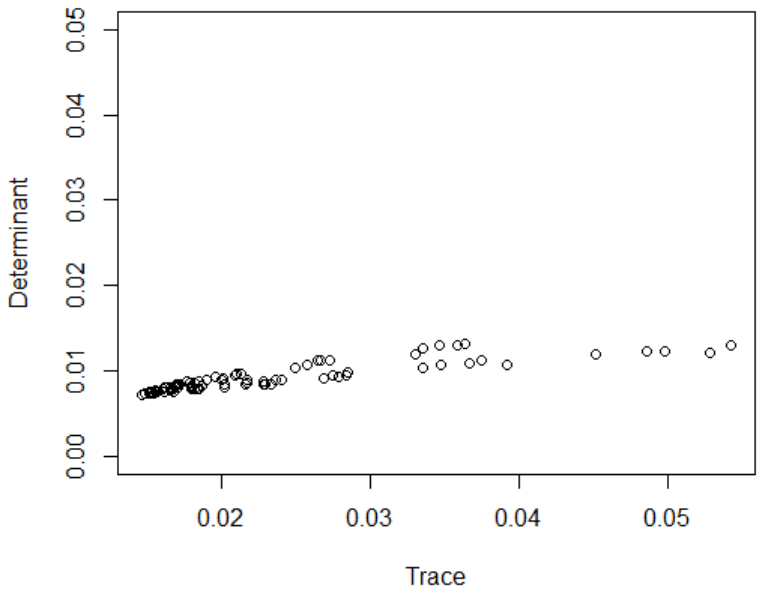


Figure 3.15: Trace Vs Determinant Comparison, Same Scale

Now that the axis are the same, we see a clear difference in the two scoring rules. What this plot indicates is that as the trace values are changing over the differing posterior distributions, the determinant values remain almost unchanged over the different distributions. This indicates that these two scoring rules are on very different scales and hence can not be compared on the same scale.

Overall, the ability to differentiate between distributions given to us from the trace scoring rule is a positive for using it to analyse the variance matrix.

3.4.2 Accuracy over Repeat Simulations

Another test to be done in order to help us compare these two scoring rules are to fix the value of p and then look at the variance of the results of both of the scores. This process will be repeated over all of the different p values and we can see the stability of the scoring rule and again give an indication to the viability of it.

3.4.2.1 Variance

The first of these checks that needs to be done on the possible scoring rules is to see how much variation there is in repeated trials when the p value is fixed in the repeats.

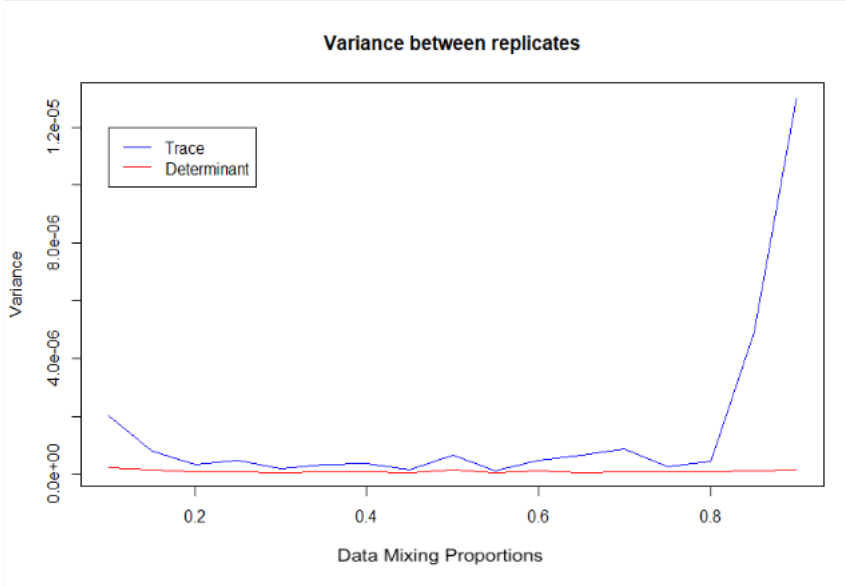


Figure 3.16: Variance of Trace and Determinant Scoring Rules over Repeat Trials

What we see is that over a large portion of the domain, from 0.2 to 0.8, the variance for both of these scoring rules is relatively consistent and also low. We do see an increase in the trace values variance at the extreme ends of the domain but however this is unlikely to be a negative as these values are not likely to be where the partition is going to be made in order to optimise the variance of the posterior distribution.

The reduced amount of variance we can see in the determinant function is also going to be a fact of the lower sensitivity found in the previous section.

3.4.2.2 Coefficient of Variation

In order to check that the variance values are still proportionally small values then we need to also look at the coefficient of variation (CV) values in the same way we looked at the variance above.

The CV is a way of measuring the variability of data while at the same time eliminating units from this measure [1]. This means we can then fairly compare the CV of different data sets even if they have different units themselves.

The units are eliminated by normalising the variance by dividing by the mean of the data points for each different p value.

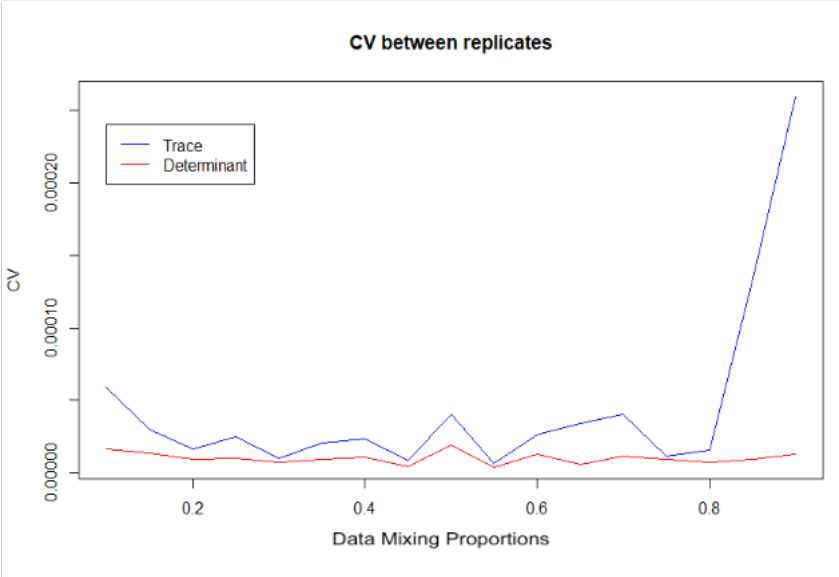


Figure 3.17: CV of Trace and Determinant Scoring Rules over Repeat Trials

We see that the corresponding values of the CV have increased by a factor of 10 but both of them have stayed proportional to each other with the same relationship. This means we can make the same conclusions that we did from the variance plot and which both of these scores are equally good choices in the more viable region of 0.2 to 0.8.

Chapter 4

Analysis of Phylogentic Tree Posterior Distribution

In this chapter we extend the analysis to GBS analysis in Chapter 3 for inferring species tree and ancestral demographics. This analysis is going to be a proof of concept and hence will be completed on a simulated data set.

We will be using the recently published package SNAPPER [18], which can be found in the BEAST2 package [2]. SNAPPER has been designed to take independent biallelic (two state) markers and to return a sample from the posterior distribution of species trees, divergence times and also ancestral population sizes. The default for SNAPPER is to take SNP data, that is that all the markers are assumed to be polymorphic (non-constant).

It was argued in [4] that ignoring constant sites led to a loss of important information about population sizes. With GBS we have a choice for each and

every ligation point of selecting a downstream site to be polymorphic or a site purely at random, irrespective if it is polymorphic or not. In terms of our simple experiment, choosing the ith site to be polymorphic is equivalent to to choosing $a^T x_i$, whereas choosing the ith without any filtering is equivalent to choosing $b^T x_i$. We can define and compute the likelihood for each site, and we face the same question as earlier, what proportion of each data type do we include?

The extraction of the different types of marker from a GBS analysis requires a significant amount of bioinformatics work which falls outside of the scope of this thesis, hence we use simulated data. The method for simulating the SNP's is based on a given species tree. Data simulation will be carried out using the program SimSnap [4]. This program requires the user to give both a species tree for the markers to be simulated on and also the number of both filtered and unfiltered marker data points required. The filtering process for the marker simulator excludes constant sites.

We use the package SNAPPER [18] in the program BEAST2 [2] in order to analyse the mixed filtered and unfiltered markers and run a MCMC to sample from the posterior distribution of tree parameters. We ran MCMC for one million iterations with a thinning of one in a thousand. SNAPPER outputs log files from each of the posterior draws and a tree file containing all of the phylogenetic tree information. The posterior draw values which are returned are the different theta values, posterior probabilities, tree likelihood values, prior values and tree heights and lengths. These posterior draw val-

ues and their respective species trees will be analysed using both visual and numeric inspection methods. To ease comparison with the earlier studies, we do not sample species tree itself, but only the parameters of divergence times and population sizes. These variables have proven difficult to infer in earlier studies, irrespective of the tree.

4.1 First Analysis: highly divergent sequences

Marker data will be simulated on a phylogeny (Figure 4.1) with 3 different species; A, B and C. The reason for not having more species in this analysis is because it reduces computation time and allows a more in-depth investigation.

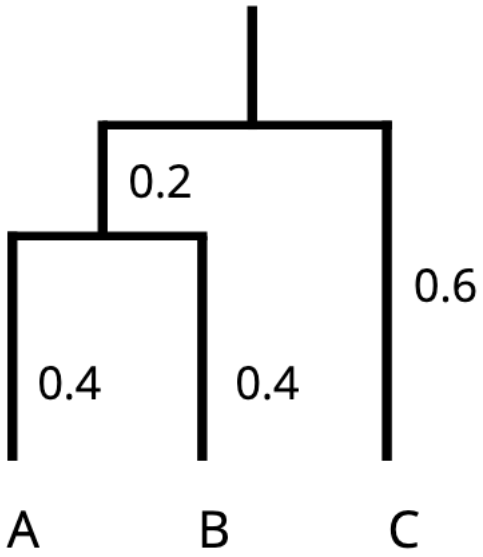


Figure 4.1: 3 Species Phylogenetic Tree

The main parameters which we are wanting to have comparable and wish to see a relationship between are the population size (θ) values along each branch which will be drawn from the posterior distribution, with the other parameters, for example tree height, staying constant.

For this analysis, different marker data sets were created with different proportions of un-filtered data included, ranging from 0% to 100%. These data sets were then ran through Snapper and the resulting posterior sample variance matrices were analysed by using the trace scoring rule.

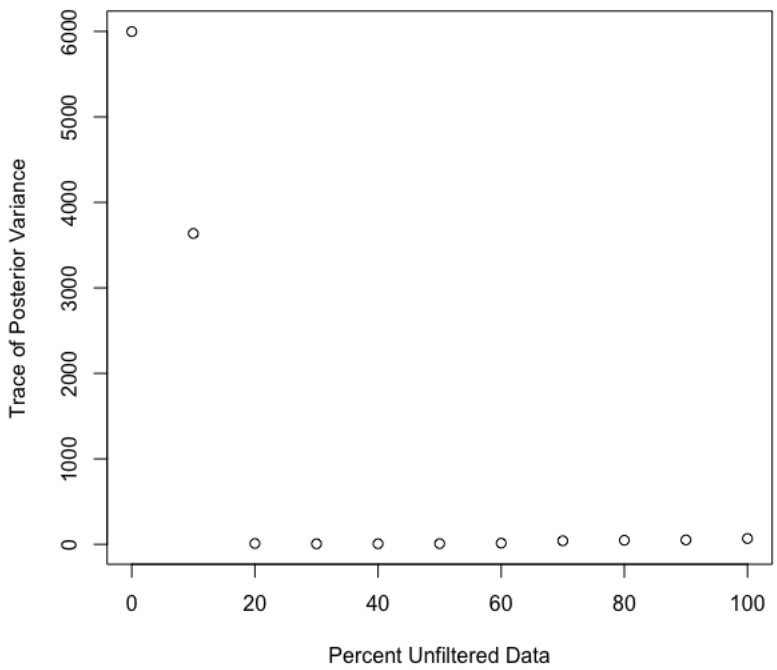


Figure 4.2: Trace Values of Estimated Posterior Variance Matrix

Figure 4.2 shows a plot of the trace of the estimated posterior variance matrices for the θ values. What we can see from this plot is that there are large variance score values present for the posterior distributions defined with 0% and 10% of unfiltered data used. In order to try and observe if there is any visible trend ignoring these potentially non convergent chain score values and plot the same data except remove the first two data points.

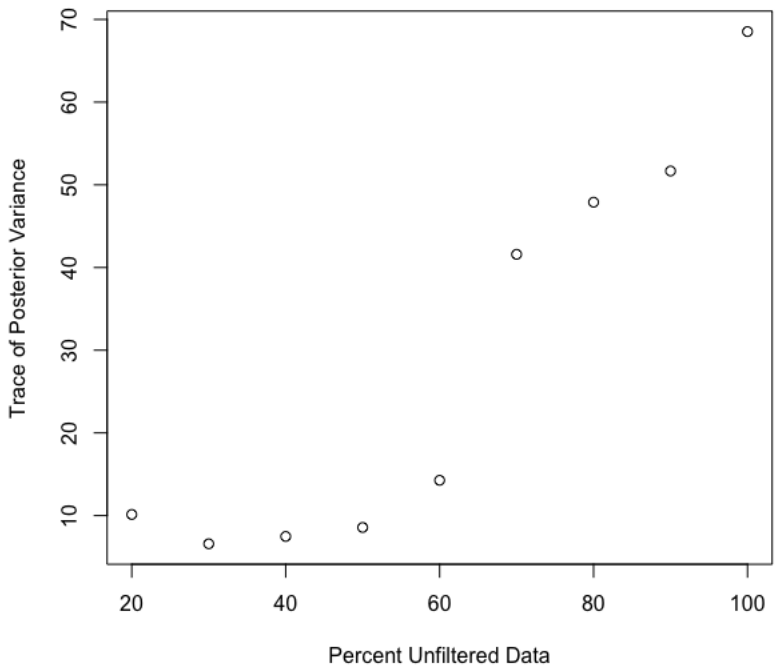


Figure 4.3: Trace Values of Estimated Posterior Variance Matrix, omitting $p \leq 10\%$

There is a quadratic-like relationship with minimum around 30% unfiltered data. In order to confirm the presence of a quadratic-like relationship and to get a more exact approximation for the minimising critical point we will fit a line of best fit with a quadratic model.

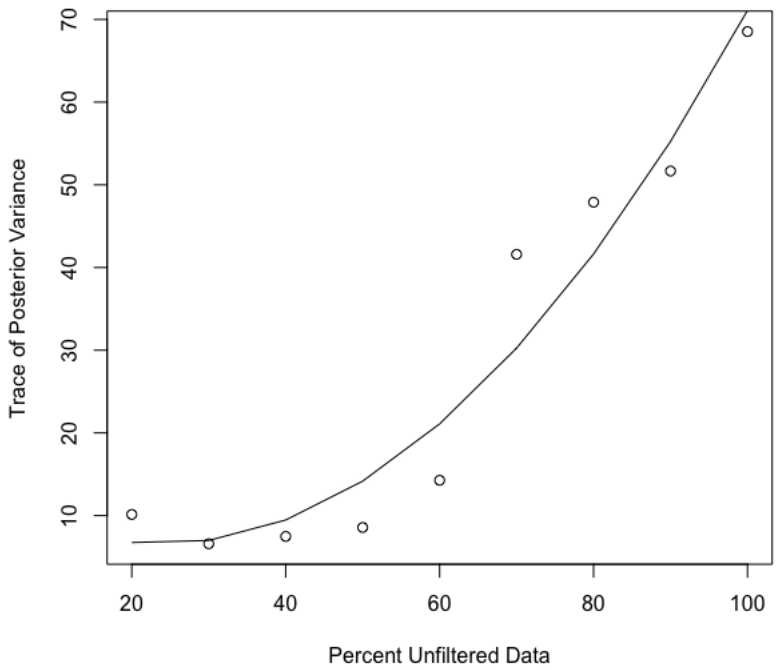


Figure 4.4: Fitted Quadratic Curve for Estimated Posterior Variance Trace Values

This curve has the following equation which can be easily minimised

$$y = 0.01114x^2 - 0.53247x + 12.93647. \tag{4.1}$$

The percentage value in order to minimise this value is

$$\frac{53247}{2228} \approx 23.899. \tag{4.2}$$

What this can imply is if these values can be trusted considering the lack of convergence in some of the chains, 24% of unfiltered data used for the data set is the optimal in order to decrease the posterior variance. This is a

significant result as it demonstrates clearly that the current practice of using 0% unfiltered data is non-optimal.

4.2 Second Analysis: less divergent sequences

The branch lengths used in the previous experiment are longer that those which would be expected in practice. Hence we conducted a second experiment with branch lengths one tenth of the previous case (Figure 4.5).

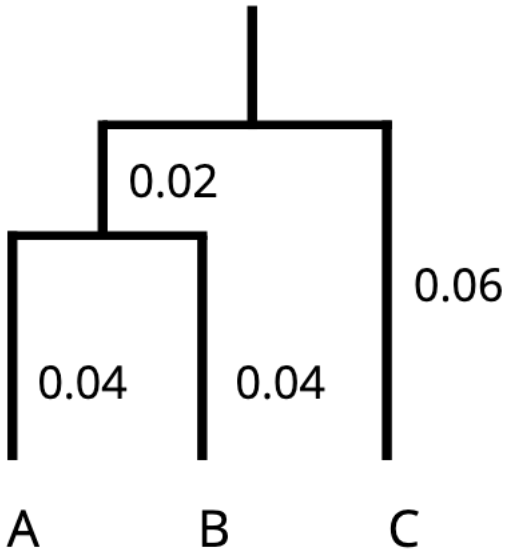


Figure 4.5: Less Divergent 3 Species Phylogenetic Tree

To improve the reliability for the estimates of each posterior variance matrix we ran three chains in parallel for each filtration percentage.

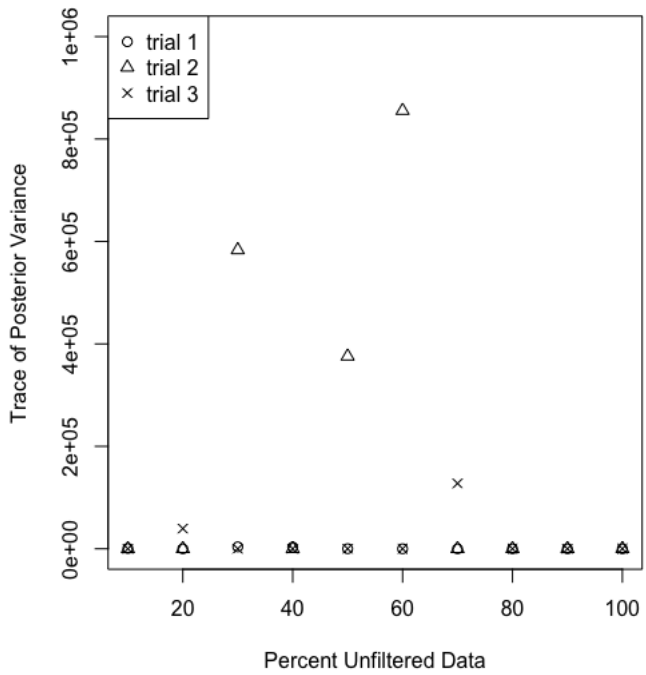


Figure 4.6: Trace Values versus Filtration Rates, with all Trials

Figure 4.6 plots the estimates of the traces for the different posterior variance matrices for each run. What we observe is that there are significant outliers, which we determined were due to lack of convergence.

In order to be able to see the true values from converged posterior draw values in more detail and find a quadratic model to approximate the pattern, we remove these non convergent trace values. As we have multiple chains for each percentage value, this will still allow us to observe a relationship.

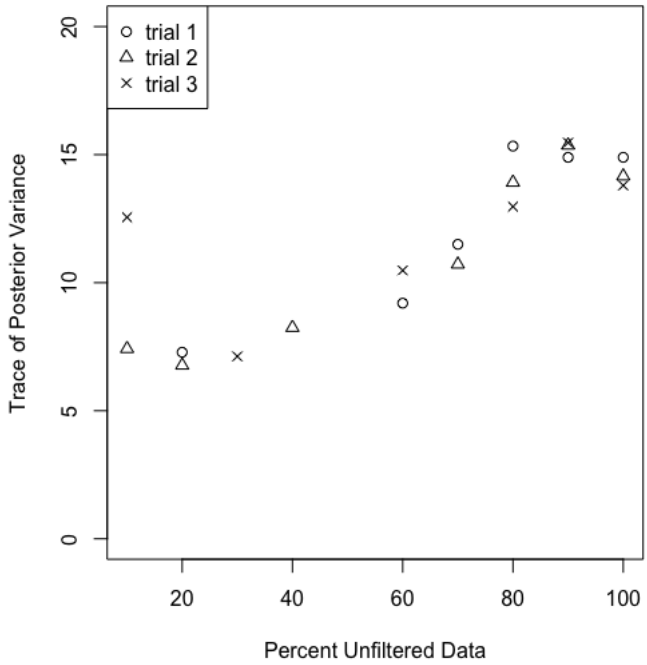


Figure 4.7: Trace Values versus Filtration Rates, with only Converged Trials

With this new plot for the trace values of the posterior variances in Figure 4.7, we start to see that there is a visible pattern. In order to analyse this pattern in more detail and try and obtain the percentage value which will minimise the trace, we will fit a quadratic curve in order to model this data.

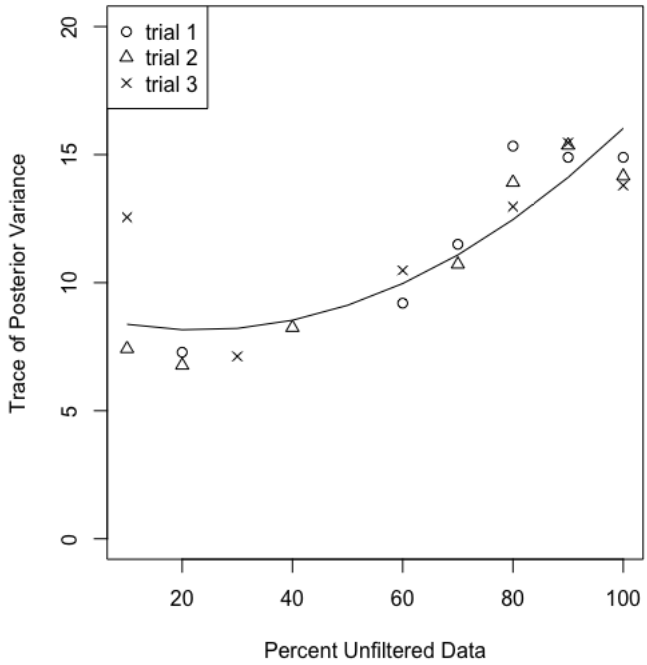


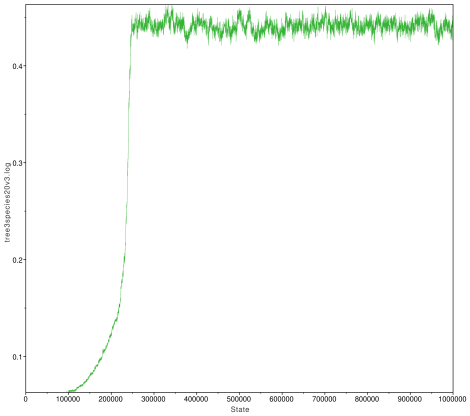
Figure 4.8: Trace Values versus Different Filtration Rates, with Quadratic Fitted Curve

This curve has the following analytic equation

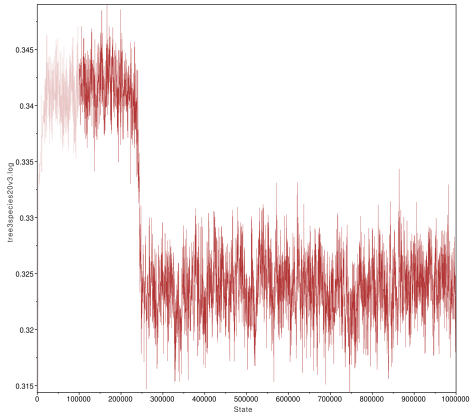
$$y = 13.289x^2 - 6.119x + 8.854 \tag{4.3}$$

The percentage value in order to minimise the posterior trace is approximately 23.0228%. The positive result with this analysis with multiple chains and the previous single chained analysis is that they have returned similar percent of unfiltered data in order to optimise. This repeated result gives us confidence that we are observing the true optimum value.

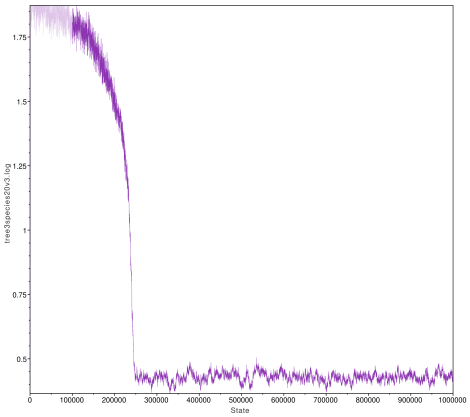
This plot and associated curve was obtained with removing the outlier points which had the extremely large trace values of their posterior variance. The most plausible reason for them having a very large sample variance trace is because after the initial burn in was removed from the draws, these chains were still not drawing from the correct distribution and had not converged yet. This theory will be confirmed visually by looking at their respective trace plots for different tree parameters.



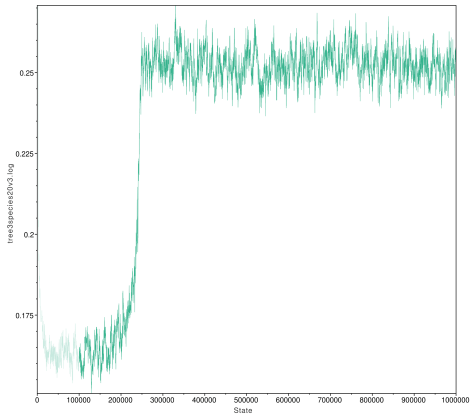
(a) Trace Plot for Theta 1



(b) Trace Plot for Theta 2

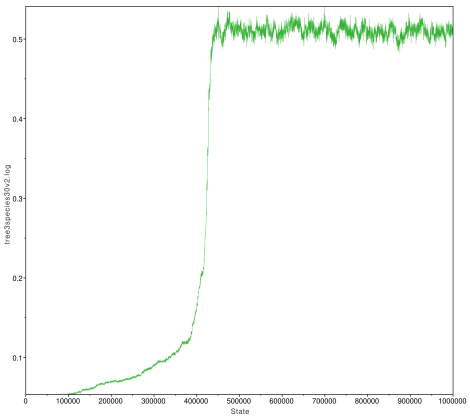


(c) Trace Plot for Theta 3

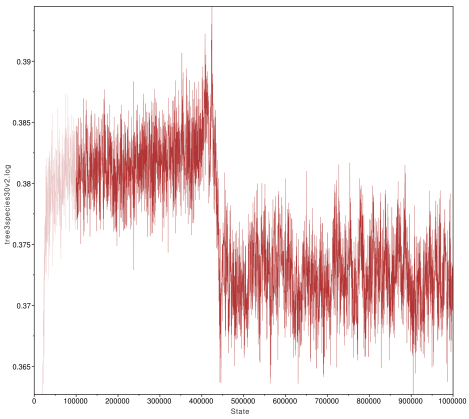


(d) Trace Plot for Theta 4

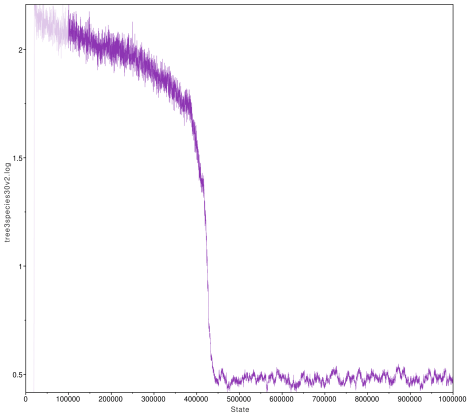
Figure 4.9: MCMC Trace Plots - 20% Unfiltered Data v3



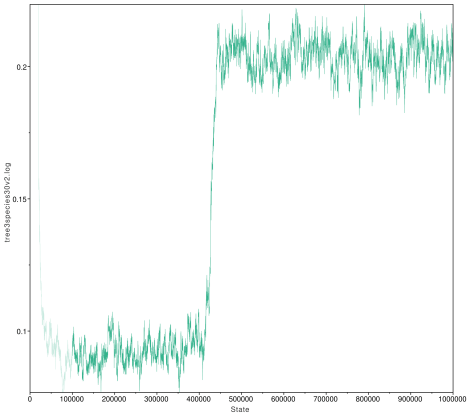
(a) Trace Plot for Theta 1



(b) Trace Plot for Theta 2



(c) Trace Plot for Theta 3



(d) Trace Plot for Theta 4

Figure 4.10: MCMC Trace Plots - 30% Unfiltered Data v2

This theory was checked and confirmed using a visual check of the trace plots of the outlier chains seen in Figure 4.9 and 4.10. What we can visually obtain is that a larger burn in percentage should be used. More specifically we can for example see that the non convergent case for 20% unfiltered data required a 30% burn in and the case for 30% unfiltered data required a 50% burn in to be removed. Once this process was completed for each of the non convergent cases, their new posterior variance trace value was obtained and then added to the plot.

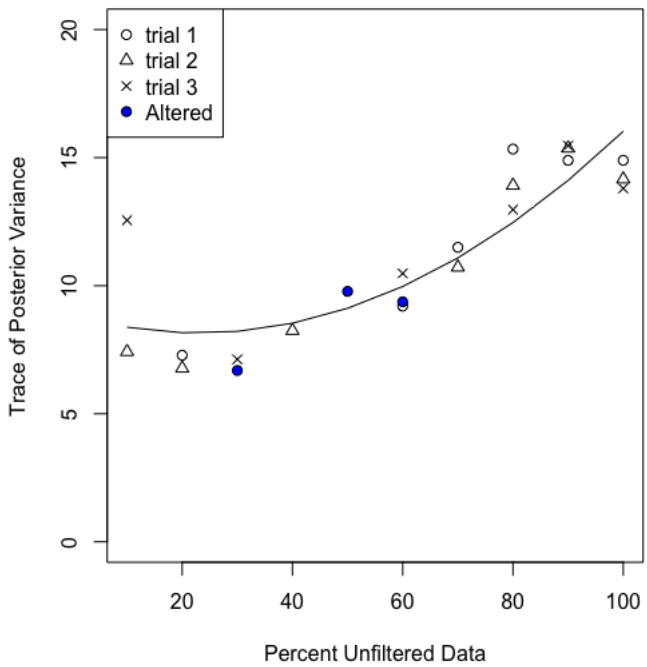
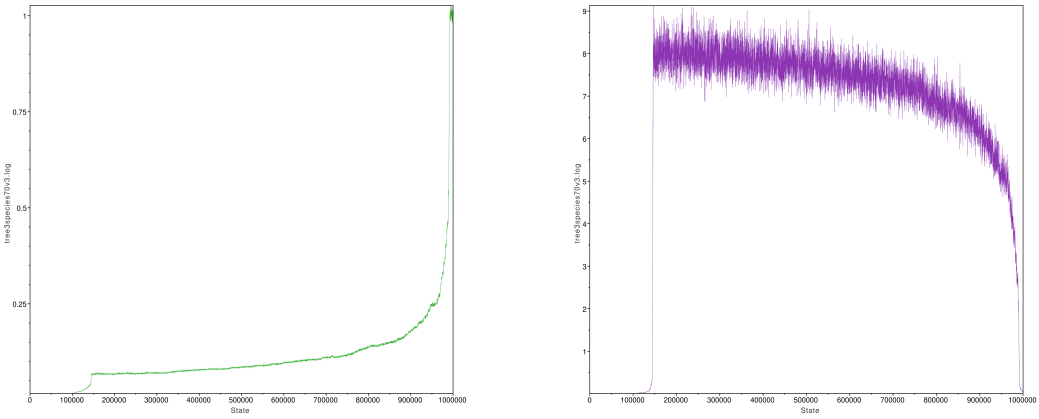


Figure 4.11: Trace Values versus Filtration Rates, all Trials Converging

What can be seen in Figure 4.11 is that once we only look at the converged parts of the outlier chains, we still see the trace value's following the same

curve and hence following the same relationship.

One note that will be made from this specific analysis is that one of the non convergent cases could not be altered in order to obtain convergence.



(a) Trace Plot for Theta 1

(b) Trace Plot for Theta 3

Figure 4.12: MCMC Trace Plots - 70% Unfiltered Data v3

Figure 4.12 gives the trace plot for a choice of 70% unfiltered data. We see from this plot that during the one million long MCMC, convergence has still not been reached from the initial starting state and hence any of the draws that were obtained during this one chain is not representative of the posterior distribution we wish to analyse. Because of this reason we ignore the results from this repeat trial as a failed MCMC.

As we have been able to conclude that including 23% of unfiltered SNP data in the MCMC analysis, we run another MCMC analysis with this optimal value and will compare different aspects of this optimal posterior distributions with the other posterior distributions.

Using the same method for the MCMC posterior distribution analysis, we obtain posterior distribution draws with the optimal 23% of unfiltered data included. The trace of the sample variance matrix was obtained for this optimal distribution and then compared with the previously obtained values.

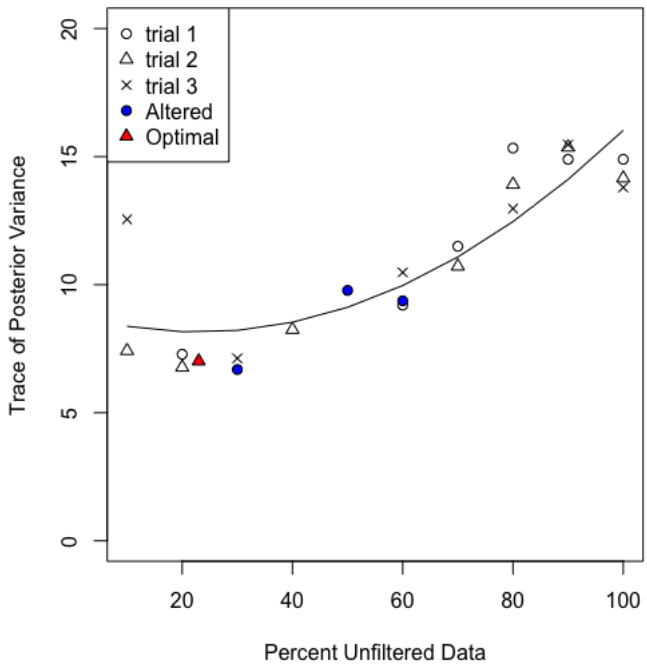


Figure 4.13: Trace Values versus Filtration Rates, including the Optimal Data Filtration

What we can see with this visualisation in Figure 4.13, is that the estimated optimal percentage is returning a near minimal trace value for the posterior variance matrix. This once again is another proof of concept for improving the posterior distribution accuracy and information provided by applying

different proportions of filtered and unfiltered data.

As mentioned for the highly divergent example, this is a significant result as it demonstrates that the current practise of using 0% unfiltered data is not the optimal method to be using.

4.2.1 Alternative Scoring Rule Check

If we want to include the covariance terms for the scoring rule analysis, we can use the determinant scoring rule to compare the estimated optimal percentage from our trace analysis. The optimal percentage of unfiltered data to include for reference was 23%.

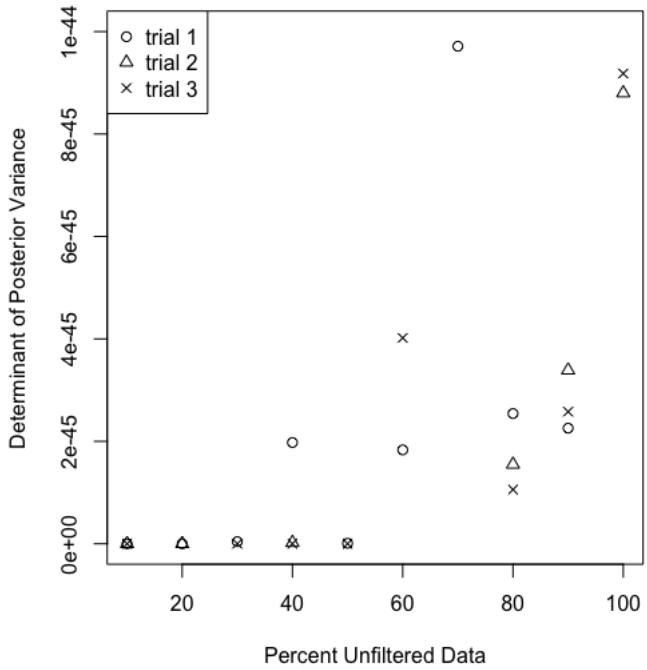


Figure 4.14: Determinant Values versus Filtration Rates

What we can see is that when we include the covariance terms of the variance matrix and it is now a multiplicative scoring rule, the plot has now changed. The range of the scoring values has decreased to smaller values and there is a new relationship which we observe over the percentages of unfiltered data. What we see is that for any percentage below 50%, the determinant value is able to be minimised and any posterior distributions created with more than 50% of unfiltered data will have a larger determinant value. A larger determinant value or a larger generalized variance value will imply larger variance and covariance terms and hence more variation from the posterior distribution draws.

With the approximated minimum of 23% falling in the range of the smaller determinant values, this confirms that our found minimum will also have a minimal generalized variance value and hence another confirmation for our found analysis.

4.3 Conclusions

This chapter describes a proof of concept investigation into whether optimisation of the posterior variance is possible and desirable. We have observed that the traditional method of always conditioning on the data being observed starting at a variable SNP is non-optimal. For this particular data which was simulated from this specific phylogenetic tree, the optimal is shown to be to include 23% of data which is not conditioned on starting from a SNP and could originate at a constant site instead.

This conclusion is then aided with the check of the determinant scoring rule as a secondary check. What we observe from this is that any of the posterior distribution's defined with 40% or less unfiltered data will have a minimal determinant score value. This agrees with the 23% minimum that was found earlier.

Clearly, there is considerable scope to expand the size and range of this in-

vestigation. That said, the MCMC analyses are computationally expensive, and large-scale analyses would require considerable computational resources and time.

Chapter 5

Conclusions

In this thesis, we set out to reduce the variance and increase the information obtained when looking at the posterior distribution of phylogenetic trees with marker data obtained from GBS.

In Chapter 2 we introduced our general definitions for filtering the data which was defined as reduced data. The simpler Gaussian distribution with reduced data is defined analytically, along with the different scoring rules for these analytically defined distributions. These rules included the trace of the posterior variance matrix, the nth root of the determinant of the posterior variance matrix and the continuous Shannon entropy of the posterior distribution. We proved that these scoring rules, which we used for data selection, are all convex and hence any minimum solved for is also the global minimum.

In Chapter 3 we moved away from analytic analysis and moved into analysis of the different posterior distributions using Monte Carlo samples. We were

able to show that we were drawing from the same posterior distribution as defined in Chapter 2. We then explored the different scoring rules of the different posterior distributions and attempted to optimise the proportion of data mixture based on these scoring rules. Plots for these different scoring rules were obtained and quadratic functions were fitted to them to allow us to optimise for the best posterior distribution. Other aspects of the scoring rules were compared to help aid with the choice of scoring rule. These aspects were the accuracy and the sensitivity of the different scoring rules.

In Chapter 4 we applied the chosen trace scoring rule to now look at the actual posterior distribution for the application. This is the posterior distribution of phylogenetic tree and their corresponding parameters defined with marker data. This marker data obtained from GBS can be chosen to be either filtered or unfiltered and then the entire data set is a mixture of these two types of data at some proportion. The significant result that was obtained was the demonstration that the current practise of including 0% unfiltered data is a non-optimal method.

Bibliography

[1] Hervé Abdi, *Coefficient of variation*, Encyclopedia of research design **1** (2010), 169–171.

[2] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond, *Beast 2: a software platform for bayesian evolutionary analysis*, PLoS Comput Biol **10** (2014), no. 4, e1003537.

[3] Stephen P Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[4] David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A Rosenberg, and Arindam RoyChoudhury, *Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis*, Molecular biology and evolution **29** (2012), no. 8, 1917–1932.

[5] Siddhartha Chib and Edward Greenberg, *Understanding the Metropolis-Hastings algorithm*, The American Statistician **49** (1995), no. 4, 327–335.

[6] Charles Darwin, *The origin of species*, PF Collier & son New York, 1909.

[7] Jason V Davis and Inderjit S Dhillon, *Differential entropic clustering of multivariate Gaussians*, Advances in Neural Information Processing Systems, 2007, pp. 337–344.

[8] Arthur P Dempster, *A generalization of Bayesian inference*, Journal of the Royal Statistical Society: Series B (Methodological) **30** (1968), no. 2, 205–232.

[9] Iddo Eliazar, *How random is a random vector?*, Annals of Physics **363** (2015), 164–184.

[10] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin, *Bayesian data analysis*, CRC press, 2013.

[11] Tilmann Gneiting and Adrian E Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association **102** (2007), no. 477, 359–378.

[12] John P Huelsenbeck and Fredrik Ronquist, *MrBayes: Bayesian inference of phylogenetic trees*, Bioinformatics **17** (2001), no. 8, 754–755.

[13] Koushlesh Ranjan, Minakshi Prasad, and Gaya Prasad, *Application of molecular and serological diagnostics in veterinary parasitology*, The Journal of Advances in Parasitology **2** (2016), 80–99.

[14] Fredrik Ronquist and John P Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models*, Bioinformatics **19** (2003), no. 12, 1572–1574.

[15] David B Rosen, *How good were those probability predictions? the expected recommendation loss (ERL) scoring rule*, Maximum Entropy and Bayesian Methods, Springer, 1996, pp. 401–408.

[16] Stephan C Schuster, *Next-generation sequencing transforms today's biology*, Nature methods **5** (2008), no. 1, 16–18.

[17] Santosh Srivastava and Maya R Gupta, *Bayesian estimation of the entropy of the multivariate gaussian*, 2008 IEEE International Symposium on Information Theory, IEEE, 2008, pp. 1103–1107.

[18] Marnus Stoltz, Boris Baeumer, Remco Bouckaert, Colin Fox, Gordon Hiscott, and David Bryant, *Bayesian Inference of Species Trees using Diffusion Models*, Systematic Biology (2020), syaa051.

[19] Yousin Suh and Jan Vijg, *SNP discovery in associating genetic variation with human disease phenotypes*, Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **573** (2005), no. 1-2, 41–53.

[20] S. Wilks, *Certain generalizations in the analysis of variance*, Biometrika **24** (1932), no. 3-4, 471–494.