Sparse Hierarchical Bayesian Models for Detecting Relevant Antigenic Sites in Virus Evolution

Vinny Davies

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy

December 2016

Abstract

Understanding how virus strains offer protection against closely related emerging strains is vital for creating effective vaccines. For many viruses, including Foot-and-Mouth Disease Virus (FMDV) and the Influenza virus where multiple serotypes often co-circulate, *in vitro* testing of large numbers of vaccines can be infeasible. Therefore the development of an *in silico* predictor of cross-protection between strains is important to help optimise vaccine choice. Vaccines will offer cross-protection against closely related strains, but not against those that are antigenically distinct. To be able to predict cross-protection we must understand the antigenic variability within a virus serotype, distinct lineages of a virus, and identify the antigenic residues and evolutionary changes that cause the variability. In this thesis we present a family of sparse hierarchical Bayesian models for detecting relevant antigenic sites in virus evolution (SABRE), as well as an extended version of the method, the extended SABRE (eSABRE) method, which better takes into account the data collection process.

The SABRE methods are a family of sparse Bayesian hierarchical models that use spike and slab priors to identify sites in the viral protein which are important for the neutralisation of the virus. In this thesis we demonstrate how the SABRE methods can be used to identify antigenic residues within different serotypes and show how the SABRE method outperforms established methods, mixed-effects models based on forward variable selection or ℓ_1 regularisation, on both synthetic and viral datasets. In addition we also test a number of different versions of the SABRE method, compare conjugate and semi-conjugate prior specifications and an alternative to the spike and slab prior; the binary mask model. We also propose novel proposal mechanisms for the Markov chain Monte Carlo (MCMC) simulations, which improve mixing and convergence over that of the established component-wise Gibbs sampler. The SABRE method is then applied to datasets from FMDV and the Influenza virus in order to identify a number of known antigenic residue and to provide hypotheses of other potentially antigenic residues. We also demonstrate how the SABRE methods can be used to create accurate predictions of the important evolutionary changes of the FMDV serotypes.

In this thesis we provide an extended version of the SABRE method, the eSABRE method, based on a latent variable model. The eSABRE method takes further into account the structure of the datasets for FMDV and the Influenza virus through the latent variable model and gives an improvement in the modelling of the error. We show how the eSABRE method outperforms the SABRE methods in simulation studies and propose a new information criterion for selecting the random effects factors that should be included in the eSABRE method; block integrated Widely Applicable Information Criterion (biWAIC). We demonstrate how biWAIC performs equally to two other methods for selecting the random effects factors and combine it with the eSABRE method to apply it to two large Influenza datasets. Inference in these large datasets is computationally infeasible with the SABRE methods, but as a result of the improved structure of the likelihood, we are able to show how the eSABRE method offers a computational improvement, leading it to be used on these datasets. The results of the eSABRE method show that we can use the method in a fully automatic manner to identify a large number of antigenic residues on a variety of the antigenic sites of two Influenza serotypes, as well as making predictions of a number of nearby sites that may also be antigenic and are worthy of further experiment investigation.

Acknowledgements

I would like to start by thanking my supervisor Prof. Dirk Husmeier for sharing with me his knowledge of all things statistics and biology. Without his help and patience I definitely could not have finished this thesis to anything like the standard it is now, and it certainly would have been far too *verbose* if it had not been for his input! I would also like to thank Dr. Richard Reeve and Dr. Will Harvey for their help with the biological elements of this thesis.

Looking beyond my Ph.D. work I would like to thank my parents and girlfriend, Sam, for their support, without their support I do not think I would have survived the length of my Ph.D. I feel I should also apologies to them for my intermittent contact and responses, I have lost count of how many times I quite rightly received texts saying ‘ring your mum’ or ‘is your phone dead’.

Outside of work and beyond family, I would like to thank the various friend I have made in my time in Glasgow, without them my time would not have been so enjoyable. To name just a few, I would like thank Gabriele for the ridiculous number of games of pool and snooker we have played (I definitely won overall!), and for making me feel better about myself by drinking half-shandies. Also my friends that can drink more than a half-shandy; Shawn for the constant invites to his flat or Pets at Home and Craig for persuading me to just have one more! Finally to my fellow Ph.D. students, in particular my various office mates, who have provided both educated and uneducated discussions throughout the course of my Ph.D., both of which I appreciate equally!

Declaration of Authorship

I, Vinny Davies, declare that this thesis titled, 'Sparse Hierarchical Bayesian Models for Detecting Relevant Antigenic Sites in Virus Evolution' and the work presented in it are my own. I confirm that where I have consulted the published work of others, this is always clearly attributed.

The content of this thesis is a result of the work carried out in my Ph.D. and this work has resulted in the following papers:

- Davies et al. (2014) Sparse Bayesian variable selection for the identification of antigenic variability in the Foot- and-Mouth Disease Virus. *Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS)*, 33:149-158.
- Davies et al. (2016a) A sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution. *Computational Statistics (Under Revision)*.
- Davies et al. (2016b) Selecting random effect components in a sparse hierarchical Bayesian model for identifying antigenic variability. In Angelini, C., Rancoita, P. M. V., and Rovetta, S., editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 14-27.

The contents of these papers was written by myself with input from Dirk Husmeier, Richard Reeve and Will Harvey. Chapter 2 takes the biological descriptions and explanations given in detail in Davies et al. (2016a). Chapter 3 uses some of the methods introduced in Davies et al. (2016a). Finally Chapters 4 and 5 provide the models and results from all of the published papers (Davies et al., 2014, 2016a,b).

Finally I must note that the phylogenetic trees in this paper were constructed and provided by Will Harvey and are presented with his permission; Figures 2.1, 2.2, 2.3, 2.4, 5.8, 5.9, 5.12 and B.1.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Influenza, more commonly known as flu, and Foot-and-Mouth Disease Virus (FMDV) both come with considerable danger for those that are infected. Influenza comes in yearly outbreaks which are estimated to result in 3-5 million cases of severe illness and about 250,00-500,000 deaths (WHO, 2009), while FMDV is endemic in sub-Saharan Africa causing regular outbreaks in the cattle there (Reeve et al., 2010). Both viruses also cause severe outbreaks of the disease. Influenza has caused three pandemics in the 20th century, Spanish Influenza (1918), Asian Influenza (1958) and Hong Kong Influenza (1968), all of which have resulted in more than a million deaths, with Spanish Influenza estimated to have killed 40-50 million people alone (WHO, 2005). FMDV, as well as being endemic in sub-Saharan Africa, has also caused major outbreaks throughout the world, with the 2001 United Kingdom (UK) outbreak estimated to have resulted in the deaths of 10 million sheep and cattle (through culling) and an economic cost of around £8 billion (BBC, 2016).

To counter the effects of the virus and prevent the spread of the disease, vaccines are usually used to protect people and animals against Influenza and FMDV. However in both cases multiple strains often co-circulate and therefore vaccines must protect the person or animal against a variety of virus strains. With the continuous evolution of virus strains, vaccines only work for a short period of time. For instance the Influenza vaccine must be updated yearly to protect against the virus strains that make up that year's 'Flu Season'. When the virus strains that make up the vaccines do not match closely enough to the currently circulating strains, the effectiveness of the vaccine is reduced and the risk for the person or animal much increased.

The reason for the ever changing vaccines is to offer protection against the ever evolving strains of a particular virus. In both Influenza and FMDV there is high genetic variability and this results in changes to the virus proteins giving new virus strains;

Chapter 2. Changes in the virus proteins, known as antigenic proteins, result in differences between the virus strains, antigenic differences, and these reduce how antigenically similar the strains are, affecting the ability of the host immune system to recognise the virus; see Section 2.1. As a consequence of this antigenic variability, vaccines are only effective against strains that are genetically related and antigenically similar to the vaccine (Mattion et al., 2004). This, along with the ever evolving virus strains, motivates the need to continuously update the vaccine, however choosing the correct virus strains to make into a vaccine can be time consuming and expensive. Understanding how cross-protection, the protection against one strain conferred by previous exposure to another strain either by infection or vaccination (Paton et al., 2005), is therefore vital for understanding the severity of an outbreak and how a particular vaccine will reduce the spread of the disease.

Vaccines will not work across different serotypes, genetically and antigenically distinct virus lineages between which there is no degree of cross-protection, and often vaccines must be made up of virus strains from multiple serotypes. However within serotypes, vaccines can offer protection against groups of antigenically similar virus strains, but not against those that are antigenically distinct. Given the importance of Influenza and FMDV it is important to understand which vaccines offer protection against which currently circulating strains. To do this we must understand how genetic changes affect antigenicity and within-serotype cross-protection. Biological experiments to identify both the antigenic proteins which cause antigenic differences and the effective vaccines is time consuming and expensive. Therefore the development of *in silico* models which can predict both antigenic residues and the likely cross-protection offered by virus strains is vital for directing these experiments in an efficient manner and reducing the number of experiments that must be carried out.

The motivation behind this work is to develop models that can predict antigenically significant residues within the different serotypes of Influenza and FMDV. Doing so can lead to the identification of these antigenic residues and help guide the selection of vaccines, mitigating the effect of the circulating virus strains. In order to do this we can use genetic data and *in vitro* measures of the antigenic variability between virus strains to understand how these genetic changes affect the ability of virus strains to cross react. The measures of antigenic variability, Virus Neutralisation (VN) titre and Haemagglutination inhibition (HI) assay, approximate the extent to which one strain confers protection against another by recording the maximum dilution at which the virus-specific antibody in a sample of antiserum from a cow (VN titre) or ferret (HI assay) exposed to one strain of the virus remains able to neutralise a sample of a second virus strain. To model antigenic variability effectively we must account for the experimental effects inherent in these processes and then link the differences in the residues and evolutionary history to the differences in the measured antigenic variability. To do this effectively we must simultaneously account for the experimental variability and select which of residues have an effect on the measured antigenic variability and are therefore likely to be antigenic residues. Previous work, e.g. Reeve et al. (2010), has used basic statistical techniques such as mixed-effects models to model antigenic variability but these methods are not statistical optimal; see Chapter 3.

To achieve improved performance we propose a family of models, Sparse hierArchical Bayesian models for detecting Relevant antigenic sites in virus Evolution (SABRE), which can simultaneously account for the experimental affects and select the residues and evolutionary changes that affect the measured antigenic variability; Chapters 4 and 5. The SABRE methods are Bayesian hierarchical models that can account for the experimental affect of the data collection process through random effects, while simultaneously selecting the significant residues and evolutionary changes through the integration of spike and slab priors (Mitchell and Beauchamp, 1988). Spike and slab priors have been shown to improve variable selection and avoid the excessive shrinkage incurred by alternative methods from Chapter 3 (Mohamed et al., 2012), while hierarchical models allow consistent inference of all parameters and hyperparameters, and inference borrows strength by the systematic sharing and combination of information (Gelman et al., 2013a).

The advantages of the SABRE methods are fully discussed and demonstrated in Chapters 4 and 5, where we show that in terms of correctly selecting variables the SABRE methods outperform the alternative methods introduced in Chapter 3; classical mixed-effects models, the mixed-effects Least Absolute Shrinkage and Selection Operator (LASSO) and the mixed-effects elastic net. We additionally explore different versions of the SABRE methods, in order to find the one that best works with our data. We provide a first comparison between the binary mask model and models based on the slab and spike prior, as well as looking at how different levels of conjugacy in the hierarchical models can affect the models performance. Chapters 4 and 5 also look at various different ways of improving the mixing and performance of the model, before finally applying a version of the SABRE method to real life FMDV and Influenza datasets. Our results, compared against those already available, show the significant improvement of the SABRE methods and the improvement they offer in modelling antigenic variability and identifying antigenic residues. Our results identify a number of previously known antigenic residues, as well as providing novel predictions of other residues that are potentially antigenic (Davies et al., 2014, 2016a,b).

One problem with the SABRE methods is that they are computationally infeasible for larger datasets meaning that data simplification must be carried out or more inaccurate methods used. To counter this drawback of the SABRE methods we have proposed the extended SABRE (eSABRE) method; see Chapter 6. The eSABRE method is based on the SABRE methods but better takes into account the structure of the data from the FMDV and Influenza serotypes that we have available. In Chapter 7 we show how this method outperforms the SABRE methods from Chapter 4 in terms of variable selection on realistic simulated datasets and therefore also the alternative methods from Chapter 3. We also show how the eSABRE methods allow us to gain a massive improvement in terms computational efficiency, meaning that using the eSABRE becomes viable on the larger datasets where the SABRE method was not. We demonstrate this on the large datasets for the Influenza serotypes, identifying known antigenic residues and providing novel predictions of potential antigenic residues.

The work of this thesis has taken on the challenges provided by antigenic variability and the biological threat that it poses. We have proposed the SABRE methods which provide a technique for understanding antigenic variability and cross-protection. We have also further explored how differences between the different SABRE methods can affect inference and shown how these methods outperform the standard methods that are used. We have then proposed the eSABRE method which takes into account the data generation process better and shown how it outperforms the SABRE methods in terms of both variable selection and computational efficiency. Finally we have applied all of the proposed models to real life FMDV and Influenza datasets and the predictions we have made will help to identify more of the antigenic residues that cause antigenic variability and in long term hopefully improve the selection of effective vaccines.

1.1 Thesis Overview

This thesis has demonstrated the effectiveness of multiple models for tackling the problems caused by antigenic variability. The structure of the thesis is in the following form: Chapter 2 provides information about the biological problem, antigenic variability, and gives details of the type of data we have available and the individual datasets used in our studies. Chapter 3 introduces and discusses established methods that are used to model antigenic variability, as well as introducing the Bayesian methods that are used to construct the models proposed in this thesis. Chapter 4 introduces the SABRE methods and Chapter 5 explores different specifications that can be used with the hierarchical models proposed. Comparisons with the methods from Chapter 3 are also given and the methods are applied to real life datasets to prove predictions of residues that are potentially antigenic. Chapter 6 provides details of the eSABRE method, while Chapter 7 shows the improvements it offers over the SABRE methods, before applying it the real life Influenza and FMDV datasets. Finally Chapter 8 provides a summary of the work that has been undertaken as part of this thesis and gives details of areas for potential further work.

Chapter 2

Data

In this chapter we will provide information about the biological problem, antigenic variability, that has inspired the work in this thesis and motivates the need for statistical models to help tackle the problem and make useful biological conclusions. In Section 2.1 we introduce the biological problem, explain what type of data is available to tackle it and discuss how the data can be used to create statistical models to help understand antigenic variability. We discuss where the data comes from and the experimental variation inherent in its collection (Section 2.1.1). We then look at how the surface structure of the viruses (Section 2.1.2) and their evolutionary histories (Section 2.1.3) can be used to understand antigenic variability.

In Sections 2.2 and 2.3, we discuss Foot-and-Mouth Disease Virus (FMDV) and the Influenza (Flu) virus, and give details of the different datasets we have available for these viruses. For each of the viruses we have datasets for different serotypes, genetically and antigenically distinct virus lineages, and we introduce these and explain what dangers the different viruses cause to human and animal populations.

The final part of the chapter, Section 2.4, discusses what information we have about the antigenic sites of FMDV and Influenza serotypes introduced in Sections 2.2 and 2.3. Section 2.4 summarises the experimental information we have about each of the viruses and explains how we can use that, as well as information from other serotypes, to make informed decisions about the plausibility of some of the biological results found by our models in Chapters 5 and 7.

2.1 Antigenic Variability

Ribonucleic acid (RNA) viruses such as FMDV and Influenza have been shown to have high genetic variability (Holland et al., 1982). This variability results in changes to the virus proteins that effect recognition by the host immune system, also known as antigenic differences. Differences in these proteins, also known as antigenic proteins, affect how antigenically similar different viruses are. As a consequence of the antigenic variability in the viruses, vaccines are only effective against field strains that are genetically related and antigenically similar to the vaccine strain (Mattion et al., 2004). This feature of FMDV and Influenza makes it important to estimate antigenic similarity among strains and therefore cross-protection, the protection against one strain conferred by previous exposure to another strain by either infection or vaccination (Paton et al., 2005). Understanding cross-protection is vital for predicting the severity of an outbreak and understanding how different vaccine strains will mitigate the spread of the disease. As the testing of new candidate vaccines is expensive, the development of an *in silico* predictor that can identify which strains are likely to give the broadest cross-protection is essential.

RNA viruses are classified into serotypes, genetically and antigenically distinct virus lineages between which there is no effective degree of cross-protection. Individual vaccines may protect against large groups of genetically diverse viruses within a serotype, however there are antigenically distinct subtypes against which the vaccines do not work. Within these serotypes are significant levels of antigenic variability, which allows us to examine the relationship between genetic and antigenic variation and to determine which protein changes affect recognition by the immune system. Given the importance of FMDV and Influenza, as well as the difficulties with vaccination caused by antigenic variation, it is vital to understand how genetic changes affect antigenicity and within-serotype cross-protection. Biological experiments to confirm the effects of genetic changes are both time consuming and expensive, therefore making accurate *in silico* predictions as to which of the changes caused the antigenic variations is important to reduce the number of experiments that must be carried out.

In order to infer the antigenic importance of specific genetic changes that have occurred during the evolution of the virus, we require a measure of the antigenic similarity of any two virus strains. Virus Neutralisation (VN) titre and Haemagglutination inhibition (HI) assay give *in vitro* measures of antigenic similarity between a protective, i.e. a potential vaccine, and a challenge strain, i.e. a potential circulating virus (Hirst, 1942; WHO, 2011). They approximate the extent to which one strain confers protection against another by recording the maximum dilution at which the virus-specific antibody in a sample of antiserum from a cow (VN titre) or ferret (HI assay) exposed to one strain of the virus (the protective strain) remains able to neutralise a sample of a second virus strain (the challenge strain). Higher titres or assay measures indicate that the antiserum still neutralises the challenge strains at greater dilution and therefore that the protective and challenge strains are more antigenically similar. The highest VN titre or HI assay measurements will be when two identical strains are used as the challenge and protective strains, with any difference between the strains causing antigenic difference and lower VN titre or HI assay measurements. Gaining an effective understanding of why certain pairs of virus strains produce higher measured antigenic variability means that we can use the genetic data of newly emerging virus strains to understand the likely cross protection offered by different vaccines.

The antigenic differences between different virus strains is caused primarily by changes in the residues on the proteins on the surface of the capsid or virus shell; see Section 2.1.2. Here changes in these residues mean that virus strains are less antigenically similar, reducing the effectiveness of vaccines. However the antigenic similarity can also be affected by how the viruses within the serotypes have evolved and this must also be considered; Section 2.1.3. Finally the measured antigenic variability can be influenced by a number of experimental factors and these can affect the accuracy of the VN titre and HI assay and so must be accounted for; see Section 2.1.1. In terms of standard mixed-effects models, to be discussed in more detail in Section 3.1.1, the variables related to the residues and evolutionary history would be considered fixed-effects variables and the experimental factors would be random-effects factors.

2.1.1 Experimental Effects

The experiments to measure the antigenic variability between any two virus strains contain experimental errors in the measured VN titre or HI assay that they produce. When modelling the VN titre or HI assay it is important to take these experimental effects into account, otherwise the way we interpret the antigenic similarity of the strains will be inaccurate. The measured VN titre or HI assay can be affected by a number of things, including which challenge strain, protective strain and antiserum were used when the data was collected, as well as the date the experiment itself was completed. In the datasets we have for FMDV and Influenza, information about the factors has been recorded and we can use this in our models in Chapters 4 and 6. However not all datasets contain all the desired information, so we are limited about which factors we can account for in some datasets. The available factors are specified for each dataset in the individual sections of Section 2.2 and 2.3.

The experimental affects need to be considered in our models for a number of reasons. Individual challenge and protective strains can have different effects on the measured VN titre or HI assay. For instance some challenge strains can be more reactive against all strains causing higher measurements, while some protective strains can have higher or lower measurements against all challenge strains regardless of antigenic similarity. The animals from which the antisera come from can similarly produce different strength antisera and this can also affect the measured VN titre or HI assay. Finally it is possible that the person doing the experiment can have an affect on the measurements and while none of the datasets contain this information, we can account for it via a proxy; the date of the experiment. Initial results from Reeve et al. (2010) suggested that the protective strain did not affect the measured VN titre or HI assay, suggesting it should not be included as a random effect factor. We have initially based our choice of random effect factors on these results, but later tested which factors should be included as random effect factors through the use of information criteria.

2.1.2 Antigenic Residues

In the outer capsid or virus shell, proteins influence antigenicity. Many areas of these proteins are exposed on the surface of the capsid and among these are antigenic regions that are recognised by the host immune system. Single amino acid substitutions (mutations) within these antigenic regions can dramatically affect recognition by the immune system. Identifying the specific amino acid residues that comprise these antigenic regions and the substitutions that cause antigenic differences is critical to understanding antigenic similarity among viruses and cross-protection within serotypes. Producing models which can rank how likely residues are to have an antigenic affect is important as it can direct the biological experiments to those residues which are most likely to affect antigenicity and are therefore the most important to understand.

The data about the residues we have for the FMDV and Influenza virus looks at whether a particular residue is different for the two virus strains for which the antigenic variability is being measured, i.e. an amino acid substitution (mutation) has occurred in the evolutionary path between the two virus strains. The data is recorded as 1 if a mutation has occurred and 0 otherwise. The inclusion of a residue's data in a model from the methods in Chapters 3, 4 and 6 indicates that the particular variable has an effect on antigenic variability and the corresponding residue is therefore predicted to be antigenic. Given the virus strains tested throughout the dataset do not change during the data collection period (all viral evolution happened before this point for the virus strains in the datasets), the measurements of the residues will remain the same for a given pair of virus strains for each VN titre or HI assay measurement that they are used to produce. This however is not the case with the evolutionary data described in Section 2.1.3. The evolutionary data only remains the same for a given challenge and protective strain, and the data will not remain the same if the challenge strain is used as the protective strain and vica versa (unlike the residue data where it will). It is this structure in the genetic and evolutionary data that provided the motivation for the model described in Chapter 6.

Various pieces of information are known about the residues of the FMDV and Influenza serotypes in Sections 2.2 and 2.3 and more information about the residues of the individual serotypes is given in Section 2.4 where we classify their plausibility of being antigenic. As a general rule, residues can be classified based on there locations, with some regions known to be antigenic or provide certain functions to the virus. Information can also be taken from other serotypes of the same virus, as in many viruses certain regions can be antigenic in all tested serotypes. For this reason residues are given by their common alignment taken from Reeve et al. (2010) and Harvey et al. (2016).

2.1.3 Evolutionary History of Viruses

Changes in the antigenic proteins described in Section 2.1.2 occur as the strains within each serotype evolved. The accumulation of these changes in geographically isolated virus lineages allows for the division of serotypes into topotypes, groups of genetically similar viruses associated with a particular geographic area (Knowles and Samuel, 2003). Strains within topotypes share a common evolutionary history that is distinct from strains within other topotypes. Accounting for the genetic differences between topotypes that have arisen due to their significantly different evolutionary paths is necessary for understanding antigenic variability (Reeve et al., 2010). Interpreting the antigenic consequences of genetic differences between topotypes can improve our understanding of the evolutionary history of serotypes, as well as the likely extent of vaccine coverage across topotypes.

When we observe antigenic differences between virus lineages that we are unable to attribute to amino acid changes at any specific residue. In these cases we wish to relate the changes to the evolutionary history of the virus. This evolutionary history is ordinarily described by a phylogenetic tree, e.g. Figure 2.1, which maps the evolution of the sampled viruses (*the leaves*) back to their most recent common ancestor (*the root*). The internal vertices of the tree (*the nodes*) then represent inferred ancestors of the sampled viruses (*the leaves*). The edges joining these nodes (*the branches*) connect ancestors and their immediate evolutionary descendants, and are each associated with a set of amino acid substitutions estimated to have occurred between the nodes they connect. Groups of leaves separated from the root by a particular branch, are said to be *a clade* defined by that branch, i.e. *virus A* and *virus B* in Figure 2.1 are a clade defined by *branch x*.

The reconstruction of phylogenetic trees is not the subject of this thesis, and therefore for the datasets in Sections 2.2 and 2.3 we have used the trees generated from the paper where the data on that serotype was originally presented. Within these trees, each branch has the potential to explain antigenic differences and these are included as fixed-effects by noting whether each branch lies between the challenge and protective strains (1) or not (0) in an indicator variable, as in Reeve et al. (2010). For each pair of strains tested, it does not make a difference which virus from the pair is the challenge or protective strain,

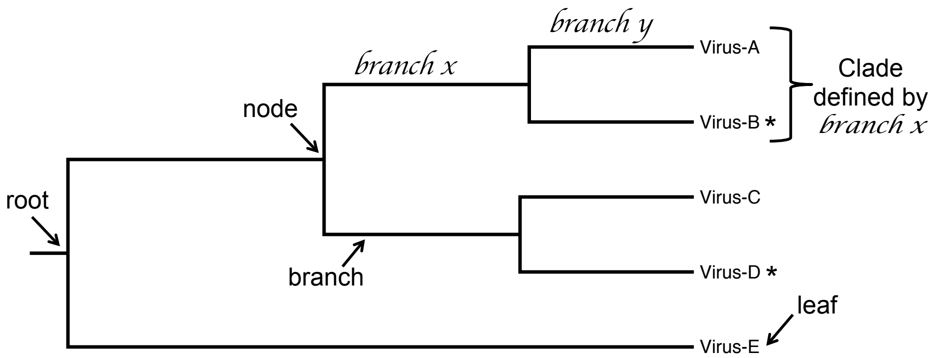


Figure 2.1: **Example Phylogentic Tree.** The phylogentic tree was created in FigTree v1.4.2. Marked on the tree are protective strains (*).

only that the branch lies between the two strains chosen. For example, in Figure 2.1, the indicator variable for *branch x* would be 1 for a comparison between any virus in the clade defined by *branch x* (*virus A or B*) and a virus outside of the clade (*viruses C, D or E*) regardless of which virus is the challenge or protective strain, and 0 otherwise. Then if there is a significant antigenic difference between *viruses A and B* and *viruses C, D and E*, the antigenic effect of *branch x* would be selected.

However, other non-antigenic properties of the virus can also affect the VN titre or HI assay measurements, and these were introduced by Davies et al. (2016a). One of those properties is that certain amino acid substitutions may increase (decrease) reactivity of the challenge strains resulting in a lower (higher) VN titre or HI assay measurements against all antisera. We call this a *reactivity effect* and include a second type of indicator variable for this type of effect. This indicator variable for *branch x* in Figure 2.1 would be 1 if the challenge strain is *virus A or B* and 0 if it is *virus C, D or E*. If challenge strains in the clade defined by *branch x* show consistently higher or lower VN titre or HI assay measurements regardless of their antigenic similarity to the protective strain, then this second type of indicator variable will be selected.

Finally, amino acid substitutions can also alter the virus so that protective strains carrying these amino acid substitutions produce antisera that have higher or lower VN titres or HI assay measurements against all challenge viruses irrespective of antigenic similarity. We call this third property an *immunogenic effect*, and include a third indicator variable for this effect. This indicator variable for *branch x* in Figure 2.1 would be 1 if the protective strain is *virus A or B* and 0 if it is *virus C, D or E*. If protective strains in the clade defined by *branch x* show consistently higher or lower VN titre or HI assay measurements regardless of their antigenic similarity to the challenge strain, then this third type of indicator variable will be selected.

While we can distinguish these three properties in theory, it is not always possible to discriminate between them in practise. For a given branch, it is only possible to define all of the properties when the clade defined by that branch includes at least one virus used as a protective strain and one as a challenge strain. Note that not all protective strains are used as challenge strains in our studies. For example, in Figure 2.1, it is possible to distinguish the three properties for *branch x* whose clade includes both a protective strain (*virus B*) and challenge strains (*viruses A and B*). However, the clade defined by *branch y* only contains a challenge strain (*virus A*) and therefore an *immunogenic effect*, an effect associated with protective strains, cannot be observed. Similarly as *virus A* is not used as a protective strain it is not possible to determine whether any variation in VN titre or HI assay measurements associated with its use as a challenge strain is the result of the antigenic distinctiveness of the virus (i.e. an antigenic change in *branch y*) or simply that the virus differs in its reactivity (i.e. a reactivity change in *branch y*).

Finally that it is worth noting the consistency of the evolutionary data for a given pair of challenge and protective strains, re-enforcing the statements made at the end of the second paragraph in Section 2.1.2. For a given pair of challenge and protective strains the variables relating to the evolutionary history of the virus will remain the same. However, unlike the residues data in Section 2.1.2, it does not remain the same when the virus strains used as the challenge and protective strains are swapped. This is a result of branches between pairs of virus strains having either reactivity or immunogenic affects depending on which virus strain is used as the challenge strain and which the protective.

2.2 Foot-and-Mouth Disease Virus

There are seven serotypes of FMDV; A, C, O, Asia 1 and South African Territories types 1, 2 and 3 (SAT1, SAT2 and SAT3). The virus is endemic in sub-Saharan Africa where six of the seven serotypes occur. Of these serotypes, SAT1 and SAT2 are responsible for the majority of FMDV outbreaks in cattle in the region and also show high levels of antigenic variability between virus strains. The significant levels of antigenic variability in these serotypes makes it important to understand cross protection between strains so that effective vaccines can be created. The high variability also allows us to examine the relationship between genetic and antigenic variation using the data provided in Reeve et al. (2010) and Maree et al. (2015).

2.2.1 SAT1 Serotype

There are two SAT1 datasets that have been available during the period in which the work for this thesis was undertaken. The original SAT1 dataset is a smaller dataset originally

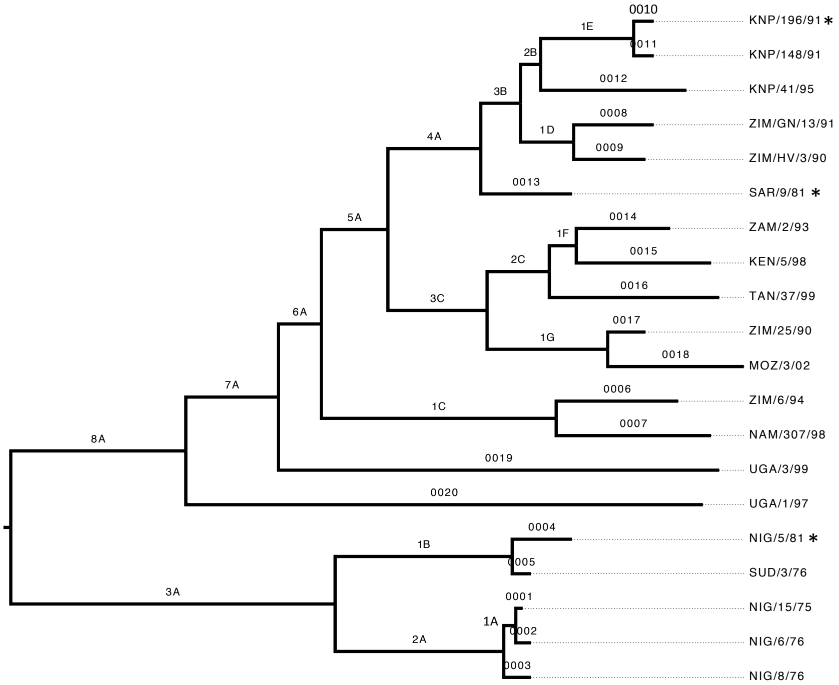


Figure 2.2: **Labelled phylogenetic tree for original SAT1 dataset described.** The labelled phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. The leaves of the phylogenetic tree, see Section 2.1.3, give the SAT1 viruses strains contained in the data, i.e. KNP/196/91. All strains are used as challenge strains and those used as protective strains are marked with a *. Branches are labelled based on their evolutionary distance from the leaves (observed virus strains). Leaf branches are denoted by numbers, while internal branches are labelled by numbers and letters, where the numbers depend on the maximum number of nodes (inferred ancestors) between this branch and any leaf which is part of the clade defined by the branch.

collected and analysed in Reeve et al. (2010) and has been available throughout all the work completed. This dataset has been used in Davies et al. (2014) and Davies et al. (2016a). Further data was collected and analysed in Maree et al. (2015), and this data became available at a later point in time and so was only analysed in Davies et al. (2016a) and Davies et al. (2016b).

The original SAT1 dataset analysed in Reeve et al. (2010) is made up of 246 VN titre measurements of comparisons between 3 protective and 20 challenge strains, where the virus strains are the leaves of the phylogenetic tree, see Section 2.1.3, in Figure 2.2. For each of these measurements, there are 754 residues in the amino acid sequence of the structural proteins. Of these, 306 are exposed on the surface of the capsid, and 137 are variable between the 20 test viruses, producing usable indicator variables to assess the antigenic effect of amino acid substitutions. The phylogenetic tree given in Figure 2.2 contains 38 branches, and it is possible to include additional variables to account for the different types of branch effect (see Section 2.1.3), resulting in 64 different indicator

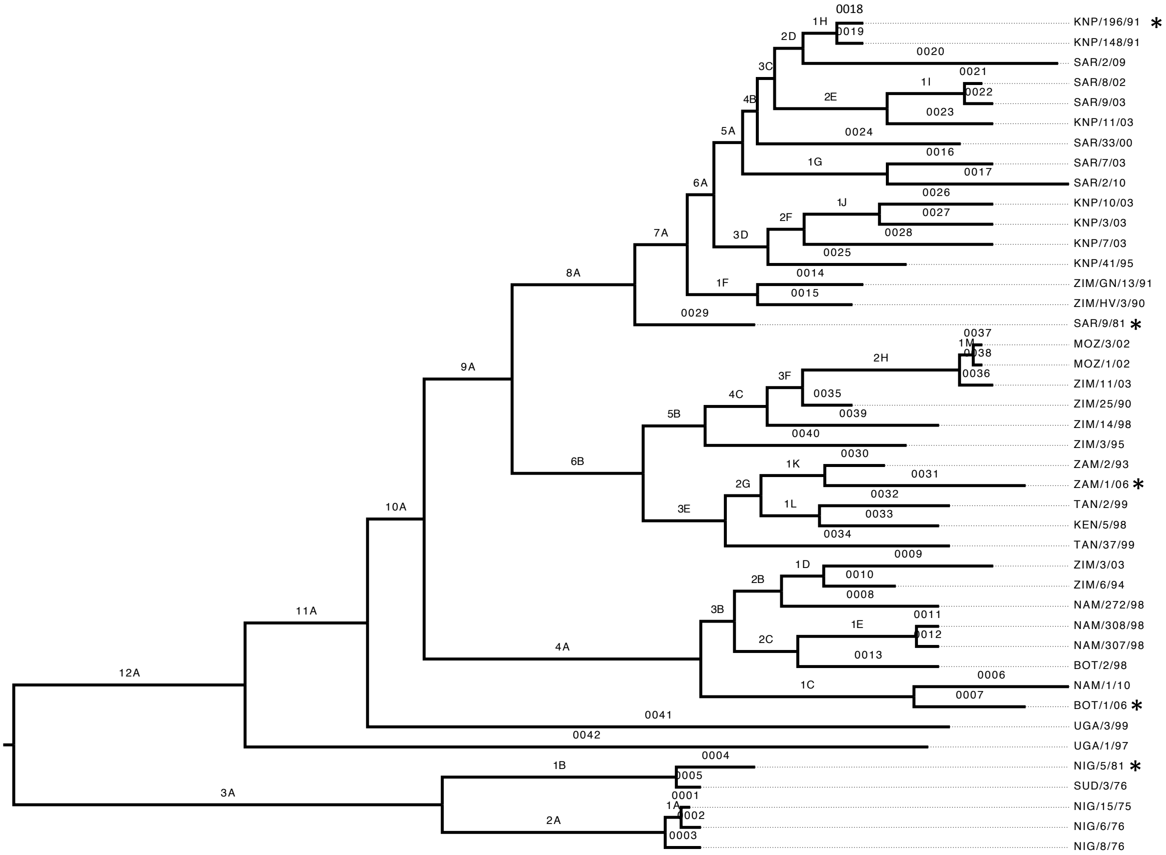


Figure 2.3: **Labelled phylogenetic tree for extended SAT1 dataset.** The labelled phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. The leaves of the phylogenetic tree, see Section 2.1.3, give the SAT1 viruses strains contained in the data, i.e. KNP/196/91. All strains are used as challenge strains and those used as protective strains are marked with a *. Branches are labelled based on their evolutionary distance from the leaves (observed virus strains). Leaf branches are denoted by numbers, while internal branches are labelled by numbers and letters, where the numbers depend on the maximum number of nodes (inferred ancestors) between this branch and any leaf which is part of the clade defined by the branch.

variables to help determine the effect of each branch and the evolution they represent. Recorded experimental effects for the original SAT1 dataset include the challenge strain, protective strain and antiserum, see Section 2.1.1, and these can be accounted for as random effects in our models in Chapter 4.

After the analysis of the original SAT1 dataset in Reeve et al. (2010), more data was collected, including additional strains and repeated experiments (Maree et al., 2015). This dataset, to be known here as the extended SAT1 dataset, includes the original SAT1 data and consists of a total of 2125 VN titre measurements with 5 protective and 42 challenge strains, where the virus strains are the leaves of the phylogenetic tree, see Section 2.1.3, in Figure 2.3. Of the 306 surface exposed sites, the amino acid sequence

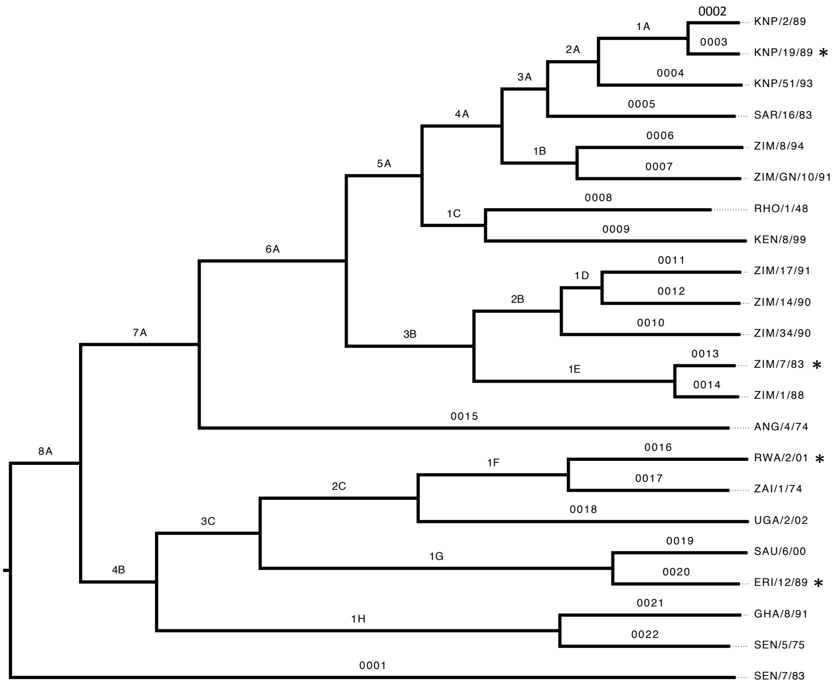


Figure 2.4: **Labelled phylogenetic tree for SAT2 dataset.** The labelled phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. The leaves of the phylogenetic tree, see Section 2.1.3, give the SAT2 viruses strains contained in the data, i.e. KNP/2/89. All strains are used as challenge strains and those used as protective strains are marked with a *. Branches are labelled based on their evolutionary distance from the leaves (observed virus strains). Leaf branches are denoted by numbers, while internal branches are labelled by numbers and letters, where the numbers depend on the maximum number of nodes (inferred ancestors) between this branch and any leaf which is part of the clade defined by the branch.

is variable between the viruses at 146. 132 variables associated with the phylogenetic tree in Figure 2.3 are also used, with the variables representing a variety of evolutionary effects (see Section 2.1.3). Recorded experimental effects for the extended SAT1 dataset include the challenge strain, protective strain, antiserum and the date of the experiment, see Section 2.1.1, and these can be accounted for as random effects in our models in Chapter 4.

2.2.2 SAT2 Serotype

The SAT2 data was originally analysed in Reeve et al. (2010) and contains 320 VN titre measurements of 4 protective and 22 challenge strains, where the virus strains are the leaves of the phylogenetic tree, see Section 2.1.3, in Figure 2.4. It contains data on 128 variable surface exposed residues and 80 variables associated with the phylogenetic tree in Figure 2.4, where the different type of evolutionary effects are taken into account (see Section 2.1.3). Recorded experimental effects for the SAT2 dataset include the challenge strain, protective strain and antiserum, see Section 2.1.1, and these can be accounted for as random effects in our models in Chapter 4.

2.3 Influenza Virus

Influenza, more commonly known as flu, is estimated to cause the death of between 250,000 and 500,000 people each year (WHO, 2009). Due to its particular danger to the old and sick, countries like the United Kingdom (UK) provide regular vaccinations for vulnerable people it an attempt to reduce the expected number of mortalities. For this reason it is vital to choose the right virus strains to be made into vaccines in order to reduce the risk of death from that year's Influenza strains. In the UK these have usually contained strains taken from three different serotypes; Influenza A (H1N1), Influenza A (H3N2) and Influenza B. We have datasets for two of these serotypes which we can use in an attempt to understand antigenic variability and the ability of different virus strains to offer cross protection.

2.3.1 Influenza A (H1N1) Serotype

H1N1 viruses entered the human population in 1977 and co-circulated with a viruses of a second influenza A subtype, H3N2, and influenza B viruses until their replacement by a novel distantly related lineage of H1N1 viruses in the 2009 swine-origin pandemic (Barr et al., 2014). During this period the influenza vaccine included a H1N1 strain which had to be updated on nine occasions in order to remain antigenically matched to, and therefore capable of protecting the human population from, circulating strains. The dataset analysed here comprises 43 H1N1 viruses collected from 1978 to 2009 that were each used as both challenge and protective strains. There are 15,693 HI assay measurements, with 279 explanatory variables, 53 surface exposed residues and 226 variables related to the phylogenetic data; the tree for an extended version of this H1N1 dataset can be found in Harvey et al. (2016). Recorded experimental effects for the H1N1 dataset include the challenge strain, protective strain and the date of the experiment, see Section 2.1.1, and these can be accounted for as random effects in our models in Chapters 4 and 6.

2.3.2 Influenza A (H3N2) Serotype

H3N2 viruses emerged in the human population in 1968 and continue to circulate to the present day. During this period H3N2 viruses have been responsible for the majority of severe illness attributed to seasonal influenza, which is in part due to the increased rate of antigenic change in these viruses relative to other influenza viruses (Barr et al., 2014).

The H3N2 dataset includes 229 viruses collected from 1968 to 2013, of which 169 were used as protective strains. There are 7,315 HI measurements with 1,777 pairs of challenge and protective strains. There are are 1,264 explanatory variables which consists of 145 surface exposed antigenic residues and 1,119 variables relating to the evolutionary history of the serotype. Finally there are recorded experimental effects for the challenge strain, the protective strain and the date of the experiment, see Section 2.1.1, and these can be accounted for as random effects in our models in Chapter 6.

2.4 Classifying Variables

Once we have used the methods in Chapters 4 and 6 to select the most statistically relevant residues, it is important to validate our results and understand how likely our results are to be biologically correct. Although knowledge of which residues are antigenically important is at least partially incomplete in all serotypes of FMDV and the Influenza virus, for validation purposes we can use previous experimental results to assign residues for each serotype (except the SAT2 FMDV serotype) to three different levels of plausibility, *proven*, *plausible* and *implausible*, based on how likely they are to be antigenic.

2.4.1 SAT1 Serotype

For the SAT1 FMDV serotype, residues are included in the experimentally *proven* group for three different reasons. Firstly we include any residues which have been experimentally validated as important within the SAT1 serotype by monoclonal antibody escape mutant studies (MAbs) (Grazioli et al., 2006). Secondly, we include those residues which are part of cords of connected experimentally validated antigenic residues for four or more different serotypes; VP1 140-169 (part of the VP1 G-H loop), VP1 200-224 (VP1 C terminus), VP2 70-82 (VP2 B-C loop) and VP3 56-61 (VP3 B-B knob) (Aktas and Samuel (2000); Barnett et al. (1989); Crowther et al. (1993a); Baxt et al. (1989); Bolwell et al. (1989); Grazioli et al. (2006); Grazioli et al. (2013); Lea et al. (1994); Kitson et al. (1990); Mateu (1995); Saiz et al. (1991); Thomas et al. (1988a); Thomas et al. (1988b)). As antigenic sites have been found in a large number of different individual locations, we include additional information from other serotypes when classifying whole loops due the similar structure of the different serotypes. Finally, we also include a number of topotype-defining branches that are known to represent significant changes in the evolutionary history (Reeve et al., 2010).

We define the *plausible* group to consist of residues from any protein loop where residues have been identified in at least one FMDV serotype, excluding those residues that are already classified as proven. Additionally, any non-topotype-defining branches of the phylogenetic trees are included in the plausible group, as it is unknown which of the remaining branches may also be significant in evolutionary history of the serotype. Finally we classify any residues not included in these groups as *implausible*.

2.4.2 SAT2 Serotype

Although knowledge of the SAT2 FMDV serotype is minimal and we do not classify residues into different levels of plausibility, for minimal validation purposes we can exploit knowledge gained from other serotypes of FMDV and previous work on the SAT2 serotype. Grazioli et al. (2006) and Crowther et al. (1993b) has found evidence for antigenicity of the following three areas of the SAT2 capsid: VP1 140-169 (part of the VP1 G-H loop), VP1 200-224 (VP1 C terminus) and VP2 70-82 (VP2 B-C loop). Many regions have also been found to be antigenic on multiple other FMDV serotypes and it is also likely that they are in SAT2.

2.4.3 Influenza A (H1N1) Serotype

For influenza viruses, the haemagglutinin (HA) surface protein is responsible for binding to host cells and is also the major target for neutralising antibodies (Skehel and Wiley, 2000). The structure of HA can be broadly be divided into the stalk domain which connects to the virus capsid and a head domain which contains the residues involved in binding to the host cell. Experimental studies have identified that the major antigenic regions of HA are exposed areas in the head of the HA protein surrounding the receptor binding site (Skehel and Wiley, 2000). For H1, these experiments have identified four antigenic sites (Caton et al., 1982), however other sites are also known to be important (McDonald et al., 2007). We classify residues as proven if they belong to any of the four antigenic sites or have other experimental support for their role in antigenicity. Other regions of the head domain are considered to be plausible residues, while residues belonging to the stalk domain are considered unlikely to play a role in antigenic change.

2.4.4 Influenza A (H3N2) Serotype

The antigenicity of human H3N2 has been explored in greater depth than H1N1 due to the greater burden of disease and faster rate of antigenic evolution in H3 viruses. Experimental studies have revealed the structure of the H3 HA and studies of antigenically drifted mutant viruses generated in the laboratory have identified five distinct antigenic sites (A-E) on the surface (Wiley and Skehel, 1987). These antigenic sites have been

Table 2.1: **Table of classification for correlated variables.** The table gives the classification of groups of completely correlated variables based on the different types included. Ticks indicate which types of variables are in the group of correlated variables. The same rules apply to proven and plausible variables, so these have been combined into one group.

Proven/Plausible	Implausible	Branch	Classification
✓		✓	Proven/Plausible
	✓	✓	Implausible
✓	✓	✓	Plausible
✓	✓		Implausible

subsequently been extended and the set of residues reported by Shih et al. (2007) are classified as proven for model validation purposes. Structural and phylogenetic analysis of the H3 HA has produced an extended set of potentially antigenic residues which are classified as proven (Bush et al., 1999), with the remaining variables classified as implausible. Additionally, we do not consider some residues where the reliability of the genetic code is questionable. While initially included in the datasets, these have been excluded when considering the selected residues.

2.4.5 Classification of Completely Correlated Variables

It is common that variables have correlation coefficients exactly equal to one. In this case we only include one of the variables in the model and use Table 2.1 to guide the classification into the proven, plausible and implausible groups.

When an amino acid substitution at a single residue only occurs once in the evolutionary history of the virus, both the residue and branch variable explain that particular mutation. In this case both variables are the same and only one variable is included in the model. That variable then retains the classification given to that residue, either proven/plausible (line 2 in Table 2.1) or implausible (line 3 in Table 2.1).

Alternatively it is also possible that several residues have an amino acid substitution at only one point in the phylogenetic tree. In this case multiple residue variables are the same as a single branch variable and it is impossible to tell which of these residues are having the antigenic effect, so again only one variable is included in the model. If all the residues have the same classification, either proven/plausible or implausible, then the variable included in the model is given that classification (lines 2 and 3 in Table 2.1), where we take proven over plausible. If there are both proven/plausible and implausible variables, we have classified them as plausible to reflect our lack of knowledge of which residue is having the antigenic effect (line 4 in Table 2.1). The only exception to this rule is when branches that are known to be significant changes in the evolutionary history are selected by the model, in which case the variable is classified as proven regardless of which residues are also selected, as we know one of these changes must be significant.

Conversely when residue variables are not the same as a branch variable, we should be able to understand better their importance in explaining antigenic variability, as the antigenic effect of the substitution at this residue has been seen at multiple points in the evolutionary history of the virus. In this case if we have selected proven/plausible and implausible variables that are identical, then we classify this selection as implausible (line 5 in Table 2.1). This is because any genuinely significant change is unlikely to occur in direct correlation with an implausible variable at multiple points in the evolutionary history of the virus. It is possible that some of these variables are proven or plausible, but it is not possible to determine this from the current data.

2.5 Discussion

In this chapter we have introduced the biological problem, antigenic variability, and explained why it is important to understand it; Section 2.1. We have also motivated the need for an *in silico* model which can predict antigenic variability and reduce the number of *in vitro* experiments required to select an effective vaccine. We have then introduced a number of FMDV and Influenza serotypes, explaining their relevance and why they cause real biological problems, before giving details of the datasets we have available to analyse these datasets; Sections 2.2 and 2.3. Finally we have provided the biological evidence which we can use to validate the predictions we make and classify the plausibility of the residues that the models in Chapters 4 and 6 select.

Chapter 3

Methods

In this chapter we introduce a number of standard methods which can be used to model antigenicity. In Section 3.1 we introduce some classical methods that can account for the experimental variation and have previously been used to model antigenicity, e.g. the mixed-effects models of Reeve et al. (2010). We also discuss alternative Frequentist methods, the Least Absolute Shrinkage and Selection Operator (LASSO) and elastic net, as well as providing extensions to these methods in the form of the mixed-effects LASSO and mixed-effects elastic net (Davies et al., 2016a; Schelldorfer et al., 2011) and detail their implementation. These methods are used as a comparison for the more complex methods introduced in Chapters 4 and 6. Section 3.2 provides details of some of the Bayesian inference techniques used in Chapters 4 and 6, in particular Markov chain Monte Carlo (MCMC). These methods are combined with the Bayesian sparsity methods from Section 3.3 to create the Bayesian models introduced in Chapters 4 and 6. Finally we look at evaluation (Section 3.4) and model selection (Section 3.5) methods, discussing methods that measure the ability of a model in terms of prediction and variable selection, as well as how to choose between different models.

3.1 Classical Methods

A variety of classical statistical methods have previously been applied in predicting antigenic variability in order to identify antigenic sites. In this section we review some of these methods and propose variations which are applicable in the context of understanding antigenic variability.

3.1.1 Mixed-Effects Models

Classical mixed-effects models are a simple method which can be used to model antigenic variability and account for the experimental variability inherent in the data, e.g. Reeve et al. (2010). In classical mixed-effects models we define the response $\mathbf{y} = (y_1, \ldots, y_N)^\top$ and denote the explanatory variables, $\mathbf{X}$, as a matrix of $J+1$ columns and N rows, where the first column is an intercept. Each column of explanatory variables, $\mathbf{X}_j$, is then given an associated regression (or fixed effects) coefficient, w_j, to control its influence on the response.

We further set the random-effects design matrix, $\mathbf{Z}$, as the matrix of indicators with N rows and $||\mathbf{b}||$ columns, where $||.||$ indicates the dimension of the vector. The random-effect coefficients are given as $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_G^\top)^\top$ and represent a vector of parameters related to each of the groups $g \in \{1, \ldots, G\}$. Each $\mathbf{b}_g$ has length $||\mathbf{b}_g||$, where $||\mathbf{b}|| = \sum_{g=1}^{G} ||\mathbf{b}_g||$, and follows a zero mean Gaussian distribution with a group dependent variance, $\mathbf{b}_g \sim \mathcal{N}(\mathbf{b}_g|\mathbf{0}, \boldsymbol{\sigma}_{b,g}^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. This leads to the random-effect coefficients having the following joint distribution $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$, where we define $\boldsymbol{\Sigma}_\mathbf{b} = diag(\boldsymbol{\sigma}_\mathbf{b}^2)$ with $\boldsymbol{\sigma}_\mathbf{b}^2 = (\sigma_{\mathbf{b},1}^2, \ldots, \sigma_{\mathbf{b},G}^2)$ where each element has length $||\mathbf{b}_g||$. See Pinheiro and Bates (2000) for more details on mixed-effects models.

We therefore define the mixed-effects model as:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}) \tag{3.1}$$

where we assign the model independent and identically distributed Gaussian errors. Using a simple application of Gaussian integrals (Bishop, 2006), we integrate over $\mathbf{b}$ to give the likelihood:

$$L(\mathbf{w}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}_\mathbf{b}|\mathbf{y}, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{b}\mathbf{Z}^\top + \sigma_\varepsilon^2\mathbf{I}). \tag{3.2}$$

In classical mixed-effects models, model comparison techniques are often used to choose which variables are included within the model. To get a sparse model, Reeve et al. (2010) used a form of forward inclusion, making an adjustment for multiple testing using the Holm-Bonferroni correction (Holm, 1979). They firstly included terms to account for the evolutionary history of the viruses based on their phylogenetic trees, before adding variables corresponding to the surface exposed residues. The residue variables were added one at a time, before checking for significance and removing so to test other variables. Variables with a p-value of less than 0.05 were said to be significant and the corresponding residue proposed to be antigenically important. This technique was used by Reeve et al. (2010) on the SAT1 and SAT2 FMDV datasets (Sections 2.2.1 and 2.2.2) by Maree et al.

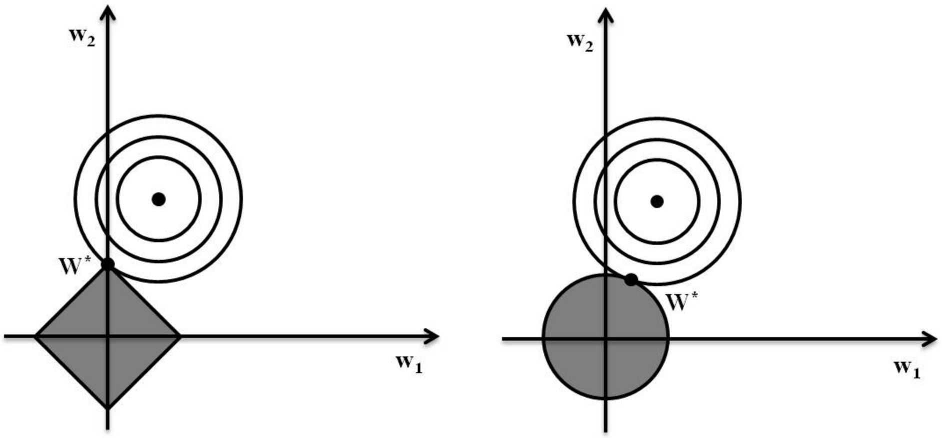


Figure 3.1: **Plot demonstrating the Sparsity caused by the LASSO Penalty.** The plot shows the contours of the unregularised error function along with the constrained region for the LASSO (λ_1, left) and ridge penalties (λ_2, right) where the optimum value of the regression parameters is given by $\mathbf{w}^*$. The LASSO gives a sparse solution in which $w_1^* = 0$. This figure is adapted from Bishop (2006).

(2015) on the extended SAT1 FMDV dataset (Section 2.2.1). Similar methods, but with further manual intervention, has been used by Harvey et al. (2016) on the H1N1 Influenza dataset (Section 2.3.1) and Harvey (2016) on the H3N2 dataset.

3.1.2 LASSO

A problem with the classical mixed-effects models of Reeve et al. (2010) is the reliance on stepwise regression techniques, which do not explore all variable configurations and can result in a non-optimal solution. A classical alternative to forward variable selection which does allow for simultaneous variable selection is the LASSO of Tibshirani (1996, 2011). The LASSO achieves its variable selection through an ℓ_1 penalty (equivalent to a Bayesian Laplace prior). In the simplest case of linear regression, this gives the following parameter estimates:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \lambda \sum_{j=1}^{J} |w_j| \right\}. \tag{3.3}$$

In the linear case this is a convex optimisation problem where a variety of fast and effective algorithms exist (e.g. Efron et al. (2004); Hastie et al. (2009)). The effect of (3.3) is to simultaneously shrink and prune parameters $\mathbf{w}$, thereby promoting a sparse model; see Bishop (2006) for examples. The degree of sparsity depends on the regularization parameter λ, which can be optimised via cross-validation or information criteria, e.g. Bayesian Information Criterion (BIC).

To see why the ℓ_1 penalty leads to a sparse model we first note that (3.3) is equivalent to minimising the unregularised sum of squares error subject to the constraint:

$$\sum_{j=1}^{J} |w_j| \leq \eta \tag{3.4}$$

for an appropriate value of the parameter η (Bishop, 2006). The reason for the sparsity can be seen by looking at Figure 3.1 which show the minimisation of the error function subject to the constraint in (3.4), the LASSO penalty in the left panel of Figure 3.1 forces one of the variables to equal zero, $w_1^* = 0$.

3.1.3 Elastic Net

A potential improvement over the LASSO is the elastic net of Zou and Hastie (2005). It has several advantages over the LASSO including the ability to select more than N variables in a $J > N$ situation, whereas the LASSO saturates to at most N variables (Zou and Hastie, 2005). More importantly for our application is that it also deals better with groups of correlated variables. While the LASSO will arbitrarily select one of the correlated variables, the penalty of the elastic net allows it to keep all of the variables in the model. See Section 2.3 of Zou and Hastie (2005) for more information on the grouping effect.

The elastic net combines ℓ_1 and ℓ_2 penalties and in the case of linear regression gives the following parameter estimates:

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \alpha\lambda \sum_{j=1}^{J} |w_j| + (1-\alpha)\lambda \sum_{j=1}^{J} |w_j|^2 \right\} \tag{3.5}$$

where λ is the penalty parameter and α controls the ratio of the ℓ_1 and ℓ_2 penalties. When $\alpha = 1$ the Elastic Net is equivalent to the LASSO and likewise ridge regression when $\alpha = 0$. We can fix $\alpha < 1$ and the problem becomes strictly convex, so we have a unique global minimum regardless of whether $\mathbf{X}$ is full rank. In practise Ruyssinck et al. (2014) have found that the choice of α is not important provided it is $0 < \alpha < 1$ and we have further explored this in the context of the mixed-effects elastic net (Section 3.1.5) in Chapter 5.

3.1.4 Mixed-Effects LASSO

An extension of the standard LASSO is the mixed-effects LASSO proposed by Schelldorfer et al. (2011), who estimate the regression coefficients $\mathbf{w}$, random-effect variances $\boldsymbol{\sigma}_{\mathbf{b}}^2$ and the variance of the noise σ_ε^2 as:

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\sigma}}_{\mathbf{b}}^2, \hat{\sigma}_\varepsilon^2) = \underset{\mathbf{w}, \boldsymbol{\sigma}_{\mathbf{b}}^2>0, \sigma_\varepsilon^2>0}{\operatorname{argmin}} \left\{ \tfrac{1}{2} \log |\mathbf{V}| + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \sum_{j=1}^{J} |w_j| \right\} \tag{3.6}$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{Z}^\top + \sigma_\varepsilon \mathbf{I}$. The mixed-effects LASSO can be combined with different information criteria to select the penalty parameter, λ. In Chapter 4 we have used BIC and the corrected Akaike Information Criterion (AICc) (Hurvich and Tsai, 1989).

A problem with the mixed-effects LASSO of Schelldorfer et al. (2011) is that the method has only been developed for one random-effect factor. In order to deal with this problem the Cartesian product of several random-effects factors can be mapped onto a single random-effect factor. However this can lead to over-estimating the complexity of the model, so we have developed our own mixed-effects LASSO which allows multiple random effect factors in order to deal with this (Davies et al., 2016a). Our method uses a conjugate gradient optimisation strategy available in R (R Core Team, 2013), but requires a tolerance that must be determined by the user. In practise we have found this easy to do, as for a sufficiently large λ and reasonably standardised data there will be a group of regressors clearly grouped around zero. The tolerance can then be set such as to force these values to zero, i.e. exclusion from the model, and other values of λ used. While this may not be as effective as the purpose-built block coordinate descent scheme proposed in Schelldorfer et al. (2011), we have found in practise that they achieve the same results.

3.1.5 Mixed-Effects Elastic Net

Like the LASSO, we can expand the elastic net into the context of a mixed-effects model (Davies et al., 2016a):

$$\begin{aligned}(\hat{\mathbf{w}}, \hat{\boldsymbol{\sigma}}_{\mathbf{b}}^2, \hat{\sigma}_\varepsilon^2) = \underset{\mathbf{w}, \boldsymbol{\sigma}_{\mathbf{b}}^2>0, \sigma_\varepsilon^2>0}{\operatorname{argmin}} \Big\{ &\tfrac{1}{2} \log |\mathbf{V}| + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &+ \alpha\lambda \sum_{j=1}^{J} |w_j| + (1-\alpha)\lambda \sum_{j=1}^{J} |w_j|^2 \Big\}\end{aligned} \tag{3.7}$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{Z}^\top + \sigma_\varepsilon \mathbf{I}$. Again we use the simple optimisation strategy we proposed for the mixed-effects LASSO in Section 3.1.4.

3.2 Bayesian Inference with Markov chain Monte Carlo

In Bayesian inference the posterior distribution, the distribution which contains all the current information about the parameters $\boldsymbol{\theta}$, is defined by Bayes theorem (Bayes, 1763). For a given model specification, we define the likelihood to be the probability of the data, $\mathfrak{D}$, given the model distribution, $p(.)$, and model parameters, $\boldsymbol{\theta}$. To get the posterior distribution the likelihood is multiplied by the prior distribution, $p(\boldsymbol{\theta})$, and normalised:

$$p(\boldsymbol{\theta}|\mathfrak{D}) = \frac{p(\mathfrak{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathfrak{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \sim p(\mathfrak{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.8}$$

Usually integrating over $\boldsymbol{\theta}$ is not possible in complex or high dimensional problems, but the posterior distribution can be sampled from $p(\mathfrak{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ using MCMC methods.

MCMC methods are a family of estimation methods used to approximate a target distribution. They are used where integration over all model parameters is not analytically tractable and can be used in Bayesian inference to sample from the posterior distribution of a given model. The idea of the method is to sample values of the parameter, θ, from approximate distributions and then correct those draws to better approximate the target posterior distribution, $p(\theta|\mathbf{y})$. Samples are drawn such that they only depend on the last value drawn and hence form a Markov chain. Doing this produces a sequence of samples (chain) which converges to a stationary distribution at time t where:

$$\theta^{t+1}|(\theta^{t} \sim p(.)) \sim p(.). \tag{3.9}$$

In Bayesian inference the stationary (or equilibrium) distribution is the posterior distribution and is independent of the starting state. Samples from this distribution will come from the target distribution (posterior).

As convergence to the stationary distribution (posterior) is not instant, we must remove the period of samples before convergence has been achieved. This section of samples is usually known as burn-in and convergence is often assessed by running multiple chains and diagnosing convergence using Potential Scale Reduction Factors (PSRFs); see Section 3.2.3. Additionally, samples from the posterior distribution can be highly autocorrelated and samples are therefore often thinned, e.g. take every ith sample, in order to get independent samples of the posterior and accurate estimates of θ.

3.2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm was introduced by Metropolis et al. (1953) and Hastings (1970), and can be used as a method to sample parameters in the posterior distribution through an acceptance and rejection step. Normally parameters are proposed individually and put through the acceptance and rejection step which is based on the ratio of the posterior and proposal distributions. In this sense parameters are gradually updated throughout the MCMC chain.

To get the ith sample of θ, θ^i, via the M-H algorithm we firstly need to propose a potential new value, θ^*. This is done through the proposal distribution $q(\theta^*|.)$. For continuous variables the proposal distribution is usually centred around the previous value of the sequence, θ^{i-1}, i.e. $q(\theta^*|\theta^{i-1})$ where $q(.)$ is a Gaussian distribution, but this is not always possible. The distribution of $q(\theta^*|.)$ can be freely chosen, but its choice affects the speed of convergence and mixing. In the second step of the M-H algorithm the proposal parameter value, θ^*, is accepted or rejected via the acceptance probability. This is given as the ratio of posterior distributions of the proposed and previous parameter values, as well as the forwards, $q(\theta^*|.)$, and backwards, $q(\theta^{i-1}|.)$, proposal densities:

$$\alpha(\theta^*, \theta^{i-1}|\mathfrak{D}) = \min\left(1, \frac{p(\theta^*|\mathfrak{D})q(\theta^{i-1}|.)}{p(\theta^{i-1}|\mathfrak{D})q(\theta^*|.)}\right). \tag{3.10}$$

The proposed parameter value, θ^*, is then accepted if $\alpha(\theta^*, \theta^{i-1}|\mathfrak{D})$ is greater than a uniform random variable u, where $u \sim \mathcal{U}[0,1]$. If the proposed parameter is accepted then we set θ^i to be equal to θ^* and if not set it such that $\theta^i = \theta^{i-1}$.

3.2.2 Gibbs Sampling

Gibbs sampling is a special case of the M-H algorithm proposed by Ripley (1979) and Geman and Geman (1984). Unlike in the M-H algorithm, θ_j is not sampled from the full posterior distribution, $p(\boldsymbol{\theta}|\mathfrak{D})$. Instead each parameter, $\theta_j \in \boldsymbol{\theta}$, is sampled from its conditional distribution, subject to $\boldsymbol{\theta}_{-j} \in \boldsymbol{\theta}$. Gibbs sampling requires the conditional distribution to follow a standard distribution and if not the sampling of θ_j should be done through the M-H algorithm. Due to the conditional distribution of θ_j following a standard form we can propose the value of θ^* from the conditional distribution. This results in $q(\theta_j|\boldsymbol{\theta}_{-j}, \mathfrak{D}) = p(\theta_j|\boldsymbol{\theta}_{-j}, \mathfrak{D})$ in (3.10), resulting in the acceptance rate equalling one, $\alpha(\theta^*, \theta^{i-1}|\mathfrak{D}) = 1$. In practise this means we can simply sample θ_j^* from the conditional distribution and immediately set $\theta_j^* = \theta_j^i$.

To use Gibbs sampling and the M-H algorithm to sample the full posterior distribution, we sample a new value for each parameter $\theta_j \in \boldsymbol{\theta}$ based on the current values of $\boldsymbol{\theta}_{-j} \in \boldsymbol{\theta}$. Each parameter is sampled from its conditional distribution $\theta_j^i \sim p(\theta_j^i|\theta_1^i, \ldots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \ldots, \theta_J^{i-1}, \mathfrak{D})$, where the parameters conditioned on take the value of their most recent sample. Where the conditional distribution is of a known form it is standard to use Gibbs sampling, although alternative proposals can be used with the M-H algorithm instead. If this is not the case then the M-H algorithm should be used. The initial values of the parameters, $\boldsymbol{\theta}^1$, are set to some arbitrary values in the correct parameter space. Under reasonable general conditions and a sufficient number of iterations, i, the algorithm will converge to the target distribution.

3.2.3 Potential Scale Reduction Factors

PSRFs are a measure which quantifies the convergence of multiple MCMC chains as introduced by Gelman and Rubin (1992). PSRFs are based on the assumption that multiple chains using the same data should have the same variation within each chain as they do between them, if this has not occurred then the chains have clearly not converged (Gelman et al., 2013a).

The calculation of the PSRF for each parameter, θ, requires m parallel sequences, each of length n. To calculate the PSRF of each of the model parameters we compute the between-sequence, B, and within-sequence, W, variances:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_j - \bar{\theta}\right)^2, \quad \text{where} \quad \bar{\theta}_j = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j \tag{3.11}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij} - \bar{\theta}_j\right)^2. \tag{3.12}$$

We can then estimate $\mathrm{Var}(\theta|\mathbf{y})$, the marginal posterior variance of the parameter, by a weighted average of W and B:

$$\widehat{\mathrm{Var}}^+(\theta|\mathbf{y}) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{3.13}$$

This quantity overestimates the marginal posterior variance of the parameter, $\mathrm{Var}(\theta|\mathbf{y})$, while W underestimates it for finite n. From this the PSRF can be calculated as follows:

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^+(\theta|\mathbf{y})}{W}} \tag{3.14}$$

where the value declines to 1 as $n \to \infty$. Large values of $\hat{R}$ indicate a lack of convergence and values of less than 1.05 or 1.1 are generally said to indicate convergence, e.g. Grzegorczyk and Husmeier (2013).

3.2.4 Joint Distribution Tests

When using different sampling schemes it is often important to check whether the MCMC sampler approximates the correct posterior distribution. Joint distribution tests as proposed by Geweke (2004) can be used to do this. The idea behind the joint distribution test is to draw D sets of model parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$ from the model's prior distributions $p_{\boldsymbol{\theta}}(.)$ and then use these parameters to generate D datasets, $\mathfrak{D}_1, \ldots, \mathfrak{D}_D$. Using the same model and prior specifications we can then use the MCMC sampler that is being tested to sample from each of the posterior distributions, $p(\boldsymbol{\theta}_d|\mathfrak{D}_d)$, of the D datasets. From each of the MCMC chains of each posterior distribution we can then take N independent samples of the model parameters, $\boldsymbol{\theta}_{d,1}, \ldots, \boldsymbol{\theta}_{d,N}$. To work out whether the MCMC samplers are sampling from the correct posterior distribution we then check whether the samples $\boldsymbol{\theta}_{i,d}$ for $i \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$ follow the prior distribution used to generate the data, $p(.)$.

$$\frac{1}{D}\sum_{d=1}^{D} p(\boldsymbol{\theta}|\mathfrak{D}_d) \approx \int p(\boldsymbol{\theta}|\mathfrak{D})p(\mathfrak{D})d\mathfrak{D} = \int p(\mathfrak{D}, \boldsymbol{\theta})d\mathfrak{D} = p(\boldsymbol{\theta}) \tag{3.15}$$

If the sampled parameters follow $p(.)$ then for a large enough D and N we can conclude that the MCMC sampler is correctly sampling the posterior.

3.3 Bayesian Sparsity Methods

A variety of methods exist in Bayesian inference for achieving a sparse model. Like with the Frequentist methods in Section 3.1 we can use ℓ_1 regularisation and similar methods exist in the Bayesian paradigm, e.g. Bayesian LASSO (Park and Casella, 2008). However ℓ_1 methods have their drawbacks as discussed below and so alternative methods have been proposed which get round some of these issues, e.g. the spike and slab prior (George and McCulloch, 1993, 1997; Mitchell and Beauchamp, 1988) and the binary mask model, e.g. Murphy (2012).

Many of these Bayesian methods have been shown to give an improvement over ℓ_1 regularisation methods in terms of variable selection and prediction (Davies et al., 2014, 2016a; Mohamed et al., 2012). One of the reasons for this is the ℓ_1 regularisation term itself, equivalent to a Laplace prior in a Bayesian context (Park and Casella, 2008). This is computationally efficient and leads to a convex optimisation problem for penalised maximum likelihood or Bayesian maximum a posteriori (MAP) inference. However, ℓ_1 regularisation gives an increased bias from shrinkage while not giving sufficient sparsity, as discussed in Chapter 13 of Murphy (2012). The Bayesian methods, such as spike and slab priors, can improve variable selection and avoid excessive shrinkage, but lead to a non-convex optimisation problem. These priors can also be integrated into Bayesian hierarchical models, as can be seen in Chapters 4 and 6, and this also gives a number of other advantages. In particular Bayesian hierarchical models allow consistent inference of all parameters and hyper-parameters, and inference borrows strength by the systematic sharing and combination of information; see Gelman et al. (2013a).

3.3.1 Spike and Slab Prior

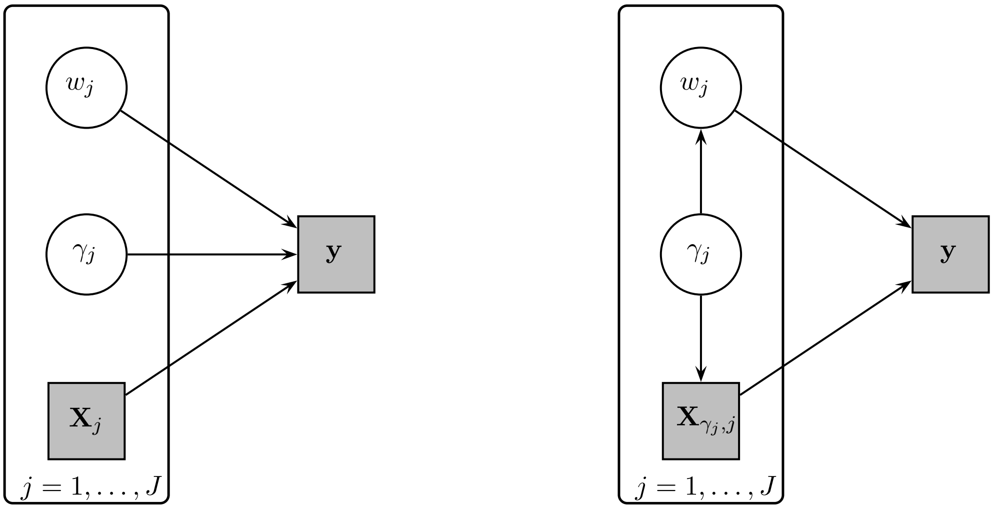
Spike and slab priors have been used in a number of different contexts and have been shown to outperform ℓ_1 methods both in terms of variable selection and out-of-sample predictive performance (Mohamed et al., 2012). They were originally proposed by Mitchell and Beauchamp (1988) as a mixture of a Gaussian distribution and a Dirac spike, but have also been used as a mixture of two Gaussian distributions (George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005). Spike and slab priors are based on the idea that the prior reflects whether the feature is relevant based on the values of a inferred vector of binary indicator parameters, $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\top \in \{0,1\}^J$. The relevance of the jth column of $\mathbf{X}$ is determined by $\gamma_j \in \{0,1\}$, where feature j is said to be relevant if $\gamma_j = 1$. In this way we expect that $w_j = 0$ if $\gamma_j = 0$, i.e. the feature is irrelevant, and conversely it should be non-zero if the variable is relevant, $w_j \neq 0$ if $\gamma_i = 1$.

The spike and slab prior of Mitchell and Beauchamp (1988) approaches this concept by assigning a conjugate Gaussian prior where the feature, w_j, is relevant, i.e. $\gamma_j = 1$, and a Dirac spike at zero where it is not:

$$p(w_j|\gamma_j, \mu_w, \sigma_w^2) = \begin{cases} \delta_0(w_j) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_j|\mu_w, \sigma_w^2) & \text{if } \gamma_j = 1. \end{cases} \tag{3.16}$$

Here we have a spike at 0 and as $\sigma_w^2 \to \infty$ the distribution, $p(w_j|\gamma_j = 1)$, approaches a uniform distribution, a slab of constant height. In this sense where $\gamma_j = 0$ the variable w_j and corresponding variable $\mathbf{X}_j$ are effectively removed from the model as demonstrated by the following example:

$$\begin{gathered} \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,2} & x_{3,3} \end{bmatrix}; \; \mathbf{X}_{\boldsymbol{\gamma}} = \begin{bmatrix} x_{1,1} & x_{1,3} \\ x_{2,1} & x_{2,3} \\ x_{3,1} & x_{3,3} \end{bmatrix}; \\ \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}; \; \mathbf{w}_{\boldsymbol{\gamma}} = \begin{bmatrix} w_1 \\ w_3 \end{bmatrix}; \; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 = 1 \\ \gamma_2 = 0 \\ \gamma_3 = 1 \end{bmatrix}. \end{gathered} \tag{3.17}$$



(a) Binary Mask Model

(b) Spike and Slab Model

Figure 3.2: **Probabilistic Graphical Models (PGMs) for the (a) binary mask and (b) spike and slab models.** The *grey* squares refer to the data, while the *white* circles refer to parameters and hyperparameters that are inferred.

The alternative spike and slab prior of George and McCulloch (1993, 1997) approximates the spike and slab prior of Mitchell and Beauchamp (1988) by replacing the Dirac spike with a highly peaked Gaussian distribution centred around zero:

$$p(w_j|\gamma_j, \mu_w, \sigma^2_{w_1}, \sigma^2_{w_2}) = \begin{cases} \mathcal{N}(w_j|0, \sigma^2_{w_1}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_j|\mu_w, \sigma^2_{w_2}) & \text{if } \gamma_j = 1. \end{cases} \tag{3.18}$$

In this case the values of the spike variance parameter is usually fixed to be very small such that $\sigma^2_{w_1} << \sigma^2_{w_2}$. The idea of fixing $\sigma^2_{w_1}$ to be small is to force any w_j where $\gamma_j = 0$ to be approximately 0, i.e. $w_j \approx 0$. In this thesis we have not explored this specification, as mathematically it is inferior to the spike and slab prior of Mitchell and Beauchamp (1988) due to the irrelevant variables only being approximately fixed to zero and the necessity to a-priori fix the value of $\sigma^2_{w_1}$.

3.3.2 Binary Mask Model

An alternative to the spike and slab prior is the binary mask model, e.g. Jow et al. (2014). Instead of the prior on the regression coefficients reflecting the relevance of the variable, in the binary mask model the indicator variables, $\boldsymbol{\gamma}$, 'mask' or hide the impact of the non-zero coefficients, $\mathbf{w}$, and explanatory variables, $\mathbf{X}$, when the variable is not selected:

$$p(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma}, \sigma_\varepsilon^2, \mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\Gamma}\mathbf{w}, \sigma_\varepsilon^2\mathbf{I}) \tag{3.19}$$

where $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma})$. This is different to the spike and slab based methods where the variables, and their corresponding coefficients, are effectively removed from the model via a 'spike' or delta prior, rather than simply masked. The difference can be seen by comparing the directed edges associated with the $\boldsymbol{\gamma}$ vertex in the Probabilistic Graphical Model (PGM) of the spike and slab model, Figure 3.2b, with the PGM of the binary mask model given in Figure 3.2a.

3.4 Evaluation Methods

To compare the different methods and model specifications that will be used in this thesis, we need to introduce a variety of different methods to evaluate them. We are firstly interested in evaluating explanatory performance, e.g the reliability of the selection of relevant explanatory variables. In this case the distinction between in-sample and out-of-sample prediction becomes obsolete, as the status of the variables does not change. The explanatory methods here are sensitivity, specificity, precision, F1-score (Section 3.4.1), Receiver Operating Characteristic (ROC) curves and Area Under the ROC curve (AU-ROC) values (Section 3.4.2). We also wish to monitor predictive performance, where the values change from case to case. To reduce over-optimism we therefore assess predictive performance out-of-sample, in our case looking at out-of-sample likelihoods and Mean Squared Errors (MSEs) of out-of-sample observations (Section 3.4.1).

3.4.1 Summary Statistics

Sensitivity, specificity, precision and F1-scores are all measures of the performance of a binary classification, e.g. the successful inclusion or exclusion of relevant or irrelevant explanatory variables. These are given in terms of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN);

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.20}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.21}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.22}$$

$$\text{F1-Score} = \frac{2TP}{2TP + FP + FN} \tag{3.23}$$

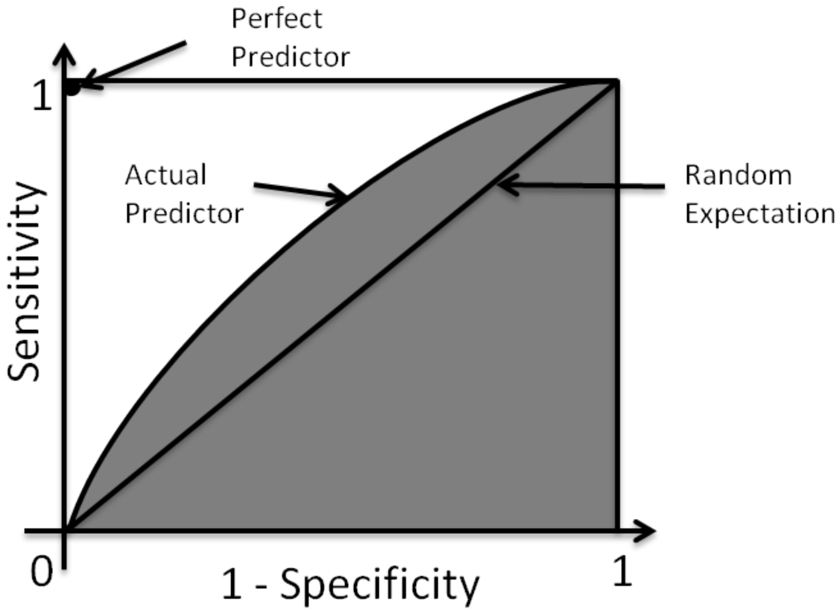


Figure 3.3: **Example ROC Curve.** A plot showing an example ROC curve, where the perfect predictor, the actual predictor and random expectation are indicated and the AUROC value is given by the shaded area.

where higher values imply improved performance. These summary statistics measure explanatory performance and are used to compare different methods in their abilities to correctly select fixed or random effects. Sensitivity and specificity are also used to create ROC curves and the resulting AUROC values in Section 3.4.2.

Predictive performance is usually calculated out-of-sample. Here we use MSEs and likelihoods of out-of-sample observations, $\mathbf{y}_{out}$, based on predicted observations, $\mathbf{y}_{pred}$, taken from the inferred parameter values, $\boldsymbol{\theta}_{inf}$, from training data, $\mathbf{y}_{obs}$. In this case the out-of-sample MSEs and mean log likelihoods are defined as follows:

$$MSE(\mathbf{y}_{out}|\mathbf{y}_{obs}(\boldsymbol{\theta}_{inf})) = \frac{1}{||\mathbf{y}_{out}||} \sum \left(\left(\mathbf{y}_{out} - \mathbf{y}_{pred}(\boldsymbol{\theta}_{inf}) \right)^2 \right) \tag{3.24}$$

$$\hat{p}_{out}(\mathbf{y}_{out}|\boldsymbol{\theta}_{inf}(\mathbf{y}_{obs})) = \frac{1}{||\mathbf{y}_{out}||} \log \left(p(\mathbf{y}_{out}|\boldsymbol{\theta}_{inf}(\mathbf{y}_{obs}) \right) \tag{3.25}$$

where $||\mathbf{y}_{out}||$ denotes the number of out-of-sample observations.

3.4.2 ROC Curves

ROC curves are an important tool for measuring the performance of a method in variable selection (e.g. Hanley and McNeil (1982); Section 5.7. of Murphy (2012)). ROC curves can be constructed when an underlying gold standard is known, e.g. in a simulation study where the relevant variables are known, and a method of ranking the importance of the variables is given, e.g. posterior probability of inclusion of a variable. To create the ROC curves we use the rankings to define inclusion thresholds between each ordered variable and plot the sensitivity, (3.20), against one minus the specificity, (3.21) for each possible threshold. Linear interpolation is then used to complete the ROC curve. An example ROC curve is given in Figure 3.3 and is marked as the ‘actual predictor’.

From the ROC curves AUROC values can then be calculated using numerical integration, where the area that makes up the AUROC value is shaded in Figure 3.3. AUROC values give a measure of global performance that is not dependant on an arbitrary threshold and like ROC curves can be used to compare the performance of different methods in terms of variable selection. Random expectation gives an AUROC value of 0.5 (‘Random expectation’ in Figure 3.3), while a value of 1 means a method offers perfect selection (‘perfect predictor’ in Figure 3.3). The higher the AUROC value, the better the method is said to have performed in terms of variable selection.

3.5 Model Selection Methods

We are also interested in choosing between different models and model specifications. To choose between competing models or model specifications we can use the Widely Applicable Information Criterion (WAIC), Watanabe (2010), or Bayesian 10-fold Cross Validation (CV), e.g. Chapter 7 of Gelman et al. (2013a).

3.5.1 Bayesian Cross Validation

Bayesian CV methods are reliable, if computationally expensive, techniques for measuring the out-of-sample performance of different models. CV methods work by partitioning the data into K groups and then analysing the predictive performance of a given model on each of the K different groups using the remainder of the data for training. In this sense CV methods estimate out-of-sample predictive performance while still making use of all of the available data.

Various CV methods can be used to analyse the performance of different models. Leave-One-Out CV (LOO-CV) uses each observation as an individual group, i.e. $K = N$, with the advantage of making maximum use of the available data at every step. However LOO-CV is computationally infeasible for many models, as it requires fitting the model N times. As a compromise 10-fold CV is often used, where $K = 10$, as it only involves fitting 10 models and this method has been used here.

To calculate the 10-fold Bayesian CV performance of a model, we apply the method to partial data, $\mathbf{y}_{-k}$, and $\mathfrak{D}_{-k}$, and use thinned samples of the model parameters, $\boldsymbol{\theta}^{\iota}$, for

$\iota \in \{1, \ldots, I\}$, from $p(\boldsymbol{\theta}|\mathbf{y}_{-k}, \boldsymbol{\mathfrak{D}}_{-k})$, to estimate the performance on the remaining data, $\mathbf{y}_k$ and $\boldsymbol{\mathfrak{D}}_k$, using the likelihood. Doing this for each of the K groups gives the 10-fold Bayesian CV performance:

$$p_{CV} = \frac{1}{K}\sum_{k=1}^{K} \log \int p(\mathbf{y}_k|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\mathfrak{D}}_{-k})d\boldsymbol{\theta} \propto \frac{1}{K}\sum_{k=1}^{K} \log \frac{1}{I}\sum_{\iota=1}^{I} p(\mathbf{y}_k|\boldsymbol{\theta}^{\iota}). \tag{3.26}$$

where $\boldsymbol{\theta}^{\iota}$ is a sample from $p(\boldsymbol{\theta}|\boldsymbol{\mathfrak{D}}_{-k})$.

3.5.2 Widely Applicable Information Criterion

WAIC (Watanabe, 2010) and Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) are both useful criteria for selecting the correct models in a Bayesian context. DIC is effectively a Bayesian version of the Akaike Information Criterion (AIC), where the posterior mean is used instead of the maximum likelihood estimate and k is replaced with a data-based bias correction (Gelman et al., 2013b):

$$p_{DIC} = 2\left(p(\mathbf{y}|\bar{\boldsymbol{\theta}}) - \frac{1}{I}\sum_{\iota=1}^{I} p(\mathbf{y}|\boldsymbol{\theta}^{\iota})\right). \tag{3.27}$$

Here the first part measures predictive performance and the second is the effective number of parameters. DIC has been shown to work well in a number of situations, however its performance becomes poor when the model used is singular, e.g. when spike and slab priors are used. In this situation the posterior mean becomes a poor representation of the posterior samples of a given parameter and the method suffers accordingly.

While DIC struggles with singular models, WAIC still remains effective for selecting the correct model and this is why we have used WAIC in this thesis rather than DIC (Gelman et al., 2013b). WAIC averages over the posterior distribution which is both desirable and allows the criterion to work with singular models. Watanabe (2010) also showed how WAIC is asymptotically equivalent to Bayesian LOO-CV. WAIC can be computed using the thinned parameter samples, $\boldsymbol{\theta}^{\iota}$, from the posterior distribution of the full dataset, $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\mathfrak{D}})$, meaning the sampling process must only be carried out once for the whole dataset:

$$p_{WAIC} = -2\sum_{i=1}^{N}\left(\log\left(\frac{1}{I}\sum_{\iota=1}^{I} p(y_i|\boldsymbol{\theta}^{\iota}, \boldsymbol{\mathfrak{D}}_i)\right) - \mathrm{Var}\left(\log(p(y_i|\boldsymbol{\theta}^{\iota}, \boldsymbol{\mathfrak{D}}_i))\right)\right) \tag{3.28}$$

where Var is the sample variance.

3.6 Discussion

In this chapter we have introduced a number of standard methods which are relevant to the methods proposed in this thesis. We have described some classical methods, Section 3.1, which will be used as a comparison to the methods proposed in Chapter 4. These include standard mixed-effects, the LASSO, elastic net and mixed-effect model versions of the LASSO and elastic net. We have also demonstrated, in Section 3.2, basic Bayesian methods and how to infer the posterior distributions of the model parameters. These techniques will be used in the methods proposed in Chapters 4 and 6. We have also introduced the Bayesian sparsity methods that will be used in Chapters 4 and 6 and discussed how they can offer an improvement over the classical methods of Section 3.1. Section 3.4 has then specified some different evaluation methods which will be used throughout Chapters 5 and 7. Finally in Section 3.5 we have looked at methods for choosing between different model specifications.

Chapter 4

Sparse Hierarchical Bayesian Models for Understanding Antigenic Variability - The Methods

In this chapter we introduce the family of Sparse hierArchical Bayesian models for detecting Relevant antigenic sites in virus Evolution (SABRE); the SABRE methods. The methods can account for the experimental variability in the data and predict antigenic variability. The SABRE methods integrate spike and slab priors into a Bayesian hierarchical model in order to select the significant variables and identify the corresponding sites in the viral protein which are important for the neutralisation of the virus.

The original SABRE method (Section 4.1), as published in Davies et al. (2014), is a Bayesian hierarchical mixed effects model, based on the Frequentist mixed effects models of Reeve et al. (2010) described in Section 3.1.1. The method aims to predict either log VN titre or log HI assay measurements (Section 2.1) based on the fixed effects, the antigenic residues and phylogenetic tree branches (Sections 2.1.2 and 2.1.3), and the random effects (Section 2.1.1). To do this effectively the original SABRE method uses spike and slab priors (Section 3.3.1) to select the relevant fixed effects and identify potential antigenic residues. The spike and slab prior is known to outperform the Least Absolute Shrinkage and Selection Operator (LASSO) in terms of variable selection (Mohamed et al., 2012) and its incorporation into a Bayesian hierarchical model allows the consistent inference of all parameters and hyper-parameters, and inference borrows strength by the systematic sharing and combination of information; see Gelman et al. (2013a).

Section 4.2 discusses a variety of potential improvements to the original SABRE method proposed in Davies et al. (2014), as discussed in Davies et al. (2016a). Firstly a separate intercept parameter is introduced (Section 4.2.1) and the addition of this creates what is known as the Semi-Conjugate SABRE method. Specifying the prior on the intercept correctly is important as it is a biologically significant parameter which gives the VN titre or HI assay when any two identical viruses are used as the challenge and protective strains, i.e. when all covariates are equal to zero. Section 4.2.2 details the Conjugate SABRE method, this gives the model increased conjugacy which introduces additional relationships into the model and provides the opportunity to improve the sampling scheme. Section 4.2.3 introduces the binary mask model (Section 3.3.2) in the context of the SABRE method, allowing us to test the difference in performance between models based on the spike and slab prior and those based on the binary mask model. Finally Section 4.2.4 looks at different specifications of random effect priors, namely it looks at the possibility of using the half-t prior proposed in Gelman (2006), something that has previously been suggested in the literature.

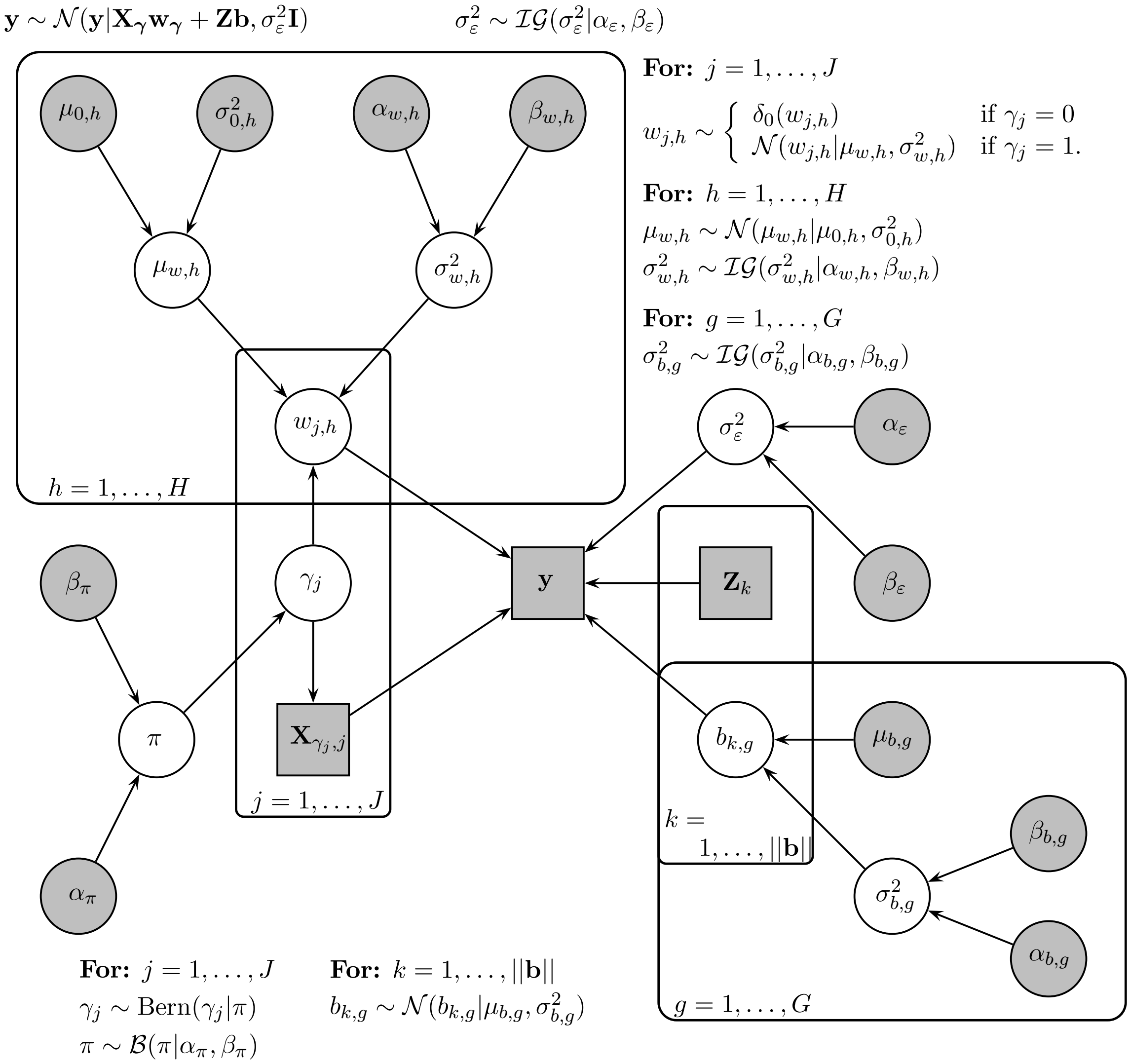
Section 4.3 discusses posterior inference for all of the SABRE methods based on the methods discussed in Section 3.2, providing the conditional distributions needed to sample from the model. Section 4.3.5 discusses in detail the sampling of the latent inclusion variables, $\boldsymbol{\gamma}$, that are used in the spike and slab priors (Section 3.3.1). In particular it looks at sampling multiple parameters via block M-H sampling, as well as exploring the more standard method of component wise Gibbs sampling, in order to find the most effective way of sampling $\boldsymbol{\gamma}$. Finally in Section 4.3.6 we discuss the conjugate sampling scheme (CSS) that can be used with the conjugate SABRE in order to potentially improve the computational efficiency.

4.1 The Original SABRE Method

The original SABRE method was proposed in Davies et al. (2014) and incorporates the spike and slab prior into a hierarchical Bayesian model. The model is shown in the PGM in Figure 4.1 and the parameters are sampled from the posterior distribution using MCMC based on the methods in Section 3.2, where the conditional distributions are given in Section 4.3.1.

4.1.1 Likelihood

The likelihood for the original SABRE method is similar to the classical mixed-effects model described in Section 3.1.1, however we include only the relevant residue and phylogenetic tree variables, X, and regressors, w. However instead of including all the variables, X, and their corresponding regression coefficient, we now only include relevant variables,



$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I})$ $\sigma^2_\varepsilon \sim \mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon)$

For: $j = 1, \ldots, J$

$$w_{j,h} \sim \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h}|\mu_{w,h}, \sigma^2_{w,h}) & \text{if } \gamma_j = 1. \end{cases}$$

For: $h = 1, \ldots, H$

$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma^2_{0,h})$

$\sigma^2_{w,h} \sim \mathcal{IG}(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h})$

For: $g = 1, \ldots, G$

$\sigma^2_{b,g} \sim \mathcal{IG}(\sigma^2_{b,g}|\alpha_{b,g}, \beta_{b,g})$

For: $j = 1, \ldots, J$

$\gamma_j \sim \text{Bern}(\gamma_j|\pi)$

$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)$

For: $k = 1, \ldots, ||\mathbf{b}||$

$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma^2_{b,g})$

Figure 4.1: **Compact representation of the original SABRE method as a PGM.** The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred.

$\mathbf{X}_{\boldsymbol{\gamma}}$, and regressors, $\mathbf{w}_{\boldsymbol{\gamma}}$:

$$p(\mathbf{y}|\mathbf{w}_{\boldsymbol{\gamma}}, \mathbf{b}, \sigma^2_\varepsilon, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}) = \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I}). \tag{4.1}$$

The relevance of variable j is determined by $\gamma_j \in \{0, 1\}$, where feature j is said to be relevant if $j = 1$. This gives $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J) \in \{0, 1\}^J$ where $\gamma_0 = 1$ is fixed meaning that there is always an intercept in the model. We then define $\mathbf{X}_{\boldsymbol{\gamma}}$ to be the matrix of relevant explanatory variables with $\sum_{j=1}^{J} \gamma_j$ columns and N rows. Similarly $\mathbf{w}_{\boldsymbol{\gamma}}$ is given as the column vector of regressors, where the inclusion of each parameter is again dependent on $\boldsymbol{\gamma}$.

4.1.2 Noise Prior

As with the classical methods described in Section 3.1, we assume additive iid Gaussian noise with variance σ_ε^2. In a Bayesian context we wish to infer σ_ε^2, so we specify the conjugate prior:

$$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2 | \alpha_\varepsilon, \beta_\varepsilon) \tag{4.2}$$

where the hyper-parameters α_ε and β_ε are fixed, as indicated by the grey nodes in Figure 4.1.

4.1.3 Spike and Slab Prior

Spike and slab priors have been used in a number of different contexts and have been shown to outperform ℓ_1 methods both in terms of variable selection and out-of-sample predictive performance (Mohamed et al., 2012). They were originally proposed by Mitchell and Beauchamp (1988) as a mixture of a Gaussian distribution and Dirac spike, but have also been used as a mixture of two Gaussians distributions; see Section 3.3.1.

The prior for $\mathbf{w}_\gamma$ is set in the manner proposed in Mitchell and Beauchamp (1988) such that it reflects whether a feature is relevant. In this way we expect that $w_{j,h} = 0$ if $\gamma_j = 0$, i.e. the feature is irrelevant, and conversely it should be non-zero if the variable is relevant, $w_{j,h} \neq 0$ if $\gamma_j = 1$. The variables are then divided into related groups $h \in \{1, \ldots, H\}$, in this case two: the intercept and the covariates. A conjugate prior is chosen when the feature is relevant:

$$p(w_{j,h} | \gamma_{j,h}, \mu_{w,h}, \sigma_{w,h}^2) = \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h} | \mu_{w,h}, \sigma_{w,h}^2) & \text{if } \gamma_j = 1. \end{cases} \tag{4.3}$$

where δ_0 is the delta function. Here we have a spike at the mean, $\mu_{w,h}$, and as $\sigma_{w,h}^2 \to \infty$ the distribution, $p(w_{j,h} | \gamma_j = 1)$, approaches a uniform distribution, a slab of constant height. For this reason, these models are often known as spike and slab models.

For mathematical convenience we then define the prior distribution of $\mathbf{w}_\gamma = (w_1, \ldots, w_J)^\top$ as:

$$\mathbf{w}_\gamma \sim \mathcal{N}(\mathbf{w}_\gamma | \mathbf{m}_{\mathbf{w}_\gamma, \gamma}, \boldsymbol{\Sigma}_{\mathbf{w}_\gamma}) \tag{4.4}$$

where $\mathbf{m}_{\mathbf{w}_\gamma,\gamma} = (\mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}_\gamma} = diag(\boldsymbol{\sigma}_{\mathbf{w}}^2)$ with $\boldsymbol{\sigma}_{\mathbf{w}}^2 = (\sigma_{w,1}^2, \ldots, \sigma_{w,1}^2, \sigma_{w,2}^2, \ldots, \sigma_{w,H}^2)^\top$. Each $\mu_{w,h}$ and $\sigma_{w,h}^2$ is repeated with length $||\mathbf{w}_{\gamma,h}||$ dependent on $\boldsymbol{\gamma}$.

Through giving each group h a separate hyper-parameter $\sigma^2_{w,h}$ in (4.3), we leave the model open to penalising the groups of variables to different degrees through the priors:

$$\sigma^2_{w,h} \sim \mathcal{IG}(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h}). \tag{4.5}$$

By choosing the same fixed hyper-parameters, $\alpha_{w,h}$ and $\beta_{w,h}$ for each h, we lose information coupling between the different groups, although this could be regained with an addition layer in the hierarchical model.

In addition to $\sigma^2_{w,h}$, we use the hyper-parameters $\mu_{w,h}$ to reflect the likely non-zero means of each group h:

$$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma^2_{0,h}) \tag{4.6}$$

where the hyper-parameters $\mu_{0,h}$ and $\sigma^2_{0,h}$ are fixed. This specification comes from the expected biological values of each regression coefficients $w_{j,h}$. In the FMDV and Influenza data we are likely to observe a comparatively large intercept with negative regression coefficients for the variables. This is a result of amino acid changes decreasing the similarity between virus strains and therefore reducing the measured VN titre or HI assay. Similarly, traversing a significant branch of the phylogenetic tree is likely to cause differences between the strains.

A prior must also be given for $\gamma_j \in \{2, \ldots, J\}$, the parameters which determine the relevance of the covariates. No prior is included for the latent indicator variable associated with the intercept, as this is a-priori fixed to 1, $\gamma_1 = 1$.

$$p(\boldsymbol{\gamma}_{2:J}|\pi) = \prod_{j=2}^{J} \text{Bern}(\gamma_j|\pi) \tag{4.7}$$

where π is the probability of the individual variable being relevant.

The value of π can either be set as a fixed hyper-parameter as in Sabatti and James (2005), where they argue that it should be determined by underlying knowledge of the problem. Alternatively it can be given a conjugate Beta prior:

$$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi) \tag{4.8}$$

as in this case, where the likely number of relevant variables cannot be easily specified *a priori*. This is a more general model, which subsumes a fixed π as a limiting case for $\alpha_\pi\beta_\pi/((\alpha_\pi + \beta_\pi)^2(\alpha_\pi + \beta_\pi + 1)) \rightarrow 0$.

4.1.4 Random-Effects Prior

In mixed-effects models the random effects, $b_{k,g}$, are usually given group dependant Gaussian priors where the group g is defined by k, i.e. $b_{k,g}$ is shorthand for b_{k,g_k}:

$$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma^2_{b,g}). \tag{4.9}$$

We define this to have a fixed mean, $\mu_{b,g} = 0$, and a common variance parameter, $\sigma^2_{b,g}$, with a conjugate Inverse-Gamma prior for each random-effects group g, as shown in Figure 4.5a:

$$\sigma^2_{b,g} \sim \mathcal{IG}(\sigma^2_{b,g}|\alpha_{b,g}, \beta_{b,g}) \tag{4.10}$$

where $\alpha_{b,g}$ and $\beta_{b,g}$ are fixed hyper-parameters for each g and we define $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \mathbf{\Sigma_b})$ where $\mathbf{\Sigma_b} = diag(\boldsymbol{\sigma}^2_{\mathbf{b}})$ with $\boldsymbol{\sigma}^2_{\mathbf{b}} = (\sigma^2_{b,1}, \dots, \sigma^2_{b,1}, \sigma^2_{b,2}, \dots, \sigma^2_{b,G})^\top$ such that each $\sigma^2_{b,g}$ is repeated with length $||\mathbf{b}_g||$.

4.2 The Alternative SABRE Methods

Various different adjustments have been applied to the original SABRE method of Davies et al. (2014), as detailed in Davies et al. (2016a). These changes have resulted in several different versions of the method, the semi-conjugate (SC), conjugate (C) and binary mask conjugate (BM) SABRE methods, and these are detailed in this section.

4.2.1 The Semi-Conjugate SABRE Method

The semi-conjugate SABRE method, as proposed in Davies et al. (2016a), changes the likelihood of the original SABRE method, (4.1) in Section 4.1.1, to accommodate a separate parameter for the biologically significant intercept parameter, w_0:

$$p(\mathbf{y}|w_0, \mathbf{w}_\gamma, \mathbf{b}, \sigma^2_\varepsilon, \mathbf{X}_\gamma, \mathbf{Z}) = \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_\gamma\mathbf{w}_\gamma + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I}). \tag{4.11}$$

The intercept, w_0 is especially important when it comes to modelling antigenic variability as it is likely that the measures of antigenicity, VN titre and HI assay, will have a high value when just the intercept has an affect. This occurs when two identical virus strains are tested against each other and the associated variables are therefore all zero.

The change of the likelihood to (4.11) means that we also require a prior on the

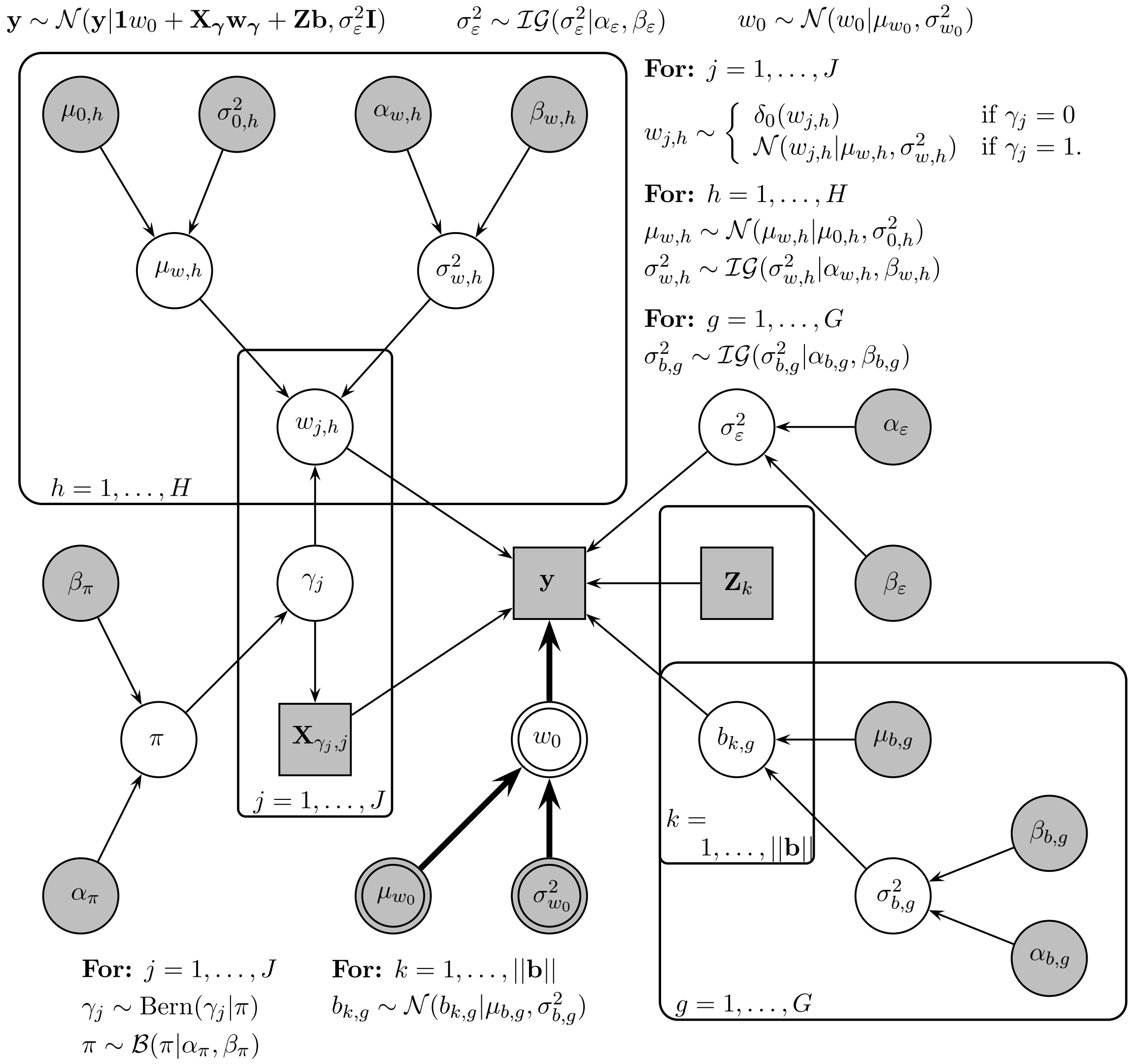

$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})$
$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)$
$w_0 \sim \mathcal{N}(w_0|\mu_{w_0}, \sigma_{w_0}^2)$
For: $j = 1, \dots, J$
$w_{j,h} \sim \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h}|\mu_{w,h}, \sigma_{w,h}^2) & \text{if } \gamma_j = 1. \end{cases}$
For: $h = 1, \dots, H$
$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{0,h}^2)$
$\sigma_{w,h}^2 \sim \mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h})$
For: $g = 1, \dots, G$
$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g})$
$\mu_{0,h}$
$\sigma_{0,h}^2$
$\alpha_{w,h}$
$\beta_{w,h}$
$\mu_{w,h}$
$\sigma_{w,h}^2$
$w_{j,h}$
$h = 1, \dots, H$
σ_ε^2
α_ε
β_π
γ_j
$\mathbf{y}$
$\mathbf{Z}_k$
β_ε
π
$\mathbf{X}_{\gamma_j,j}$
w_0
$b_{k,g}$
$\mu_{b,g}$
$j = 1, \dots, J$
$k = 1, \dots, ||\mathbf{b}||$
$\beta_{b,g}$
α_π
μ_{w_0}
$\sigma_{w_0}^2$
$\sigma_{b,g}^2$
$\alpha_{b,g}$
For: $j = 1, \dots, J$
$\gamma_j \sim \text{Bern}(\gamma_j|\pi)$
$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)$
For: $k = 1, \dots, ||\mathbf{b}||$
$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2)$
$g = 1, \dots, G$

Figure 4.2: **Compact representation of the semi-conjugate SABRE method as a PGM.** The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred. The PGM shows the addition of nodes and edges connecting w_0, μ_{w_0} and $\sigma_{w_0}^2$ into the model, something not seen in the original SABRE method in Figure 4.1.

intercept, w_0:

$$w_0 \sim \mathcal{N}(w_0|\mu_{w_0}, \sigma_{w_0}^2). \tag{4.12}$$

We treat the intercept differently from the remaining regressors, wishing to use vague prior settings so as not to penalise this term and effectively make the model scale invariant (Hastie et al., 2009). The difference between the semi-conjugate SABRE method and the original SABRE method can be seen graphically by comparing the PGMs in Figures 4.2 and 4.1, where Figure 4.2 shows the addition of nodes and edges connecting w_0, μ_{w_0} and $\sigma_{w_0}^2$ into the model.

For mathematical convenience we then define the prior distribution of $\mathbf{w}^*_{\boldsymbol{\gamma}} = (w_0, \mathbf{w}_{\boldsymbol{\gamma}}^\top)^\top$ as:

$$\mathbf{w}^*_{\boldsymbol{\gamma}} \sim \mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}} | \mathbf{m}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \tag{4.13}$$

where $\mathbf{m}_{\boldsymbol{\gamma}} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \ldots, \sigma^2_{w,1}, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$. Each $\mu_{w,h}$ and $\sigma^2_{w,h}$ is repeated with length $||\mathbf{w}_{\boldsymbol{\gamma},h}||$ dependent on $\boldsymbol{\gamma}$.

4.2.2 The Conjugate SABRE Method

The conjugate SABRE method of Davies et al. (2016a) makes the SABRE method conjugate rather than semi conjugate, as it is in the semi-conjugate and original SABRE methods (Sections 4.2.1 and 4.1). The idea of conjugate Bayesian models is discussed in detail in Chapter 3 of Gelman et al. (2013a), but in general the idea is to introduce extra links between the parameters in the model to increase information sharing. For the conjugate SABRE method we add relationships between w_0, $\mathbf{w}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\mu}_{\mathbf{w}} = (\mu_{w,1}, \ldots, \mu_{w,H})^\top$ with the error variance σ^2_ε. Adding these additional relationship increases information sharing and means that the error variance in terms of model fit is reflected in the distribution of the regression coefficients and associated mean. In addition to this increased information sharing, conjugate models also have a computational advantage as the sampling can be improved through using collapsed Gibbs sampling, as will be described in Section 4.3.6. The additional conjugacy of the conjugate SABRE method can be seen by looking at its PGM in Figure 4.3 and comparing it with that of the semi-conjugate SABRE method in Figure 4.2.

Adding the increased conjugacy requires the replacement of three of the equations from the semi-conjugate SABRE method. We must firstly replace the distribution of the intercept parameter, w_0, from (4.12):

$$w_0 \sim \mathcal{N}(w_0 | \mu_{w_0}, \sigma^2_{w_0}\sigma^2_\varepsilon). \tag{4.14}$$

We must also adjust the spike and slab prior in the model, (4.3), so that the distribution has increased conjugacy for the relevant variables, i.e. when $\gamma_j = 1$:

$$p(w_{j,h} | \gamma_j, \mu_{w,h}, \sigma^2_{w,h}, \sigma^2_\varepsilon) = \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h} | \mu_{w,h}, \sigma^2_{w,h}\sigma^2_\varepsilon) & \text{if } \gamma_j = 1 \end{cases} \tag{4.15}$$

which means we must also replace (4.13) with the following notationally convenient dis-

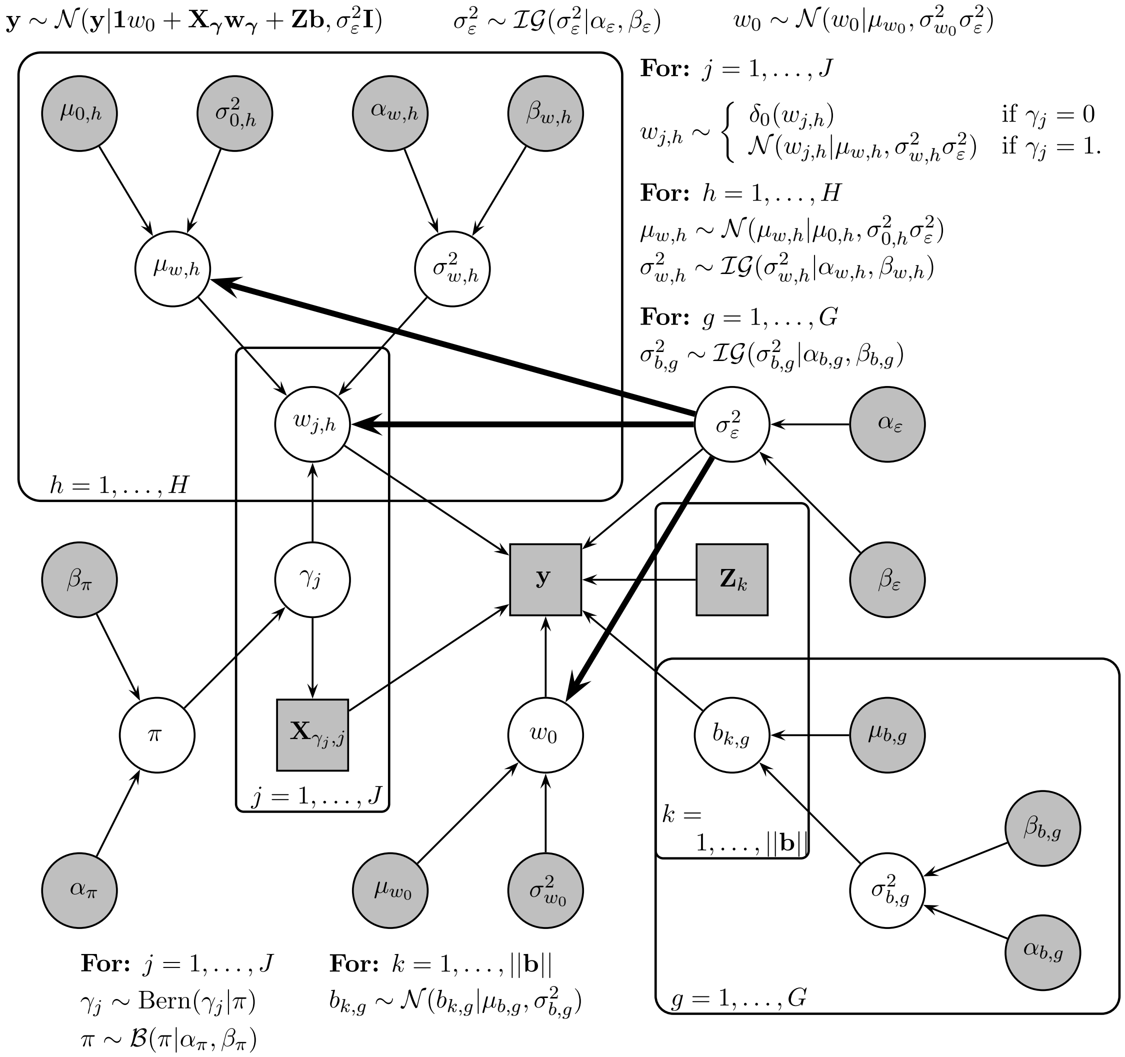

$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})$
$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)$
$w_0 \sim \mathcal{N}(w_0|\mu_{w_0}, \sigma_{w_0}^2\sigma_\varepsilon^2)$
For: $j = 1, \dots, J$
$w_{j,h} \sim \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h}|\mu_{w,h}, \sigma_{w,h}^2\sigma_\varepsilon^2) & \text{if } \gamma_j = 1. \end{cases}$
For: $h = 1, \dots, H$
$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{0,h}^2\sigma_\varepsilon^2)$
$\sigma_{w,h}^2 \sim \mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h})$
For: $g = 1, \dots, G$
$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g})$
$\mu_{0,h}$
$\sigma_{0,h}^2$
$\alpha_{w,h}$
$\beta_{w,h}$
$\mu_{w,h}$
$\sigma_{w,h}^2$
$w_{j,h}$
σ_ε^2
α_ε
$h = 1, \dots, H$
β_π
γ_j
$\mathbf{y}$
$\mathbf{Z}_k$
β_ε
π
$\mathbf{X}_{\gamma_j,j}$
w_0
$b_{k,g}$
$\mu_{b,g}$
$j = 1, \dots, J$
$k = 1, \dots, ||\mathbf{b}||$
$\beta_{b,g}$
α_π
μ_{w_0}
$\sigma_{w_0}^2$
$\sigma_{b,g}^2$
$\alpha_{b,g}$
For: $j = 1, \dots, J$
$\gamma_j \sim \text{Bern}(\gamma_j|\pi)$
$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)$
For: $k = 1, \dots, ||\mathbf{b}||$
$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2)$
$g = 1, \dots, G$

Figure 4.3: **Compact representation of the conjugate SABRE method as a PGM.** The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred. The difference between this PGM and that of the semi-conjugate SABRE method in Figure 4.2 can be seen by noting the extra, highlighted, edges between w_0, $w_{j,h}$ and $\mu_{w,h}$ and the error variance σ_ε^2.

tribution:

$$\mathbf{w}_{\boldsymbol{\gamma}}^* \sim \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}^*|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}) \tag{4.16}$$

Finally we must change the distribution of mean parameter of the regression coefficients from (4.6) to the following prior distribution:

$$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{0,h}^2\sigma_\varepsilon^2). \tag{4.17}$$

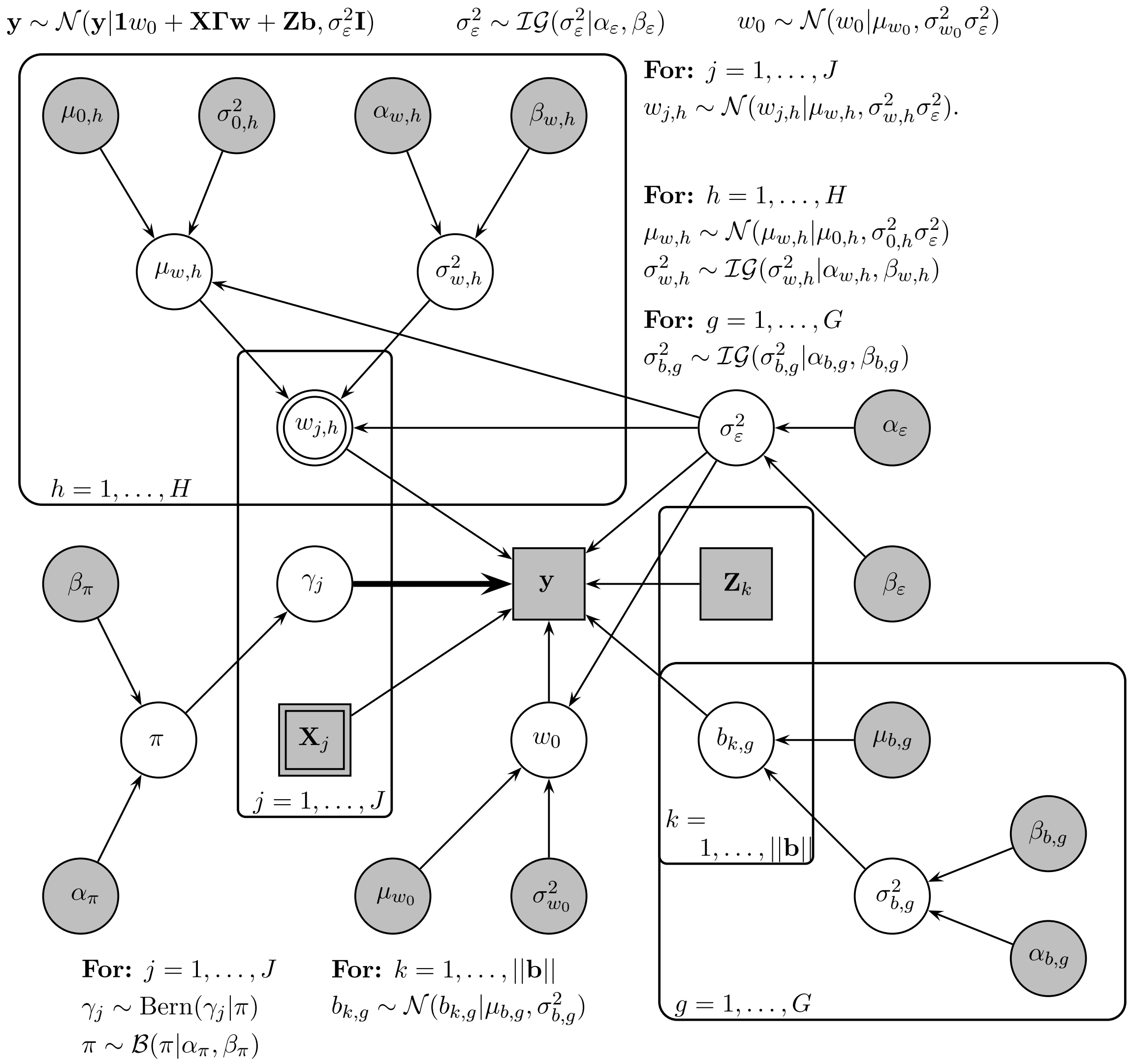

$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X\Gamma w} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})$
$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)$
$w_0 \sim \mathcal{N}(w_0|\mu_{w_0}, \sigma_{w_0}^2\sigma_\varepsilon^2)$
For: $j = 1, \dots, J$
$w_{j,h} \sim \mathcal{N}(w_{j,h}|\mu_{w,h}, \sigma_{w,h}^2\sigma_\varepsilon^2)$.
For: $h = 1, \dots, H$
$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{0,h}^2\sigma_\varepsilon^2)$
$\sigma_{w,h}^2 \sim \mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h})$
For: $g = 1, \dots, G$
$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g})$
$\mu_{0,h}$
$\sigma_{0,h}^2$
$\alpha_{w,h}$
$\beta_{w,h}$
$\mu_{w,h}$
$\sigma_{w,h}^2$
$w_{j,h}$
σ_ε^2
α_ε
$h = 1, \dots, H$
β_π
γ_j
$\mathbf{y}$
$\mathbf{Z}_k$
β_ε
π
$\mathbf{X}_j$
w_0
$b_{k,g}$
$\mu_{b,g}$
$j = 1, \dots, J$
$k =$
$1, \dots, ||\mathbf{b}||$
$\beta_{b,g}$
$\sigma_{b,g}^2$
α_π
μ_{w_0}
$\sigma_{w_0}^2$
$\alpha_{b,g}$
For: $j = 1, \dots, J$
$\gamma_j \sim \text{Bern}(\gamma_j|\pi)$
$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)$
For: $k = 1, \dots, ||\mathbf{b}||$
$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2)$
$g = 1, \dots, G$

Figure 4.4: **Compact representation of the binary mask conjugate SABRE method as a PGM.** The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred. Compared to the PGM of the conjugate SABRE method, Figure 4.3, the nodes here have a different structure as depicted in Figure 3.2.

4.2.3 The Binary Mask Conjugate SABRE Method

The binary mask conjugate SABRE provides an alternative to the conjugate SABRE method by using the binary mask model (Section 3.3.2), rather than the spike and slab prior (Section 3.3.1) (Davies et al., 2016a). In the binary mask model the indicator variables, $\boldsymbol{\gamma}$, 'mask' the impact of the regression coefficients rather than removing them from the model as in a spike and slab prior based model. To get the likelihood of the binary mask conjugate model we replace the likelihood of the conjugate and semi-conjugate SABRE methods, (4.11), with a binary mask version:

$$p(\mathbf{y}|w_0, \mathbf{w}, \boldsymbol{\gamma}, \mathbf{b}, \sigma_\varepsilon^2, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X\Gamma w} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I}) \tag{4.18}$$

where $\mathbf{\Gamma} = diag(\boldsymbol{\gamma})$. The differences can be seen by comparing the PGM of the binary mask conjugate SABRE method in Figure 4.4 with that of the conjugate SABRE method in Figure 4.3. Alternatively the difference can be seen by looking at Figure 3.2. The binary mask conjugate SABRE method will be compared with the conjugate and semi-conjugate SABRE methods in Section 5.3.

Despite the different model specification given in (4.18), most of the prior distributions given in the main paper remain the same. The only prior that changes is that of $w_{j,h}$, which is now given by:

$$w_{j,h} \sim \mathcal{N}(w_{j,h} | \mu_{w,h}, \sigma^2_{w,h} \sigma^2_{\varepsilon}) \tag{4.19}$$

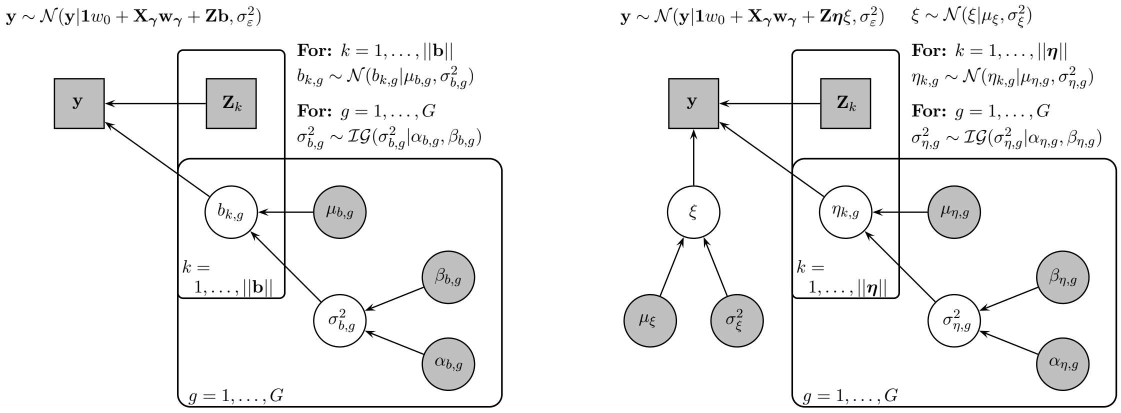
replacing (4.15) and resulting in the following multivariate prior for $\mathbf{w}^* = (w_0,\, \mathbf{w}^\top)^\top$:

$$\mathbf{w}^* \sim \mathcal{N}(\mathbf{w}^* | \mathbf{m}, \sigma^2_{\varepsilon} \mathbf{\Sigma}_{\mathbf{w}^*}) \tag{4.20}$$

where $\mathbf{m} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\mathbf{\Sigma}_{\mathbf{w}^*} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \ldots, \sigma^2_{w,1},\, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$. Each of the components $\mu_{w,h}$ and $\sigma^2_{w,h}$ is repeated with length $||\mathbf{w}_h||$ and unlike with the slab and spike prior their lengths do not depend on $\boldsymbol{\gamma}$.

4.2.4 Alternative Random Effect Priors

The final possible improvement to the SABRE methods is to try an alternative random effects prior to that described in Section 4.1.4. One such alternative is the folded-non-central-t prior distribution described in Gelman (2006), which gives a redundant multiplicative reparameterisation to the model in Figure 4.5a. This prior has several potential advantages over the Inverse-Gamma prior. Firstly it is considered to be a prior that better represents non-informativeness. While the posterior distribution can be sensitive to the fixed hyper-parameter settings of an Inverse-Gamma prior, the impact is reduced when the folded-non-central-t prior is used. In that case the posterior distribution does not have a sharp peak at zero unlike with an vague Inverse-Gamma prior, reducing problems with underestimating the variance. Secondly, Gelman (2006) found that the folded-non-central-t prior results in a more realistic posterior distribution of $\sigma^2_{b,g}$ when there are only a few random effects (usually less than 8) in each group g. The author showed that the posterior distribution reflected the marginal distribution well at its low end, but removed its unrealistically heavy tail; see Figure 2 in Gelman (2006). Doing this ensures that $\sigma^2_{b,g}$ is not overestimated and does not lead to non-optimal shrinkage of $\mathbf{b}_g$. Finally the over-parameterisation can improve sampling by reducing the dependence between parameters


$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_\gamma \mathbf{w}_\gamma + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2)$
For: $k = 1, \ldots, ||\mathbf{b}||$
$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2)$
For: $g = 1, \ldots, G$
$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g})$
$\mathbf{y}$
$\mathbf{Z}_k$
$b_{k,g}$
$\mu_{b,g}$
$\beta_{b,g}$
$\sigma_{b,g}^2$
$\alpha_{b,g}$
$k = 1, \ldots, ||\mathbf{b}||$
$g = 1, \ldots, G$
$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_\gamma \mathbf{w}_\gamma + \mathbf{Z}\boldsymbol{\eta}\xi, \sigma_\varepsilon^2)$
$\xi \sim \mathcal{N}(\xi|\mu_\xi, \sigma_\xi^2)$
For: $k = 1, \ldots, ||\boldsymbol{\eta}||$
$\eta_{k,g} \sim \mathcal{N}(\eta_{k,g}|\mu_{\eta,g}, \sigma_{\eta,g}^2)$
For: $g = 1, \ldots, G$
$\sigma_{\eta,g}^2 \sim \mathcal{IG}(\sigma_{\eta,g}^2|\alpha_{\eta,g}, \beta_{\eta,g})$
$\mathbf{y}$
$\mathbf{Z}_k$
ξ
$\eta_{k,g}$
$\mu_{\eta,g}$
$\beta_{\eta,g}$
$\sigma_{\eta,g}^2$
$\alpha_{\eta,g}$
μ_ξ
σ_ξ^2
$k = 1, \ldots, ||\boldsymbol{\eta}||$
$g = 1, \ldots, G$

(a) Inverse-Gamma Prior

(b) Half-t Prior

Figure 4.5: **PGMs for the two different specifications of the hierarchical random-effects model.** (a) Classical random-effects model using Gaussian and Inverse-Gamma priors. (b) Half-t prior specified in a hierarchical manner, as suggested by Gelman (2006). The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred.

in the hierarchical model leading to improved MCMC convergence (Gelman, 2004).

The redundant multiplicative reparameterisation used for this prior specification sets $\mathbf{b} = \boldsymbol{\eta}\xi$ and is given by the following conjugate priors and shown in Figure 4.5b:

$$\eta_{k,g} \sim \mathcal{N}(\eta_{k,g}|\mu_{\eta,g}, \sigma_{\eta,g}^2) \tag{4.21}$$

$$\xi \sim \mathcal{N}(\xi|\mu_\xi, \sigma_\xi^2) \tag{4.22}$$

where μ_ξ and σ_ξ^2 are fixed for identifiability, $\mu_{\eta,g} = 0$, $\eta_{k,g}$ is shorthand for η_{k,g_k} and each $b_{k,g} = \xi\eta_{k,g}$. Following Gelman (2006), we fix $\mu_\xi = 0$ which leads to the half-t distribution. We then set a prior on $\sigma_{\eta,g}^2$:

$$\sigma_{\eta,g}^2 \sim \mathcal{IG}(\sigma_{\eta,g}^2|\alpha_{\eta,g}, \beta_{\eta,g}) \tag{4.23}$$

where $\alpha_{\eta,g}$ and $\beta_{\eta,g}$ are fixed hyper-parameters. In terms of standard mixed-effects models, the variance is given by $\sigma_{b,g}^2 = \xi^2\sigma_{\eta,g}^2$. For convenience we define $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ when $\mu_{\eta,g} = 0$ for all g and where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = diag(\boldsymbol{\sigma}_{\boldsymbol{\eta}}^2)$ with $\boldsymbol{\sigma}_{\boldsymbol{\eta}}^2 = (\sigma_{\eta,1}^2, \ldots, \sigma_{\eta,1}^2, \sigma_{\eta,2}^2, \ldots, \sigma_{\eta,G}^2)^\top$ where each $\sigma_{\eta,g}^2$ is repeated with length $||\boldsymbol{\eta}_g||$. In this thesis we implement the folded-non-central-t prior into an alternative version of the conjugate SABRE method and compare them in Section 5.3 (Davies et al., 2016a).

4.3 Posterior Inference

In order to explore the posterior distributions of the different SABRE methods described in Sections 4.1 and 4.2 we use an MCMC algorithm as introduced in Section 3.2. Having generally chosen conjugate priors means that we can mainly use Gibbs sampling (Section 3.2.2) to sample to majority of parameters in all of the SABRE methods. The only exception is $\boldsymbol{\gamma}$, although it is possible to use component-wise Gibbs sampling with a small adaptation; see Section 4.3.5. Additionally we sample the intercept and regression parameters together and define $\mathbf{w}^*_{\boldsymbol{\gamma}} = (w_0, \mathbf{w}_{\boldsymbol{\gamma}}^\top)^\top$, $\mathbf{X}^*_{\boldsymbol{\gamma}} = (\mathbf{1}, \mathbf{X}_\gamma)$, $\mathbf{m}_{\boldsymbol{\gamma}} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \ldots, \sigma^2_{w,1}, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$. Each $\mu_{w,h}$ and $\sigma^2_{w,h}$ is repeated with length $||\mathbf{w}_{\boldsymbol{\gamma},h}||$ dependent on $\boldsymbol{\gamma}$, as indicated below (4.13).

In this section we give the conditional distributions of all those parameters that are amenable to Gibbs sampling as well as conditional distributions for $\boldsymbol{\gamma}$. For readability we do not mathematically derive these distributions in this section and instead they are given in Appendix A.1. For convenience, we denote $\boldsymbol{\theta}$ to be a vector of all parameters and hyperparameters. The distributions required to sample the original SABRE methods are given in Section 4.3.1, with the required changes for the alternative SABRE methods given in Sections 4.3.2, 4.3.3 and 4.3.4. After the distributions are given, Section 4.3.5 then looks in detail at how to effectively sample $\boldsymbol{\gamma}$ as this is the only model parameter that is not sampled effectively with any form of Gibbs sampling. Finally, Section 4.3.6 details the conjugate sampling strategy that can be used with the conjugate and binary mask conjugate SABRE methods.

4.3.1 Original SABRE Method

The posterior distributions for the model parameters of the original SABRE method which can be sampled via Gibbs sampling are given as follows, where the analytical derivations are given in Appendix A.1.1:

$$\mathbf{w}_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{-\mathbf{w}_{\boldsymbol{\gamma}}}, \mathfrak{D} \sim \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\mathbf{X}_{\boldsymbol{\gamma}}^\top(\mathbf{y}-\mathbf{Z}\mathbf{b})/\sigma^2_\varepsilon + \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}_{\boldsymbol{\gamma}}}\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}) \tag{4.24}$$

$$\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathfrak{D} \sim \mathcal{N}(\mathbf{b}|\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top(\mathbf{y}-\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}})/\sigma^2_\varepsilon, \mathbf{V}_{\mathbf{b}}) \tag{4.25}$$

$$\sigma^2_{b,g}|\boldsymbol{\theta}_{-\sigma^2_{b,g}}, \mathfrak{D} \sim \mathcal{IG}(\sigma^2_{b,g}|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}_g^\top\mathbf{b}_g) \tag{4.26}$$

$$\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}}, \mathfrak{D} \sim \mathcal{N}(\mu_{w,h}|V^{-1}_{\mu_\gamma,h}(\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h})/\sigma^2_{w,h} + \mu_{0,h}/\sigma^2_{0,h}), V_{\mu_\gamma,h}) \tag{4.27}$$

$$\sigma^2_{w,h}|\boldsymbol{\theta}_{-\sigma^2_{w,h}}, \mathfrak{D} \sim \mathcal{IG}(\sigma^2_{w,h}|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h} + \tfrac{1}{2}\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{w,h})^2) \tag{4.28}$$

$$\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathfrak{D} \sim \mathcal{IG}(\sigma^2_\varepsilon|N/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}\Sigma(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})^2) \tag{4.29}$$

$$\pi|\boldsymbol{\theta}_{-\pi}, \mathfrak{D} \sim \mathcal{B}(\pi|\alpha_\pi + \textstyle\sum\boldsymbol{\gamma}, \beta_\pi + J - \sum\boldsymbol{\gamma}) \tag{4.30}$$

where we sample $\sigma^2_{b,g}$, $\mu_{w,h}$ and $\sigma^2_{w,h}$ for each g and h respectively. We also define $\mathbf{V}_{\mathbf{w}_\gamma} = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma/\sigma^2_\varepsilon + \boldsymbol{\Sigma}_\mathbf{w}^{-1})^{-1}$, $\mathbf{V}_\mathbf{b} = (\mathbf{Z}^\top\mathbf{Z}/\sigma^2_\varepsilon + \boldsymbol{\Sigma}_\mathbf{b}^{-1})^{-1}$ and $V_{\mu_\gamma,h} = ((||\mathbf{w}_{\gamma,h}||/\sigma^2_{w,h})^{-1} + (\sigma^2_{0,h})^{-1})^{-1}$ for notational simplicity. These distributions can be sampled in any order, with each update using the most recent sample of the conditioned parameters; see Section 3.2.2.

Sampling $\boldsymbol{\gamma}$ is more difficult, as it does not naturally form a standard distribution. Methods for achieving this are discussed in more detail in Section 4.3.5, however in order to do this we need a conditional distribution:

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\gamma}, \mathfrak{D}) \propto \text{Bern}(\boldsymbol{\gamma}|\pi) \int \mathcal{N}(\mathbf{y}|\mathbf{X}_\gamma \mathbf{w}_\gamma + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon \mathbf{I}) \mathcal{N}(\mathbf{w}_\gamma|\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w}) d\mathbf{w}_\gamma \tag{4.31}$$

$$\propto \pi^{\sum\gamma}(1-\pi)^{J-\sum\gamma}\mathcal{N}(\mathbf{y}|\mathbf{X}_\gamma\boldsymbol{\mu}_\mathbf{w} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I} + \mathbf{X}_\gamma\boldsymbol{\Sigma}_\mathbf{w}\mathbf{X}_\gamma^\top) \tag{4.32}$$

where there are J variables. Here we have used a collapsing step as in Sabatti and James (2005), integrating out $\mathbf{w}_\gamma$ through the application of standard Gaussian integrals (Bishop, 2006) to reduce the computational requirements. The normalisation constant is not required in (4.31) and (4.32) as it cancels out in all of the methods discussed in Section 4.3.5.

4.3.2 Semi-Conjugate SABRE Method

To get the conditional distributions of the model parameters for the semi-conjugate SABRE methods we begin with the conditional distributions for the original SABRE method and replace (4.24), (4.25) and (4.29) with the following equation which have been derived in Appendix A.1.2:

$$\mathbf{w}^*_\gamma|\boldsymbol{\theta}_{-\mathbf{w}^*_\gamma}, \mathfrak{D} \sim \mathcal{N}(\mathbf{w}^*_\gamma|\mathbf{V}_{\mathbf{w}_\gamma}\mathbf{X}_\gamma^\top(\mathbf{y} - \mathbf{Z}\mathbf{b})/\sigma^2_\varepsilon + \mathbf{V}_{\mathbf{w}_\gamma}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}_\gamma}\boldsymbol{\mu}_\mathbf{w}, \mathbf{V}_{\mathbf{w}^*_\gamma}) \tag{4.33}$$

$$\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathfrak{D} \sim \mathcal{N}(\mathbf{b}|\mathbf{V}_\mathbf{b}\mathbf{Z}^\top(\mathbf{y} - \mathbf{X}^*_\gamma\mathbf{w}^*_\gamma)/\sigma^2_\varepsilon, \mathbf{V}_\mathbf{b}) \tag{4.34}$$

$$\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathfrak{D} \sim \mathcal{IG}(\sigma^2_\varepsilon|N/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}\textstyle\sum(\mathbf{y} - \mathbf{X}^*_\gamma\mathbf{w}^*_\gamma - \mathbf{Z}\mathbf{b})^2) \tag{4.35}$$

where we define $\mathbf{V}_{\mathbf{w}^*_\gamma} = (\mathbf{X}^{*,\top}_\gamma\mathbf{X}^*_\gamma/\sigma^2_\varepsilon + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*})^{-1}$ for notational simplicity.

To sample $\boldsymbol{\gamma}$ for the semi-conjugate SABRE method we again use collapsing steps (Sabatti and James, 2005), however in this instance we integrate out both $\mathbf{w}_\gamma$ and π. While it was also possible to integrate out π in the original SABRE method, we did not do this in Davies et al. (2014) and therefore we have not integrated π out in Section 4.3.1 either. Integrating over $\mathbf{w}_\gamma$ and π then leaves the following conditional distribution for $\boldsymbol{\gamma}$:

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathfrak{D}) \propto \int \beta(\pi|\alpha_\pi, \beta_\pi)\, \mathrm{Bern}(\boldsymbol{\gamma}|\pi)& \\
\mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})d\pi d\mathbf{w}_{\boldsymbol{\gamma}}& \qquad (4.36)\\
\propto \tfrac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}\mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma^2_\varepsilon\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}})& \qquad (4.37)
\end{aligned}
$$

which replace (4.31) and (4.32) and can be sampled using the methods from Section 4.3.5.

4.3.3 Conjugate SABRE Method

The conditional distributions of the conjugate SABRE method are similar to those of the semi-conjugate SABRE method and their derivations can be found in Appendix A.1.3. To sample the model parameters of the conjugate SABRE method, we use the method for the semi-conjugate SABRE method but replace (4.33), (4.27), (4.28) and (4.35) with the following distributions:

$$
\begin{aligned}
\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{-\mathbf{w}^*_{\boldsymbol{\gamma}}}, \mathfrak{D} &\sim \mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}(\mathbf{y} - \mathbf{Zb}) + \mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}},2}) && (4.38)\\
\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}}, \mathfrak{D} &\sim \mathcal{N}(\mu_{w,h}|V^{-1}_{\mu_\gamma,h}(\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h})/\sigma^2_{w,h} + \mu_{0,h}/\sigma^2_{0,h}), \sigma^2_\varepsilon V_{\mu_\gamma,h,2}) && (4.39)\\
\sigma^2_{w,h}|\boldsymbol{\theta}_{-\sigma^2_{w,h}}, \mathfrak{D} &\sim \mathcal{IG}(\sigma^2_{w,h}|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h} + \tfrac{1}{2\sigma^2_\varepsilon}\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{w,h})^2) && (4.40)\\
\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathfrak{D} &\sim \mathcal{IG}(\sigma^2_\varepsilon|(N + ||\mathbf{w}^*_{\boldsymbol{\gamma}}|| + H)/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}R_{\sigma^2_\varepsilon}) && (4.41)
\end{aligned}
$$

where we define $\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}},2} = (\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})^{-1}$ and $R_{\sigma^2_\varepsilon} = (\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Zb})^\top(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Zb}) + (\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}})^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}(\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}}) + \sum_{h=1}^{H}(\mu_{w,h} - \mu_{0,h})^2/\sigma^2_{0,h}$ for notational simplicity.

To sample $\boldsymbol{\gamma}$ in the conjugate SABRE method we use the same method as the semi-conjugate SABRE method but changing the distribution of $\mathbf{w}^*_{\boldsymbol{\gamma}}$, replacing (4.13) with (4.16):

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathfrak{D}) \propto \int \beta(\pi|\alpha_\pi, \beta_\pi)\, \mathrm{Bern}(\boldsymbol{\gamma}|\pi)& \\
\mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})d\pi d\mathbf{w}_{\boldsymbol{\gamma}}& \qquad (4.42)\\
\propto \tfrac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}\mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma^2_\varepsilon[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}]).& \qquad (4.43)
\end{aligned}
$$

which replaces (4.36) and (4.37) in the sampling strategy. We can also use the CSS when sampling the conjugate SABRE method and this is discussed in Section 4.3.6.

Finally we discuss the conditional distributions of the conjugate SABRE method when the half-t prior is used instead of the standard Inverse-Gamma prior. In order to do this we set $\mathbf{b} = \boldsymbol{\eta}\xi$ and $\sigma^2_{b,g} = \xi^2\sigma^2_{\eta,g}$ in (4.38), (4.41) and (4.43) of the sampling strategy for the conjugate SABRE method. We can then sample $\boldsymbol{\eta}$, ξ and $\sigma^2_{\eta,g}$ from their conditional distributions, replacing (4.34) and (4.26):

$$\begin{aligned}
\boldsymbol{\eta}|\boldsymbol{\theta}_{-\boldsymbol{\eta}},\boldsymbol{\mathfrak{D}} &\sim \mathcal{N}(\boldsymbol{\eta}|\tfrac{\xi}{\sigma^2_\varepsilon}\mathbf{V}_{\boldsymbol{\eta}}\mathbf{Z}^\top(\mathbf{y}-\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}),\mathbf{V}_{\boldsymbol{\eta}}) && (4.44)\\
\xi|\boldsymbol{\theta}_{-\xi},\boldsymbol{\mathfrak{D}} &\sim \mathcal{N}(\xi|V_\xi[\tfrac{\mu_\xi}{\sigma^2_\xi}+\tfrac{1}{\sigma^2_\varepsilon}\boldsymbol{\eta}^\top\mathbf{Z}^\top(\mathbf{y}-\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}})],V_\xi) && (4.45)\\
\sigma^2_{\eta,g}|\boldsymbol{\theta}_{-\sigma^2_{\eta,g}},\boldsymbol{\mathfrak{D}} &\sim \mathcal{IG}(\sigma^2_{\eta,g}|||\boldsymbol{\eta}_g||/2+\alpha_{\eta,g},\beta_{\eta,g}+\tfrac{1}{2}\boldsymbol{\eta}_g^\top\boldsymbol{\eta}_g) && (4.46)
\end{aligned}$$

where $\mathbf{V}_{\boldsymbol{\eta}} = (\frac{\xi^2}{\sigma^2_\varepsilon}\mathbf{Z}^\top\mathbf{Z}+\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\eta}})^{-1}$ and $V_\xi = (\frac{1}{\sigma^2_\xi}+\frac{1}{\sigma^2_\varepsilon}\boldsymbol{\eta}^\top\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\eta})^{-1}$.

4.3.4 Binary Mask Conjugate SABRE Method

Changing from models that use spike and slab priors, Section 3.3.1, to a binary mask model, Section 3.3.2, causes a number of changes to the conditional distributions. This is a result of a change in the likelihood, (4.18), and the prior on $w_{j,h}$, (4.19), and means that only the conditional distributions of $\sigma^2_{b,g}$ and π remain the same as the conjugate SABRE method. We give the other conditional distributions as follows:

$$\begin{aligned}
\mathbf{w}^*|\boldsymbol{\theta}_{-\mathbf{w}^*},\mathbf{X}^*,\mathbf{Z},\mathbf{y} &\sim \mathcal{N}(\mathbf{w}^*|\mathbf{V}_{\mathbf{w}^*}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}(\mathbf{y}-\mathbf{Zb})+\mathbf{V}_{\mathbf{w}^*}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*}\mathbf{m},\sigma^2_\varepsilon\mathbf{V}_{\mathbf{w}^*}) && (4.47)\\
\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}},\mathbf{X}^*,\mathbf{Z},\mathbf{y} &\sim \mathcal{N}(\mathbf{b}|\mathbf{V}_\mathbf{b}\mathbf{Z}^\top(\mathbf{y}-\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^*)/\sigma^2_\varepsilon,\mathbf{V}_\mathbf{b}) && (4.48)\\
\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}},\mathbf{X}^*,\mathbf{Z},\mathbf{y} &\sim \mathcal{N}(\mu_{w,h}|V^{-1}_{\mu,h}(\Sigma(\mathbf{w}_h)/\sigma^2_{w,h}+\mu_{0,h}/\sigma^2_{0,h}),\sigma^2_\varepsilon V_{\mu,h}) && (4.49)\\
\sigma^2_{w,h}|\boldsymbol{\theta}_{-\sigma^2_{w,h}},\mathbf{X}^*,\mathbf{Z},\mathbf{y} &\sim \mathcal{IG}(\sigma^2_{w,h}|\ ||\mathbf{w}_h||/2+\alpha_{w,h},\beta_{w,h}+\tfrac{1}{2\sigma^2_\varepsilon}\Sigma(\mathbf{w}_h-\mathbf{1}\mu_{w,h})) && (4.50)\\
\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon},\mathbf{X}^*,\mathbf{Z},\mathbf{y} &\sim \mathcal{IG}(\sigma^2_\varepsilon|(N+||\mathbf{w}^*||+H)/2+\alpha_\varepsilon,\beta_\varepsilon+\tfrac{1}{2}R_{\sigma^2_\varepsilon,2}) && (4.51)
\end{aligned}$$

where we sample $\sigma^2_{b,g}$, $\mu_{w,h}$ and $\sigma^2_{w,h}$ for each g and h respectively. We also define $\mathbf{V}_{\mathbf{w}^*} = (\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}\mathbf{X}^*\boldsymbol{\Gamma}\ +\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*})^{-1}$ and $R_{\sigma^2_\varepsilon,2} = (\mathbf{y}-\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^*-\mathbf{Zb})^\top(\mathbf{y}-\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^*-\mathbf{Zb})+(\mathbf{w}^*-\mathbf{m})^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*}(\mathbf{w}^*\ -\mathbf{m})+(\boldsymbol{\mu}_\mathbf{w}-\boldsymbol{\mu}_\mathbf{0})^\top\boldsymbol{\Sigma}_\mathbf{0}^{-1}(\boldsymbol{\mu}_\mathbf{w}-\boldsymbol{\mu}_\mathbf{0})$ for notational simplicity.

Finally, to sample $\boldsymbol{\gamma}$ we collapse over $\mathbf{w}$ and π to give the following conditional distribution, replacing (4.42) and (4.43):

$$\begin{aligned}
p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}},\boldsymbol{\mathfrak{D}}) \propto \int \beta(\pi|\alpha_\pi,\beta_\pi)\,\mathrm{Bern}(\boldsymbol{\gamma}|\pi)&\\
\mathcal{N}(\mathbf{y}|\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^*+\mathbf{Zb},\sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*|\mathbf{m},\sigma^2_\varepsilon\boldsymbol{\Sigma}_{\mathbf{w}^*})d\pi d\mathbf{w}_{\boldsymbol{\gamma}}& \quad (4.52)\\
\propto \tfrac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}\mathcal{N}(\mathbf{y}|\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m}+\mathbf{Zb},\sigma^2_\varepsilon[\mathbf{I}+\mathbf{X}^*\boldsymbol{\Gamma}^*\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}]).& \quad (4.53)
\end{aligned}$$

4.3.5 Sampling the Latent Inclusion Variables, γ

Sampling $\boldsymbol{\gamma}$ is more difficult, as it does not naturally take a distribution of standard form. However we can still get a valid conditional distribution and use a variety of techniques to sample from it. Multiple methods have been proposed for sampling the latent variables, $\boldsymbol{\gamma}$. Here we look at two of these in particular; the component-wise Gibbs sampling approach and a block M-H step. In the latter we can propose changes to multiple parameters simultaneously for a computational improvement.

A component-wise Gibbs sampler can be used to consecutively sample each γ_j from $\boldsymbol{\gamma}$ in a random order dependent on the current state, c, of all the other γs, $\boldsymbol{\gamma}^c_{-j} = (\gamma^c_1, \ldots, \gamma^c_{j-1}, \gamma^c_{j+1}, \ldots, \gamma^c_J)$. We can define the conditional distribution of the ith iteration of γ_j to be a Bernoulli distribution with probability:

$$p(\gamma_j = 1 | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\gamma}^c_{-j}, \boldsymbol{\mathfrak{D}}, \mathbf{y}) = \frac{a}{a+b}, \tag{4.54}$$

where we define $a \propto p(\gamma_j = 1, \boldsymbol{\gamma}^c_{-j} | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}, \mathbf{y})$ and $b \propto p(\gamma_j = 0, \boldsymbol{\gamma}^c_{-j} | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}, \mathbf{y})$ using the appropriate conditional distribution of $\boldsymbol{\gamma}$.

The alternative, block M-H sampling can improve mixing and convergence through proposing sets, S, of latent indicator variables, $\boldsymbol{\gamma}_S$, simultaneously, where $\boldsymbol{\gamma}_S$ denotes a column vector of all the γ_js where $j \in S$ and $\boldsymbol{\gamma}_{-S}$ its compliment. The proposals are then accepted with the following acceptance rate:

$$\alpha(\boldsymbol{\gamma}^*_S, \boldsymbol{\gamma}^c_S | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}\mathbf{y}, \boldsymbol{\gamma}^c_{-S}) := \min\left\{ \frac{q(\boldsymbol{\gamma}^c_S | \pi_{prop}) p(\boldsymbol{\gamma}_S = \boldsymbol{\gamma}^*_S, \boldsymbol{\gamma}^c_{-S} | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}, \mathbf{y})}{q(\boldsymbol{\gamma}^*_S | \pi_{prop}) p(\boldsymbol{\gamma}_S = \boldsymbol{\gamma}^c_S, \boldsymbol{\gamma}^c_{-S} | \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}, \mathbf{y})}, 1 \right\} \tag{4.55}$$

where $q(.)$ is a proposal density and is set to be: $q(\boldsymbol{\gamma}^*_S | \pi_{prop}) = \prod_{j \in S} \text{Bern}(\gamma^*_j | \pi_{prop})$, where π_{prop} is a fixed tuning parameter. Proposed moves for independent sets of randomly ordered inclusion parameters, $\boldsymbol{\gamma}^*_S$, are then accepted if $\alpha(\boldsymbol{\gamma}^*_S, \boldsymbol{\gamma}^c_S | \, \boldsymbol{\theta}_{-\gamma}, \boldsymbol{\mathfrak{D}}, \mathbf{y}, \boldsymbol{\gamma}^c_{-S})$ is greater than a uniform random variable $u \sim \mathfrak{U}[0,1]$, until updates have been proposed for all the latent indicator variables.

4.3.6 Conjugate Sampling Strategy

Collapsing can lead to improved mixing and convergence, e.g. Andrieu and Doucet (1999). We take advantage of the induced conjugacy to sample the parameters $\boldsymbol{\gamma}$, $\mathbf{w}^*_{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}_{\mathbf{w}} = (\mu_{w,1}, \ldots, \mu_{w,H})^\top$, σ^2_ε and π as a series of collapsed distributions rather than through Gibbs sampling:

$$p(\boldsymbol{\gamma}, \mathbf{w}^*_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_\varepsilon, \pi) \tag{4.56}$$

$$= p(\boldsymbol{\gamma}) p(\pi | \boldsymbol{\gamma}) p(\sigma^2_\varepsilon | \pi, \boldsymbol{\gamma}) p(\boldsymbol{\mu}_{\mathbf{w}} | \sigma^2_\varepsilon, \pi, \boldsymbol{\gamma}) p(\mathbf{w}^*_{\boldsymbol{\gamma}} | \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_\varepsilon, \pi, \boldsymbol{\gamma}) \tag{4.57}$$

$$= p(\boldsymbol{\gamma}) p(\pi | \boldsymbol{\gamma}) p(\sigma^2_\varepsilon | \boldsymbol{\gamma}) p(\boldsymbol{\mu}_{\mathbf{w}} | \sigma^2_\varepsilon, \boldsymbol{\gamma}) p(\mathbf{w}^*_{\boldsymbol{\gamma}} | \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_\varepsilon, \boldsymbol{\gamma}) \tag{4.58}$$

where the conditionality on $\boldsymbol{\theta}'$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{y}$ has been dropped and the simplification from (4.57) to (4.58) follows from the conditional independence relations shown in Figure 4.3, exploiting the fact that π is d-separated from the remaining parameters in the argument via $\boldsymbol{\gamma}$. These distributions are achieved by collapsing over parameters as derived in Appendix A.

4.4 Discussion

In this chapter we have proposed a family of sparse hierarchical Bayesian models for detecting relevant antigenic sites in virus evolution (SABRE) should offer an improvement over the classical mixed-effects model, the mixed-effects LASSO and the mixed-effects elastic net. There are four reason that we should see an improvement when the methods are compared in Chapter 5. The proposed hierarchical modelling framework with slab-and-spike prior (1) avoids the bias inherent in LASSO-type methods, (2) genuinely and consistently achieve sparsity, (3) properly accounts for uncertainty at all levels of inference, and (4) borrows strength from information coupling, whereby all parameters are systematically and iteratively inferred in the context of all other parameters. In some more detail: (1) The shrinkage effect inherent in the ℓ_1 penalty term introduces a bias by which the regression parameters are systematically underestimated. This bias is avoided with the slab and spike prior that we use. (2) The LASSO is known to only give sparse solutions at the MAP (maximum a posteriori) configuration, but not when sampling parameters from the posterior distribution. From a Bayesian perspective, the MAP is methodologically inconsistent, as it is not guaranteed to represent the region in parameter space with the highest probability mass. The spike-and-slab prior, which we use, avoids this methodological inconsistency and achieves sparsity in a sound Bayesian inference context. (3) In our hierarchical Bayesian models, all sources of uncertainty are properly accounted for. The higher-level hyperparameters have their own distributions, which are systematically inferred from the data. In contrast, the regularisation parameters of the established methods are typically fixed, set e.g. by cross-validation, but without taking their uncertainty into account (see also Chapter 5 in Gelman et al. (2013a) for a more detailed discussion). (4) In our approach, we explicitly model all dependencies among the variables, and inference is carried out within the context of the whole system. This systematically borrows strength from information coupling and avoids the piecemeal approach of established methods.

There are two fundamentally different approaches to variable selection in Bayesian hierarchical models: the slab-and-spike prior, whereby the influence of an input variable is controlled via the prior distribution of its associated regression parameters, and the binary mask model, where variables are put through a binary multiplicative filter. The difference is depicted in Figures 4.3 and 4.4, or alternatively in Figure 3.2. Which method is better? Standard textbooks, like Murphy (2012), describe both methods (see Chapter 13), but do not offer a comparative evaluation, and in the literature, authors rather arbitrarily tend to opt for one method or another (see e.g. Heydari et al. (2016)). We have proposed two version of the SABRE method in order to allow us to carry out a systematic comparison to properly quantify the difference in terms of accuracy and computational efficiency between the two approaches in Chapter 5. We have also provided a way of systematically evaluated the influence of the prior, comparing a conjugate with a non-conjugate prior, as depicted by Figures 4.3 and 4.2, and we have assessed its influence systematically in terms of accuracy, computational efficiency, and formal model selection preference in Chapter 5. The conjugate and binary mask conjugate also allow the use the conjugate sampling scheme proposed in Section 4.3.6, which potentially offers improved computational efficiency through the use of collapsed Gibbs sampling, something we test in Chapter 5

Chapter 5

Sparse Hierarchical Bayesian Models for Understanding Antigenic Variability - The Analysis

In this chapter we show how the SABRE methods introduced in Chapter 4 outperform the alternative methods discussed in Chapter 3. We introduce the simulated and real datasets that will be used to show this (Section 5.1) and detail the computational procedures needed to produce the results (Section 5.2). The results for the simulated datasets compare the SABRE methods, as well as the methods from Chapter 3, against each other in terms of variable selection and out-of-sample performance. The results show that the SABRE methods offer a clear improvement in terms of model selection over the methods described in Chapter 3, with the SABRE methods all performing roughly equally. Additionally Section 5.3.3 looks at using Bayesian 10-fold CV and Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) to select the correct random effect specification, quantifying the difference in performance (Davies et al., 2016b).

Finally Sections 5.4, 5.5 and 5.6 give the results for a number of real FMDV and Influenza datasets looking at how well the various methods do in classifying variables (based on Section 2.4) as well as discussing the biological results in terms of antigenic residues and significant evolutionary changes in the phylogenetic trees. The results given in these sections, as well as Appendix B, show that the SABRE methods identify a number of known antigenic residues, as well as making novel predictions about other potentially antigenic residues.

5.1 Data

Detailed descriptions of the different FMDV and Influenza datasets used in this thesis are given in Sections 2.2 and 2.3 of Chapter 2. In this section we detail the simulated datasets that are used to test the methods described in Sections 4.1 and 4.2 against each other and those described in Chapter 3. We also add a few extra details on the real life datasets that are specific to this chapter of the thesis.

5.1.1 Initial Simulation Study

Davies et al. (2014) used 20 datasets in their simulation study, simulated with both fixed and random effects. All of the datasets were given 30 variable, with 10 of the datasets given one group of random effects and the remaining sets given two groups. Each of the variables was then given a regression parameter. Half of each group were given small negative regressors drawn from $\mathbf{w_1} \sim \mathcal{N}(-0.2, 0.01)$ and the other half $\mathbf{w_2} \sim \mathcal{N}(0, 0.0025)$. Each response y_i was then generated from the model with each of the perturbed regressors $\tilde{w}_{h,i} \sim \mathcal{N}(w_{h,i}, 0.007)$, where $h \in \{1, 2\}$. This was done 200 times with additive Gaussian noise from $\mathcal{N}(0, 0.04)$ given to each response. Half of the data was used for training and the remaining for testing.

5.1.2 Extended Simulation Study

In Davies et al. (2016a) we simulated 9 sets of simulated data each with 100 datasets with 100 measurements for training and 900 for testing. We varied the number of variables, $||\mathbf{w}|| \in \{40, 60, 80\}$, and the size of the error, $\sigma_\varepsilon^2 \in \{0.01, 0.1, 0.3\}$, to test the methods under different circumstances. Additionally we added two groups of random effects to each dataset to represent experimental variation, both with 8 levels.

To reflect the fact that we expect many of the variables to have no influence on the response we drew a probability π from $\mathcal{U}(0.2, 0.4)$ for each dataset. With this probability, each of the variables in the dataset was then given a regressor simulated from $\mathcal{U}(-0.4, -0.2)$ and zero otherwise, remembering that we expect the variables to have a negative effect as any mutational changes will reduce the response, VN titre. Each response y_i was then generated with an intercept of 10 and with $\mathcal{N}(0, 0.02)$ iid additive Gaussian noise given to each response.

5.1.3 Final Simulation Study

The simulation study of Davies et al. (2016b) compared WAIC and 10-fold Bayesian CV by generating 20 datasets each with 500 observations and 50 possible variables. The data was generated with 10 viruses, with every virus used as both the challenge and protective strains and for any given pair of challenge and protective strains the variables remain identical as in the real FMDV and Influenza datasets. Possible random effects were the protective and challenge strains and 2 generic random effects with 8 levels. The random effects were given a variance of zero, i.e. set to be irrelevant, with probability 0.5.

5.1.4 Original SAT1 Data

The original SAT1 dataset was analysed by Reeve et al. (2010) and information about the dataset can be found in Section 2.2.1. To analyse the dataset we log transformed the VN titre measurements following Reeve et al. (2010). For the results given in Section 5.4.1 we used the challenge strain and antiserum as random effects. Variables related to the phylogenetic tree were added but only to reflect where the branch lay between the chosen challenge and protective strain, e.g. *branch* effects by the definitions in Section 2.1.3, rather than any of the more complex phylogenetic effects described in Section 2.1.3. Instead of classifying variables with correlation 1 in groups as discussed in Section 2.4, Davies et al. (2014) instead used a strategy based on prior knowledge to exclude the less biologically relevant variables with correlation 1.[1] This resulted in the original SAT1 dataset analysed in Section 5.4.1 only containing 107 variables in total; we call this the reduced SAT1 dataset.

Section 5.4.2 used 138 variables in total with only one of the variables that were completely correlated included, but with the classification being based on all of the completely correlated variables as specified in Section 2.4. Multiple types of branches were included to account for the phylogenetic tree as discussed in Section 2.1.3, rather than just the *branch* effects as in Section 5.4.1. The original SAT1 results of Section 5.3.2 use just challenge strain and antiserum as random effect groups based on the results of Reeve et al. (2010).

5.1.5 Extended SAT1 Data

The extended SAT1 dataset is an extended version of the original SAT1 dataset (Section 5.1.4) of Reeve et al. (2010) collected and analysed by Maree et al. (2015). We

[1] Davies et al. (2014) included all proven variables based on the classification in Section 2.4, then added the branches of the phylogentic tree and finally excluded any plausible or implausible variables which made the matrix singular; see Davies et al. (2014) for details.

have again log transformed the data following Maree et al. (2015) and have included multiple types of phylogentic effects; see Section 2.1.3. Random effects were included in Section 5.4.3 to account for the challenge strain, antiserum and date of the experiment based on the results of Maree et al. (2015); see Section 2.2.1.

5.1.6 SAT2 Data

The SAT2 dataset was originally analysed by Reeve et al. (2010) and is described in Section 5.1.6. The VN titre measurements were again log transformed and we have included multiple types of phylogentic effects; see Section 2.1.3. Random effects were included in Section 5.5 to account for the challenge strain and antiserum based on the results of Reeve et al. (2010); see Section 2.2.2.

5.1.7 H1N1 Data

Harvey et al. (2016) used a H1N1 dataset that contained 506 challenge strains and 43 protective strains. Here we have used a slightly smaller dataset in order to fully account for the effect of the phylogentic structure. The dataset used here contains 15,693 HI assay measurements with 43 challenge and 43 protective strains. As this full dataset is too large to analyse using the conjugate SABRE method we have summarised the data to just be 570 mean HI assay measurement for each combination of challenge and protective strains. For each pair of challenge and protective strains the 279 explanatory variables, 53 surface exposed residues and 226 variables related to the phylogenetic data, remain the same. Doing this however means we cannot use the date of the experiment as a random effect and additionally the dataset does not contain antiserum data, meaning we have only used the challenge strain as random effects in Section 5.6.

5.2 Computational Inference

Our code has been implemented in *R* (R Core Team, 2013), using the packages *lme4* (Bates et al., 2013) and *lmmlasso* (Schelldorfer et al., 2011) for the comparison with standard and LASSO mixed-effects models. For the mixed-effects models, as in Reeve et al. (2010), forward inclusion was used adjusting for multiple testing using the Holm-Bonferroni correction.

For the MCMC chains we sampled 10,000 iterations for the simulated datasets, with varying numbers of iterations for the real data as required to get convergence. This was determined by running 4 chains for each model and computing the PSRF (Gelman and Rubin, 1992) from the within-chain and between-chain variances (Plummer et al., 2006). We take a PSRF ≤ 1.05 as a threshold for convergence and terminate the burn-in when this is consistently satisfied for 95% of the variables. In general, the fixed hyper-parameters, shown as grey nodes in Figures 4.1, 4.2, 4.3 and 4.4, were set to give a vague distribution for the flexible (hyper-)parameters, shown as white nodes. The only exception was the prior on π, defined in (4.8), which was set to be weakly informative such that $\alpha_\pi = 1$ and $\beta_\pi = 4$, except in Section 5.3.1 where the parameters were set to be $\alpha_\pi = 1$ and $\beta_\pi = 1$. Setting the parameters to be weakly informative, $\alpha_\pi = 1$ and $\beta_\pi = 4$, corresponds to prior knowledge that only a small number of residues or branches have a significant antigenic effect.

The following hyper-parameters are fixed to give vague distributions: $\alpha_{b,g} = \beta_{b,g} = \alpha_{\eta,g} = \beta_{\eta,g} = 0.001$ and $\mu_{b,g} = \mu_{\eta,g} = 0$ for all g, $\alpha_{w,h} = \beta_{w,h} = 0.001$, $\mu_{0,h} = 0$ and $\sigma^2_{0,h} = 100$ for all h, $\mu_\xi = 0$, $\sigma^2_\xi = 100$, $\mu_{w0} = max(\mathbf{y})$, $\sigma^2_{w0} = 100$ and $\alpha_\varepsilon = \beta_\varepsilon = 0.001$. The only unusual choice is $\mu_{w0} = max(\mathbf{y})$ which follows from us expecting a high intercept with the regression coefficients then having a negative effect on the response. This is a result of strains having high reactivity with themselves, and any changes making the strains less similar, reducing their reactivity. The only exception to this is in the original SABRE method where intercept is treated as the only member of the first group of fixed-effects. Here we set $\alpha_{w,1} = 1.501$ to give a finite mean and variance for the prior distribution of $\sigma^2_{w,1}$. Although this is not a vague prior, we have tested a number of other values and found that this specification has little effect on the results.

To analyse the best proposal method we tested the component-wise Gibbs sampler and several specifications of the Metropolis-Hastings sampler on the several datasets (Section 5.4.5). For the reduced SAT1 dataset used by Davies et al. (2014) (Section 5.1.4) we tested the component-wise Gibbs sampler and proposed the inclusion or exclusion of variables in groups of 4, 8, 16, 32 and 64 with the block Metropolis-Hastings sampler. We analysed convergence by monitoring the percentage of variables with a PSRF ≤ 1.1 as in Grzegorczyk and Husmeier (2013) (Davies et al., 2014). For the full SAT1, extended SAT1 and H1N1 dataset we again used the component-wise Gibbs sampler but proposed the inclusion or exclusion of variables in groups of 5, 10, 15, 20 and 30 with the block Metropolis-Hastings sampler. We analysed convergence by monitoring the percentage of variables with a PSRF ≤ 1.05, similar to Grzegorczyk and Husmeier (2013) (Davies et al., 2016a).

For selecting variables in the mixed-effects LASSO and elastic net we used BIC as in Schelldorfer et al. (2011). For the SABRE methods there are a variety of techniques that have been used in the literature to choose a cut-off. Often a cut-off of 0.5 is used and this has been shown to be the best predictive model under strict conditions (Barbieri and Berger, 2004). Alternatively the top $J\hat{\pi}$ ranked variables have been taken, where J is the number of variables and $\hat{\pi}$ is the posterior mean of π, defined in (4.7) and (4.8), i.e. the global probability of variables being included in the model.

5.3 Results for the Simulation Studies

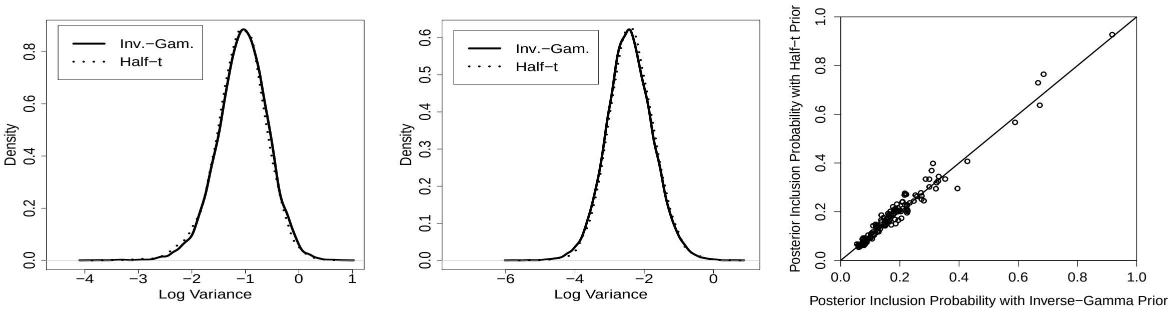
To summarise, we have introduced a hierarchical Bayesian modelling framework (called SABRE) for selecting relevant antigenic sites in viral evolution. There are two fundamentally different approaches to variable selection: the slab and spike prior, whereby the influence of an input variable is controlled via the prior distribution of its associated regression parameters, and the binary mask model, where variables are put through a binary multiplicative filter. There are also different prior distributions one can choose: a conjugate prior, and a semi-conjugate prior. This gives us four variants of the proposed modelling framework, including the original SABRE method which does not include an intercept parameter:

- The original SABRE method, with slab and spike prior
- The conjugate SABRE method, with slab and spike prior
- The semi-conjugate SABRE method, with slab and spike prior
- The binary mask SABRE method.

These four variants are depicted as probabilistic graphical models in Figures 4.1, 4.3, 4.2 and 4.4. We have compared their performance with that of two established methods from the literature: the mixed-effects model with stepwise variable selection, and the mixed-effects LASSO. Since there are indications from the literature that the elastic net offers an improvement over the LASSO, we have also modified the mixed-effects LASSO model from the literature (Schelldorfer et al., 2011) by a novel mixed-effects elastic net model. This gives us three classical methods for comparison:

- Mixed-effects model with stepwise variable selection
- Mixed-effects LASSO model
- Mixed-effects elastic net model.

We have applied and assessed the proposed methods with a three-pronged approach. Firstly, we have tested them on a large set of synthetic benchmark data, where the true structure of the model is known, and it is therefore straightforward to quantify the accuracy of inference. This is discussed here in Section 5.3 and contains results Davies



(a) Challenge (b) Antiserum (c) Inclusion Probabilities

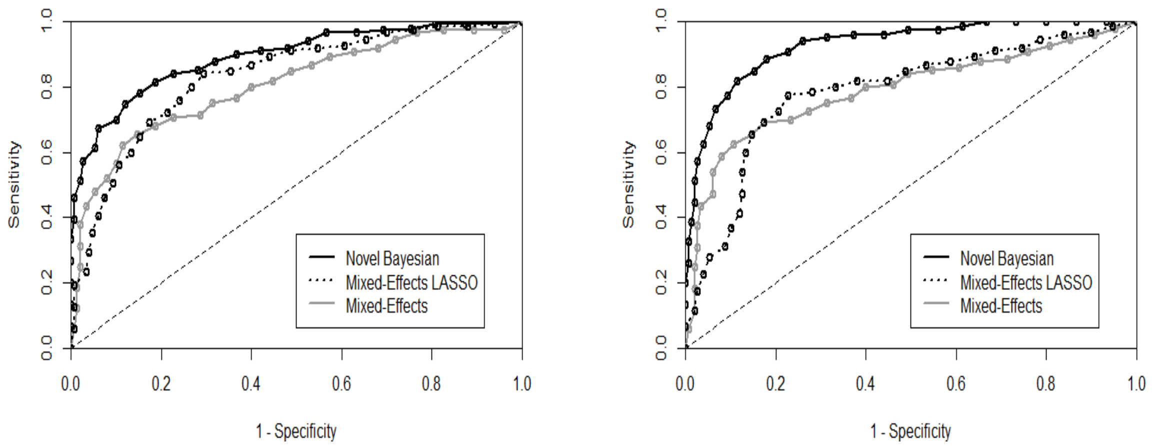
Figure 5.1: **Gaussian Kernel density estimation plots of random effects variances and a comparison of posterior inclusion probabilities.** Gaussian kernel density estimation plots are shown for the sampled posterior densities of the log random effect variance. This is given for the two groups of random effects, (a) challenge strain and (b) antiserum, under a vague Inverse-Gamma prior (solid) and the half-t prior (dotted) proposed in Gelman (2006). (c) Plot showing the comparative posterior inclusion probability for each variable for the two models.

et al. (2014), Davies et al. (2016a) and Davies et al. (2016b) in Sections 5.3.1, 5.3.2 and 5.3.3 respectively. Secondly, we have applied the methods to real data for which partial biological prior knowledge is known, which can be used to partially assess the model predictions. These findings are presented in Section 5.4. Finally, in Sections 5.5 and 5.6, we present novel applications to new data, from the less well known FMDV serotype, SAT2, and as well as from seriously reduced version of the H1N1 Influenza dataset where it is not relevant to compare our results against those obtained from a larger dataset. Here the purpose of our study is new hypothesis generation.

As part of the extended simulation study of Davies et al. (2016a) given in Section 5.3.2 we also tested the choice of random effects prior, comparing the Inverse-Gamma prior (Section 4.1.4) with the half-t prior prior proposed in Gelman (2006) (Section 4.2.4). Figures 5.1a and 5.1b show posterior samples of the log variance of the two random-effects groups from the conjugate SABRE method applied to the SAT2 dataset (Section 5.1.6) comparing the half-t and Inverse-Gamma priors, and shows no notable differences. Similarly Figure 5.1c shows that the inclusion probabilities for the two competing models are approximately the same. Based on these findings, we only report the results obtained with the conjugate Inverse-Gamma prior throughout this section.

5.3.1 Initial Simulation Study

Figure 5.2 shows ROC curves (Section 3.4.2) for the classical mixed-effects models, the mixed-effects LASSO and the original SABRE method. For two random effects groups



(a) One Random-Effect Group

(b) Two Random-Effect Groups

Figure 5.2: **ROC Curves for the Initial Simulation Study data described in Section 5.1.1.** ROC curves are given for the original SABRE method (black), the mixed-effects LASSO (black dotted) and classical mixed-effects (grey) (Davies et al., 2014). The original SABRE method is given in the figure as the 'Novel Bayesian' method. The simulated data was generated with (a) one and (b) two random effect groups; see section 5.1.1.

(Figure 5.2b), the original SABRE method, AUROC = 0.93, consistently outperforms the mixed-effects LASSO, AUROC = 0.79, and standard mixed-effects model, AUROC = 0.79. This is presumably a consequence of the fact that the mixed-effects LASSO of Schelldorfer et al. (2011), is defined for a single random effect. To deal with two random effects, we need to map the matrix of random effect combinations into a vector of substitute single random effects, which may render the model over-complex and hence susceptible to over-fitting. When the analysis of the simulated data in Davies et al. (2014) was carried out, a mixed-effects LASSO with the ability to handle multiple random effects did not exist. For data with a single random effect (Figure 5.2a), the original SABRE method still achieves a greater AUROC value, 0.89, than the LASSO, 0.83, and standard mixed effects model, 0.81.

In addition to the comparison of AUROC values, we also looked at the predictive performance. For the data with 2 groups of random effects the original SABRE method got a mean out-of-sample log-likelihood of -113.8, outperforming the mixed-effect LASSO of Schelldorfer et al. (2011) with BIC, -160.8, and AICc, -163.3, and the standard mixed-effect model, -127.7. Similar results were also achieved for the data with 1 random effect group, with the models achieving a mean out-of-sample log-likelihoods of -99.9, -104.2, -105.9 and -112.4, respectively.

5.3.2 Extended Simulation Study

Table 5.1 compares the different methods in terms of variable selection, WAIC score (Watanabe, 2010), predictive performance and fixed effects coefficients inference using the simulated datasets described in Section 5.1.2. To measure variable selection we have ranked the covariates in terms of their significance or influence. For the Bayesian methods, the ranking is defined by the marginal posterior probabilities of inclusion. For the alternative methods, we explain the way the ranking is obtained below. Since for the simulated data the true covariates are known, this ranking can be used to produce a ROC curve (e.g. Hanley and McNeil (1982); Section 5.7. of Murphy (2012)), where for all possible values of the inclusion threshold, the sensitivity or recall (the relative proportion of true positive covariates: TP/(TP+FN)) is plotted against the complementary specificity (the relative proportion of false positive covariates: FP/(FP+TN))[2]. By numerical integration we obtain the AUROC value as a global measure of accuracy, where larger values indicate a better performance, starting from AUROC = 0.5 to indicate random expectation, to AUROC = 1 for perfect variable identification; see Section 3.4.2.

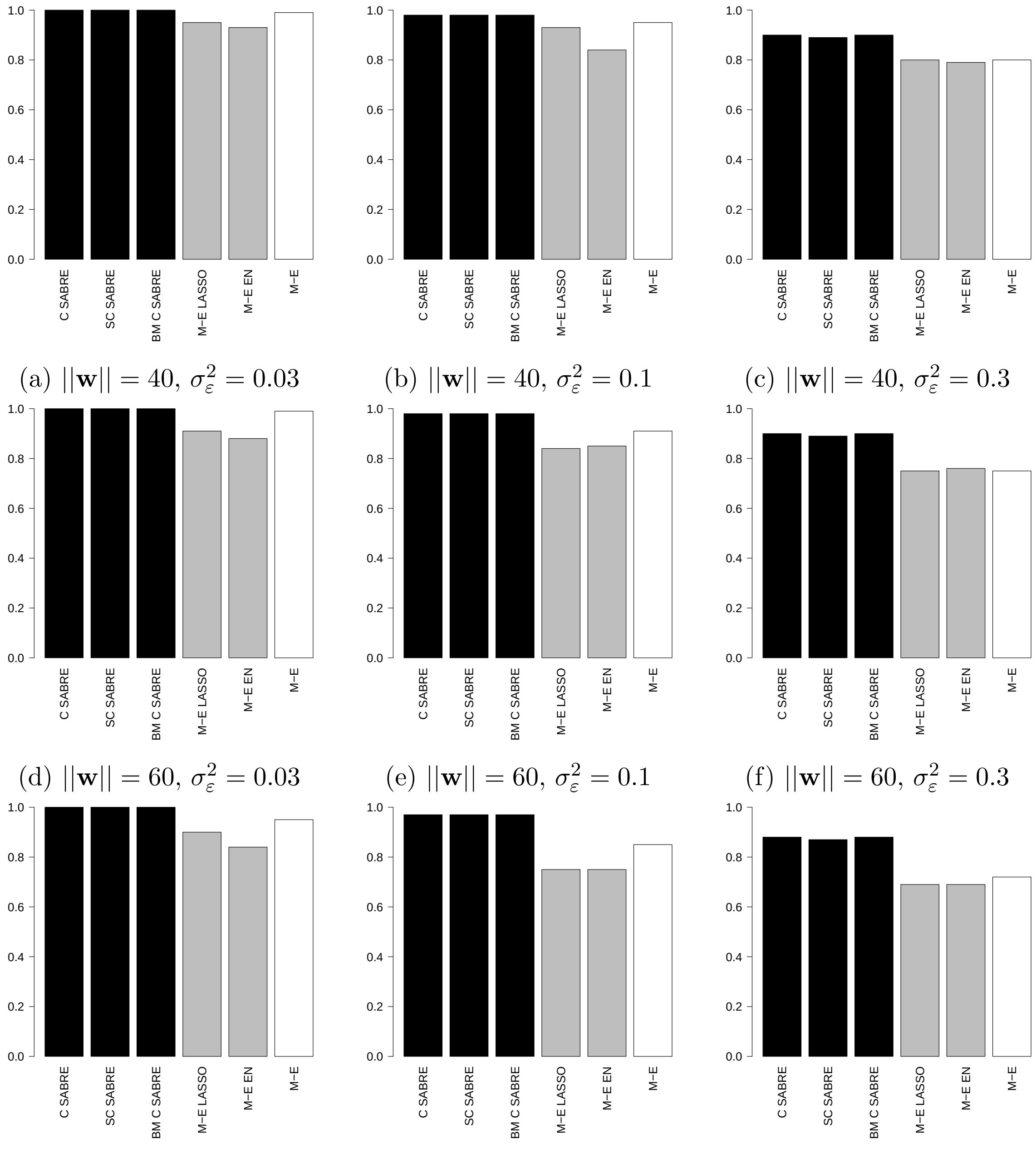
In addition to ranking the covariates to get ROC curves for the SABRE methods, we also need to rank the alternative established methods for a comparison. For the classical mixed-effects models this is done by removing the significance threshold and ranking the edges by order of inclusion. For the mixed-effects LASSO and elastic net we predicted models for a variety of different penalty parameters, λ, to create the so called LASSO path and create a ranking based on when variables become 0. For the mixed-effects elastic net we only show the results for $\alpha = 0.3$ following Ruyssinck et al. (2014), however the remaining results are available in Section B.1. Alternative AUROC values based on using model selection and then ranking the variables based on the absolute values of the regression coefficients (Aderhold et al., 2014), as well as other results, are also available in Section B.1.

Table 5.1 also measures the accuracy of predicting out of sample observations, $\mathbf{y_{out}}$, and the fixed effects coefficients, $\mathbf{w}$ in terms of MSEs. For the Bayesian methods, the predictions are made by sampling from the model and then choosing which variables are included based on taking the top $J \times \hat{\pi}$ variables with the highest inclusion probabilities. The model is then sampled with just those variables set to be included and the estimates calculated. For the mixed-effects LASSO, mixed-effects elastic net and classical mixed effects models the regression coefficients can be taken from the chosen model. The random effects coefficients can then be calculated using the best linear unbiased estimator and predictions of the out of sample observations, $\mathbf{y_{out}}$, made.

[2] TP: true positive count, FP: false positive count, TN: true negative count, FN: false negative count

Table 5.1: **Table of Simulation Study Results for the data described in Section 5.1.2.** The table gives results for the Conjugate, Semi-Conjugate and Binary Mask (BM) Conjugate SABRE methods, the mixed-effects LASSO, the mixed-effects (M-E) elastic net with $\alpha = 0.3$ and the classical mixed-effects models applied to the simulated data described in Section 5.1.2. The table gives the mean AUROC value based on ordering the variables, the MSEs of the out-of-sample observations, $\mathbf{y_{out}}$, the MSEs of the fixed effects coefficients, $\mathbf{w}$, and the mean WAIC scores for each method. An extended version of these results is given in Tables B.1-B.6.

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
AUROC	Conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88
	Semi-Conjugate SABRE	1	0.98	0.89	1	0.98	0.89	1	0.97	0.87
	BM Conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88
	Mixed-Effects LASSO	0.95	0.93	0.80	0.91	0.84	0.74	0.90	0.75	0.69
	M-E Elastic Net ($\alpha = 0.3$)	0.93	0.84	0.79	0.88	0.85	0.76	0.84	0.75	0.69
	Mixed-Effects Models	0.99	0.95	0.80	0.99	0.91	0.75	0.95	0.85	0.72
MSE($\mathbf{y_{out}}$)	Conjugate SABRE	0.15	0.22	0.49	0.18	0.30	0.57	0.26	0.36	0.63
	Semi-Conjugate SABRE	0.16	0.23	0.48	0.18	0.29	0.57	0.24	0.35	0.63
	BM Conjugate SABRE	0.16	0.22	0.49	0.18	0.29	0.56	0.24	0.36	0.62
	Mixed-Effects LASSO	0.06	0.22	0.59	0.13	0.40	0.75	0.31	0.56	1.37
	M-E Elastic Net ($\alpha = 0.3$)	0.06	0.18	0.60	0.11	0.34	0.75	0.31	0.65	1.81
	Mixed-Effects Models	0.08	0.23	0.53	0.16	0.37	0.68	0.32	0.50	0.77
MSE($\mathbf{w}$)	Conjugate SABRE	0.019	0.019	0.025	0.017	0.021	0.024	0.021	0.022	0.024
	Semi-Conjugate SABRE	0.021	0.022	0.022	0.017	0.020	0.025	0.019	0.020	0.025
	BM Conjugate SABRE	0.020	0.018	0.022	0.016	0.019	0.023	0.019	0.022	0.025
	Mixed-Effects LASSO	0.003	0.017	0.046	0.009	0.034	0.060	0.020	0.024	0.071
	M-E Elastic Net ($\alpha = 0.3$)	0.004	0.010	0.045	0.007	0.022	0.052	0.020	0.038	0.112
	Mixed-Effects Models	0.008	0.020	0.032	0.015	0.031	0.041	0.033	0.040	0.044
WAIC	Conjugate SABRE	-309.7	-173.2	-100.4	-314.0	-172.2	-100.8	-309.8	-172.8	-103.1
	Semi-Conjugate SABRE	-308.7	-170.5	-96.8	-312.1	-171.2	-98.5	-310.5	-171.4	-101.3
	BM Conjugate SABRE	-309.7	-173.5	-98.7	-313.9	-171.9	-101.3	-310.4	-172.0	-103.3

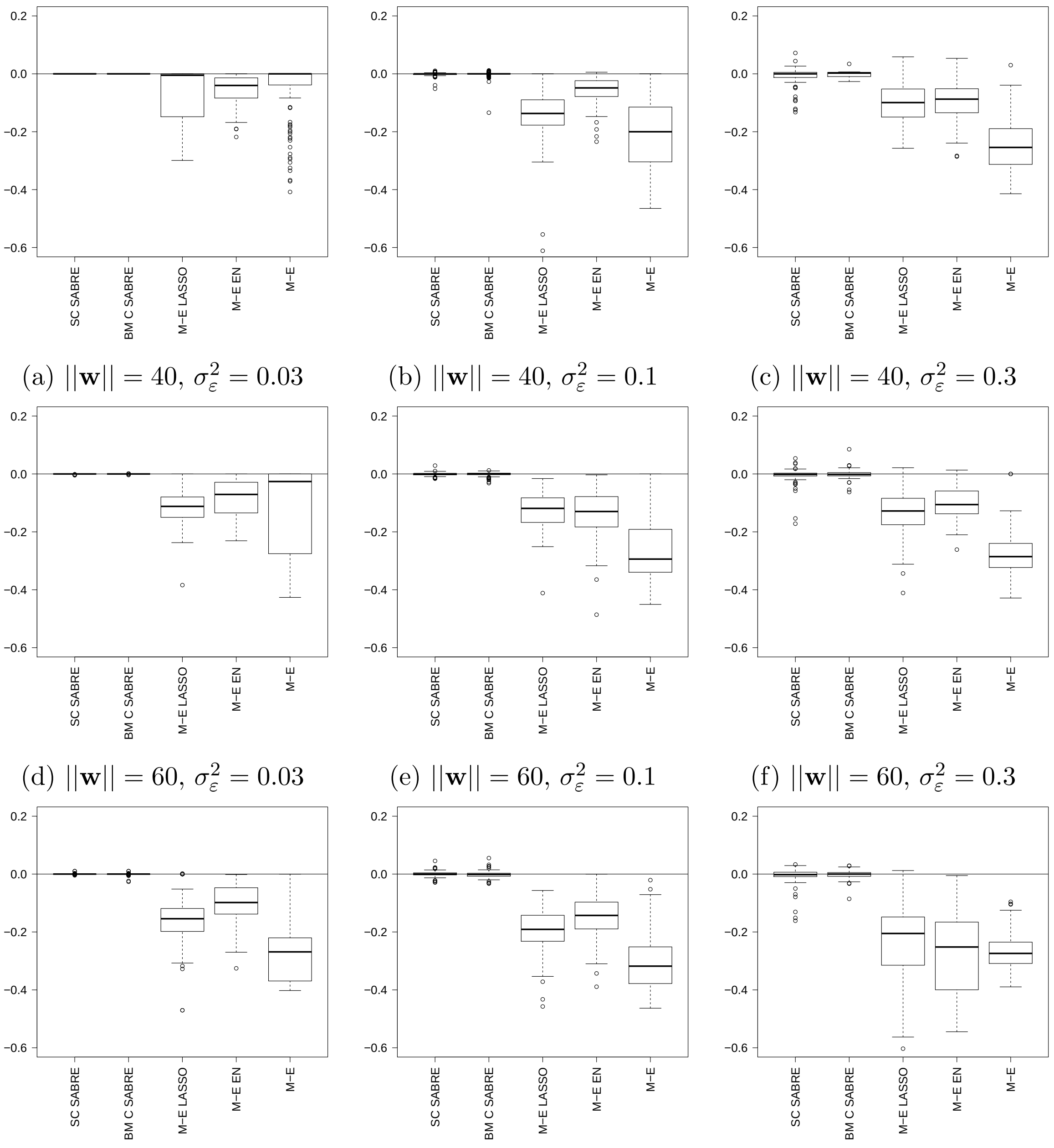


(a) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.03$ (b) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.1$ (c) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.3$

(d) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.03$ (e) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.1$ (f) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.3$

(g) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.03$ (h) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.1$ (i) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.3$

Figure 5.3: **Bar plot of AUROC values from the Simulation Study Results in Table 5.1.** The bar plots gives AUROC values for the Conjugate (C), Semi-Conjugate (SC) and Binary Mask Conjugate (BM C) SABRE methods (black bars), the mixed-effects (M-E) LASSO, the mixed-effects elastic net (M-E EN) with $\alpha = 0.3$ (both grey bars) and standard mixed-effects (M-E) models (white bars) applied to the simulated data described in Section 5.1.2.



(a) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.03$ (b) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.1$ (c) $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.3$

(d) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.03$ (e) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.1$ (f) $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.3$

(g) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.03$ (h) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.1$ (i) $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.3$

Figure 5.4: **Box plots of the difference in AUROC values for each method in comparison to the conjugate SABRE method.** The box plots give the difference in AUROC values for each of the methods after the AUROC value of the conjugate SABRE method has been subtracted for the appropriate dataset. Negative values indicate that the conjugate method has outperformed the alternative method. Each box plot contains 100 datasets as described in Section 5.1.2. The alternative methods are the Semi-Conjugate (SC) and Binary Mask Conjugate (BM C) SABRE methods, the mixed-effects (M-E) LASSO, the mixed-effects elastic net (M-E EN) with $\alpha = 0.3$ and the classical mixed-effects models (M-E).

In terms of variable selection, the AUROC values shown in Figure 5.3 and Table 5.1 show that all the SABRE methods outperform the alternative methods; the mixed-effects LASSO, the mixed effects elastic net and the classical mixed effects models. This is achieved across all datasets and is highlighted in Figure 5.4, which compares the difference in AUROC values obtained by the different methods and that of the conjugate SABRE method.A negative score signifies a reduction in performance compared to the conjugate SABRE method. Figure 5.4 shows that the conjugate SABRE method performs significantly better than the mixed-effects LASSO, the mixed-effects elastic net and the classical mixed-effects models in all sets of data.

The performance in terms of predicting out of sample observations and inferring fixed effects coefficients shown in Table 5.1 again shows the SABRE methods outperforming the alternative methods in most cases. Table 5.1 shows a huge improvement for the SABRE methods in all cases except where both the error variance and number of variables is small. This is especially the case with the mixed-effects LASSO and the mixed-effects elastic net where the reliance on ℓ_1 regularisation causes a bias which affects both the inference of the fixed effects coefficients and the variable selection, as well as subsequently the out of sample predictions. The alternative methods do outperform the SABRE methods in some sets of data where the number of variables is small and the error variance is low, but this is only in 2 out of 9 sets of data. The reason for these counter intuitive results is the model selection technique used with the SABRE methods, as in both of the sets of data where the improvement is shown the SABRE methods achieve mean AUROC values of 1, better than the alternative methods.[3]

We have also explored multiple different versions of the SABRE method, namely the semi-conjugate (Figure 4.2), conjugate (Figure 4.3) and binary mask conjugate (Figure 4.4) SABRE methods[4]. As far as we are aware the quantitative comparison between a spike and slab based method and a binary mask based one is the first of its kind. Our results given in Table 5.1, as well as Figures 5.3 and 5.4, show a strong similarity in performance between the methods. The comparison of AUROC values given in Figure 5.4 clearly shows a large overlap in both method's variable selection performance and this is backed up by the paired t-tests given in Tables B.4-B.6. Identifying that these methods give similar results is important, as in practise both methods are discussed and used throughout the literature, e.g Jow et al. (2014); Murphy (2012).

We have also compared the conjugate and semi-conjugate SABRE models, as depicted

[3] By choosing the $J \times \hat{\pi}$ variables with the highest marginal probability of inclusion, we have chosen the wrong number of variables resulting in a mismatch between the inferred fixed-effects coefficients and their true values.

[4] We do not have a comparison with the original SABRE method, as it does correctly specify the biologically significant intercept parameter.

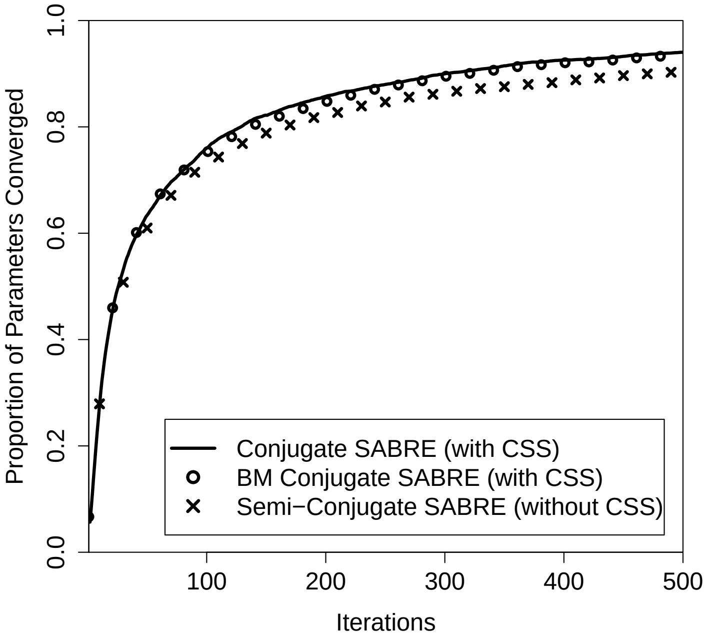


Figure 5.5: **Convergence diagnostics comparing the sampling performance of different versions of the SABRE method.** Convergence diagnostics for the conjugate SABRE method with the collapsed sampling scheme (CSS) (solid line), the semi-conjugate SABRE method without CSS (crosses) and the BM conjugate SABRE method with CSS (circles). The lines show the proportion of parameters converged (PSRF$<$ 1.05) versus the number of iteration of the 4 MCMC chains. The proportion is based on all of the simulated datasets from Section 5.1.2.

in Figure 4.3 and 4.2. Overall, our results, shown in Table 5.1and Tables B.1-B.6, suggest that the two methods perform similarly across the wide range of simulated data sets. A paired t-test, summarised in Tables B.4-B.6, identifies two data sets ($||\mathbf{w}|| = 40$, $\sigma_{\varepsilon}^2 = 0.3$; $||\mathbf{w}|| = 60$, $\sigma_{\varepsilon}^2 = 0.3$) where the conjugate SABRE model outperforms the semi-conjugate SABRE model. Formal model selection based on WAIC also shows a slight, but significant preference for the conjugate model (see Table B.6).

The final contribution of our simulation study is to test whether the use of the collapsed sampling scheme in conjunction with increased conjugacy achieves an improvement in terms of MCMC mixing and convergence. Figure 5.5 indicates that a slight improvement is achieved with the conjugate SABRE model over the semi-conjugate one. However, this difference is not statistically significant, as becomes clear when considering the confidence intervals (not shown in Figure 5.5 to avoid clutter). This finding suggests that the major bottleneck in the MCMC sampling scheme is caused by the latent variables $\boldsymbol{\gamma}$ rather than the regression parameters.

Table 5.2: **Results comparing the model selection performance of WAIC compared to 10-fold Bayesian CV on the simulated datasets described in Section 5.1.3.** The mean and 95% confidence intervals are given in terms of correctly including or excluding random effect components in the simulated datasets described in Section 5.1.3.

	10-fold Bayesian CV	WAIC
Sensitivity	0.91 (0.85,0.97)	0.78 (0.69,0.87)
Specificity	0.63 (0.52,0.73)	0.77 (0.68,0.86)
Predictive Accuracy	0.79 (0.70,0.88)	0.78 (0.68,0.87)
F1-Score	0.83 (0.75,0.91)	0.80 (0.71,0.88)

5.3.3 Final Simulation Study

To analyse the performance of WAIC in comparison to 10-fold Bayesian CV, Davies et al. (2016b) looked at how accurate each method was at correctly selecting the random effect components used to generate the datasets simulated in Section 5.1.3. Both methods were applied to each of the 16 possible models for each dataset and selected the best model in each case. The ability of the best models to correctly include or exclude the random effect components that were used or not used to generate each of the datasets was then analysed, where Table 5.2 gives the results in terms of sensitivity, specificity, predictive accuracies and F-scores; see Section 3.4.1.

The results of Table 5.2 show that WAIC performs similarly to 10-fold Bayesian CV in terms of correctly selecting random effect components. While 10-fold Bayesian CV gets an increased sensitivity, WAIC has a better specificity and both perform similarly in their predictive accuracy and F1-score. However WAIC is much more computationally effective and to run the MCMC simulations for the WAIC took on average 87 minutes, as opposed to 761 minutes for 10-fold Bayesian CV.

Using a spike and slab prior to include or exclude all random effect coefficients, $\mathbf{b}_g$, from a particular random effect component, g, is an alternative to both WAIC and 10-fold Bayesian CV. While WAIC and 10-fold Bayesian CV would be applied to each combination of random effect components separately, spike and slab priors would only require one model to be fitted. However, using spike and slab priors for selecting the random effects will come at a large computational cost. Some of the random effect components from the FMDV datasets contain between 30 to 50 different levels and this would mean including or excluding 30 to 50 parameters simultaneously at each proposal step of the MCMC sampling scheme. This is likely to lead to poor mixing as the difference in log-likelihood for the inclusion and exclusion of a random effect component is likely to be large. Poor mixing leads to the possibility of not sampling the optimal combination

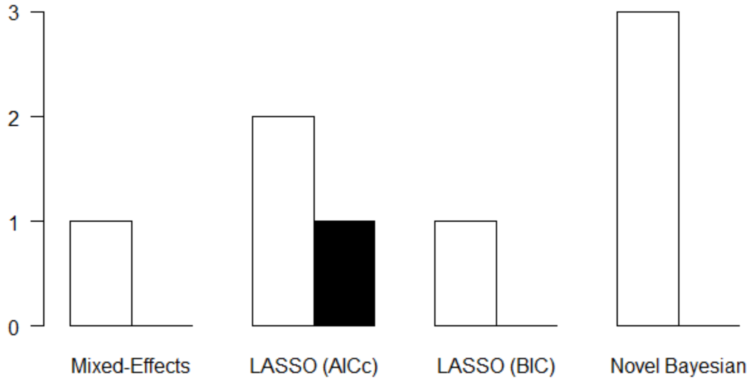
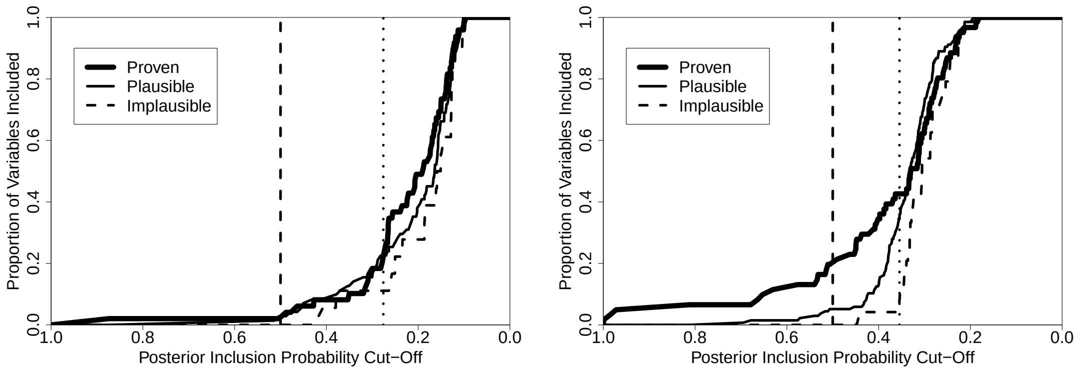


Figure 5.6: **Bar plot showing the results for the reduced SAT1 dataset in Davies et al. (2014).** The bar plot shows proven residues (white) and implausible residues (black) for the mixed-effects model results of Reeve et al. (2010), the mixed-effects LASSO using AICc and BIC (Schelldorfer et al., 2011) and the original SABRE method (given here 'novel Bayesian').

of fixed and random effects, as the proposals will struggle to move between different combinations of random effect components. Therefore in order to ensure the optimal selection of fixed and random effects is found it would be necessary to sample the model for a large number of iterations. Due to the computational inefficiency of this inter-model approach, we have used an intra-model approach and run MCMC simulations for a relatively small number of models in parallel to compute WAIC and 10-fold Bayesian CV scores for each plausible candidate model separately.

5.4 Results for the SAT1 Datasets

Both SAT1 datasets have been analysed using classical mixed-effects models. Originally Reeve et al. (2010) analysed the original SAT1 dataset (Section 5.1.4) and Maree et al. (2015) investigated an extended version of this dataset (Section 5.1.5). We have used our method on each of these datasets in order to identify a number of candidate residues which could be considered important for understanding antigenic variability. Knowledge of which residues are antigenically important is partially incomplete. Therefore, for validation purposes, residues were assigned to three different groups, proven, plausible and implausible, based on how likely they are to be antigenic based on experimental results; see Section 2.4.



(a) Original SAT1

(b) Extended SAT1

Figure 5.7: **Proportion of categorised SAT1 variables included based on different cut-off values for posterior inclusion probability.** The graph shows the proportion of the experimentally proven (thick solid line), plausible (solid line) and implausible (dashed line) variables based on a cut-off value for the posterior inclusion probability. The variables were classified into groups based on the method outlined in Section 2.4. Cut-offs are marked at 0.5 posterior inclusion probability (vertical dashed line) and the posterior inclusion probability equivalent to the top $J\hat{\pi}$ variables with the highest posterior inclusion probabilities (vertical dotted line).

5.4.1 Reduced SAT1 Dataset

The reduced SAT1 dataset described in Section 5.1.4 was analysed in Davies et al. (2014) with the original SABRE method. With respect to the evaluation of the prediction, we need to point out that the original SABRE method is the only one that could be applied in a fully automatic manner. The forward-variable selection technique used in Reeve et al. (2010) drew on biological prior knowledge to design an effective variable selection schedule, and the optimisation algorithm for the mixed-effects LASSO, as implemented in the software of Schelldorfer et al. (2011), failed due to ill-conditioned (i.e. quasi-singular) matrices.

To cope with the latter problem, we applied the mixed-effects LASSO as follows: in the first instance, we included all proven residues (as informed by Section 2.4.1). We then included any branches of the phylogenetic tree that did not prevent the matrix inversion as explanatory variables. The plausible and implausible residues were then added, before being iteratively excluded until the matrix inversion no longer ran into numerical problems. We need to point out that this strategy uses prior knowledge that would usually not be available and is not required for the proposed Bayesian method. However for a fair comparison, we used this reduced set of 107 variable for all methods.

For performance evaluation, we have concentrated on the prediction of the relevant residues, which indicate areas of the virus protein that are targeted by the immune system,

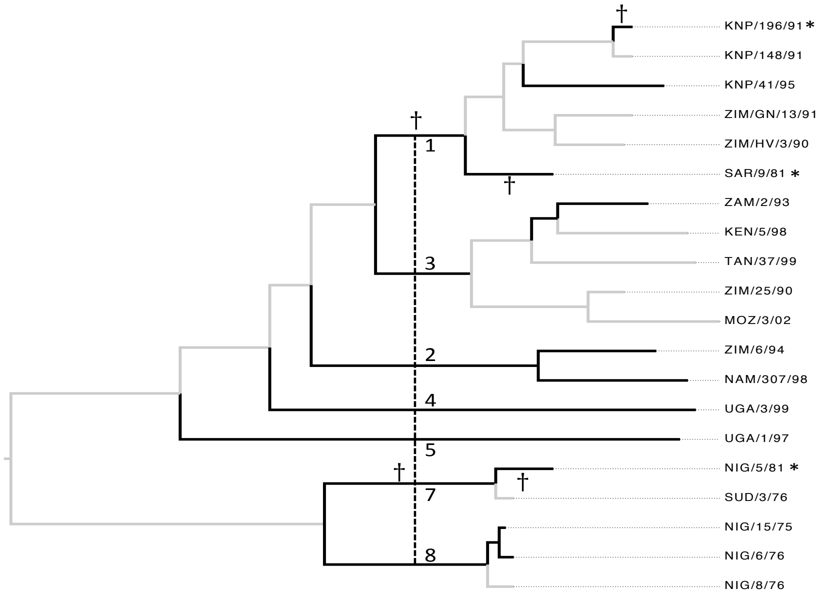


Figure 5.8: **Phylogenetic tree indicating significant branches in the evolutionary history of the SAT1 serotype based on the original SAT1 dataset in Section 5.1.4.** Phylogenetic trees were created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*) and topotype defining branches (dashed vertical line). Branches inferred by the conjugate SABRE method are highlighted (black). Symbols indicate whether this was inferred to be a change in virus antigenicity (†), virus reactivity (‡) or virus immunogenicity (§). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance was taken to be the $J\hat{\pi}$ variables with the highest marginal inclusion probability, where the branches chosen are given in Table B.8.

where mutations potentially allow the virus to escape the host immune response. For evaluation we used the classification scheme described in Section 2.4.1. The predictions are shown in Figure 5.6. It can be seen that the original SABRE method finds no implausible variables, while also showing an increased number of proven variables.

5.4.2 Original SAT1 Dataset

The analysis of the original SAT1 dataset with the conjugate SABRE method in Davies et al. (2016a) has resulted in the identification of 29 residues or branches of importance based on taking the top $J\hat{\pi}$ variables with the highest marginal posterior inclusion probabilities. 9 of the selected residues and 2 of the branches are classified as proven, at the expense of only 1 implausible variable. A full list of selected variables can be found in Table B.7. The proportion of the differently classified variables at different cut-off points is shown in Figure 5.7a. The proven residues include several that have been validated using MAbs in the SAT1 serotype (Grazioli et al., 2006), as well as others from the VP2

B-C loop, VP1 G-H loop and VP1 C-terminus (the end of the VP1 protein) and we have focused on these proven residues in our analysis. The classifications of the variables are taken from Section 2.4.

The residues that have been experimentally validated in the SAT1 serotype are VP3 71 and VP3 77 in the VP3 B-C loop and VP1 144 and VP1 149 in the VP1 G-H loop (Grazioli et al., 2006). Additionally in the VP1 G-H loop, an antigenic loop in every FMDV serotype (Bolwell et al., 1989; Crowther et al., 1993b; Grazioli et al., 2013, 2006; Kitson et al., 1990; Lea et al., 1994) known to distract the host immune systems, the conjugate SABRE method has also identified VP1 143 and VP1 150. These residues are next to the experimentally validated residues in the protein alignment and confirm that the VP1 G-H loop is a highly antigenic part of the SAT1 serotype.

In addition to the residues in the VP3 B-C and VP1 G-H loops, the conjugate SABRE method has additionally selected VP2 74 in the VP2 B-C loop, as well as VP1 216 and VP1 219 in the VP1 C-terminus. The VP2 B-C loop is antigenic in all serotypes and contains the highly antigenic VP2 72 residue, which has been experimentally validated in all of the FMDV serotypes except SAT2 (Aktas and Samuel, 2000; Crowther et al., 1993a; Grazioli et al., 2013, 2006; Kitson et al., 1990; Lea et al., 1994; Saiz et al., 1991). The VP1 C-terminus has been proven to be antigenic in all but the Asia1 serotype, although it is almost certainly antigenic there also (Aktas and Samuel, 2000; Baxt et al., 1989; Grazioli et al., 2006; Mateu, 1995).

Figure 5.8 shows the model predictions for the antigenically significant branches based on using just the branch variables from the original SAT1 dataset. Here we have identified all of the branches known to divide topotypes (Reeve et al., 2010), as well as a number of other branches. Several of the branches, including two topotype defining branches, have been specifically identified as reactivity, immunogenic or antigenic changes, an improvement over previously used models.

5.4.3 Extended SAT1 Dataset

The analysis of the extended SAT1 dataset (Section 5.1.5) with the conjugate SABRE method in Davies et al. (2016a) resulted in selecting 76 variables, which included 24 proven residues, 4 important branches in the evolutionary history and only 2 implausible residues. A full list of the selected variables can again be found in Table B.9 and the proportion of proven, plausible and implausible residues selected at different cut-offs is shown in Figure 5.7b here. The improved results over Section 5.4.2 show the advantage of getting a larger dataset through testing an increased number of strains under a variety of different experimental conditions.

The conjugate SABRE method has identified 11 residues in the highly variable VP1

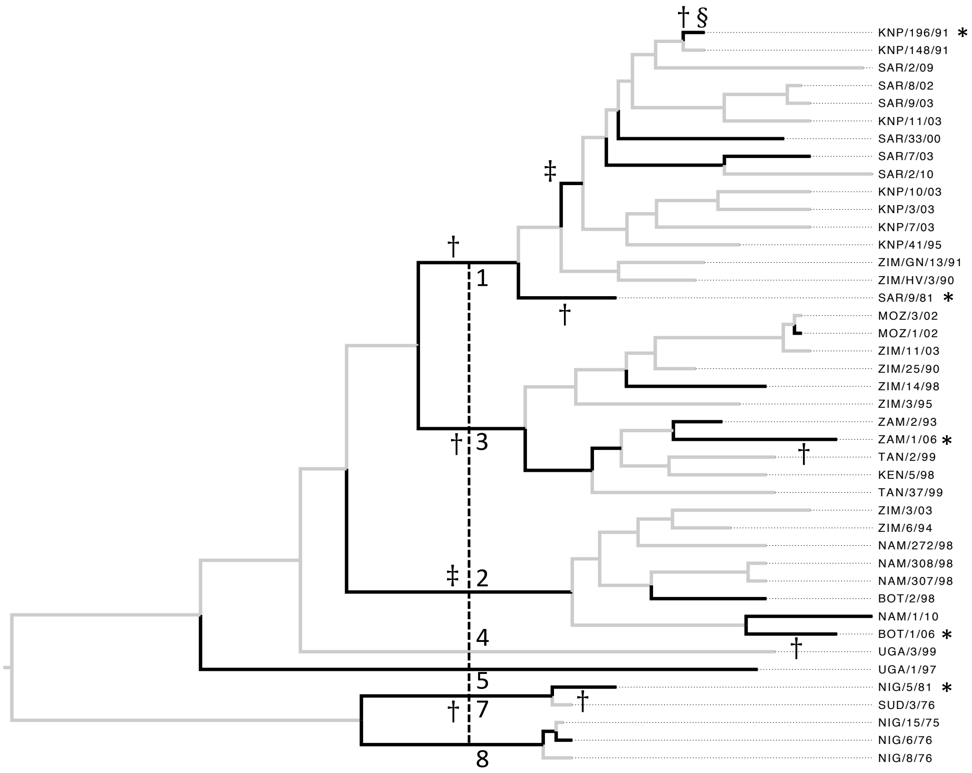


Figure 5.9: **Phylogenetic trees indicating significant branches in the evolutionary history of the SAT1 serotype.** Phylogenetic trees were created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*) and topotype defining branches (dashed vertical line). Branches inferred by the conjugate SABRE method are highlighted (black). Symbols indicate whether this was inferred to be a change in virus antigenicity (†), virus reactivity (‡) or virus immunogenicity (§). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance was taken to be 0.5 highest marginal inclusion probability, where the branches chosen are given in Table B.10.

G-H loop (VP1 142, VP1 143, VP1 144, VP1 147, VP1 148, VP1 149, VP1 150, VP1 155, VP1 156, VP1 163 and VP1 164). Finding this many significant residues in this highly antigenic region while keeping the number of implausible residues low shows that the model is working effectively.

Additionally, like with the original SAT1 dataset in Section 5.4.2, the conjugate SABRE method has selected VP2 74 from the VP2 B-C loop. However in addition it has also selected VP2 72 which is antigenic in all FMDV serotypes and VP2 79 which has been experimentally validated in the A, O, Asia1 and SAT2 serotypes (Grazioli et al., 2013, 2006; Mateu, 1995). The conjugate SABRE model also again selects several residues from the VP1 C-terminus; VP1 209, VP1 211 and VP1 218.

The final proven residues are from the VP3 B-B knob or have been experimentally validated specifically in the SAT1 serotype (Grazioli et al., 2006). In the VP3 B-B knob the conjugate SABRE method has identified VP3 58 (serotypes A, O, C and Asia1) and VP3 61 (serotype A) (Grazioli et al., 2006; Lea et al., 1994; Mateu, 1995). From those residues which have specifically been validated in the SAT1 serotype, again VP3 71 and VP3 77 from the VP3 B-C loop have been selected. However for the extended SAT1 dataset, the conjugate SABRE method has also selected VP3 138, which was also found in Reeve et al. (2010), from VP3 E-F loop.

As well as finding some branches in our overall model (including 4 topotype defining branches identified as representing significant evolutionary changes *a priori*), we have also compiled a model based only on branches to help us understand the evolutionary history of the serotype. The results of this model are given in Figure 5.9, where the seven branches known to define topotypes are indicated by the vertical line. In order to produce more interpretable results, where larger groups of strains are not separated by a significant evolutionary change (selected branch), we have used a cut-off of 0.5. The full results using a $J\hat{\pi}$ cut-off are given in Figure B.1. The results given in Figure 5.9 show that we have been able to identify all but one of the topotype defining branches, while the other is found when the $J\hat{\pi}$ cut-off is used. We have also been able to specify whether the evolutionary changes have affected virus antigenicity, immunogenicity or reactivity, helping us to further understand the underlying biological processes.

5.4.4 Comparison with Previous Work

To compare the results of the SABRE method against the mixed-effects models used in Reeve et al. (2010) and Maree et al. (2015), we examine which categories (proven, plausible or implausible) the various residues selected fall into. Note that to do this we ignore any branch terms that do not directly correspond to a residue term. The full results for variables selected can be found in Tables B.7 and B.9. For comparison, the results of Maree et al. (2015) are given in Table B.13, as the results of the equivalent study are not given in the original paper.

For the original SAT1 dataset, Reeve et al. (2010) selected 0 proven, 0 plausible and 0 implausible residues using the method described in Section 3.1.1 (i.e. when the Holm-Bonferroni correction was used). These results compare to 1 proven, 1 plausible and 0 implausible residues when the conjugate SABRE method was used and selecting any residue variables with a marginal posterior inclusion probability of greater than or equal to 0.5.[5] We have also looked at how well the methods do before selecting an implausible variable or before a p-value of greater than 0.05 (before the Holm-Bonferroni correction

[5] The power can be further improved (12 proven and 9 plausible residues) by inferring the selection threshold and selecting the top $J\hat{\pi}$ variables, at the expense of the selection of 1 implausible residue.

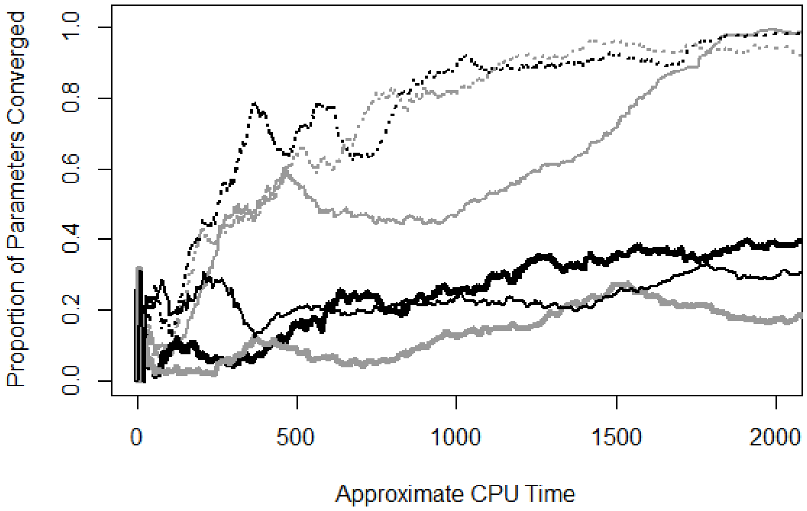


Figure 5.10: **Convergence diagnostics for the reduced SAT1 dataset used in Davies et al. (2014) and described in Section 5.1.4.** The lines show the proportion of parameters that have converged (PSRF ≤ 1.1) when using component-wise Gibbs sampling (black) and Metropolis-Hastings sampling proposing 4 (grey), 8 (black dashed), 16 (grey dashed), 32 (black thick) and 64 (grey thick) inclusion parameters simultaneously.

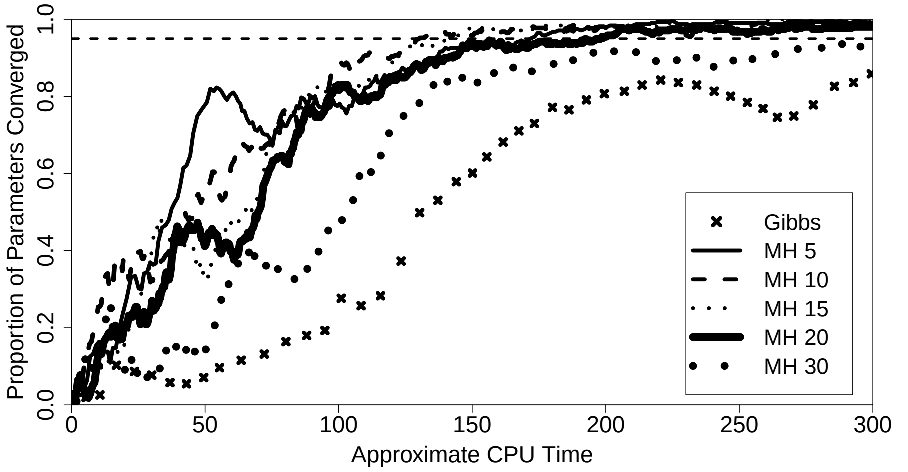
was used) was reached (in Reeve et al. (2010) the variable selection process was stopped as soon as a 0.05 p-value was reached). In this situation again the conjugate SABRE method offers an improvement, selecting 5 proven, 5 plausible and 0 implausible residues compared to 1, 1 and 0 respectively for the standard-mixed effects models. The difference in these results shows an advantage for the conjugate SABRE method over the standard mixed-effects models.

In the extended SAT1 dataset, Maree et al. (2015) used the method of Reeve et al. (2010) to select 5 proven, 0 plausible and 0 implausible residues, or 8, 1 and 0, respectively, if the method continued until selecting the first implausible residue. The conjugate SABRE method selected 11 proven, 3 plausible and 0 implausible residues when taking any variables with marginal posterior inclusion probabilities of greater than or equal to 0.5, or 15, 4 and 0, respectively, before selecting the first implausible residue.[6] It can again be seen that the power of the proposed conjugate SABRE method has improved over the method of Reeve et al. (2010).

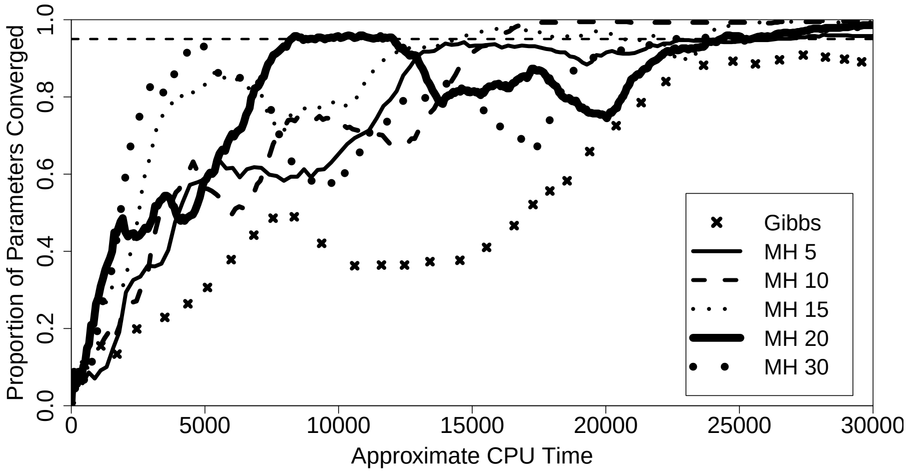
5.4.5 Sampling of Latent Indicators

Figures 5.10 and 5.11 compare component-wise Gibbs sampling against block Metropolis-Hastings sampling (both described in Section 4.3.5) in terms of speed of convergence. To

[6] The power can be further improved (24 proven and 15 plausible residues) by inferring the selection threshold and selecting the top $J\hat{\pi}$ variables, at the expense of the selection of 2 implausible residues.



(a) Original SAT1



(b) Extended SAT1

Figure 5.11: **Convergence diagnostics for the original and extended SAT1 datasets described in Section 5.1.** The lines show the proportion of parameters that have converged (PSRF < 1.05) versus the average CPU time (second) when using component-wise Gibbs sampling (crosses) and Metropolis-Hastings sampling proposing 5 (solid), 10 (dashed), 15 (dotted), 20 (thick solid) and 30 (thick dotted) inclusion parameters simultaneously.

do this we ran 4 chains for the component-wise Gibbs sampler and each of the variations of the Metropolis-Hastings sampler, monitoring the PSRFs for each parameter in the different methods. Figures 5.10 and 5.11 show the proportion of parameters with PSRFs < 1.1 (Figure 5.10) or PSRFs < 1.05 (Figure 5.11) in each case compared with the CPU time taken to get that number of samples. The higher the proportion of parameters with PSRFs lower than the required value (1.1 or 1.05), the better the method is said to have performed (Grzegorczyk and Husmeier, 2013).

Figure 5.10 compares convergence speed of different methods of proposing $\boldsymbol{\gamma}$ on the reduced SAT1 dataset used in Davies et al. (2014); see Section 5.1.4. The results, based on monitoring whether the PSRFs were less than 1.1, show that proposing a larger proportion of 8 (7.5%) or 16 (15%) binary selection hyperparameters, $\boldsymbol{\gamma}$, simultaneously in a block Metropolis-Hastings scheme achieves faster convergence than component-wise Gibbs sampling, despite the higher rejection probability (recall that Gibbs sampling has an acceptance probability of 1). This suggests that component-wise Gibbs sampling should not always be the default choice.

The results from Figure 5.11 support the advantage of a block Metropolis-Hastings sampler over a component-wise Gibbs sampler as shown in Figure 5.10, where following Davies et al. (2016a) convergence was determined by monitoring the percentage of variables with a PSRF ≤ 1.05. In all of the datasets the block Metropolis-Hastings samplers have outperformed the component-wise Gibbs sampler, with the exception of when more than 40 or 50 variables were sampled at a time (not shown in the diagrams for clarity). This shows that even sampling a reasonably large number of variables simultaneously, where the acceptance rate is likely to be low, can still yield a notable improvement. The results[7] in Figures 5.10 and 5.11 suggest that as a rule of thumb, sampling about 10 of the variables at a time will lead to effective sampling with the quickest convergence

5.5 Results for the SAT2 Dataset

For the SAT2 dataset, very little knowledge is available on how mutational changes affect antigenic variability, and no significant variables have been found in previous *in silico* work (Reeve et al., 2010). We have therefore applied our conjugate SABRE method as a tool for new hypothesis generation; see Table B.11 for the full results. For partial validation of our results, we exploit the fact that previous work by Grazioli et al. (2006) and Crowther et al. (1993b) has found evidence for antigenicity of the following three

[7] The best performing samplers in Figure 5.11 are as follows: Metropolis-Hastings samplers with 10 (7.2%) or 15 (10.9%) variables at a time for the original SAT1 dataset, with 10 (4.5%) or 15 (6.8%) variables at a time for the extended SAT1 dataset and 5 (1.8%) and 10 (3.6%) variables at a time for the H1N1 dataset.

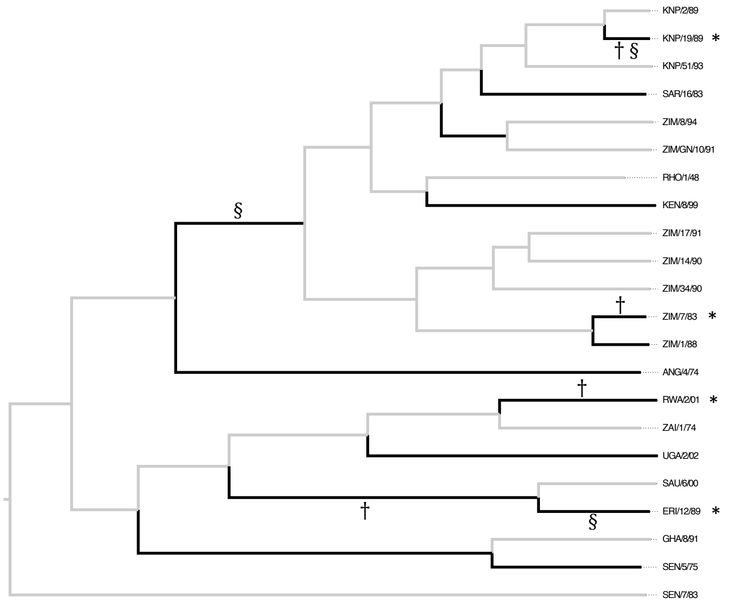


Figure 5.12: **Phylogenetic tree indicating significant branches in the evolutionary history of the SAT2 serotype based on the SAT2 dataset in Section 5.1.6.** The phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*). Branches associated with a change in virus phenotype are highlighted (black). Symbols indicate whether this was inferred to be a change in virus antigenicity (†), virus reactivity (none-identified) or virus immunogenicity (§). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance was taken to be the $J\hat{\pi}$ variables with the highest marginal inclusion probability, where the branches chosen are given in Table B.12.

areas: VP1 140-169 (part of the VP1 G-H loop), VP1 200-224 (VP1 C terminus) and VP2 70-82 (VP2 B-C loop).

Firstly in the VP2 B-C loop, the SABRE method has identified 5 residues that are antigenic; VP2 71, VP2 72, VP2 78, VP2 79 and VP2 80 (Grazioli et al., 2013, 2006; Kitson et al., 1990; Lea et al., 1994; Saiz et al., 1991). Of these VP2 78 has been experimentally identified using MAbs (Grazioli et al., 2006). Additionally VP2 72 is known to be antigenic in all other serotypes and these results suggest it is also antigenic in the SAT2 serotype (Grazioli et al., 2013, 2006; Mateu, 1995).

The second region in which antigenically significant residues have been found is in the VP1 G-H loop. The VP1 G-H loop is known to be a highly variable distracter site designed to confuse the host immune system (Crowther et al., 1993b) and is antigenic in all of the FMDV serotypes. In this loop, the conjugate SABRE method has specifically identified VP1 144 and VP1 166, where it is notable that VP1 166 lies directly between several residues that have been experimentally validated in the SAT2 serotype using MAbs (Crowther et al., 1993b).

The final known antigenic region that has been identified by the conjugate SABRE method is part of the VP1 C-terminus, the end of the VP1 protein. In the VP1 C-terminus we have identified VP1 207, VP1 208, VP1 209, VP1 210 and VP1 211 which are part of a region known to be antigenic in all FMDV serotypes except Asia1 (Aktas and Samuel, 2000; Grazioli et al., 2006; Lea et al., 1994; Saiz et al., 1991). With the conjugate SABRE method identifying all these neighbouring residues, it suggests that this section of the protein is a highly antigenic part of the SAT2 serotype.

Figure 5.12 gives the phylogentic tree for the SAT2 serotype with the predicted significant evolutionary changes. Unlike the SAT1 serotype, there is no prior knowledge of which residues and branches are antigenically relevant and we therefore apply our method to generate genuinely new hypotheses. The results presented give our best prediction for the significant branches and show a couple of potentially interesting groupings which could represent functional groups for the SAT2 serotype.

5.6 Results for the H1N1 Dataset

The analysis of the H1N1 dataset described in Section 5.1.7 selected 62 variables including 11 proven residues, 3 plausible residues and 5 implausible residues. A full list of the selected variables can again be found in Table B.14. Of the proven residues, one was identified on the RBS, position 187 (on the H1 common alignment) from the Sb antigenic site, and 4 others nearby; positions 130, 153, 189 and 190. Of those nearby, two occurred close together on the Sb antigenic site (189 and 190) and another on the Sa antigenic site (153). The other proven residue close to the RBS (130) is not part of an antigenic site but is known to be the location of a major antigenic change (Harvey et al., 2016).

The other proven residues selected come from two of the other antigenic sites; Ca and Cb. Positions 69, 72 and 74 are all found on Cb antigenic site, while positions 139, 141 and 142 are found on the Ca antigenic site. Additionally two of the plausible residues are also found nearby the Ca antigenic site. The remaining plausible residue (252) is part of the head domain and therefore considered plausible. The implausible residues selected cannot easily be explained but those selected may be partially a result of reducing the dataset (see Section 2.3.1). The one implausible residue that can be explained however is position 43 which by chance has a strong correlation with a known antigenic site (Harvey et al., 2016) rationalising its selection.

We have not constructed a separate estimate of the antigenicity of the branches of the H1N1 like we did for the FMDV datasets. We have not done this due to the phylogenetic tree of the H1N1 serotype being large and difficult to interpret. Additionally the H1N1 serotype is subject to rapid antigenic drift (Harvey et al., 2016) and therefore any inference would have less relevance. Finally we have not done a comparison with the results of Harvey et al. (2016) as they used a much larger dataset with more challenge strains and so any comparison would not be relevant.

5.7 Discussion

We have addressed the problem of identifying the residues within the SAT1 and SAT2 serotypes of FMDV and Influenza A (H1N1) that are responsible for changes in antigenic variability. This allows us to identify which residues must remain the same in order for two strains to cross react and for one strain to potentially be used as an effective vaccine against another. Identifying such residues can reduce the number of strains that must be tested as a vaccine, potentially reducing the time and cost associated with the selection procedure.

We have tested the family of SABRE methods introduced in Chapter 4 and shown how they offer improvement over the classical mixed-effects model, the mixed-effects LASSO and the mixed-effects elastic net as a result of the differences discussed in Section 5.7; see Section 5.3. We have additionally examined to fundamentally different approaches to variable selection in Bayesian hierarchical models: the slab-and-spike prior and the binary mask model; see Section 3.3. Our results given in Table 5.1 and displayed in Figures 5.3 and 5.4 show that the difference between these methods is negligible. We have also evaluated the difference between using a conjugate and semi-conjugate prior, as depicted by Figures 4.3 and 4.2. The differences in accuracy are negligible (see e.g. Figure 5.4). The conjugate model has slightly better computational efficiency (Figure 5.5), but this difference is not significant; this finding indicates that the bottleneck in the computational procedure is the sampling of the latent variables rather than the regression parameters. The conjugate model shows a slight but significant improvement over the non-conjugate model in the model selection scores based on WAIC, as seen from Tables 5.1 and B.6, but this has little immediate impact on the variable selection. Overall, our findings demonstrate a remarkable robustness of the proposed hierarchical modelling framework with respect to minor model modifications, which boosts our confidence in the predictions and in the variable ranking.

Further to this we have investigated the sampling of latent inclusion variables. We have shown that by proposing multiple variables simultaneously through Metropolis-Hastings sampling it is possible to give a significant computational improvement over the conventional component-wise Gibbs sampler (Figures 5.10 and 5.11). We have shown this improvement in a number of different datasets and have offered a general rule of thumb that proposing 10 variables at a time will lead to good mixing within MCMC chains for a variety of different datasets.

Through the use of this new model with the improved sampling techniques we have been able to identify an increased number of known antigenic sites in the SAT1 serotype of FMDV (Grazioli et al., 2006) compared to Reeve et al. (2010) and Maree et al. (2015), while incurring no (for the default selection threshold 0.5) or only a very small number (for the inferred selection threshold $J\hat{\pi}$) implausible residues. Very little biological knowledge exists about the SAT2 serotype, and a previous in silico application has failed to make any predictions at all (Reeve et al., 2010). To our knowledge, our study is the first time that specific new hypotheses about genetic-antigenic associations have been made with an *in silico* model based on the currently available data. Additionally we have provided an insight into the evolutionary history of the SAT serotypes (Figures 5.8, 5.9 and 5.12) and have provided a novel way of interpreting the biological effects of these virus mutations. Finally we have identified a number of significant antigenic sites in the H1N1 Influenza virus based on a reduced dataset and provided new hypotheses for this virus.

Chapter 6

A Sparse Hierarchical Bayesian Latent Variable Model for Understanding Antigenic Variability - The Methods

While the SABRE method offers consistent parameter inference and improved variable selection leading novel biological predictions, it does not fully take into account the data generation process. The structure of the data, discussed in Section 2.1, is a result of the same pair of challenge and protective strains being used to create multiple VN titre or HI assay measurements. As a result, the genetic and evolutionary data described in Sections 2.1.2 and 2.1.3 will be the same for any two measurements where the same challenge and protective strains are used. Modelling this structure more accurately is important and doing so should lead to more accurate biological results than those achieved by both the alternative methods in Chapter 3 and SABRE methods in Chapters 4 and 5.

In the work described in the current chapter, we describe an extended version of the conjugate SABRE method, the extended SABRE (eSABRE) method, which can properly account for the structure of the data while still retaining the attractive properties of the SABRE methods discussed and tested in Chapters 4 and 5. The eSABRE method introduces a latent variable structure into the mixed-effects model likelihood previously used in the SABRE methods in order to properly account for the data structure described in Chapter 2. In Section 6.1.1 we introduce the likelihood for the eSABRE method, with the remainder of the Section 6.1 defining the prior distributions of the model. In general the prior distributions for the eSABRE method follow those of the conjugate SABRE method (Section 4.2.2), but with adjustments and additions to fit in with the new latent variable likelihood described in Section 6.1.1.

As a result of using similar prior distributions to the conjugate SABRE method, the posterior inference of the eSABRE method in Section 6.2 roughly follows that of the conjugate SABRE method and we have used the conjugate sampling scheme proposed in Section 4.3.6. The differences in the posterior inference does however indicate one important advantage of the eSABRE method; its increased computational efficiency for larger datasets. As a result of the improved likelihood of the eSABRE method in Section 6.1.1, the sampling of the latent indicators, $\boldsymbol{\gamma}$, become less computationally onerous. This is massively advantageous as the sampling of $\boldsymbol{\gamma}$ was identified as the computational bottleneck of the SABRE methods in Chapter 4. The reduction in computational complexity comes from reducing the complexity of calculating the conditional distribution of $\boldsymbol{\gamma}$ by making it dependant on the inferred mean VN titre or HI assay for each pair of challenge and protective strains, rather than all of the individual VN titre and HI assay measurements. There are less pairs of challenge and protective strains then there are VN titre and HI assay measurements in all of the FMDV and Influenza datasets. This reduction in computational complexity is possible as a result of the latent variable structure of the likelihood introduced in Section 6.1.1 and explained further in Section 6.2.

Finally, in addition to proposing the eSABRE method, the current chapter also looks at methods for selecting random effects factors as we did previously in Chapters 4 and 5. As the latent variable likelihood for the eSABRE methods is specified as the product of two distributions, it is possible that alternative model selection techniques may offer an improvement over those proposed in Section 3.5 and tested in Chapter 5. Here we introduce a variation of the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010), block integrated WAIC (biWAIC) based on integrated WAIC (iWAIC) as proposed in Li et al. (2015). biWAIC takes into account the specific structure of the model and integrates over the latent variables. We have described how this converges to a particular form of Cross Validation (CV) and in Chapter 7 we use a simulation study to compare it to Bayesian 10-fold integrated CV (iCV) and non-integrated WAIC (nWAIC), a method which naively applies WAIC to the part of the latent variable likelihood containing the response, $\mathbf{y}$.

6.1 The eSABRE Method

The eSABRE method is based on the conjugate SABRE method from Section 4.2.2 in Chapter 4 (Davies et al., 2016a) but with a likelihood that better takes into account the data structure described in Chapter 2. The change in the structure is given in Section 6.1.1 and the remaining sections define the prior distributions of the eSABRE method, keeping to those used for the conjugate SABRE method as close as possible. Finally, the model is shown as a PGM in Figure 6.1 and the parameters are sampled from the posterior distribution using MCMC based on the methods described in Section 3.2.

6.1.1 Latent Variable Based Likelihood

The conjugate SABRE method described in Chapter 4 used the following likelihood, also given in (4.11), similar to classical mixed-effects models (Davies et al., 2016a):

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I}). \tag{6.1}$$

In (6.1), the response, log HI assay or log VN titre, is given by $\mathbf{y} = (y_1, \ldots, y_N)^\top$. The random-effects design matrix, $\mathbf{Z}$, is set to be a the matrix of indicators with N rows and $||\mathbf{b}||$ columns, where $||.||$ indicates the length of the vector and $\mathbf{b}$ is a column vector of random-effect coefficients. The explanatory variables, $\mathbf{X}$, are given as a matrix of $J+1$ columns and N rows and contain indicators of mutational changes at different residues or information on the phylogenetic structure where the first column is a column full of ones for the intercept. Of the explanatory variables, $\mathbf{X}$, only the relevant variables, $\mathbf{X}_{\boldsymbol{\gamma}}$, are included in (6.1) dependant on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\top \in \{0,1\}^J$. The relevance of the jth column of $\mathbf{X}$ is determined by $\gamma_j \in \{0,1\}$, where feature j is said to be relevant if $\gamma_j = 1$. Similarly $\mathbf{w}_{\boldsymbol{\gamma}}$ is given as the column vector of regressors, where the inclusion of each parameter is dependent on $\boldsymbol{\gamma}$.

While (6.1) gives a general model which can be used in a variety of different contexts, it does not completely account for the structure of the data used to model antigenic variability described in Chapter 2. The structure from the experiments means that any observations from the same challenge and protective strains will have the same explanatory variables. However it is worth noting that a given pair of viruses will give different explanatory variables if the strains used as challenge and protective strains are switched. As a result of this structure, we can introduce latent variables, $\boldsymbol{\mu}_{\mathbf{y}}$, into the model, where each $\mu_{\mathbf{y},p}$ represents the inferred underlying HI assay measurement of any given pair of challenge and protective strains, p.

The introduction of the latent variables, $\boldsymbol{\mu}_{\mathbf{y}}$, into the models results in the following distribution for $\mathbf{y}$:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{Z}\mathbf{b}, \sigma_y^2\mathbf{I}) \tag{6.2}$$

where $\mathbf{M}$ is a design matrix which ensures that each y has the underlying inferred VN titre or HI assay measurement, $\mu_{\mathbf{y},p}$, for its given pair of challenge and protective strains,

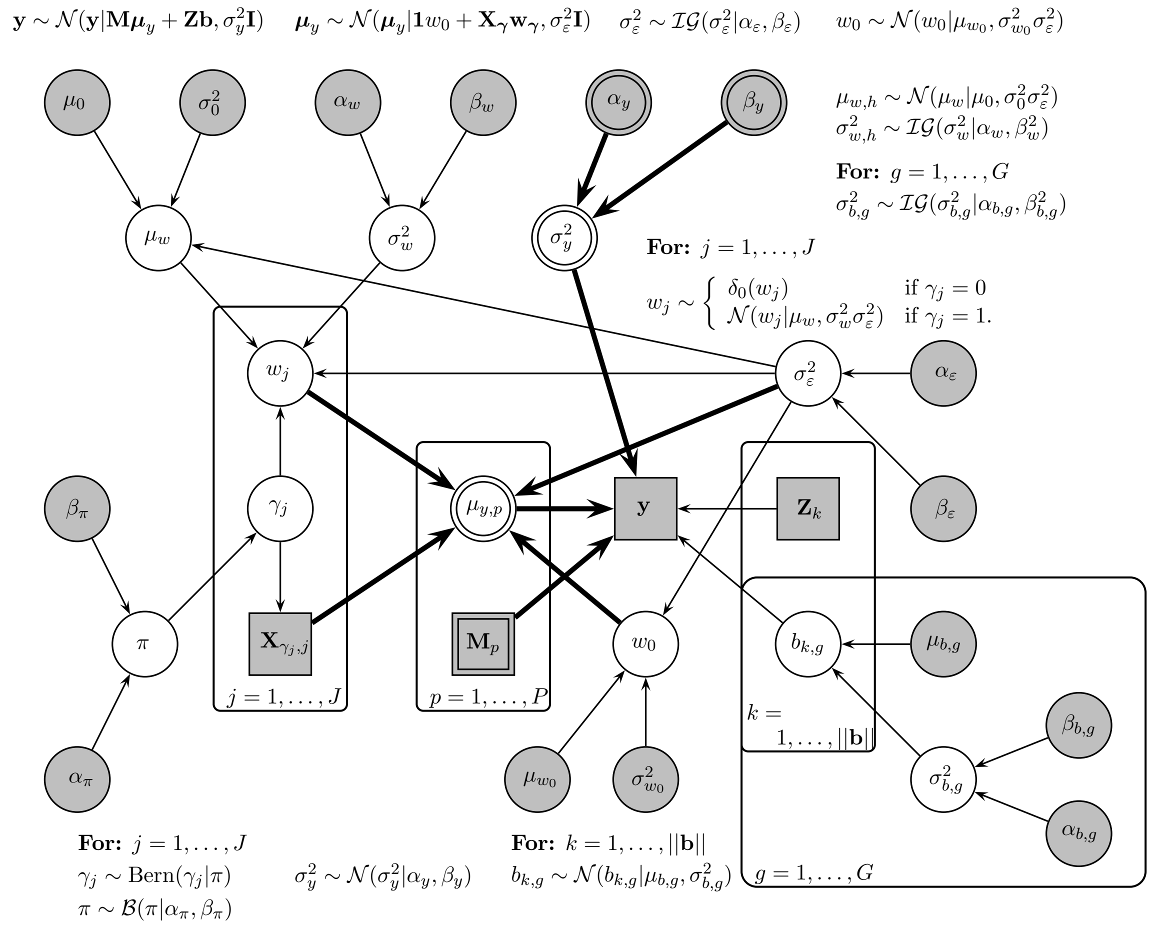

$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_y + \mathbf{Z}\mathbf{b}, \sigma_y^2\mathbf{I})$
$\boldsymbol{\mu}_y \sim \mathcal{N}(\boldsymbol{\mu}_y|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\mathbf{I})$
$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)$
$w_0 \sim \mathcal{N}(w_0|\mu_{w_0}, \sigma_{w_0}^2\sigma_\varepsilon^2)$
μ_0
σ_0^2
α_w
β_w
α_y
β_y
$\mu_{w,h} \sim \mathcal{N}(\mu_w|\mu_0, \sigma_0^2\sigma_\varepsilon^2)$
$\sigma_{w,h}^2 \sim \mathcal{IG}(\sigma_w^2|\alpha_w, \beta_w^2)$
For: $g = 1, \ldots, G$
$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g}^2)$
μ_w
σ_w^2
σ_y^2
For: $j = 1, \ldots, J$
$w_j \sim \begin{cases} \delta_0(w_j) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_j|\mu_w, \sigma_w^2\sigma_\varepsilon^2) & \text{if } \gamma_j = 1. \end{cases}$
w_j
σ_ε^2
α_ε
β_π
γ_j
$\mu_{y,p}$
$\mathbf{y}$
$\mathbf{Z}_k$
β_ε
π
$\mathbf{X}_{\gamma_j,j}$
$\mathbf{M}_p$
w_0
$b_{k,g}$
$\mu_{b,g}$
$j = 1, \ldots, J$
$p = 1, \ldots, P$
$k =$
$1, \ldots, ||\mathbf{b}||$
$\beta_{b,g}$
α_π
μ_{w_0}
$\sigma_{w_0}^2$
$\sigma_{b,g}^2$
$\alpha_{b,g}$
For: $j = 1, \ldots, J$
$\gamma_j \sim \text{Bern}(\gamma_j|\pi)$
$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)$
$\sigma_y^2 \sim \mathcal{N}(\sigma_y^2|\alpha_y, \beta_y)$
For: $k = 1, \ldots, ||\mathbf{b}||$
$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2)$
$g = 1, \ldots, G$

Figure 6.1: **Compact representation of the eSABRE method as a PGM.** The *grey* circles and squares refer to the fixed hyperparameters and data respectively, while the *white* circles refer to parameters and hyperparameters that are inferred. The main differences with the conjugate SABRE method given in Figure 4.3 can be seen by noting the addition of the latent variables, $\mu_{y,p}$, between w_j and $\mathbf{y}$, the addition of nodes and edges connecting σ_y^2, α_y and β_y, and the edges connecting σ_ε^2 and w_0 to $\mu_{y,p}$ rather than $\mathbf{y}$.

p. The random effects factors are added into this part of the likelihood as some of these factors, e.g. the date of the experiment, affect measurements at the individual level, i.e. they are different for each y; see Section 2.1.1 for details on the random effects factors.

We then wish to infer the values of the VN titre or HI assay measurements of the pairs of challenge and protective strains, $\boldsymbol{\mu}_{\mathbf{y}}$, based on the differences in the protein structure and evolutionary history of the virus described in Sections 2.1.2 and 2.1.3:

$$\boldsymbol{\mu}_{\mathbf{y}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\mathbf{I}). \tag{6.3}$$

As with the SABRE methods in Chapter 4, we only wish to use the relevant explanatory variables, $\mathbf{X}_{\boldsymbol{\gamma}}$, and corresponding regression coefficients, $\mathbf{w}_{\boldsymbol{\gamma}}$. We also include an intercept parameter, w_0 as we expect high underlying HI assay measurements when the two virus strains used are the same, i.e. the explanatory variables are equal to zero. The full model is given graphically in Figure 6.1.

The eSABRE method's latent variable likelihood, given in (6.2) and (6.3), has two major advantages over the likelihood of the conjugate SABRE method, given in (6.1). Firstly it allows us to better attribute the error to the correct part of the model. In the VN titre and HI assay measurements some of the error comes from variability within the experiments, e.g. getting multiple different results for the same pair of challenge and protective strains once the experimental conditions have been taken into account, and this is modelled by σ_y^2. Other errors will come from the model fit, e.g. our model not completely replicating the true underlying biological process, and this is given by σ_ε^2. Attributing the error better means our model matches better with the data collection process and should result in more accurate results.

The second advantage of the eSABRE is massively improved computational performance. For example to analyse the H1N1 dataset would take the SABRE method weeks or months to sample the required number of iterations to achieve convergence and a reasonable sample size after burn-in, the eSABRE method is able to achieve the result in less than a day. The improvement is a result of reducing the computation required to calculate the posterior distribution of $\boldsymbol{\gamma}$. In essence, through the introduction of latent variables the eSABRE method reduces the posterior distribution of $\boldsymbol{\gamma}$ to a multivariate Gaussian distribution of dimension $||\boldsymbol{\mu}_\mathbf{y}||$, $||\boldsymbol{\mu}_\mathbf{y}|| = 570$ in the H1N1 dataset, as opposed to dimension $||\mathbf{y}||$, $||\mathbf{y}|| = 15,693$ in the H1N1 dataset, in the SABRE method. This is a result of the d-separation of $\mathbf{y}$ and $\boldsymbol{\gamma}$ via $\boldsymbol{\mu}_\mathbf{y}$ in Figure 6.1. Similar results are also likely for the H3N2 dataset, although the times required would be much larger.

6.1.2 Noise and Intercept Priors

Unlike the SABRE methods in Chapter 4, the eSABRE method contains two types of error rather than one to better reflect the error coming from the data collection process. The first part of the error is given by σ_y^2 in (6.2). This error term represents the variation seen in the measurements collected from the same pair of challenge and protective strains:

$$\sigma_y^2 \sim \mathcal{IG}(\sigma_y^2|\alpha_y, \beta_y) \tag{6.4}$$

where the hyper-parameters α_y and β_y are fixed, as indicated by the grey nodes in Figure 6.1. As with the SABRE methods in Chapter 4 we have used conjugate priors where possible, so we can use Gibbs sampling to sample as many parameters as possible.

The other error comes from the second part of the likelihood, (6.2), and is given by

σ_ε^2:

$$\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2 | \alpha_\varepsilon, \beta_\varepsilon) \tag{6.5}$$

where the hyper-parameters α_ε and β_ε are fixed. This represents the error between the inferred underlying HI assay or VN titre measurements for each pair of challenge and protective strains and what can be explained by the fixed effects, $\mathbf{w}_{\boldsymbol{\gamma}}^*$. σ_ε^2 is also included in the distributions for w_0, $\mathbf{w}_{\boldsymbol{\gamma}}$ and μ_w (defined in (6.6) and Section 6.1.3) following the conjugate SABRE method described in Section 4.2.2. The advantage of this information sharing is that the error variance in terms of model fit is reflected in the distribution of the regression coefficients and a potential computational advantage can also be obtained through collapsed Gibbs sampling; see Davies et al. (2016a).

Additionally we also require a prior on our intercept:

$$w_0 \sim \mathcal{N}(w_0 | \mu_{w_0}, \sigma_{w_0}^2 \sigma_\varepsilon^2). \tag{6.6}$$

As discussed in Section 4.2.1, we treat the intercept differently from the remaining regressors, wishing to use vague prior settings so as not to penalise this term and effectively make the model scale invariant (Hastie et al., 2009).

6.1.3 Spike and Slab Priors

As with the conjugate SABRE method, we use spike and slab priors as proposed by Mitchell and Beauchamp (1988) and described in Section 3.3.1. Again the idea of the spike and slab prior is that the prior reflects whether the feature is relevant based on the values of $\boldsymbol{\gamma}$. In this way we expect that $w_j = 0$ if $\gamma_j = 0$, i.e. the feature is irrelevant, and conversely it should be non-zero if the variable is relevant, $w_j \neq 0$ if $\gamma_j = 1$. With the eSABRE method the effects of the spike and slab prior are seen on the estimate of $\boldsymbol{\mu}_{\mathbf{y}}$ rather than $\mathbf{y}$ itself as in the SABRE methods, with $\boldsymbol{\mu}_{\mathbf{y}}$ then affecting the estimate of $\mathbf{y}$. This can be seen by comparing Figures 4.3 and 6.1. Following the conjugate SABRE method, we again add σ_ε^2 into the distribution for further conjugacy:

$$w_j \sim \begin{cases} \delta_0(w_j) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_j | \mu_w, \sigma_w^2 \sigma_\varepsilon^2) & \text{if } \gamma_j = 1 \end{cases} \tag{6.7}$$

for $j \in 1, \ldots, J$ and where δ_0 is the delta function. The prior for the variance of the parameter is then given by:

$$\sigma_w^2 \sim \mathcal{IG}(\sigma_w^2 | \alpha_w, \beta_w). \tag{6.8}$$

where α_w and β_w are fixed; see Figure 6.1.

As with the conjugate SABRE method, we again assign a flexible parameter for the mean of the regression coefficients, $\mathbf{w}_\gamma$:

$$\mu_w \sim \mathcal{N}(\mu_w | \mu_0, \sigma_0^2 \sigma_\varepsilon^2) \tag{6.9}$$

where the hyper-parameters μ_0 and σ_0^2 are fixed and σ_ε^2 is again included in the variance for further conjugacy. The need for a flexible vale of μ_w is due to our biological understanding of the problem, with the model likely to have a high intercept, w_0, and only negative regression coefficients; see Section 2.1.

The final part of the spike and slab prior is to define the prior for the latent binary indicators, $\boldsymbol{\gamma}$. For this we assign Bernoulli prior for each γ_j with probability π, with the probability π itself given a prior following a conjugate Beta distribution:

$$p(\boldsymbol{\gamma}|\pi) = \prod_{j=1}^{J} \text{Bern}(\gamma_j|\pi) \tag{6.10}$$

$$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi) \tag{6.11}$$

where α_π and β_π are fixed, as indicated by the grey nodes in Figure 6.1.

6.1.4 Random-Effects Priors

For the random effects priors we use the same priors as with the conjugate SABRE method. We do not consider the folded-non-central-t prior distribution described in Gelman (2006) and tested here in Section 5.3 (Davies et al., 2016a). The results of Figure 5.1 showed that the prior did not offer any advantage in the context of the SABRE methods and therefore we have not used it here.

As with mixed-effects models and the SABRE methods we give the random effects coefficients, $b_{k,g}$, group dependant Gaussian priors where the group is defined by k, i.e. $b_{k,g}$ is shorthand for b_{k,g_k}:

$$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2). \tag{6.12}$$

where we again fix $\mu_{b,g} = 0$ with the group dependant variance parameter, $\sigma_{b,g}^2$, given a conjugate Inverse-Gamma prior:

$$\sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g}) \tag{6.13}$$

where $\alpha_{b,g}$ and $\beta_{b,g}$ are fixed hyper-parameters for each g. Again, as in Section 4.1.4, we define $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}})$ where $\boldsymbol{\Sigma}_{\mathbf{b}} = diag(\boldsymbol{\sigma}^2_{\mathbf{b}})$ with $\boldsymbol{\sigma}^2_{\mathbf{b}} = (\sigma^2_{b,1}, \ldots, \sigma^2_{b,1}, \sigma^2_{b,2}, \ldots, \sigma^2_{b,G})^\top$ such that each $\sigma^2_{b,g}$ is repeated with length $||\mathbf{b}_g||$.

6.2 Posterior Inference

To explore the posterior distribution of the eSABRE method we have used an MCMC algorithm; see Section 3.2. As with the SABRE methods in Chapter 4, we have chosen conjugate priors where possible meaning that we can use Gibbs sampling for most of the model parameters; see Section 3.2.2. The distributions needed for sampling are given here and are derived in Section A.2, where we again use a slight abuse of notation and denote $\boldsymbol{\theta}'$ as all other parameters that are not on the left of the conditioning bar. The only parameter that we cannot use Gibbs sampling with is $\boldsymbol{\gamma}$ and this is discussed in Section 6.2.1.

$$\boldsymbol{\mu}_{\mathbf{y}}|\boldsymbol{\theta}_{-\boldsymbol{\mu}_{\mathbf{y}}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{V}_{\mathbf{y}}(\mathbf{M}^\top(\mathbf{y} - \mathbf{Z}\mathbf{b})/\sigma^2_y + \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}/\sigma^2_\varepsilon), \mathbf{V}_{\mathbf{y}}) \tag{6.14}$$

$$\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{-\mathbf{w}^*_{\boldsymbol{\gamma}}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{m}_\gamma, \sigma^2_\varepsilon\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \tag{6.15}$$

$$\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mathbf{b}|\tfrac{1}{\sigma^2_y}\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}}), \mathbf{V}_{\mathbf{b}}) \tag{6.16}$$

$$\mu_w|\boldsymbol{\theta}_{-\mu_w}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mu_w|V_{\mu_w}(\mathbf{1}\mathbf{w}_{\boldsymbol{\gamma}}/\sigma^2_w + \mu_0/\sigma^2_0), \sigma^2_\varepsilon V_{\mu_w}) \tag{6.17}$$

$$\sigma^2_y|\boldsymbol{\theta}_{-\sigma^2_y}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma^2_y|\ ||\mathbf{y}||/2 + \alpha_y, \tfrac{1}{2}(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{Z}\mathbf{b})) \tag{6.18}$$

$$\sigma^2_w|\boldsymbol{\theta}_{-\sigma^2_w}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma^2_w|\ ||\mathbf{w}_{\boldsymbol{\gamma}}||/2 + \alpha_w, \tfrac{1}{2\sigma^2_\varepsilon}(\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{I}\mu_w)^\top(\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{I}\mu_w)) \tag{6.19}$$

$$\sigma^2_{b,g}|\boldsymbol{\theta}_{-\sigma^2_{b,g}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma^2_{b,g}|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}_g^\top\mathbf{b}_g) \tag{6.20}$$

$$\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma^2_\varepsilon|(||\boldsymbol{\mu}_{\mathbf{y}}|| + ||\mathbf{w}^*_{\boldsymbol{\gamma}}|| + 1)/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}R_{\sigma^2_\varepsilon}) \tag{6.21}$$

$$\pi|\boldsymbol{\theta}_{-\pi}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \beta(\pi|\ \alpha_\pi + ||\boldsymbol{\gamma}||, \beta_\pi + J - ||\boldsymbol{\gamma}|). \tag{6.22}$$

where we sample $\sigma^2_{b,g}$ for each g. We also define $\mathbf{V}_{\mathbf{y}} = (1/\sigma^2_\varepsilon\mathbf{I} + \mathbf{M}^\top\mathbf{M}/\sigma^2_y)^{-1}$, $\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = (\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})^{-1}$, $\mathbf{V}_{\mathbf{b}} = (\frac{1}{\sigma^2_y}\mathbf{Z}^\top\mathbf{Z} + \boldsymbol{\Sigma}^{-1}_{\mathbf{b}})^{-1}$, $V_{\mu_w} = (1/\sigma^2_0 + ||\mathbf{w}_{\boldsymbol{\gamma}}||/\sigma^2_w)^{-1}$ and $R_{\sigma^2_\varepsilon} = (\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{X}^*_\gamma\mathbf{w}^*_\gamma)^\top(\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{X}^*_\gamma\mathbf{w}^*_\gamma) + (\mathbf{w}^*_\gamma - \mathbf{m}_\gamma)^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}(\mathbf{w}^*_\gamma - \mathbf{m}_\gamma) + (\mu_{\mathbf{w}} - \mu_0)^\top(\mu_{\mathbf{w}} - \mu_0)/\sigma^2_0$ for notational simplicity.

Following Davies et al. (2016a) we have again used collapsing in an attempt to improve mixing and convergence, e.g. Andrieu and Doucet (1999). As in Section 4.3.6 this is achieved through a series of collapsed distributions for $\boldsymbol{\gamma}$, $\mathbf{w}^*_{\boldsymbol{\gamma}}$, μ_w, σ^2_ε and π:

$$\begin{aligned} p(\boldsymbol{\gamma}, \mathbf{w}^*_{\boldsymbol{\gamma}}, \mu_w, \sigma^2_\varepsilon, \pi) &= p(\boldsymbol{\gamma})p(\pi|\boldsymbol{\gamma})p(\sigma^2_\varepsilon|\pi, \boldsymbol{\gamma})p(\mu_w|\sigma^2_\varepsilon, \pi, \boldsymbol{\gamma})p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mu_w, \sigma^2_\varepsilon, \pi, \boldsymbol{\gamma}) && (6.23) \\ &= p(\boldsymbol{\gamma})p(\pi|\boldsymbol{\gamma})p(\sigma^2_\varepsilon|\boldsymbol{\gamma})p(\mu_w|\sigma^2_\varepsilon, \boldsymbol{\gamma})p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mu_w, \sigma^2_\varepsilon, \boldsymbol{\gamma}) && (6.24) \end{aligned}$$

where the conditionality on $\boldsymbol{\theta}'$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{y}$ has been dropped and the simplification from

(6.23) to (6.24) follows from the conditional independence relations shown in Figure 6.1, exploiting the fact that π is d-separated from the remaining parameters in the argument via $\boldsymbol{\gamma}$. These distributions are achieved by collapsing over parameters as derived in Section A.2.

6.2.1 Sampling the Latent Indicators

In the SABRE methods of Chapter 4, sampling $\boldsymbol{\gamma}$ is both difficult, as a result of it not naturally taking a distribution of standard form, and computationally expensive. However a conditional distribution can still be obtained and Davies et al. (2016a) used collapsing methods following Sabatti and James (2005), as described in Section 4.3.5, to achieve faster mixing and convergence as follows:

$$p(\gamma|\boldsymbol{\theta}_{-\gamma}, \mathbf{X}^*_\gamma, \mathbf{Z}, \mathbf{y}) \propto \int p(\gamma|\boldsymbol{\theta}_{-\gamma}, \mathbf{X}^*_\gamma, \mathbf{Z}, \mathbf{y}) d\mu_w d\mathbf{w}^*_\gamma d\pi d\sigma^2_\varepsilon \tag{6.25}$$

where using the likelihood for the conjugate SABRE method given in (6.1) and the priors described in Sections 4.1 and 4.2.

However with the likelihood for the conjugate SABRE method given in (6.1), as well as the likelihoods for the other SABRE methods, the computational cost of computing (6.25) becomes dependant inverting a $||\mathbf{y}|| \times ||\mathbf{y}||$ matrix. For the FMDV datasets this is not problematic, as $||\mathbf{y}||$ is relatively small. However with the H1N1 and H3N2 datasets, where $||\mathbf{y}|| = 15,693$ and $||\mathbf{y}|| = 7,315$ respectively, calculating any distribution where a $||\mathbf{y}|| \times ||\mathbf{y}||$ matrix inversion is repeatedly required becomes infeasible.

It is at this point that the latent variable likelihood given in (6.2) and (6.3) shows its huge computational advantage over the SABRE methods discussed in Chapter 4; see Table 7.1 for an example of the computational savings. As in the conjugate SABRE method, (6.25), we use collapsing methods and integrate over μ_w, $\mathbf{w}^*_\gamma$, π and σ^2_ε. However while in the SABRE method this gives a computational dependence on $||\mathbf{y}||$, $||\mathbf{y}|| = 15,693$ for the H1N1 dataset, for the eSABRE method we get a computational dependence on $||\boldsymbol{\mu}_\mathbf{y}||$:

$$p(\gamma|\boldsymbol{\theta}_{-\gamma}, \mathbf{X}^*_\gamma, \boldsymbol{\mu}_\mathbf{y}) \propto \int p(\gamma|\boldsymbol{\theta}_{-\gamma}, \mathbf{X}^*_\gamma, \boldsymbol{\mu}_\mathbf{y}) d\mu_w d\mathbf{w}^*_\gamma d\pi d\sigma^2_\varepsilon. \tag{6.26}$$

The dependency on $\boldsymbol{\mu}_\mathbf{y}$ rather than $\mathbf{y}$ is a result of (6.2) not containing $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$ therefore does not need to be included in (6.26). The dependence on $||\boldsymbol{\mu}_\mathbf{y}||$ rather than $||\mathbf{y}||$ is where the main computational cost reduction occurs, as in the H1N1 dataset $||\boldsymbol{\mu}_\mathbf{y}|| = 570$ is much smaller $||\mathbf{y}||$ making the computational cost of computing (6.26) far less than (6.25). Further collapsing is possible within the sampling step for $\boldsymbol{\gamma}$ in the eSABRE method, i.e.

collapsing over $\boldsymbol{\mu}_{\mathbf{y}}$. However despite the potentially improved sampling available per iteration by doing this, the increased computational cost of calculating (6.26) at each step would far outweigh any gains that would be made.

Based on the results of Section 5.4.5 taken from Davies et al. (2014) and Davies et al. (2016a) we have chosen to sample $\boldsymbol{\gamma}$ via a block Metropolis-Hastings step. In those studies it was found that block Metropolis-Hastings sampling was the method that offered the quickest convergence of the parameters based on CPU time. The only difference here is that we have a posterior distribution of dimension $||\boldsymbol{\mu}_{\mathbf{y}}||$ rather than $||\mathbf{y}||$

6.3 Selection of Random Effect Components

There are various methods that can be used to select the random effects that should be used within the model, here we look at Bayesian integrated CV (iCV), e.g. Vehtari and Ojanen (2012), and several variations of WAIC (Watanabe, 2010).

6.3.1 Integrated Cross Validation

Bayesian CV methods are reliable, if computationally expensive, techniques for measuring the out-of-sample performance of different models. Bayesian iCV is a special version of CV which works well in latent variable models. Bayesian iCV integrates over the latent variables, in this case $\boldsymbol{\mu}_{\mathbf{y}}$, to give the following utility function for k-fold Bayesian iCV:

$$p_{iCV} = \frac{1}{K}\sum_{k=1}^{K}\log\frac{1}{I}\sum_{\iota=1}^{I}p(\mathbf{y}_k|\boldsymbol{\theta}^{\iota}) \tag{6.27}$$

where the distribution $p(\mathbf{y}_k|\boldsymbol{\theta}^{\iota})$ comes from integrating over $\boldsymbol{\mu}_{\mathbf{y}}$ in the distribution given by the product of (6.2) and (6.3). The parameter samples, $\boldsymbol{\theta}^{\iota}$, are taken from the eSABRE method applied to $\mathbf{y}_{-k}$, $\mathbf{X}_{-k}$, $\mathbf{Z}_{-k}$ and $\mathbf{M}_{-k}$.

6.3.2 Block Integrated WAIC

WAIC, as proposed in Watanabe (2010) and defined here in Section 3.5.2, is a natural method for selecting the correct model when the underlying model is singular, i.e models with a non-identifiable parameterisation, such as the SABRE method. WAIC has been proven to be asymptotically equivalent to Bayesian leave-one-out CV (LOO-CV) in Watanabe (2010) and is computed as follows from posterior samples $\boldsymbol{\theta}^{\iota}$ for $\iota \in \{1, \ldots, I\}$:

$$p_{WAIC} = -2\sum_{i=1}^{N}\left(\log\left(\frac{1}{I}\sum_{\iota=1}^{I}p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,i}, \mathbf{Z}_i)\right) - \mathrm{Var}\left(\log(p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,i}, \mathbf{Z}_i))\right)\right). \tag{6.28}$$

where Var is the sample variance. WAIC can be used for a wide variety of problems, however it is only justifiable for problems where the observed data are independently distributed with a population distribution, e.g. the SABRE method where the joint likelihood is given by (6.1).

To make WAIC more applicable to latent variable models such as the eSABRE method, Li et al. (2015) introduced two alternative versions of WAIC; non-integrated WAIC (nWAIC) and integrated WAIC (iWAIC). nWAIC applies WAIC to the predictive density of the observed variables, $\mathbf{y} = (y_1, \ldots, y_N)$, conditional on the model parameters, $\boldsymbol{\theta}$, and the potentially correlated latent variables, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_N)$:

$$p_{nWAIC} = -2\sum_{i=1}^{N}\left(\log\left(\frac{1}{I}\sum_{\iota=1}^{I} p(y_i|\boldsymbol{\theta}^{\iota}, \psi_i^{\iota}, \mathbf{Z}_i)\right) - \operatorname{Var}\left(\log(p(y_i|\boldsymbol{\theta}^{\iota}, \psi_i^{\iota}, \mathbf{Z}_i))\right)\right) \tag{6.29}$$

where $\boldsymbol{\theta}^{\iota}$ and ψ_i^{ι} are sampled via MCMC and Var is the sample variance. In the proposed eSABRE method, taking just the likelihood for y_i from (6.2) would be the distribution corresponding to $p(y_i|\boldsymbol{\theta}^{\iota}, \psi_i^{\iota}, \mathbf{Z}_i)$ and would seem unlikely to completely satisfy the independence assumptions of WAIC based methods.

nWAIC also does not fully account for the mismatch in the model fit of the latent variables, i.e. how well the latent variables are predicted by the fixed effects. Li et al. (2015) therefore proposed iWAIC:

$$p_{iWAIC} = -2\sum_{i=1}^{N}\left(\log\left(\frac{1}{I}\sum_{\iota=1}^{I} p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\boldsymbol{\gamma},i}, \mathbf{Z}_i, \boldsymbol{\psi}_{\textbf{-i}}^{\iota})\right) - \operatorname{Var}\left(\log(p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\boldsymbol{\gamma},i}, \mathbf{Z}_i, \boldsymbol{\psi}_{\textbf{-i}}^{\iota}))\right)\right) \tag{6.30}$$

where Var is the sample variance and the distribution used is given by $p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\boldsymbol{\gamma},i}, \mathbf{Z}, \boldsymbol{\psi}_{\textbf{-i}}^{\iota})$ $= \int p(y_i|\boldsymbol{\theta}^{\iota}, \boldsymbol{\psi}_{\textbf{-i}}^{\iota}, \psi_i, \mathbf{Z})p(\psi_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\boldsymbol{\gamma}})d\psi_i$, the marginal likelihood based on taking both parts of the likelihood of the latent variable model and integrating over the latent variable ψ_i corresponding to y_i.

The proposed version of iWAIC does not however work with the eSABRE method. This is a result of each observation, y_i, not having its own corresponding latent variable, ψ_i. Instead any two observations, y_1 and y_2, from the same pair of challenge and protective strains, p, will have the same latent variable, i.e. $\psi_1 = \psi_2 = \mu_{\mathbf{y},p}$. Under this model, i.e. where $\rho(\psi_1, \psi_2) = 1$, it is mathematically intractable to integrate over $\psi_1 = \mu_{\mathbf{y},p}$ without integrating over $\psi_2 = \mu_{\mathbf{y},p}$, something which is required in order to calculate $p(y_i|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\boldsymbol{\gamma},i}, \mathbf{Z}_i, \boldsymbol{\psi}_{\textbf{-i}})$ as needed for (6.30). We must therefore either use nWAIC given by (6.29) or find an alternative.

In this current work we proposed biWAIC for latent variable models with latent variables that are either completely correlated or have no correlation. While WAIC, nWAIC and iWAIC rely on using independent distributions for each y_i, biWAIC instead uses a distribution for independent groups of observations $\mathbf{y}_p$, given by $\mathbf{y}_p : y_i$ where $p_i = p$. Given this notation we can then compute biWAIC as follows:

$$p_{biWAIC} = -2\sum_{p=1}^{P}\left(\log\left(\frac{1}{I}\sum_{\iota=1}^{I} p(\mathbf{y}_p|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,p}, \mathbf{Z}_p)\right) - \text{Var}\left(\log(p(\mathbf{y}_p|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,p}, \mathbf{Z}_p))\right)\right) \tag{6.31}$$

where Var is the sample variance and the distribution used is given by $p(\mathbf{y}_p|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,p}, \mathbf{Z}) = \int p(\mathbf{y}_p|\boldsymbol{\theta}^{\iota}, \mu_{y,p}, \mathbf{Z})\ p(\mu_{y,p}|\boldsymbol{\theta}^{\iota}, \mathbf{X}_{\gamma,p})d\mu_{y,p}$ where the two distributions that are part of the marginalisation are taken from (6.2) and (6.3).

As well as being applicable to the eSABRE method and particular specifications of latent variable models, biWAIC also has some useful asymptotic properties. Previously Watanabe (2010) has shown that WAIC is asymptotically equivalent to Bayesian LOO-CV. While biWAIC is not asymptotically equivalent to LOO-CV, based on the same concept it is asymptotically equivalent to Bayesian leave-one-group-out CV (LOGO-CV). We define LOGO-CV as the cross validation method where observations are divided into P independent groups based on the latent structure, as opposed to n groups of single observations for LOO-CV or k groups for k-fold CV.

6.4 Discussion

In this chapter we have introduced the eSABRE method and discussed how it can offer improved performance over the SABRE methods discussed in Chapter 4 and 5. In Section 6.1 we have described how the model can take into account the data generation process to improve modelling and variable selection performance, and have specified the change in likelihood needed to achieve this; Section 6.1.1. In Section 6.2 we have then described how the change in likelihood given in Section 6.1.1 can potentially lead to significantly improved computational efficiency and given the conditional distributions. Finally in Section 6.3 we have discussed methods for selecting the random effect components in the eSABRE method and have proposed an alternative criterion, biWAIC, which may better take into account the latent variable structure of the eSABRE method and other similar methods.

Chapter 7

A Sparse Hierarchical Bayesian Latent Variable Model for Understanding Antigenic Variability - The Analysis

In this chapter we test the effectiveness of the eSABRE method proposed in Chapter 6, as well as a newly proposed information criterion; block integrated WAIC (biWAIC). We firstly introduce the data in Section 7.1 where we describe the simulated datasets we have used to demonstrate the improvements offered by the eSABRE method over the conjugate SABRE method (Section 4.2.2). We additionally describe the real life Influenza datasets that the eSABRE method has been applied to, before Section 7.2 describes the computational inference.

Section 7.3 looks at the results of the simulation studies. The results show the improvement offered by the eSABRE method over the conjugate SABRE method when the simulated data is generated from a more biologically realistic model. The results from Table 7.1 show that the eSABRE method is robust to increases in the error related to model fit (see Section 6.1.1) and outperforms the conjugate SABRE method across all datasets. Table 7.1 additionally demonstrates the computational efficiency of the eSABRE method compared to the SABRE method when the number of observations increases, something which is important when it comes to applying the model to the real life Influenza datasets. Table 7.2 also shows how that the eSABRE method gives improved variable selection over the conjugate SABRE method when more realistic simulation studies are used. Finally, Table 7.3, Figure 7.2 and Figure 7.3 compare the performance of non-integrated WAIC (nWAIC), biWAIC and 10-fold Bayesian integrated CV (iCV) in terms of correctly selecting random effect factors. The results show that all three of the methods perform similarly, with the biWAIC offering an alternative threshold for the inclusion of random effect factors as a result of fully accounting for the latent variable likelihood.

Section 7.4 compares the performance of the eSABRE method against the conjugate SABRE method in terms of correctly identifying antigenic residues from the H1N1 Influenza serotype. The results firstly demonstrate how it is possible to apply the eSABRE method to the full H1N1 dataset, whereas for computational feasibility the conjugate SABRE method had to be applied to a reduced H1N1 dataset in Chapter 5. While the results show similar amounts of proven antigenic residues based on the classifications in Section 2.4.3, the eSABRE method reduces the number of implausible residues selected. In the H3N2 dataset our results identify a large number of antigenic residues from three of the five known antigenic regions. Additionally we propose other plausible residues that appear to be antigenic and may require further experimental investigation.

7.1 Data

Detailed descriptions of the H1N1 and H3N2 Influenza datasets are given in Section 2.3 of Chapter 2. In this section we describe the simulated datasets that are used to test the effectiveness of the eSABRE and conjugate SABRE methods described in Section 6.1 and Chapter 4, and add a few details on the real life datasets that are specific to this chapter of the thesis.

7.1.1 Non-FMDV Simulated Data

To initially test the eSABRE and conjugate SABRE methods we generated 3 datasets with a reasonably small number of variables. These 3 datasets (Simulated Dataset 1 (SD1), SD2 and SD3) are based on the same structure as the H1N1 and FMDV datasets with a varied number of random effect factors based on Section 2.1.1. In each of the datasets 2000 observations were simulated from 55 pairs of challenge and protective strains (10 viruses which are designated as both challenge and protective strains) with 50 possible fixed effects and 4 possible random effect components (including the challenge and protective strains). The random effects coefficients are generated from a zero mean Gaussian distribution with each component having a fixed variance drawn from $U(0.2, 0.5)$. Fixed effects, w_j, were given non-zero effects generated from a uniform distribution, $U(-0.4, -0.2)$, with probability $\pi \sim U(0.2, 0.4)$. σ_y^2 and σ_ε^2 were both set to be 0.033, 0.1 and 0.3 respectively for the three simulated datasets.

7.1.2 FMDV Simulated Data

To make the simulation studies more realistic we wanted to make simulated datasets based on the H1N1 and H3N2 Influenza datasets described in Sections 2.3. However using the conjugate SABRE method to analyse datasets of this size is computational prohibitive. Therefore instead we have created 20 simulated datasets based on the extended SAT1 FMDV dataset used in Maree et al. (2015) and Davies et al. (2016a); Section 2.2.1. These datasets were created to be the same size as the FMDV datasets using the maximum a-posteriori parameter estimates of the eSABRE method applied to the FMDV dataset, but with varied error in the underlying model, $\sigma_\varepsilon^2 \in \{0.02, 0.2, 0.5\}$, and different mean regression parameters, $\mu_w \in \{-0.1, -0.3, -0.5\}$, so as to highlight the differences in performance of the two models under different circumstances. Following Maree et al. (2015) we used 3 random effect components; the challenge strain, the date of the experiment and the antiserum.

7.1.3 Simulated Data for Model Selection

Finally, to compare nWAIC, biWAIC and 10-fold Bayesian iCV, we have generated 9 sets of 20 datasets with up to 4 random effects; the challenge strain, the protective strain and two generic random effect factors. The datasets were generated with 50 possible fixed effects and up to 4 random effect factors included with probability 0.5. Of the 9 sets of datasets, 3 contain 10 virus strains, where each virus strain has been used as a protective and challenge strain, meaning there are 55 pairs of challenge and protective strains. Following the same set up, 3 of the sets of datasets include 30 virus strains (465 pairs) and the other 3 have 45 virus strains (1035 pairs). Within each of these sets of 3 datasets, the model error, σ_ε^2, was varied to be either 0.1, 0.3 or 0.5.

7.1.4 Influenza Data

Both of H1N1 and H3N2 are described in Section 2.3 and we have used the full datasets described here. In each case we have used biWAIC to choose the random effect factors that are included in the models analysed in Sections 7.4 and 7.5.

7.2 Computational Inference

To test model convergence for both the simulated and real datasets we ran 4 chains for each model and then computed the PSRF (Gelman and Rubin, 1992) from the within-chain and between-chain variances. We took the threshold of convergence to be a PSRF

Table 7.1: **Table of AUROC values and CPU time for the eSABRE and the conjugate SABRE methods applied to the non-FMDV based simulated datasets.** The table gives the AUROC values and CPU times per 1,000 iterations (seconds) for the eSABRE and conjugate SABRE methods, where the results for the conjugate SABRE method are given in brackets. The result come from when the methods were applied to the non-FMDV simulated datasets (SD1, SD2 and SD3) described in Section 7.1.1 with varied numbers of observations.

Obs.	**AUROC Values**			**CPU Time Per 1,000 Iterations**		
	SD1	**SD2**	**SD3**	**SD1**	**SD2**	**SD3**
500	0.98 (0.90)	0.90 (0.77)	0.82 (0.64)	25 (497)	25 (867)	47 (444)
1000	0.98 (0.83)	0.91 (0.70)	0.82 (0.59)	29 (6,931)	26 (5,623)	36 (5,546)
2000	0.98 (0.75)	0.92 (0.61)	0.83 (0.58)	32 (35,231)	25 (32,243)	43 (20,904)

≤ 1.1 and terminated the burn-in phase when this was satisfied for 95% of the variables. The fixed hyperparameters were set the same for both the eSABRE and conjugate SABRE methods such that $\boldsymbol{\alpha_b} = \boldsymbol{\beta_b} = (0.001, \ldots, 0.001)$, $\alpha_w = \beta_w = \alpha_y = \beta_y = \alpha_\varepsilon = \beta_\varepsilon = 0.001$, $\mu_0 = 0$, $\sigma_0^2 = 100$, $w_0 = max(y)$, $\alpha_\pi = 1$ and $\beta_\pi = 4$ following Davies et al. (2016a).

7.3 Results for the Simulation Studies

Table 7.1 gives the AUROC values for the eSABRE and conjugate SABRE (Section 4.2.2) methods applied to the non-FMDV simulated datasets from Section 7.1.1; SD1, SD2, SD3. For each combination of dataset and number of observations, the eSABRE method offers an improvement in terms of global variable selection performance over the SABRE method. This improvement is a result of the latent variable structure of the eSABRE method which better reflects the data generation process, where the difference in the methods can be seen by comparing the PGMs in Figures 4.3 and 6.1. Table 7.1 also shows the effect of deviating from this data collection process. For the SD1 dataset where both of the error variances in the data generation process are small, $\sigma_y^2 = \sigma_\varepsilon^2 = 0.033$, the conjugate SABRE method gives similar results to the eSABRE method. However as the error variances get larger, e.g. SD2 and SD3, the eSABRE method offers a much clearer improvement over the SABRE method. This is a result of the conjugate SABRE and eSABRE methods becoming identical models as $\sigma_\varepsilon^2 \to 0$. Given the large variance in HI assay measurement for any given pair of challenge and protective strains in the H1N1 and H3N2 datasets, this improvement is vital.

Another notable result from Table 7.1 is the reduction in performance in terms of AUROC values of the conjugate SABRE method (Section 4.2.2) as the number of observations increases. This is an unexpected result as we would expect more data to provide more information to the model, resulting in a better selection of variables in the models

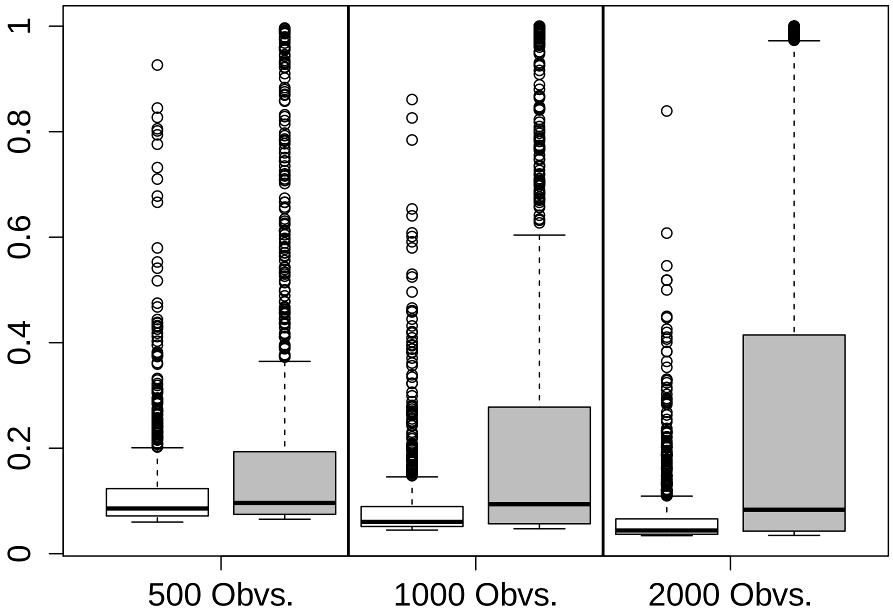


Figure 7.1: **Box plots showing the effect of non-iid Gaussian noise on a model assuming iid Gaussian noise.** The box plots show the probability of an irrelevant variable being included in a model for data with iid Gaussian noise (white) against the probabilities for a model with noise based on FMDV and Influenza Data (grey).

and higher AUROC values. The reason for this strange result is a consequence of the mismatch between the data generation process where errors come in two forms, σ_ε^2 and σ_y^2, and the model which only directly accounts for the error in $\mathbf{y}$ coming from σ_y^2.

To demonstrate that the strange reduction in performance of the conjugate SABRE method is a result of the mismatch between the data and the model we have completed a small simulation study with linear models. We have generated groups of datasets with 500, 1,000 and 2,000 observations generated from a linear model with each group containing 2000 datasets. For each of these groups, half the datasets have observations generated with iid noise, e.g. just σ_y^2, and the other half with correlated errors based on the structure of the FMDV and Influenza data, e.g. both σ_y^2 and σ_ε^2. Additionally each of the datasets contains two variables, one relevant, $\mathbf{x}_r$, and one irrelevant, $\mathbf{x}_{ir}$. We have then calculate the marginal likelihood of each of the four possible models, where we have fixed σ_w^2 and marginalised out σ_y^2 and $\mathbf{w}$, to give the probability of the irrelevant variable being included in the final model, $\mathcal{M}$, as follows:

$$\mathbb{P}(\mathbf{x}_{ir} \in \mathcal{M}) = \frac{p(\mathbf{y}|\mathbf{x}_{ir}) + p(\mathbf{y}|\mathbf{x}_{ir}, \mathbf{x}_r)}{p(\mathbf{y}|.) + p(\mathbf{y}|\mathbf{x}_{ir}, \mathbf{x}_r) + p(\mathbf{y}|\mathbf{x}_r) + p(\mathbf{y}|\mathbf{x}_{ir}, \mathbf{x}_r)}. \tag{7.1}$$

Figure 7.1 gives box plots of the probability of the irrelevant variable, $\mathbf{x}_{ir}$, being included in the final model for each of the datasets from our small simulation study. The

Table 7.2: **Table of AUROC values for the eSABRE and the conjugate SABRE methods when applied to the FMDV based simulated datasets.** The table gives AUROC values for the eSABRE and conjugate SABRE methods, where the results for the conjugate SABRE method are given in brackets, when applied to the FMDV based simulated datasets described in Section 7.1.2.

		σ_ε^2		
		0.02	0.2	0.5
μ_w	-0.1	0.67 (0.69)	0.67 (0.60)	0.63 (0.57)
	-0.3	0.72 (0.71)	0.70 (0.61)	0.67 (0.58)
	-0.5	0.75 (0.72)	0.74 (0.64)	0.73 (0.57)

box plots show the affect on the probabilities caused by the different types of noise and varied amounts of observations. Figure 7.1 shows that as the number of observations increases the chance of the irrelevant variable being included decreases for the iid noise, as would be expected. However for the non-iid noise based on the FMDV and Influenza datasets, the results show an increase in the probability of the irrelevant variable being included as the number of observations increases, indicating that the noise mismatch is what causes the strange results in Table 7.1.

Finally, Table 7.1 shows the improvement the eSABRE method offers over the conjugate SABRE method in terms of computational efficiency. Table 7.1 shows how the SABRE method becomes vastly more computationally expensive as the number of observations increases, while the require CPU hardly changes for the eSABRE method if the number of pairs of challenge and protective strains remains the same. This improvement in terms of computational efficiency explains why it is viable to use the eSABRE method on the H1N1 dataset for example, where $||y|| = 15,693$ and $P = 570$, but not the conjugate SABRE method or any of the other SABRE methods described in Chapter 4.

Table 7.2 shows the effectiveness of the eSABRE method on larger more realistic datasets (Section 7.1.2) based on the real life FMDV data from Reeve et al. (2010). Like Table 7.1, the results of Table 7.2 again show the eSABRE method clearly outperforming the conjugate SABRE method across all of the simulated datasets from Section 7.1.2. The results show that as the model error in the simulated data increases, the conjugate SABRE seriously drops off in performance while the eSABRE method remains reasonably consistent. Like with the results of Table 7.1, the difference in performance is again caused by the mismatch between the conjugate SABRE and the underlying generation process which the eSABRE method matches more closely.

To compare the methods described in Section 6.3, nWAIC, biWAIC and Bayesian 10-fold iCV, we have compared their performance in terms of correctly selecting random effect factors on the datasets from Section 7.1.3. The results are given in Table 7.3 and

Table 7.3: **Table of results looking at the random effects factor selection performance of the methods described in Section 6.3.** The table gives results in terms of the successful selection or exclusion of random effects factors when using the methods described in Section 6.3, nWAIC, biWAIC and Bayesian 10-fold iCV, on parameter samples from the eSABRE method applied to the simulated data from Section 7.1.3. The results given are sensitivity, specificity and F-scores and are displayed in an alternative manner in Figures 7.2 and 7.3.

	P	σ_ε^2	nWAIC	biWAIC	Bayesian 10-fold iCV
Sensitivity	55	0.1	0.90	0.97	0.92
	55	0.3	0.92	0.90	0.89
	55	0.5	0.78	0.71	0.93
	465	0.1	0.97	0.94	0.85
	465	0.3	0.86	0.84	0.86
	465	0.5	0.95	0.90	0.86
	1035	0.1	0.93	0.71	0.98
	1035	0.3	0.91	0.79	0.87
	1035	0.5	0.90	0.66	0.74
Specificity	55	0.1	0.68	0.56	0.15
	55	0.3	0.70	0.60	0.41
	55	0.5	0.59	0.54	0.26
	465	0.1	0.45	0.60	0.66
	465	0.3	0.49	0.63	0.63
	465	0.5	0.37	0.56	0.53
	1035	0.1	0.32	0.60	0.47
	1035	0.3	0.33	0.52	0.33
	1035	0.5	0.39	0.55	0.29
F-Score	55	0.1	0.80	0.80	0.65
	55	0.3	0.88	0.84	0.79
	55	0.5	0.72	0.66	0.70
	465	0.1	0.70	0.75	0.73
	465	0.3	0.70	0.74	0.75
	465	0.5	0.73	0.76	0.72
	1035	0.1	0.73	0.69	0.80
	1035	0.3	0.77	0.74	0.75
	1035	0.5	0.60	0.54	0.60

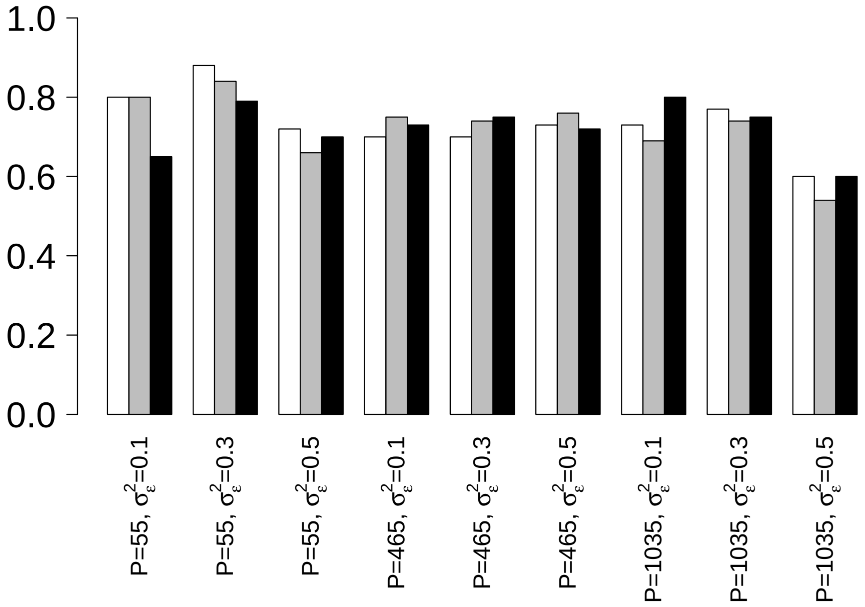


Figure 7.2: **Bar plot of F1-Scores given in Table 7.3.** The bar plot compares the F1-scores for nWAIC (white), biWAIC (grey) and Bayesian 10-fold iCV (black) in terms of correctly selecting random effect components for the dataset described in Section 7.1.3. The figure takes the results from Table 7.3.

are displayed visually in Figures 7.2 and 7.3.

The results in Table 7.3 show that all of the methods, nWAIC, biWAIC and Bayesian 10-fold iCV, perform similarly in terms of overall selection accuracy. The similarly is best demonstrated by looking at the F1-scores, which offer a more general assessment of performance than looking at specificity and sensitivity individually. The F1-scores from Table 7.3 can also be seen in Figure 7.2 where the results are shown as box plots. With the results from Table 7.3 and Figure 7.2 suggesting that the information criteria, nWAIC and biWAIC, give similar selection performance to Bayesian 10-fold iCV, it is reasonable to use one of the criteria on the Influenza dataset in Section 7.4 and 7.5, where Bayesian 10-fold iCV will be computationally onerous.

While suggesting that the methods perform similarly overall, Table 7.3 also indicates that the methods operate with different thresholds, meaning that on average some methods include more random effect factors than others. This can be seen by looking at the sensitivities and specificities of nWAIC, biWAIC and Bayesian 10-fold iCV in Table 7.3 or alternatively by looking at Figure 7.3. Figure 7.3 plots the sensitivities achieved by the different methods on each set of datasets against the 1 minus specificities and shows that the biWAIC method operates at a higher threshold for inclusion, meaning that it selects less random effect factors in the model on average. This can be seen by noting

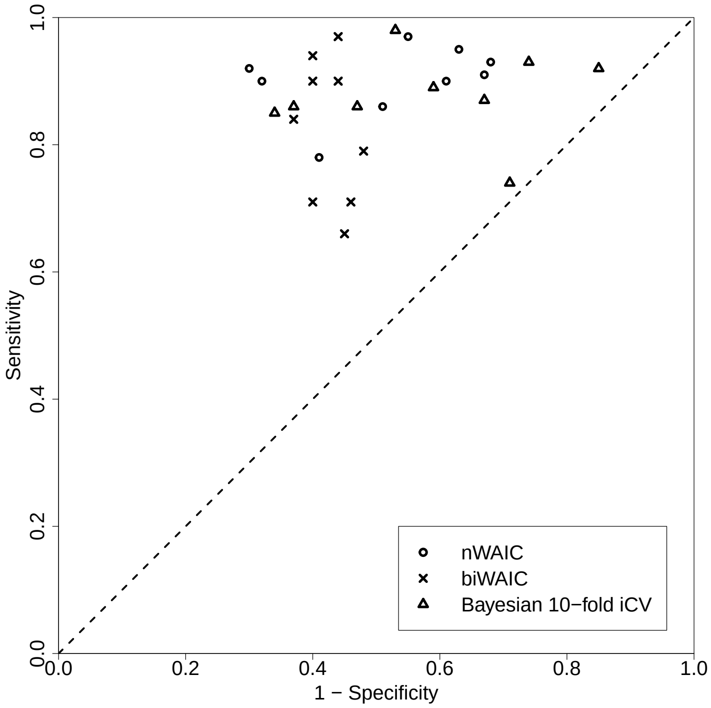


Figure 7.3: **Plot of sensitivities and 1 minus specificities for the results given in Table 7.3.** The plot compares nWAIC (circles), biWAIC (crosses) and Bayesian 10-fold iCV (triangles) in terms of correctly selecting random effect components for the dataset described in Section 7.1.3. The figure takes the results from Table 7.3 and plots the sensitivities against the 1 minus specificities, i.e. as single point from a ROC curve.

the lower sensitivities and higher specificities in Figure 7.3 or Table 7.3.

The reason for the difference between nWAIC and biWAIC in terms of the average number of random effect factors included is a result of the distribution from which they measure the sample means and variances needed to calculate the criterion. nWAIC, (6.29), takes its sample means and variances based on only the distribution of $\mathbf{y}$, (6.2), the distribution which contains the random effects specification. biWAIC, (6.31), however takes its sample means and variances from the marginalised distribution of $\mathbf{y}$ where $\boldsymbol{\mu}_{\mathbf{y}}$ has been integrated out as detailed in Section 6.3.2. As a result, like Bayesian 10-fold iCV, biWAIC takes into account both the model fit of $\mathbf{y}$ and $\boldsymbol{\mu}_{\mathbf{y}}$.

Taking into account both distributions of the latent variable likelihood, (6.2) and (6.3), better assesses the fit of the model and prevents the overfitting of the first distribution of the latent variable likelihood, (6.2). The results for nWAIC show that not accounting for (6.3) as well as (6.2) leads to unrealistically high sensitivities and low specificities. It is interesting however that we do not see a similar threshold with Bayesian 10-fold iCV which also takes into account both parts of the latent variable likelihood. This is a consequence of the different thresholds given by criteria based on WAIC and those based on CV. We observed this in Table 5.2 when we compared WAIC and Bayesian CV.

7.4 Results for the H1N1 Dataset

We have applied the eSABRE method to the H1N1 dataset using the 8 possible combinations of random effect components. The biWAIC score was then calculated for each of the models, Section 6.3, with the model with the best biWAIC score containing the challenge strain and the date of the experiment as random effect components. biWAIC was chosen to select the best model based on feasibility, it is far more computational efficient than 10-fold Bayesian iCV, and the results from Table 7.3. Full results for the variables selected by the eSABRE method are given in Table B.15, and in Table B.14 for those selected by the conjugate SABRE method based on the reduced dataset described in Section 5.1.7.

Having selected the model with the best selection of random effects, we have then compared the results in terms of variable selection to those achieved by the SABRE method on a reduced H1N1 dataset in Section 5.6. We do no compare our results with those of Harvey et al. (2016) as those results were achieved on a larger dataset, see Section 2.3.1, using a non-automated version of mixed-effects. Using the eSABRE method we have selected 5 proven, 1 plausible and 1 implausible based on choosing a marginal inclusion probability of 0.5, or 10 proven, 5 plausible and 2 implausible based on taking the $\hat{\pi}J$ variable with the highest marginal inclusion probabilities. These results compare to 5 proven, 1 plausible and 2 implausible or 11, 2 and 3 for the conjugate SABRE method based on the same criteria. The results show the methods performing reasonably similarly, however the eSABRE offers an improvement in terms of not selecting as many implausible residues. The classification of these results is based on our biological knowledge of the H1N1 serotype from Section 2.4.

Of the 10 proven residues, we have identified one residue on the Residual Binding Site (RBS) as in Section 5.6 when using the conjugate SABRE method, residue 187 on the H1 common alignment. Residue 187 is part of the the Sb antigenic site and we have also identified two other nearby residues (189 and 190) on the same antigenic site. The other proven residues come from the Ca (141, 142 and 170), Cb (69, 72 and 74) and Sa (130) antigenic sites which also contain 4 of the plausible residues predicted to be antigenic by the eSABRE method. These should potentially be investigated experimentally to determine whether they are indeed antigenic residues. The final plausible residue is related to a mutation resulting in one of the tested viruses and its potential antigenicity can be attributed to 4 different residues, 3 of which are proven and one of which is implausible[8], and it is possible that some of these residues may be antigenic.

7.5 Results for the H3N2 Dataset

As with the H1N1 dataset in Section 7.4, we have applied the eSABRE method and biWAIC to the H3N2 dataset from Section 2.3.2 with 8 different combinations of random effect components. biWAIC has indicated that the best possible model is the one that contains all of the possible random effects factors; the challenge strain, the protective strain and the date of the experiment. The full results for the eSABRE method applied to the H3N2 dataset described in Section 2.3.2 are given in Table B.16. We do not compare our results with those of Harvey (2016), as while they have used classical mixed-effects models, they used a piecemeal approach which required manual intervention to guide the selection procedure. The eSABRE method can be applied in a fully automatic manner.

The results of our analysis of the H3N2 dataset from Section 2.3.2 using the eSABRE method and biWAIC has resulted in the selection of 10 proven, 3 plausible and 2 implausible residues, given here by their common alignments; see (Harvey et al., 2016). We have ruled out one implausible residue based on the information given in Section 2.4.4. Of the proven residues, we have identified 8 in the highly variable antigenic site B (155, 158, 159, 164, 189, 183, 193,197), and among these are residues known to part of the residual binding site (Harvey, 2016). In addition we have also identified 2 other residues in the C and E antigenic regions, 276 and 262 respectively. Of the plausible sites, one gives an antigenic effect that could be explained by either a branch, an implausible residue or a proven residue. While we have no specific evidence, it is highly likely that this antigenic effect is a result of the the proven residue on the antigenic site E (75). The other two plausible residues (279 and 212) come from areas close to the C and D antigenic sites, with 212 next to a proven antigenic residue in the alignment and potentially worthy of further investigation.

7.6 Discussion

In this chapter we have tested and analysed the eSABRE method proposed in Chapter 6. We have tested it against the conjugate SABRE method proposed in Chapter 4 and shown how it offers improved performance on a variety of different simulated datasets; Tables 7.1 and 7.2. The results in Table 7.1 also demonstrate the computational improvement offered by the eSABRE methods, as discussed in Chapter 6, and give examples as

[8] We classify this variable as plausible based on line 4 of Table 2.1

to where the biggest computational improvements can be seen. In addition to testing the eSABRE method, we have also looked at the best way of selecting random effects coefficients. Table 7.3 and Figure 7.2 show that biWAIC, as proposed in Chapter 6, performs equally well in terms of selecting the correct random effect factors as two more established methods. Figure 7.3 then demonstrates how the biWAIC criterion properly accounts for the entire latent variable distribution resulting in a more realistic number of random effects factors being included.

Sections 7.4 and 7.5 demonstrate how the eSABRE method, together with biWAIC, can be effectively applied to large real life Influenza datasets. In Section 7.4 we show how the improvement in computational efficiency demonstrated in Table 7.1 allows us to make use of the full H1N1 dataset rather than a reduced version as was required for the conjugate SABRE method in Chapter 5. The results from using the full H1N1 dataset and properly accounting for the error in the data collection process through the eSABRE method show an improvement in the selection of antigenic variables in the H1N1 datasets. Finally Section 7.5 applies the eSABRE method and biWAIC to the H3N2 dataset, identifying a number of proven and plausible antigenic residues, at the expense of a small number of implausible residues.

Chapter 8

Conclusions and Further Work

The aim of this thesis has been to create models that can address the problems caused by antigenic variability. Based on this objective we have created models, Section 8.1, which can use biological measure of antigenic variability to link genetic and phylogenetic changes to significant antigenic changes. We have proposed a family of models, the SABRE methods, to this end, Section 8.1.1, and demonstrated the improved performance they offer over the standard methods used. We have then extended this method and proposed a new model, the eSABRE method, which gives an improvement in results, Section 8.1.2, and can provide accurate biological prediction on large datasets; see Section 8.2. The following sections summarise the work that has been completed in this thesis and Section 8.3 gives proposals for further work in the area.

8.1 Methodological Advances

In general, the methodological work from this thesis can be broken down into two parts; the SABRE methods and the eSABRE method. The work related to the SABRE methods in Section 8.1.1 is taken from Chapters 4 and 5, but this section also includes methods proposed in Davies et al. (2016a) which were detailed in Chapter 3. The eSABRE method was proposed and evaluated in Chapters 6 and 7, and is summarised here in Section 8.1.2.

8.1.1 The SABRE Methods

In Section 4.1 we introduced the original SABRE method, Figure 4.1, as proposed in Davies et al. (2014). The SABRE method is a Bayesian hierarchical mixed-effects model which can simultaneously account for the experimental effects of the data collection process, and select the residues and evolutionary changes that affect the measured antigenic variability. To select variables the SABRE method uses spike and slab priors, which have been shown to give improved variable selection over methods based on ℓ_1 regularisation (Mohamed et al., 2012). We have demonstrated this improvement here through both simulated and real life studies, Chapter 5, and have given a detailed explanation of the reasons for this improvement in Section 4.4. To summarise, the improvement is a result of (1) avoiding the bias inherent in ℓ_1 regularisation based methods, (2) the method giving genuine and consistent sparsity, (3) properly accounting for uncertainty, and (4) through borrowing strength from information coupling through the hierarchical structure seen in Figure 4.1.

In the remainder of Chapter 4, we have investigated potential changes to the original SABRE method that might lead to improved variable selection and sampling. We have proposed three additional versions of the SABRE method; the semi-conjugate SABRE method (Section 4.2.1), the conjugate SABRE method (Section 4.2.2) and the binary mask conjugate SABRE method (Section 4.2.3). In Chapter 5 we have compared these methods against each other and a number of alternative methods including the additional methods proposed in Davies et al. (2016a) and described in Chapter 3. The new alternative methods extend the previously proposed mixed-effects LASSO (Schelldorfer et al., 2011) to allow the specification of multiple random effects factors and propose the alternative mixed-effects elastic net.

The semi-conjugate SABRE method given in Figure 4.2, improves the original SABRE method by properly modelling the biologically significant intercept parameter. The intercept is important as it gives the VN titre or HI assay measurement when a virus is used as both the challenge and protective strain. The conjugate SABRE method given in Figure 4.3 then increases the conjugacy of the semi-conjugate SABRE method by adding additional edges between the error variance, σ_ε^2, and some of the parameters associated with the regression coefficients, $\mathbf{w}_\gamma^*$ and $\mu_{w,h}$. The conjugate SABRE method also allows for the possibility of improving the sampling scheme through collapsing; Section 4.3.6. We have compared the semi-conjugate and conjugate SABRE methods in terms of accuracy, computational efficiency, and formal model selection preference in Table 5.1. The results show that the differences in accuracy are negligible, Figure 5.4. Similarly there is no significant difference in terms of computational efficiency, Figure 5.5, indicating that the sampling of the latent indicator variables, $\boldsymbol{\gamma}$, is the computational bottleneck of the SABRE methods. In terms of model selection, WAIC showed a significant difference in favour of the conjugate SABRE method, Table 5.1, but this has little impact on the variable selection accuracy. Overall the similarity of the results supports the robustness of the SABRE methods and its reliability in making predictions.

Chapter 5 also tested the difference between a model based on the binary mask model and one using spike and slab priors; see Figure 3.2 in Section 3.3. While both methods are discussed and used in the literature (Murphy, 2012), our work represents the first quantification of the difference in performance between the two methods. We have proposed the binary mask conjugate SABRE method, Figure 4.4, and tested it against the conjugate SABRE method, Figure 4.3. Our systematic comparison quantifies the differences between these methods in terms of accuracy and computational efficiency, and found the differences to be negligible. Quantifying this result is important, as both approaches have been used as variable selection methods in the literature, with authors tending to arbitrarily chose one method or the other, e.g. Davies et al. (2014), Heydari et al. (2016).

The work in Chapters 4 and 5 also looks at the computational bottleneck of the SABRE methods, the sampling of $\boldsymbol{\gamma}$. We have investigated the possibility of sampling $\boldsymbol{\gamma}$ through a block Metropolis-Hastings sampler rather than the more commonly used component-wise Gibbs sampler; Section 4.3.5. Our results in Section 5.4.5 show the computational improvement offered by the block Metropolis-Hastings sampler. The results, shown in Figures 5.10 and 5.11, indicate that sampling around 10 latent indicators at time offers the most computational efficient sampling scheme.

Finally, we have demonstrated the conjugate SABRE method on real life FMDV and Influenza datasets from Chapter 2. Our results find a number of known antigenic residues and significant evolutionary changes, discussed in Section 8.1.2, and show that the SABRE methods are accurate *in silico* methods that can be used to identify antigenic residues and provide an effective way of modelling antigenic variability.

8.1.2 The Extended SABRE Method

In Chapter 6 we proposed the eSABRE method given in Figure 6.1. The eSABRE method replaces the likelihood of the conjugate SABRE, (6.1), with one based on a latent variable model, (6.2) and (6.3), which better accounts for the data generation process described in Chapter 2. The eSABRE method takes into account the fact that for any given pair of challenge and protective strains the fixed effects will remain the same and modelling this properly leads to an improvement in terms of model accuracy by fulling account for the error inherent in the data collection process. The method also has the advantage that $\boldsymbol{\gamma}$ is d-separated from $\mathbf{y}$ via $\boldsymbol{\mu}_{\mathbf{y}}$ in Figure 6.1, offering an improvement in computational efficiency in the sampling of $\boldsymbol{\gamma}$; see Section 6.2.1.

In addition to the eSABRE method, we have also looked at different ways of selecting the random effect factors in the eSABRE method. We have considered Bayesian 10-fold integrated CV (iCV), a CV based method that integrates over the latent variables, $\boldsymbol{\mu}_{\mathbf{y}}$, to fully account for both parts of the latent variable likelihood, (6.2) and (6.3). We have compared this against the previously proposed non-integrated WAIC (nWAIC) (Li et al., 2015), which naively applies WAIC to the part of the latent variable likelihood containing the observations, (6.2). In addition we proposed our own criterion, block integrated WAIC (biWAIC), based on integrated WAIC of Li et al. (2015), which integrates over the latent variables, $\boldsymbol{\mu}_{\mathbf{y}}$, to give a criterion which fully accounts for both distributions of the latent variable likelihood of the eSABRE method.

In Chapter 7 we have tested the eSABRE method against the SABRE method and biWAIC against nWAIC and Bayesian 10-fold iCV. The results of the simulation studies in Section 7.3 show that the eSABRE method outperforms the conjugate SABRE method both in terms of variable efficiency and variable selection accuracy. Table 7.3 and Figure 7.2 additionally showed that biWAIC, nWAIC and Bayesian 10-fold iCV all performed similarly in terms of correctly selecting random effect factors in the models and in Figure 7.3 we have demonstrated the effect of accounting for the fit of the full latent variable model in biWAIC. Finally we have demonstrated how the eSABRE method and biWAIC can be applied to the Influenza datasets from Section 2.3 to provide relevant biological results in a situation where, due to the size of the datasets, applying the SABRE methods of Chapter 4 would be computationally infeasible.

8.2 Biological Advances

In terms of direct biological improvements, in Section 2.1.3 we have proposed new methods for understanding how evolutionary changes effect antigenicity. Previous methods, e.g. Reeve et al. (2010) and Davies et al. (2014), included the branches of the phylogenetic trees to account for any changes in the measured VN titre or HI assay that could not be explained by the mutational changes. However where a particular branch separates two virus strains which have been used as both challenge and protective strains, we can include additional variables in the model and give a biological understanding of the potential reasons for their inclusion. To summarise, we can include branch variables to explain the effect amino acid substitutions at a particular phylogenetic branch have on the challenge and protective strains carrying those amino acid substitutions, see Section 2.1.3 for further details, with branches also included to explain general antigenic effects not described by the mutational changes. We have demonstrated this approach on the FMDV datasets and have made predictions of the antigenically significant evolutionary changes in both the SAT1 and SAT2 serotypes; Figures 5.8, 5.9 and 5.12. In the SAT1 serotype, where prior knowledge of these changes is available, we have identified a number of topotype defining branches and the biological effect they are having; Figures 5.8 and 5.9.

The improved variable selection and modelling accuracy of the methods proposed in Chapters 4 and 6 has resulted in more biological accurate predictions of the antigenic residues and in the datasets for the FMDV serotypes we have identified a number of known and potential antigenic residues. In the SAT1 serotype, using the extended SAT1 dataset, we have been able to demonstrate the improved ability of the conjugate SABRE to select antigenic residues over the previous work using mixed-effects models. We were able to identify significantly more known antigenic residues than Maree et al. (2015), as well as make a number of predictions of other residues that are potentially antigenic. In the SAT2 serotype we made the first *in silico* prediction of potentially antigenic residues, with Reeve et al. (2010) unable to identify any significant residues. Within these prediction we were able to identify a number of potentially antigenic regions in need of further biological experimentation.

In the Influenza datasets we were able to demonstrate the effectiveness of the eSABRE method at properly accounting for the error inherent in the data collection process and make use of the full H1N1 and H3N2 datasets from Section 2.3 through the computational improvement the method offers. Our results on the H1N1 dataset show that we have identified a number of known antigenic residues from the residual binding site and each of the four known antigenic regions for that serotype. In the H3N2 dataset we have again identified a large number of proven variables at the cost of only a small number of implausible ones. We have identified residues from the residual binding site, as well as from three of the main antigenic regions. We have also proposed additional residues as antigenic in nearby areas of the virus shell.

8.3 Further Work

The models created and tested in this thesis give an accurate way of predicting antigenic variability in order to identify antigenic residues and have been shown to work effectively in both the FMDV and Influenza datasets. However increased accuracy and biological understanding can be gained by creating extended models which can better approximate the complex biological problem that we are modelling. From the biological viewpoint, it would be valuable to extend the models to better account for four aspects of the biological process associated with antigenic variability; (1) make better use of the genetic code of the virus strains, (2) link the effects of the residues to their location on the virus shell, (3) account better for the uncertainty in the phylogenetic trees, and (4) link the effects of the evolution and residues together in more realistic manner. Additionally, from the statistic methodology perspective, (5) it would be useful to improve the sampling of the latent binary variables, $\boldsymbol{\gamma}$, in order to gain faster parameter convergence in any models which could extend the eSABRE method.

At present our datasets, described in Chapter 3, consist only of indicators of mutational changes that occur without any regard to the type of mutation; see Section 2.1.2. This is addressed in Maree et al. (2015), (1), where the variables included indicate the change in the genetic code. Adding this information into the models will allow us to differentiate between different antigenic changes and enable us to better understand the biological processes involved.

Including more information relating to the genetic code will lead to more information, and therefore more variables, relating to the mutations being included in the models. It may therefore be necessary to add additional information sharing between the latent indicator variables, $\boldsymbol{\gamma}$, to avoid selecting variables whose correlation with changes in antigenicity are only through random chance, (2). Latent Gaussian processes can be used to model this, where inference can be achieved in a variety of ways, e.g. Filippone et al. (2013). The use of latent Gaussian processes would allow us to introduce correlations between mutations of the same type or mutations occurring in similar location on the surface of the virus shell. This can potentially allow us to identify which types of mutation are important, and give us the ability to identify complete antigenic regions rather than just individual residues.

We can also improve our model by better accounting for the uncertainty of the phylogenetic tree, (3). In this thesis we used single phylogentic trees taken from the original publication of the FMDV and Influenza datasets (Harvey et al., 2016; Maree et al., 2015; Reeve et al., 2010). In these papers, multiple trees were tested based on different biological models with the best one selected using Bayes factors; see Section 2.3.2 in Harvey (2016) for details. While choosing the best phylogenetic tree via Bayes factors may give a good estimate of the true evolution of the serotypes, it does not account for the uncertainty in this choice. Sampling different trees within our models provides one way of accounting for this uncertainty, however this is likely to be computationally infeasible and an approach based on model averaging may be more feasible.

While the eSABRE method (Chapter 6) better models the biological processes of antigenic variation then the SABRE methods (Chapter 4), the eSABRE does not fully account for the changes causing antigenic differences. Both the eSABRE and SABRE methods treat the residues and evolutionary changes as equally likely to cause changes in antigenicity, however this is an approximation of how the changes in antigenicity occur, (4). In fact, the mutational changes are what is used to create the phylogenetic trees in the first places, with the trees designed to best explain the genetic differences in the residues. Therefore a more realistic model should see the mutational changes at the residues explaining the antigenic effects of the phylogenetic branch terms, with the branch terms in turn explaining the the mean VN titre or HI assay measurement of each pair of challenge and protective strains, $\boldsymbol{\mu}_{\mathbf{y}}$. This would in essence require another layer in the likelihood, with the likelihood being given in the form $p(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}})p(\boldsymbol{\mu}_{\mathbf{y}}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\mathbf{w})$, where $\boldsymbol{\phi}$ represents the phylogenetic branch terms and $\mathbf{w}$ the residue terms.

To implement any of the biological changes suggested above in the eSABRE method would likely require an improvement in the sampling strategy to make the changes feasible, (5). In this thesis we have identified that the sampling of the latent indicator variables, $\boldsymbol{\gamma}$, is the computational bottleneck of both the SABRE and eSABRE methods and so we would need to design an improved proposal method beyond the block Metropolis-Hastings samplers tested in Section 4.3.5. For continuous variables, methods such as the Delayed Rejection Adaptive Metropolis (DRAM) algorithm of Haario et al. (2006) have been proposed to take into account the posterior correlations between the variables in the proposal scheme via a multivariate Gaussian distribution inferred from the accepted parameter vector. Finding a similar method to this for binary variables would be useful for achieving faster parameter convergence in the more complex, computationally onerous models proposed above.

Appendix A

Posterior Distributions

In this appendix we derive the conditional distributions from Section 4.3 and 6.2 needed to sample the parameters of the SABRE and eSABRE methods.

A.1 SABRE Methods

The conditional distribution derived here are laid out in a similar way to Section 4.3. In Section A.1.1 we give the conditional distributions needed to sample the parameters of the original SABRE method, with only the subsequent changes needed to adjust these distributions given for the semi-conjugate, conjugate and binary mask conjugate methods in Sections A.1.2, A.1.3 and A.1.4 respectively. Finally the conditional distributions needed for the collapsing scheme described in Section 4.3.6 are given for both the conjugate and binary mask conjugate SABRE method in Section A.1.5.

A.1.1 Original SABRE Method

Using standard results for conditional Gaussian distributions, e.g. Bishop (2006), and Figure 4.1, we can calculate the conditional distributions of $\mathbf{w}_\gamma$, $\mathbf{b}$ and $\mu_{w,h}$ for the original SABRE method, where we define $\boldsymbol{\theta}$ to be a vector of all the parameters and hyperparameters:

$$\begin{aligned} p(\mathbf{w}_\gamma|\boldsymbol{\theta}_{-\mathbf{w}_\gamma}, \mathbf{X}_\gamma, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{X}_\gamma\mathbf{w}_\gamma + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}_\gamma|\mathbf{m}_{\mathbf{w}_\gamma,\gamma}, \boldsymbol{\Sigma}_{\mathbf{w}_\gamma}) && \text{(A.1)} \\ &\propto \mathcal{N}(\mathbf{w}_\gamma|\mathbf{V}_{\mathbf{w}_\gamma}\mathbf{X}_\gamma^\top(\mathbf{y}-\mathbf{Z}\mathbf{b})/\sigma_\varepsilon^2 + \mathbf{V}_{\mathbf{w}_\gamma}\boldsymbol{\Sigma}_{\mathbf{w}_\gamma}^{-1}\boldsymbol{\mu}_\mathbf{w}, \mathbf{V}_{\mathbf{w}_\gamma}) && \text{(A.2)} \end{aligned}$$

where we define $\mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}} = (\mathbf{X}_{\boldsymbol{\gamma}}^{\top}\mathbf{X}_{\boldsymbol{\gamma}}/\sigma_{\varepsilon}^2 + \boldsymbol{\Sigma}_{\mathbf{w}}^{-1})^{-1}$,

$$
\begin{aligned}
p(\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_{\varepsilon}^2\mathbf{I})\mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}) && \text{(A.3)} \\
&\propto \mathcal{N}(\mathbf{b}|\mathbf{V}_{\mathbf{b}}\mathbf{Z}^{\top}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}})/\sigma_{\varepsilon}^2, \mathbf{V}_{\mathbf{b}}) && \text{(A.4)}
\end{aligned}
$$

where we define $\mathbf{V}_{\mathbf{b}} = (\mathbf{Z}^{\top}\mathbf{Z}/\sigma_{\varepsilon}^2 + \boldsymbol{\Sigma}_{\mathbf{b}}^{-1})^{-1}$, and

$$
\begin{aligned}
p(\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma},h}|\mathbf{1}\mu_{w,h}, \sigma_{\varepsilon}^2\sigma_{\mathbf{w}_{\boldsymbol{\gamma}},h}^2\mathbf{I})\mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{\varepsilon}^2\sigma_{0,h}^2) && \text{(A.5)} \\
&\propto \mathcal{N}(\mu_{w,h}|V_{\mu_{\gamma},h}(\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h})/\sigma_{w,h}^2 + \mu_{0,h}/\sigma_{0,h}^2), \sigma_{\varepsilon}^2 V_{\mu_{\gamma},h}) && \text{(A.6)}
\end{aligned}
$$

where $V_{\mu_{\gamma},h} = ((||\mathbf{w}_{\boldsymbol{\gamma},h}||/\sigma_{w,h}^2)^{-1} + (\sigma_{0,h}^2)^{-1})^{-1}$.

We can then calculate the conditional distributions of the variance parameters:

$$
\begin{aligned}
p(\sigma_{w,h}^2|\boldsymbol{\theta}_{-\sigma_{w,h}^2}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma},h}|\mathbf{1}\mu_{w,h}, \sigma_{\varepsilon}^2\sigma_{w,h}^2\mathbf{I})\mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h}) && \text{(A.7)} \\
&\propto \mathcal{IG}(\sigma_{w,h}^2|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h} + \tfrac{1}{2\sigma_{\varepsilon}^2}\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{\boldsymbol{\gamma},h})^2) && \text{(A.8)}
\end{aligned}
$$

where we sample for each h separately,

$$
\begin{aligned}
p(\sigma_{b,g}^2|\boldsymbol{\theta}_{-\sigma_{b,g}^2}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{b}_g|\mathbf{0}, \sigma_{b,g}^2\mathbf{I})\mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g}) && \text{(A.9)} \\
&\propto \mathcal{IG}(\sigma_{b,g}^2|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}_g^{\top}\mathbf{b}_g) && \text{(A.10)}
\end{aligned}
$$

where we sample for each g separately, and

$$
\begin{aligned}
p(\sigma_{\varepsilon}^2|\boldsymbol{\theta}_{-\sigma_{\varepsilon}^2}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\sim \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_{\varepsilon}^2\mathbf{I})\mathcal{IG}(\sigma_{\varepsilon}^2|\alpha_{\varepsilon}, \beta_{\varepsilon}) && \text{(A.11)} \\
&\propto \mathcal{IG}(\sigma_{\varepsilon}^2|N/2 + \alpha_{\varepsilon}, \beta_{\varepsilon} + \tfrac{1}{2}\Sigma(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})^2) && \text{(A.12)}
\end{aligned}
$$

.

We can then get the conditional distribution of π as follows:

$$
\begin{aligned}
p(\pi|\boldsymbol{\theta}_{-\pi}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \left\{\prod_{j=1}^{J}\text{Bern}(\gamma_j|\pi)\right\}\mathcal{B}(\pi|\alpha_{\pi}, \beta_{\pi}) && \text{(A.13)} \\
&\propto \mathcal{B}(\pi|\alpha_{\pi} + \Sigma\boldsymbol{\gamma}, \beta_{\pi} + J - \Sigma\boldsymbol{\gamma}) && \text{(A.14)}
\end{aligned}
$$

and finally, via the application of standard Gaussian integrals, we have the distribution for $\boldsymbol{\gamma}$ as derived in Section 4.3.1:

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) \propto \text{Bern}(\boldsymbol{\gamma}|\pi) \int \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})d\mathbf{w}_{\boldsymbol{\gamma}} \tag{A.15}$$

$$\propto \pi^{\sum\boldsymbol{\gamma}}(1-\pi)^{J-\sum\boldsymbol{\gamma}}\mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\mathbf{w}} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I} + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{X}_{\boldsymbol{\gamma}}^\top). \tag{A.16}$$

.

A.1.2 Semi-Conjugate SABRE Method

The differences between the original SABRE method and the semi-conjugate SABRE method can be seen by comparing Figures 4.1 and 4.2 in Chapter 4. To get the conditional distributions for the semi-conjugate SABRE method we start with those given for the original SABRE method in Section A.1.1 and replace (A.2), (A.4), (A.12) and (A.16) with the distributions given below:

$$p(\mathbf{w}_{\boldsymbol{\gamma}}^*|\boldsymbol{\theta}_{-\mathbf{w}_{\boldsymbol{\gamma}}^*}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^* + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}^*|\mathbf{m}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}) \tag{A.17}$$

$$\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}^*|\mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\mathbf{X}_{\boldsymbol{\gamma}}^\top(\mathbf{y} - \mathbf{Zb})/\sigma_\varepsilon^2 + \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}}^{-1}\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}) \tag{A.18}$$

where we define $\mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}^*} = (\mathbf{X}_{\boldsymbol{\gamma}}^{*,\top}\mathbf{X}_{\boldsymbol{\gamma}}^*/\sigma_\varepsilon^2 + \boldsymbol{\Sigma}_{\mathbf{w}^*}^{-1})^{-1}$,

$$p(\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^* + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}) \tag{A.19}$$

$$\propto \mathcal{N}(\mathbf{b}|\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^*)/\sigma_\varepsilon^2, \mathbf{V}_{\mathbf{b}}) \tag{A.20}$$

where we again define $\mathbf{V}_{\mathbf{b}} = (\mathbf{Z}^\top\mathbf{Z}/\sigma_\varepsilon^2 + \boldsymbol{\Sigma}_{\mathbf{b}}^{-1})^{-1}$,

$$p(\sigma_\varepsilon^2|\boldsymbol{\theta}_{-\sigma_\varepsilon^2}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^* + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})\mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon) \tag{A.21}$$

$$\propto \mathcal{IG}(\sigma_\varepsilon^2|N/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}\Sigma(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^* - \mathbf{Zb})^2), \tag{A.22}$$

and finally the conditional distribution for $\boldsymbol{\gamma}$ original derived in Section 4.3.2

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto \int \beta(\pi|\alpha_\pi, \beta_\pi)\,\text{Bern}(\boldsymbol{\gamma}|\pi)$$

$$\mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{w}_{\boldsymbol{\gamma}}^* + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}^*|\mathbf{m}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*})d\pi d\mathbf{w}_{\boldsymbol{\gamma}} \tag{A.23}$$

$$\propto \tfrac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}\mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \sigma_\varepsilon^2\mathbf{I} + \mathbf{X}_{\boldsymbol{\gamma}}^*\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}\mathbf{X}_{\boldsymbol{\gamma}}^{*\top}). \tag{A.24}$$

A.1.3 Conjugate SABRE Method

The differences between the semi-conjugate SABRE method and the conjugate SABRE method can be seen by comparing Figures 4.2 and 4.3 in Chapter 4. To get the conditional distributions for the conjugate SABRE method we start with those used for the semi-conjugate SABRE method in Sections A.1.1 and A.1.2. We then replace (A.18), (A.6), (A.8), (A.22) and (A.24) with the following conditional distributions:

$$
\begin{aligned}
p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{-\mathbf{w}^*_{\boldsymbol{\gamma}}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \qquad \text{(A.25)}\\
&\propto \mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}(\mathbf{y} - \mathbf{Z}\mathbf{b}) + \mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \qquad \text{(A.26)}
\end{aligned}
$$

where $\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = (\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})^{-1}$,

$$
\begin{aligned}
p(\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma},h}|\mathbf{1}\mu_{w,h}, \sigma^2_\varepsilon\sigma^2_{\mathbf{w}_{\boldsymbol{\gamma}},h}\mathbf{I})\mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma^2_\varepsilon\sigma^2_{0,h}) \qquad \text{(A.27)}\\
&\propto \mathcal{N}(\mu_{w,h}|V_{\mu_{\boldsymbol{\gamma}},h}(\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h})/\sigma^2_{w,h} + \mu_{0,h}/\sigma^2_{0,h}), \sigma^2_\varepsilon V_{\mu_{\boldsymbol{\gamma}},h}) \qquad \text{(A.28)}
\end{aligned}
$$

where $V_{\mu_{\boldsymbol{\gamma}},h} = ((||\mathbf{w}_{\boldsymbol{\gamma},h}||/\sigma^2_{w,h})^{-1} + (\sigma^2_{0,h})^{-1})^{-1}$, and

$$
\begin{aligned}
p(\sigma^2_{w,h}|\boldsymbol{\theta}_{-\sigma^2_{w,h}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma},h}|\mathbf{1}\mu_{w,h}, \sigma^2_\varepsilon\sigma^2_{w,h}\mathbf{I})\mathcal{IG}(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h}) \qquad \text{(A.29)}\\
&\propto \mathcal{IG}(\sigma^2_{w,h}|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h} + \tfrac{1}{2\sigma^2_\varepsilon}\Sigma(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{\boldsymbol{\gamma},h})^2) \qquad \text{(A.30)}
\end{aligned}
$$

where we sample for each h separately.

We can then find the distribution for σ^2_ε, defining $\boldsymbol{\mu_0} = (\mu_{0,1}, \dots, \mu_{0,H})^\top$ and $\boldsymbol{\Sigma_0} = diag(\sigma^2_{0,1}, \dots, \sigma^2_{0,H})$:

$$
\begin{aligned}
p(\sigma^2_\varepsilon|\boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})\\
&\qquad \times \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}|\boldsymbol{\mu}_0, \sigma^2_\varepsilon\boldsymbol{\Sigma_0})\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon) \qquad \text{(A.31)}\\
&\propto \mathcal{IG}(\sigma^2_\varepsilon|(N + ||\mathbf{w}^*_{\boldsymbol{\gamma}}|| + H)/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}R_{\sigma^2_\varepsilon}). \qquad \text{(A.32)}
\end{aligned}
$$

where H is the number of groups of regressors and

$$
\begin{aligned}
R_{\sigma^2_\varepsilon} &= (\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})\\
&\quad + (\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}})^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}(\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}}) + (\boldsymbol{\mu}_{\mathbf{w}} - \boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}(\boldsymbol{\mu}_{\mathbf{w}} - \boldsymbol{\mu}_0) \qquad \text{(A.33)}
\end{aligned}
$$

In order to improve mixing and convergence, Davies et al. (2014) used a collapsing step over $\mathbf{w}^*_{\boldsymbol{\gamma}}$ when sampling $\boldsymbol{\gamma}$, via the application of standard Gaussian integrals, e.g. Bishop (2006), following Sabatti and James (2005). Doing this should result in an improvement in computational efficiency and we have therefore also integrated over π here via an application of Beta-Bernoulli models:

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) \propto \int p(\boldsymbol{\gamma}, \pi, \mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) d\mathbf{w}^*_{\boldsymbol{\gamma}} d\pi \tag{A.34}$$

$$\propto \int p(\boldsymbol{\gamma}|\pi)p(\pi)p(\mathbf{y}|\mathbf{w}^*_{\boldsymbol{\gamma}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})p(\mathbf{w}^*_{\boldsymbol{\gamma}})d\mathbf{w}^*_{\boldsymbol{\gamma}} d\pi \tag{A.35}$$

$$\propto \int \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon \boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \left\{\prod_{j=1}^{J} \text{Bern}(\gamma_j|\pi)\right\} \mathcal{B}(\pi|\alpha_\pi, \beta_\pi)d\mathbf{w}^*_{\boldsymbol{\gamma}} d\pi \tag{A.36}$$

$$\propto \frac{\Gamma(||\boldsymbol{\gamma}|| + \alpha_\pi)\Gamma(J - ||\boldsymbol{\gamma}|| + \beta_\pi)}{\Gamma(J + \alpha_\pi + \beta_\pi)} \int \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon \boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})d\mathbf{w}_{\boldsymbol{\gamma}} \tag{A.37}$$

$$\propto \frac{\Gamma(||\boldsymbol{\gamma}|| + \alpha_\pi)\Gamma(J - ||\boldsymbol{\gamma}|| + \beta_\pi)}{\Gamma(J + \alpha_\pi + \beta_\pi)} \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma^2_\varepsilon[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}]). \tag{A.38}$$

In addition to the conditional distributions for the standard conjugate SABRE method, we also need to calculate the conditional distributions for the half-t random-effect priors as follows:

$$p(\boldsymbol{\eta}|\boldsymbol{\theta}_{-\boldsymbol{\eta}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\boldsymbol{\eta}\xi, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \tag{A.39}$$

$$\propto \mathcal{N}(\boldsymbol{\eta}|\tfrac{\xi}{\sigma^2_\varepsilon}\mathbf{V}_{\boldsymbol{\eta}}\mathbf{Z}^\top(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}), \mathbf{V}_{\boldsymbol{\eta}}) \tag{A.40}$$

where $\mathbf{V}_{\boldsymbol{\eta}} = (\frac{\xi^2}{\sigma^2_\varepsilon}\mathbf{Z}^\top\mathbf{Z} + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\eta}})^{-1}$.

$$p(\xi|\boldsymbol{\theta}_{-\xi}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\boldsymbol{\eta}\xi, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\xi|\mu_\xi, \sigma^2_\xi) \tag{A.41}$$

$$\propto \mathcal{N}(\xi|V_\xi[\tfrac{\mu_\xi}{\sigma^2_\xi} + \tfrac{1}{\sigma^2_\varepsilon}\boldsymbol{\eta}^\top\mathbf{Z}^\top(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}})], V_\xi) \tag{A.42}$$

where $V_\xi = (\frac{1}{\sigma^2_\xi} + \frac{1}{\sigma^2_\varepsilon}\boldsymbol{\eta}^\top\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\eta})^{-1}$.

$$p(\sigma^2_{\eta,g}|\boldsymbol{\theta}_{-\sigma^2_{\eta,g}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\boldsymbol{\eta}_g|\mathbf{0}, \sigma^2_{\eta,g}\mathbf{I})\mathcal{IG}(\sigma^2_{\eta,g}|\alpha_{\eta,g}, \beta_{\eta,g}) \tag{A.43}$$

$$\propto \mathcal{IG}(\sigma^2_{\eta,g}|||\boldsymbol{\eta}_g||/2 + \alpha_{\eta,g}, \beta_{\eta,g} + \tfrac{1}{2}\boldsymbol{\eta}_g^\top\boldsymbol{\eta}_g) \tag{A.44}$$

where we sample for each g separately. These distributions replace (A.20) and (A.10) in the sampling scheme of the conjugate SABRE method described above, and we additionally set $\mathbf{b} = \boldsymbol{\eta}\xi$ and $\sigma_{b,g}^2 = \xi^2\sigma_{\eta,g}^2$ in the other conditional distributions.

A.1.4 Binary Mask Conjugate SABRE Method

The differences between the conjugate SABRE method and the binary mask conjugate SABRE method can be seen by comparing Figures 4.3 and 4.4 in Chapter 4. While the models are reasonably similar the conditional distributions are not with only the distributions for $\sigma_{b,g}^2$ and π remaining the same; (A.10) and (A.14). Here we give the remaining distributions required for the binary mask conjugate SABRE method:

$$\begin{aligned} p(\mathbf{w}^*|\boldsymbol{\theta}_{-\mathbf{w}^*}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}\boldsymbol{\Gamma}\mathbf{w} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*|\mathbf{m}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*}) && \text{(A.45)}\\ &\propto \mathcal{N}(\mathbf{w}^*|\mathbf{V}_{\mathbf{w}^*}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}(\mathbf{y} - \mathbf{Z}\mathbf{b}) + \mathbf{V}_{\mathbf{w}^*}\boldsymbol{\Sigma}_{\mathbf{w}^*}^{-1}\mathbf{m}, \sigma_\varepsilon^2\mathbf{V}_{\mathbf{w}^*}) && \text{(A.46)} \end{aligned}$$

where we define $\mathbf{V}_{\mathbf{w}^*} = (\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}\mathbf{X}^*\boldsymbol{\Gamma} \ + \boldsymbol{\Sigma}_{\mathbf{w}^*}^{-1})^{-1}$,

$$\begin{aligned} p(\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}\boldsymbol{\Gamma}\mathbf{w} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}) && \text{(A.47)}\\ &\propto \mathcal{N}(\mathbf{b}|\tfrac{1}{\sigma_\varepsilon^2}\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top(\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^*), \mathbf{V}_{\mathbf{b}}) && \text{(A.48)} \end{aligned}$$

where we define $\mathbf{V}_{\mathbf{b}} = (\frac{1}{\sigma_\varepsilon^2}\mathbf{Z}^\top\mathbf{Z} + \boldsymbol{\Sigma}_{\mathbf{b}}^{-1})^{-1}$,

$$\begin{aligned} p(\mu_{w,h}|\boldsymbol{\theta}_{-\mu_{w,h}}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_h|\mathbf{1}\mu_{w,h}, \sigma_\varepsilon^2\sigma_{\mathbf{w}_\gamma,h}^2\mathbf{I})\mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_\varepsilon^2\sigma_{0,h}^2) && \text{(A.49)}\\ &\propto \mathcal{N}(\mu_{w,h}|V_{\mu,h}^{-1}(\Sigma(\mathbf{w}_h)/\sigma_{w,h}^2 + \mu_{0,h}/\sigma_{0,h}^2), \sigma_\varepsilon^2 V_{\mu,h}) && \text{(A.50)} \end{aligned}$$

where we define $V_{\mu,h} = ((||\mathbf{w}_h||/\sigma_{w,h}^2)^{-1} + (\sigma_{0,h}^2)^{-1})^{-1}$ and sample separately for each h,

$$\begin{aligned} p(\sigma_{w,h}^2|\boldsymbol{\theta}_{-\sigma_{w,h}^2}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_h|\mathbf{1}\mu_{w,h}, \sigma_\varepsilon^2\sigma_{\mathbf{w}_\gamma,h}^2\mathbf{I})\mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h}) && \text{(A.51)}\\ &\propto \mathcal{IG}(\sigma_{w,h}^2|\ ||\mathbf{w}_h||/2 + \alpha_{w,h}, \beta_{w,h} + \tfrac{1}{2\sigma_\varepsilon^2}(\mathbf{w}_h - \mathbf{1}\mu_{w,h})^\top(\mathbf{w}_h - \mathbf{1}\mu_{w,h})) && \text{(A.52)} \end{aligned}$$

where we again sample separately for each h, and

$$\begin{aligned} p(\sigma_\varepsilon^2|\boldsymbol{\theta}_{-\sigma_\varepsilon^2}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}\boldsymbol{\Gamma}\mathbf{w} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*|\mathbf{m}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*})\\ &\qquad \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}|\boldsymbol{\mu}_0, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{0}})\mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon) && \text{(A.53)}\\ &\propto \mathcal{IG}(\sigma_\varepsilon^2|(N + ||\mathbf{w}^*|| + H)/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}R_{\sigma_\varepsilon^2}) && \text{(A.54)} \end{aligned}$$

where $R_{\sigma_\varepsilon^2} = (\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^* - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^* - \mathbf{Z}\mathbf{b}) + (\mathbf{w}^* - \mathbf{m})^\top\boldsymbol{\Sigma}_{\mathbf{w}^*}^{-1}(\mathbf{w}^* - \mathbf{m}) + (\boldsymbol{\mu}_\mathbf{w} - \boldsymbol{\mu}_\mathbf{0})^\top\boldsymbol{\Sigma}_\mathbf{0}^{-1}(\boldsymbol{\mu}_\mathbf{w} - \boldsymbol{\mu}_\mathbf{0})$.

Finally the distribution of $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) \propto \int & \beta(\pi|\alpha_\pi, \beta_\pi)\,\mathrm{Bern}(\boldsymbol{\gamma}|\pi) \\ & \mathcal{N}(\mathbf{y}|\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{w}^* + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*|\mathbf{m}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*})d\pi d\mathbf{w}_{\boldsymbol{\gamma}} \qquad \text{(A.55)} \\ \propto & \tfrac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}\mathcal{N}(\mathbf{y}|\mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2[\mathbf{I} + \mathbf{X}^*\boldsymbol{\Gamma}^*\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}]) \qquad \text{(A.56)} \end{aligned}$$

as originally defined in Section 4.3.4.

A.1.5 Conjugate Sampling Scheme

In the conjugate and binary mask conjugate model we can make use of the conjugate sampling strategy proposed in Section 4.3.6. In the conjugate sampling scheme, the conditional distribution of $\boldsymbol{\gamma}$ is found by integrating over both σ_ε^2 and $\boldsymbol{\mu}_\mathbf{w}$as well as those parameters marginalised Section A.1.3 and A.1.4; $\mathbf{w}^*_{\boldsymbol{\gamma}}$ and π. This collapsing is possible due to the conjugate prior specification of $\mathbf{w}^*_{\boldsymbol{\gamma}}$ and $\boldsymbol{\mu}_\mathbf{w}$ in both methods; see Figures 4.3 and 4.4. This step is not feasible in either the original SABRE method or the semi-conjugate SABRE method.

The distribution of $\boldsymbol{\gamma}$ for the conjugate SABRE method is given as follows:

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) &\propto \int p(\boldsymbol{\gamma}, \pi, \sigma_\varepsilon^2, \mathbf{w}^*_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_\mathbf{w}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})d\boldsymbol{\mu}_\mathbf{w}d\mathbf{w}^*_{\boldsymbol{\gamma}}d\pi d\sigma_\varepsilon^2 \qquad \text{(A.57)} \\ &\propto \int p(\boldsymbol{\gamma}|\pi)p(\pi)p(\mathbf{y}|\mathbf{w}^*_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_\mathbf{w}, \sigma_\varepsilon^2)p(\boldsymbol{\mu}_\mathbf{w})p(\sigma_\varepsilon^2)d\boldsymbol{\mu}_\mathbf{w}d\mathbf{w}^*_{\boldsymbol{\gamma}}d\pi d\sigma_\varepsilon^2 \qquad \text{(A.58)} \\ &\propto C_\pi \int \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})\mathcal{N}(\boldsymbol{\mu}_\mathbf{w}|\boldsymbol{\mu}_0, \sigma_\varepsilon^2\boldsymbol{\Sigma}_\mathbf{0}) \\ &\qquad \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)d\boldsymbol{\mu}_\mathbf{w}d\mathbf{w}^*_{\boldsymbol{\gamma}}d\sigma_\varepsilon^2 \qquad \text{(A.59)} \\ &\propto C_\pi \int \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}}, \sigma_\varepsilon^2[\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]) \\ &\qquad \mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)d\mathbf{w}^*_{\boldsymbol{\gamma}}d\sigma_\varepsilon^2 \qquad \text{(A.60)} \\ &\propto C_\pi \int \mathcal{N}(\mathbf{y}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}[\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]\mathbf{X}^{*,\top}_{\boldsymbol{\gamma}}])\mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon)d\sigma_\varepsilon^2 \qquad \text{(A.61)} \\ &\propto C_\pi|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}}[\beta_\varepsilon + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Z}\mathbf{b})^\top\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Z}\mathbf{b})]^{-(N/2+\alpha_\varepsilon)} \qquad \text{(A.62)} \end{aligned}$$

where $C_\pi = \frac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = [\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}[\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}]$, $\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} = (\mu_{w_0}, \mu_{0,1}, \ldots, \mu_{0,1}, \mu_{0,2}, \ldots, \mu_{0,H})^\top$ with each $\mu_{0,h}$ repeated with length $||\mathbf{w}_{\boldsymbol{\gamma},h}||$ dependent on $\boldsymbol{\gamma}$, and

$\mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}$ is a block diagonal matrix of $(0, \sigma_{0,1}^2, \sigma_{0,2}^2, \ldots, \sigma_{0,H}^2)$ where the square blocks have length $1, ||\mathbf{w}_{\boldsymbol{\gamma},1}||, \ldots,\ ||\mathbf{w}_{\boldsymbol{\gamma},H}||$ respectively.

We can use the Woodbury identity and the extended Sylvester's determinant theorem to speed up the computations and give the following conditional posterior distribution:

$$\begin{aligned}\log\ p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto{}& \log\Gamma(||\boldsymbol{\gamma}|| + \alpha_\pi) + \log\Gamma(J - ||\boldsymbol{\gamma}|| + \beta_\pi)\\ &- \log\Gamma(J + \alpha_\pi + \beta_\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}| - \tfrac{1}{2}\log|[\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]^{-1} + \mathbf{X}_{\boldsymbol{\gamma}}^{*\top}\mathbf{X}_{\boldsymbol{\gamma}}^*|\\ &- (\tfrac{N}{2} + \alpha_\varepsilon)\log(\beta_\varepsilon + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Zb})^\top\\ &[\mathbf{I} - \mathbf{X}_{\boldsymbol{\gamma}}^*([\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]^{-1} + \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{X}_{\boldsymbol{\gamma}}^{*\top})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^{*\top})](\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Zb})).\end{aligned} \tag{A.63}$$

This was also done with the conditional distribution of $\boldsymbol{\gamma}$ for the original and semi-conjugate SABRE methods in Sections A.1.1 and A.1.2.

In addition to the conditional distribution of $\boldsymbol{\gamma}$ we must also derive distributions for σ_ε^2 and $\boldsymbol{\mu}_\mathbf{w}$. We do not need to derive conditional distributions for $\mathbf{w}_{\boldsymbol{\gamma}}$ and π as they are identical to those given in (A.26) and (A.14).

$$p(\sigma_\varepsilon^2|\boldsymbol{\gamma}, \boldsymbol{\theta}_{-\sigma_\varepsilon^2}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} + \mathbf{Zb}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\boldsymbol{\gamma}})\mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon) \tag{A.64}$$

$$\propto \mathcal{IG}(\sigma_\varepsilon^2|||\mathbf{y}||/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Zb})^\top\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}^*\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} - \mathbf{Zb})) \tag{A.65}$$

where the first distribution is taken from the derivation of the conditional distribution of $\boldsymbol{\gamma}$.

$$p(\boldsymbol{\mu}_\mathbf{w}|\sigma_\varepsilon^2, \boldsymbol{\gamma}, \boldsymbol{\theta}_{-\boldsymbol{\mu}_\mathbf{w}}, \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y})$$

$$\propto \mathcal{N}(\mathbf{y}|\mathbf{1}\mu_{w_0} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}}\boldsymbol{\mu}_\mathbf{w} + \mathbf{Zb}, \sigma_\varepsilon^2[\mathbf{I} + \mathbf{X}_{\boldsymbol{\gamma}}^*\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}\mathbf{X}_{\boldsymbol{\gamma}}^{*\top}])\mathcal{N}(\boldsymbol{\mu}_\mathbf{w}|\boldsymbol{\mu}_\mathbf{0}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_\mathbf{0}) \tag{A.66}$$

$$\propto \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}|\mathbf{V}_{\boldsymbol{\mu}_{\boldsymbol{\gamma},\mathbf{w}}}[\boldsymbol{\Sigma}_\mathbf{0}^{-1}\boldsymbol{\mu}_\mathbf{0} + \mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}}^\top\mathbf{X}_{\boldsymbol{\gamma}}^\top[\mathbf{I} + \mathbf{X}_{\boldsymbol{\gamma}}^*\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}\mathbf{X}_{\boldsymbol{\gamma}}^{*\top}]^{-1}(\mathbf{y} - \mathbf{1}\mu_{w_0} - \mathbf{Zb})], \sigma_\varepsilon^2\mathbf{V}_{\boldsymbol{\mu}_{\boldsymbol{\gamma},\mathbf{w}}}) \tag{A.67}$$

where the first distribution is again taken from the derivation of the conditional distribution of $\boldsymbol{\gamma}$ and $\mathbf{V}_{\boldsymbol{\mu}_{\boldsymbol{\gamma},\mathbf{w}}} = [\boldsymbol{\Sigma}_\mathbf{0}^{-1} + \mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}}^\top\mathbf{X}_{\boldsymbol{\gamma}}^\top[\mathbf{I} + \mathbf{X}_{\boldsymbol{\gamma}}^*\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}^*}\mathbf{X}_{\boldsymbol{\gamma}}^{*\top}]^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}}]^{-1}$. $\mathbf{M}_{\boldsymbol{\mu}}$, required for (A.72), is a matrix of indicators where each element $m_{\boldsymbol{\mu},j,h}$ is 1 for any $w_{j,h}$ in group h and 0 otherwise, where $\mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}}$ only includes the relevant elements dependent on $\boldsymbol{\gamma}$. For example:

$$\mathbf{M}_{\boldsymbol{\mu}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \;;\; \mathbf{M}_{\boldsymbol{\gamma},\boldsymbol{\mu}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \;;\; \mathbf{w} = \begin{bmatrix} w_{1,1} \\ w_{2,1} \\ w_{3,2} \\ w_{4,2} \\ w_{5,2} \end{bmatrix} \;;\; \mathbf{w}_{\boldsymbol{\gamma}} = \begin{bmatrix} w_{1,1} \\ w_{3,2} \\ w_{5,2} \end{bmatrix} \;;\; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 = 1 \\ \gamma_2 = 0 \\ \gamma_3 = 1 \\ \gamma_4 = 0 \\ \gamma_5 = 1 \end{bmatrix}. \quad \text{(A.68)}$$

We can calculate the log conditional distribution of $\boldsymbol{\gamma}$ for the binary mask conjugate SABRE method the same way we did for the conjugate SABRE method:

$$\begin{aligned} \log p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) \propto \log \int & p(\boldsymbol{\gamma}|\pi)p(\pi)p(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\gamma}, \mathbf{b}, \sigma_\varepsilon^2, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})p(\mathbf{w}^*|\boldsymbol{\mu}_{\mathbf{w}}, \sigma_\varepsilon^2) \\ & p(\boldsymbol{\mu}_{\mathbf{w}}|\sigma_\varepsilon^2)p(\sigma_\varepsilon^2)d\boldsymbol{\mu}_{\mathbf{w}}d\mathbf{w}^*_{\boldsymbol{\gamma}}d\pi d\sigma_\varepsilon^2 \qquad \text{(A.69)} \end{aligned}$$

$$\begin{aligned} &\propto \log\Gamma(||\boldsymbol{\gamma}|| + \alpha_\pi) + \log\Gamma(J - ||\boldsymbol{\gamma}|| + \beta_\pi) \\ &- \log\Gamma(J + \alpha_\pi + \beta_\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{w}^*} + \mathbf{V}_{\mathbf{0}}| - \tfrac{1}{2}\log|[\boldsymbol{\Sigma}_{\mathbf{w}^*} + \mathbf{V}_{\mathbf{0}}]^{-1} + \boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}\mathbf{X}\boldsymbol{\Gamma}^*| \\ &- (\tfrac{N}{2} + \alpha_\varepsilon)\log(\beta_\varepsilon + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m}_{\mathbf{0}} - \mathbf{Z}\mathbf{b})^\top \\ &[\mathbf{I} - \mathbf{X}^*\boldsymbol{\Gamma}^{*\top}([\boldsymbol{\Sigma}_{\mathbf{w}^*} + \mathbf{V}_{\mathbf{0}}]^{-1} + \mathbf{X}^*\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top})^{-1}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^*](\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m}_{\mathbf{0}} - \mathbf{Z}\mathbf{b})). \quad \text{(A.70)} \end{aligned}$$

where $\mathbf{m}_{\mathbf{0}} = (\mu_{w_0}, \mu_{0,1}, \ldots, \mu_{0,1}, \mu_{0,2}, \ldots, \mu_{0,H})^\top$ with each $\mu_{0,h}$ repeated with length $||\mathbf{w}_h||$ not dependant on $\boldsymbol{\gamma}$ and $\mathbf{V}_{\mathbf{0}}$ is a block diagonal matrix of $(0, \sigma_{0,1}^2, \sigma_{0,2}^2, \ldots, \sigma_{0,H}^2)$ where the square blocks have length $1, ||\mathbf{w}_1||, \ldots ||\mathbf{w}_H||$ respectively.

Finally we can calculate the collapsing steps for the conditional distributions of σ_ε^2 and $\boldsymbol{\mu}_{\mathbf{w}}$:

$$\begin{aligned} & p(\sigma_\varepsilon^2|\boldsymbol{\gamma}_{-\sigma_\varepsilon^2}, \boldsymbol{\theta}_{-\sigma_\varepsilon^2}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) \\ & \propto \mathcal{IG}(\sigma_\varepsilon^2|||\mathbf{y}||/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m}_{\mathbf{0}} - \mathbf{Z}\mathbf{b})^\top\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}(\mathbf{y} - \mathbf{X}^*\boldsymbol{\Gamma}^*\mathbf{m}_{\mathbf{0}} - \mathbf{Z}\mathbf{b})) \qquad \text{(A.71)} \\ & p(\boldsymbol{\mu}_{\mathbf{w}}|\sigma_\varepsilon^2, \boldsymbol{\gamma}, \boldsymbol{\theta}_{-\boldsymbol{\mu}_{\mathbf{w}}}, \mathbf{X}^*, \mathbf{Z}, \mathbf{y}) \propto \\ & \quad \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}|\mathbf{V}_{\boldsymbol{\mu}_{\mathbf{w}}}[\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}\boldsymbol{\mu}_{\mathbf{0}} + \mathbf{M}_{\boldsymbol{\mu}}^\top\boldsymbol{\Gamma}^\top\mathbf{X}^\top[\mathbf{I} + \mathbf{X}^*\boldsymbol{\Gamma}^*\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}]^{-1}(\mathbf{y} - \mathbf{1}\mu_{w_0} - \mathbf{Z}\mathbf{b})], \sigma_\varepsilon^2\mathbf{V}_{\boldsymbol{\mu}_{\mathbf{w}}}) \\ & \qquad \text{(A.72)} \end{aligned}$$

where $\mathbf{V}_{\boldsymbol{\mu}_{\mathbf{w}}} = [\boldsymbol{\Sigma}_{\mathbf{0}}^{-1} + \mathbf{M}_{\boldsymbol{\mu}}^\top\mathbf{X}\boldsymbol{\Gamma}^\top[\mathbf{I} + \mathbf{X}^*\boldsymbol{\Gamma}^*\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Gamma}^{*\top}\mathbf{X}^{*\top}]^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{M}_{\boldsymbol{\mu}}]^{-1}$.

A.2 eSABRE Method

The conditional distributions for the eSABRE method in Chapter 6 can again be found by using some of the basic results from standard textbooks, e.g. Murphy (2012), where we define $\mathbf{X}^*_{\boldsymbol{\gamma}} = (\mathbf{1}, \mathbf{X}_{\boldsymbol{\gamma}})$, $\mathbf{m}_{\boldsymbol{\gamma}} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \ldots, \sigma^2_{w,1}, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$.

Using standard results for conditional Gaussian distributions and Figure 6.1, we can calculate the conditional distributions for $\boldsymbol{\mu}_{\mathbf{y}}$, $\mathbf{w}^*_{\boldsymbol{\gamma}}$, $\mathbf{b}$ and μ_w , where we define $\boldsymbol{\theta}$ to be a vector of all the parameters and hyperparameters::

$$p(\boldsymbol{\mu}_{\mathbf{y}}|\boldsymbol{\theta}_{-\boldsymbol{\mu}_{\mathbf{y}}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{Z}\mathbf{b}, \sigma^2_y\mathbf{I})\mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\mathbf{I}) \tag{A.73}$$

$$\propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{V}_{\mathbf{y}}(\mathbf{M}^\top(\mathbf{y} - \mathbf{Z}\mathbf{b})/\sigma^2_y + \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}/\sigma^2_\varepsilon), \mathbf{V}_{\mathbf{y}}) \tag{A.74}$$

where $\mathbf{V}_{\mathbf{y}} = (1/\sigma^2_\varepsilon\mathbf{I} + \mathbf{M}^\top\mathbf{M}/\sigma^2_y)^{-1}$.

$$p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{-\mathbf{w}^*_{\boldsymbol{\gamma}}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \tag{A.75}$$

$$\propto \mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}) \tag{A.76}$$

where $\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = (\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})^{-1}$.

$$p(\mathbf{b}|\boldsymbol{\theta}_{-\mathbf{b}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{Z}\mathbf{b}, \sigma^2_y\mathbf{I})\mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_b) \tag{A.77}$$

$$\propto \mathcal{N}(\mathbf{b}|\tfrac{1}{\sigma^2_y}\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}}), \mathbf{V}_{\mathbf{b}}) \tag{A.78}$$

where $\mathbf{V}_{\mathbf{b}} = (\frac{1}{\sigma^2_y}\mathbf{Z}^\top\mathbf{Z} + \boldsymbol{\Sigma}^{-1}_{\mathbf{b}})^{-1}$.

$$p(\mu_w|\boldsymbol{\theta}_{-\mu_w}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\mathbf{1}\mu_w, \sigma^2_\varepsilon\sigma^2_w\mathbf{I})\mathcal{N}(\mu_w|\mu_0, \sigma^2_0\sigma^2_\varepsilon) \tag{A.79}$$

$$\propto \mathcal{N}(\mu_w|V_{\mu_w}(\mathbf{1}\mathbf{w}_{\boldsymbol{\gamma}}/\sigma^2_w + \mu_0/\sigma^2_0), \sigma^2_\varepsilon V_{\mu_w}) \tag{A.80}$$

where $V_{\mu_w} = (1/\sigma^2_0 + ||\mathbf{w}_{\boldsymbol{\gamma}}||/\sigma^2_w)^{-1}$.

We can then calculate the conditional distributions of the variance parameters:

$$p(\sigma^2_y|\boldsymbol{\theta}_{\sigma^2_y}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{Z}\mathbf{b}, \sigma^2_y\mathbf{I})\mathcal{IG}(\sigma^2_y|\alpha_y, \beta_y) \tag{A.81}$$

$$\propto \mathcal{IG}(\sigma^2_y| \ ||\mathbf{y}||/2 + \alpha_y, \tfrac{1}{2}(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{Z}\mathbf{b})) \tag{A.82}$$

$$
\begin{aligned}
p(\sigma_w^2|\boldsymbol{\theta}_{-\sigma_w^2}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\mathbf{I}\mu_w, \sigma_\varepsilon^2\sigma_w^2\mathbf{I})\mathcal{IG}(\sigma_w^2|\alpha_w, \beta_w) && \text{(A.83)}\\
&\propto \mathcal{IG}(\sigma_w^2|\ ||\mathbf{w}_{\boldsymbol{\gamma}}||/2 + \alpha_w, \tfrac{1}{2\sigma_\varepsilon^2}(\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{I}\mu_w)^\top(\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{I}\mu_w)) && \text{(A.84)}
\end{aligned}
$$

$$
\begin{aligned}
p(\sigma_{b,g}^2|\boldsymbol{\theta}_{-\sigma_{b,g}^2}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) &\propto \mathcal{N}(\mathbf{b}_g|\mathbf{0}, \sigma_{b,g}^2\mathbf{I})\mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g}) && \text{(A.85)}\\
&\propto \mathcal{IG}(\sigma_{b,g}^2|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}_g^\top\mathbf{b}_g) && \text{(A.86)}
\end{aligned}
$$

where we sample for each g separately.

$$
\begin{aligned}
&p(\sigma_\varepsilon^2|\boldsymbol{\theta}_{-\sigma_\varepsilon^2}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y})\\
&\qquad \propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})\mathcal{N}(\mu_{\mathbf{w}}|\mu_0, \sigma_\varepsilon^2\sigma_0^2)\mathcal{IG}(\sigma_\varepsilon^2|\alpha_\varepsilon, \beta_\varepsilon) && \text{(A.87)}\\
&\qquad \propto \mathcal{IG}(\sigma_\varepsilon^2|(||\boldsymbol{\mu}_{\mathbf{y}}|| + ||\mathbf{w}^*_{\boldsymbol{\gamma}}|| + 1)/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}R_{\sigma_\varepsilon^2}). && \text{(A.88)}
\end{aligned}
$$

where we give $R_{\sigma_\varepsilon^2}$ as:

$$
\begin{aligned}
R_{\sigma_\varepsilon^2} &= (\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}})^\top(\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}})\\
&\quad + (\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}})^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}(\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}}) + (\mu_{\mathbf{w}} - \mu_0)^\top(\mu_{\mathbf{w}} - \mu_0)/\sigma_0^2 && \text{(A.89)}
\end{aligned}
$$

Finally we calculate the distribution for π:

$$
\begin{aligned}
p(\pi|\boldsymbol{\theta}_{-\pi}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) &\propto \left\{\prod_{j=1}^{J}\text{Bern}(\gamma_j|\pi)\right\}\mathcal{B}(\pi|\alpha_\pi, \beta_\pi) && \text{(A.90)}\\
&\propto \beta(\pi|\ \alpha_\pi + ||\boldsymbol{\gamma}||, \beta_\pi + J - ||\boldsymbol{\gamma}|). && \text{(A.91)}
\end{aligned}
$$

A.2.1 Sampling $\boldsymbol{\gamma}$

In order to sample $\boldsymbol{\gamma}$ we use collapsing methods as detailed in Section 6.2. Following the method proposed in Davies et al. (2016a) we integrate over μ_w, $\mathbf{w}^*_{\boldsymbol{\gamma}}$, π, and σ_ε^2, however in the case of the eSABRE method are left with a conditional distribution that includes $\boldsymbol{\mu}_{\mathbf{y}}$ but not $\mathbf{y}$, leading to the increased computational efficiency discussed and tested in Chapters 6 and 7:

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) &\propto \int p(\boldsymbol{\gamma}, \pi, \sigma_\varepsilon^2, \mathbf{w}^*_{\boldsymbol{\gamma}}, \mu_w|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})d\mu_w d\mathbf{w}^*_{\boldsymbol{\gamma}}d\pi d\sigma_\varepsilon^2 && \text{(A.92)}\\
\propto \int p(\boldsymbol{\gamma}|\pi)p(\pi)&p(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{w}^*_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2, \mathbf{X}^*_{\boldsymbol{\gamma}})p(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mu_w, \sigma_\varepsilon^2)p(\mu_w)p(\sigma_\varepsilon^2)d\mu_w d\mathbf{w}^*_{\boldsymbol{\gamma}}d\pi d\sigma_\varepsilon^2 && \text{(A.93)}
\end{aligned}
$$

$$\propto C_\pi \int \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}|\mathbf{X}^*_\boldsymbol{\gamma}\mathbf{w}^*_\boldsymbol{\gamma}, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\mathbf{w}^*_\boldsymbol{\gamma}|\mathbf{m}_\boldsymbol{\gamma}, \sigma^2_\varepsilon \boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}})\mathcal{N}(\mu_w|\mu_0, \sigma^2_\varepsilon\sigma^2_0)$$

$$\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon)d\mu_w d\mathbf{w}^*_\boldsymbol{\gamma} d\sigma^2_\varepsilon \tag{A.94}$$

$$\propto C_\pi \int \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}|\mathbf{X}^*_\boldsymbol{\gamma}\mathbf{w}^*_\boldsymbol{\gamma}, \sigma^2_\varepsilon \mathbf{I})\mathcal{N}(\mathbf{w}^*_\boldsymbol{\gamma}|\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}}, \sigma^2_\varepsilon [\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}])$$

$$\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon)d\mathbf{w}^*_\boldsymbol{\gamma} d\sigma^2_\varepsilon \tag{A.95}$$

$$\propto C_\pi \int \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}|\mathbf{X}_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}}, \sigma^2_\varepsilon [\mathbf{I} + \mathbf{X}_\boldsymbol{\gamma}[\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]\mathbf{X}^\top_\boldsymbol{\gamma}])\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon)d\sigma^2_\varepsilon \tag{A.96}$$

$$\propto C_\pi |\boldsymbol{\Sigma}_\boldsymbol{\gamma}|^{-\frac{1}{2}}[\beta_\varepsilon + \tfrac{1}{2}(\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})^\top \boldsymbol{\Sigma}^{-1}_\boldsymbol{\gamma}(\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})]^{-(N/2+\alpha_\varepsilon)} \tag{A.97}$$

where $C_\pi = \frac{\Gamma(||\boldsymbol{\gamma}||+\alpha_\pi)\Gamma(J-||\boldsymbol{\gamma}||+\beta_\pi)}{\Gamma(J+\alpha_\pi+\beta_\pi)}$, $\boldsymbol{\Sigma}_\boldsymbol{\gamma} = [\mathbf{I}+\mathbf{X}^*_\boldsymbol{\gamma}[\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}}+\mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]\mathbf{X}^{*\top}_\boldsymbol{\gamma}]$, $\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}} = (\mu_{w_0}, \mu_0, \ldots, \mu_0)^\top$ with μ_0 repeated with length $||\mathbf{w}_\boldsymbol{\gamma}||$ dependent on $\boldsymbol{\gamma}$, and $\mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}$ is a block diagonal matrix of $(0, \sigma^2_0)$ where the square blocks have length 1 and $||\mathbf{w}_\boldsymbol{\gamma}||$ respectively.

We can again use the Woodbury identity and the extended Sylvester's determinant theorem to speed up the computations and give the following conditional posterior distribution:

$$\begin{aligned}\log\ p(\boldsymbol{\gamma}|\boldsymbol{\theta}_{-\boldsymbol{\gamma}}, \mathbf{X}^*_\boldsymbol{\gamma}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) &\propto \log\Gamma(||\boldsymbol{\gamma}|| + \alpha_\pi) + \log\Gamma(J - ||\boldsymbol{\gamma}|| + \beta_\pi) \\ &- \log\Gamma(J + \alpha_\pi + \beta_\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}| - \tfrac{1}{2}\log|[\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]^{-1} + \mathbf{X}^{*\top}_\boldsymbol{\gamma}\mathbf{X}^*_\boldsymbol{\gamma}| \\ &- (\tfrac{N}{2} + \alpha_\varepsilon)\log(\beta_\varepsilon + \tfrac{1}{2}(\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})^\top \\ &[\mathbf{I} - \mathbf{X}^*_\boldsymbol{\gamma}([\boldsymbol{\Sigma}_{\mathbf{w}^*_\boldsymbol{\gamma}} + \mathbf{V}_{\boldsymbol{\gamma},\mathbf{0}}]^{-1} + \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{X}^{*\top}_\boldsymbol{\gamma})^{-1}\mathbf{X}^{*\top}_\boldsymbol{\gamma})](\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})).\end{aligned} \tag{A.98}$$

A.2.2 Collapsing Within Conditional Distributions

In order to sample the eSABRE method via the collapsing scheme suggested in Section 6.2 we must derive the collapsed conditional distributions for σ^2_ε and μ_w. The conditional distribution of $\boldsymbol{\gamma}$ is derived in Section A.2.1, while (A.75) and (A.90) in Section A.2 give the distributions for π and $\mathbf{w}^*_\boldsymbol{\gamma}$. The conditional distribution for σ^2_ε can then be derived as follows:

$$p(\sigma^2_\varepsilon|\boldsymbol{\gamma}, \boldsymbol{\theta}_{-\sigma^2_\varepsilon}, \mathbf{X}^*_\boldsymbol{\gamma}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \propto \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}|\mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}}, \sigma^2_\varepsilon\boldsymbol{\Sigma}_\boldsymbol{\gamma})\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon) \tag{A.99}$$

$$\propto \mathcal{IG}(\sigma^2_\varepsilon|||\boldsymbol{\mu}_\mathbf{y}||/2 + \alpha_\varepsilon, \beta_\varepsilon + \tfrac{1}{2}(\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})^\top\boldsymbol{\Sigma}^{-1}_\boldsymbol{\gamma}(\boldsymbol{\mu}_\mathbf{y} - \mathbf{X}^*_\boldsymbol{\gamma}\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}})) \tag{A.100}$$

where the first distribution is taken from results in Section A.2.1 and the definitions of $\mathbf{m}_{\boldsymbol{\gamma},\mathbf{0}}$ and $\boldsymbol{\Sigma}_\boldsymbol{\gamma}$ are given in Section A.2.1. Finally we can give the conditional distribution of μ_w as follows:

$$
\begin{aligned}
&p(\mu_w|\sigma_\varepsilon^2, \boldsymbol{\gamma}, \boldsymbol{\theta}_{-\mu_w}\mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{M}, \mathbf{Z}, \mathbf{y}) \\
&\qquad \propto \int \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\mathbf{I})\mathcal{N}(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}})\mathcal{N}(\mu_w|\mu_0, \sigma_\varepsilon^2\sigma_0^2)d\mathbf{w}^*_{\boldsymbol{\gamma}} &\text{(A.101)}\\
&\qquad \propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}|\mathbf{1}\mu_{w_0} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{1}\mu_w, \sigma_\varepsilon^2[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*,\top}_{\boldsymbol{\gamma}}])\mathcal{N}(\mu_w|\mu_0, \sigma_0^2) &\text{(A.102)}\\
&\qquad \propto \mathcal{N}(\mu_w|V_{\mu_w}[\mu_0/\sigma_0^2 + \mathbf{1}^\top\mathbf{X}^\top_{\boldsymbol{\gamma}}[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*,\top}_{\boldsymbol{\gamma}}]^{-1}(\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{1}\mu_{w_0})], \sigma_\varepsilon^2 V_{\mu_w}) &\text{(A.103)}
\end{aligned}
$$

where $\mathbf{V}_{\boldsymbol{\mu}_{\boldsymbol{\gamma},\mathbf{w}}} = [1/\sigma_0^2 + \mathbf{1}^\top\mathbf{X}^\top_{\boldsymbol{\gamma}}[\mathbf{I} + \mathbf{X}^*_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}]^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{1}]^{-1}$.

Appendix B

Further Results

In this appendix we give extended results for the simulation studies in Chapter 5. We also give complete lists of results for the FMDV and Influenza datasets we have analysed in Chapters 5 and 7, these results include the common alignments of the individual residues and branches that were selected (Davies et al., 2016a).

B.1 Extended Simulation Study Results

The tables given in this section relate to the work completed in Section 5.3 of Chapter 5. The tables are adapted from Davies et al. (2016a) and contain result for for different values of α for the elastic net and alternative measures of performance to those discussed in Section 5.3. For completeness and comparability, many of the related results from Section 5.3 are also given here.

Table B.1: **Table of Extended Simulation Study Results - Part 1.** The table gives results for the Conjugate, Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and the classical mixed-effects models applied to the simulated data described in Section 5.1.2. The table gives the mean AUROC value based on ordering the variables (OV) and model selection (MS).

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
AUROC Values (OV)	Conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88
	Semi-Conjugate SABRE	1	0.98	0.89	1	0.98	0.89	1	0.97	0.87
	BM Conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88
	Mixed-Effects LASSO	0.95	0.93	0.80	0.91	0.84	0.74	0.90	0.75	0.69
	M-E Elastic Net ($\alpha = 0.2$)	0.97	0.83	0.74	0.90	0.79	0.73	0.85	0.77	0.66
	M-E Elastic Net ($\alpha = 0.3$)	0.93	0.84	0.79	0.88	0.85	0.76	0.84	0.75	0.69
	M-E Elastic Net ($\alpha = 0.4$)	0.92	0.90	0.80	0.93	0.87	0.76	0.87	0.72	0.69
	M-E Elastic Net ($\alpha = 0.5$)	0.92	0.92	0.81	0.93	0.88	0.75	0.89	0.72	0.69
	M-E Elastic Net ($\alpha = 0.6$)	0.93	0.92	0.81	0.94	0.87	0.74	0.90	0.73	0.71
	M-E Elastic Net ($\alpha = 0.8$)	0.94	0.93	0.80	0.93	0.86	0.71	0.90	0.74	0.69
	Mixed-Effects Models	0.99	0.95	0.80	0.99	0.91	0.75	0.95	0.85	0.72
AUROC Values (MS)	Conjugate SABRE	-	-	-	-	-	-	-	-	-
	Semi-Conjugate SABRE	-	-	-	-	-	-	-	-	-
	BM Conjugate SABRE	-	-	-	-	-	-	-	-	-
	Mixed-Effects LASSO	0.85	0.72	0.57	0.72	0.61	0.53	0.72	0.63	0.54
	M-E Elastic Net ($\alpha = 0.2$)	0.68	0.71	0.69	0.68	0.65	0.61	0.74	0.64	0.56
	M-E Elastic Net ($\alpha = 0.3$)	0.73	0.72	0.63	0.68	0.66	0.61	0.74	0.65	0.57
	M-E Elastic Net ($\alpha = 0.4$)	0.77	0.73	0.59	0.70	0.66	0.58	0.73	0.64	0.56
	M-E Elastic Net ($\alpha = 0.5$)	0.80	0.72	0.59	0.71	0.66	0.56	0.74	0.63	0.56
	M-E Elastic Net ($\alpha = 0.6$)	0.83	0.70	0.62	0.71	0.64	0.55	0.73	0.63	0.55
	M-E Elastic Net ($\alpha = 0.8$)	0.84	0.69	0.58	0.71	0.64	0.54	0.75	0.62	0.57
	Mixed-Effects Models	0.94	0.79	0.65	0.87	0.71	0.62	0.77	0.67	0.61

Table B.2: **Table of Extended Simulation Study Results - Part 2.** The table gives results for the Conjugate, Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and the classical mixed-effects models applied to the simulated data described in Section 5.1.2. The table gives the MSEs of the out-of-sample observations, $\mathbf{y_{out}}$, and the MSEs of the fixed effects coefficients, $\mathbf{w}$.

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
MSE($\mathbf{y_{out}}$)	Conjugate SABRE	0.15	0.22	0.49	0.18	0.30	0.57	0.26	0.36	0.63
	Semi-Conjugate SABRE	0.16	0.23	0.48	0.18	0.29	0.57	0.24	0.35	0.63
	BM Conjugate SABRE	0.16	0.22	0.49	0.18	0.29	0.56	0.24	0.36	0.62
	Mixed-Effects LASSO	0.06	0.22	0.59	0.13	0.40	0.75	0.31	0.56	1.37
	M-E Elastic Net ($\alpha = 0.2$)	0.06	0.18	0.55	0.12	0.32	0.76	0.38	0.61	1.57
	M-E Elastic Net ($\alpha = 0.3$)	0.06	0.18	0.60	0.11	0.34	0.75	0.31	0.65	1.81
	M-E Elastic Net ($\alpha = 0.4$)	0.06	0.19	0.62	0.12	0.37	0.80	0.35	0.65	2.13
	M-E Elastic Net ($\alpha = 0.5$)	0.06	0.20	0.82	0.11	0.38	0.79	0.28	0.60	1.93
	M-E Elastic Net ($\alpha = 0.6$)	0.06	0.26	0.84	0.11	0.40	0.82	0.29	0.55	0.93
	M-E Elastic Net ($\alpha = 0.8$)	0.06	0.22	0.81	0.12	0.41	0.84	0.27	0.71	0.94
	Mixed-Effects Models	0.08	0.23	0.53	0.16	0.37	0.68	0.32	0.50	0.77
MSE($\mathbf{w}$)	Conjugate SABRE	0.019	0.019	0.025	0.017	0.021	0.024	0.021	0.022	0.024
	Semi-Conjugate SABRE	0.021	0.022	0.022	0.017	0.020	0.025	0.019	0.020	0.025
	BM Conjugate SABRE	0.020	0.018	0.022	0.016	0.019	0.023	0.019	0.022	0.025
	Mixed-Effects LASSO	0.003	0.017	0.046	0.009	0.034	0.060	0.020	0.024	0.071
	M-E Elastic Net ($\alpha = 0.2$)	0.004	0.010	0.039	0.008	0.020	0.043	0.026	0.035	0.093
	M-E Elastic Net ($\alpha = 0.3$)	0.004	0.010	0.045	0.007	0.022	0.052	0.020	0.038	0.112
	M-E Elastic Net ($\alpha = 0.4$)	0.003	0.013	0.047	0.007	0.026	0.065	0.023	0.036	0.132
	M-E Elastic Net ($\alpha = 0.5$)	0.003	0.014	0.049	0.007	0.029	0.062	0.018	0.035	0.118
	M-E Elastic Net ($\alpha = 0.6$)	0.003	0.016	0.049	0.007	0.031	0.065	0.018	0.032	0.063
	M-E Elastic Net ($\alpha = 0.8$)	0.003	0.017	0.049	0.007	0.032	0.069	0.017	0.039	0.063
	Mixed-Effects Models	0.008	0.020	0.032	0.015	0.031	0.041	0.033	0.040	0.044

Table B.3: **Table of Extended Simulation Study Results - Part 3.** The table gives results for the Conjugate, Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and the classical mixed-effects models applied to the simulated data described in Section 5.1.2. The table gives the MSEs of the random effects coefficients, $\mathbf{b}$, and the mean WAIC scores for each method.

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
MSE(b)	Conjugate SABRE	0.019	0.025	0.032	0.020	0.025	0.040	0.026	0.027	0.039
	Semi-Conjugate SABRE	0.020	0.026	0.033	0.020	0.024	0.040	0.023	0.028	0.039
	BM Conjugate SABRE	0.020	0.025	0.035	0.020	0.024	0.042	0.025	0.029	0.038
	Mixed-Effects LASSO	0.020	0.032	0.058	0.060	0.042	0.076	0.036	0.104	0.143
	M-E Elastic Net ($\alpha = 0.2$)	0.021	0.027	0.054	0.040	0.032	0.053	0.039	0.056	0.099
	M-E Elastic Net ($\alpha = 0.3$)	0.021	0.030	0.063	0.029	0.036	0.072	0.031	0.067	0.116
	M-E Elastic Net ($\alpha = 0.4$)	0.019	0.029	0.066	0.037	0.052	0.068	0.037	0.072	0.136
	M-E Elastic Net ($\alpha = 0.5$)	0.020	0.031	0.112	0.026	0.050	0.072	0.033	0.056	0.146
	M-E Elastic Net ($\alpha = 0.6$)	0.021	0.033	0.105	0.033	0.049	0.080	0.031	0.084	0.071
	M-E Elastic Net ($\alpha = 0.8$)	0.019	0.035	0.103	0.039	0.064	0.078	0.029	0.129	0.076
	Mixed-Effects Models	0.015	0.025	0.034	0.019	0.027	0.045	0.029	0.033	0.042
WAIC	Conjugate SABRE	-309.7	-173.2	-100.4	-314.0	-172.2	-100.8	-309.8	-172.8	-103.1
	Semi-Conjugate SABRE	-308.7	-170.5	-96.8	-312.1	-171.2	-98.5	-310.5	-171.4	-101.3
	BM Conjugate SABRE	-309.7	-173.5	-98.7	-313.9	-171.9	-101.3	-310.4	-172.0	-103.3

Table B.4: **Table of P-Values for the Simulation Study Results - Part 1.** The table gives the results for paired t-tests where the Conjugate SABRE is compared against each of the other methods; the Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and classical mixed-effects models. The table gives the p-values for comparing the mean AUROC value based on ordering the variables (OV) and model selection (MS).

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
AUROC Values (OV)	Semi-Conjugate SABRE	1	0.056	0.004	0.080	0.272	0.043	0.356	0.559	0.065
	BM Conjugate SABRE	1	0.182	0.469	0.612	0.160	0.886	0.289	0.257	0.185
	Mixed-Effects LASSO	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.2$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.3$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.4$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.5$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.6$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.8$)	0	0	0	0	0	0	0	0	0
	Mixed-Effects Models	0	0	0	0	0	0	0	0	0
AUROC Values (MS)	Semi-Conjugate SABRE	-	-	-	-	-	-	-	-	-
	BM Conjugate SABRE	-	-	-	-	-	-	-	-	-
	Mixed-Effects LASSO	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.2$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.3$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.4$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.5$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.6$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.8$)	0	0	0	0	0	0	0	0	0
	Mixed-Effects Models	0	0	0	0	0	0	0	0	0

Table B.5: **Table of P-Values for the Simulation Study Results - Part 2.** The table gives the results for paired t-tests where the Conjugate SABRE is compared against each of the other methods; the Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and classical mixed-effects models. The table gives the p-values for comparing the MSEs of the out-of-sample observations, $\mathbf{y_{out}}$ and the MSEs of the fixed effects coefficients, $\mathbf{w}$

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3	$\sigma_\varepsilon^2 =$ 0.03	$\sigma_\varepsilon^2 =$ 0.1	$\sigma_\varepsilon^2 =$ 0.3
MSE($\mathbf{y_{out}}$)	Semi-Conjugate SABRE	0.075	0.165	0.046	0.770	0.567	0.588	0.158	0.212	0.611
	BM Conjugate SABRE	0.443	0.978	0.833	0.979	0.329	0.370	0.138	0.990	0.169
	Mixed-Effects LASSO	0	0.926	0	0.001	0.006	0	0.262	0	0
	M-E Elastic Net ($\alpha = 0.2$)	0	0	0	0	0.193	0	0.003	0	0
	M-E Elastic Net ($\alpha = 0.3$)	0	0	0	0	0.758	0	0.859	0	0
	M-E Elastic Net ($\alpha = 0.4$)	0	0.032	0	0	0.633	0	0.029	0	0.353
	M-E Elastic Net ($\alpha = 0.5$)	0	0.108	0	0	0.330	0	0.673	0.003	0.020
	M-E Elastic Net ($\alpha = 0.6$)	0	0.470	0	0	0.090	0	0.742	0.002	0
	M-E Elastic Net ($\alpha = 0.8$)	0	0.982	0	0	0.082	0	0.488	0	0
	Mixed-Effects Models	0	0.032	0	0.118	0	0	0.984	0	0
MSE($\mathbf{w}$)	Semi-Conjugate SABRE	0.197	0.586	0.289	0.687	0.186	0.437	0.259	0.107	0.163
	BM Conjugate SABRE	0.437	0.842	0.927	0.925	0.131	0.536	0.209	0.662	0.124
	Mixed-Effects LASSO	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.2$)	0	0	0	0	0	0	0	0.006	0
	M-E Elastic Net ($\alpha = 0.3$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.4$)	0	0	0	0	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.5$)	0	0	0	0	0	0	0	0.009	0
	M-E Elastic Net ($\alpha = 0.6$)	0	0	0	0	0	0	0	0.001	0
	M-E Elastic Net ($\alpha = 0.8$)	0	0	0	0	0	0.018	0	0.022	0
	Mixed-Effects Models	0	0	0	0	0	0.001	0	0.094	0.336

Table B.6: **Table of P-Values for the Simulation Study Results - Part 3.** The table gives the results for paired t-tests where the Conjugate SABRE is compared against each of the other methods; the Semi-Conjugate and BM Conjugates SABRE methods, the mixed-effects LASSO, the mixed-effects elastic net with $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ and classical mixed-effects models. The table gives the p-values for comparing the MSEs of the random effects coefficients, $\mathbf{b}$, and the mean WAIC scores with the conjugate SABRE for each method.

	Method	$\|\|\mathbf{w}\|\| = 40$			$\|\|\mathbf{w}\|\| = 60$			$\|\|\mathbf{w}\|\| = 80$		
		$\sigma^2_\varepsilon =$ 0.03	$\sigma^2_\varepsilon =$ 0.1	$\sigma^2_\varepsilon =$ 0.3	$\sigma^2_\varepsilon =$ 0.03	$\sigma^2_\varepsilon =$ 0.1	$\sigma^2_\varepsilon =$ 0.3	$\sigma^2_\varepsilon =$ 0.03	$\sigma^2_\varepsilon =$ 0.1	$\sigma^2_\varepsilon =$ 0.3
MSE(b)	Semi-Conjugate SABRE	0.256	0.193	0.299	0.634	0.220	0.567	0.073	0.837	0.584
	BM Conjugate SABRE	0.381	0.616	0.465	0.564	0.112	0.919	0.285	0.526	0.127
	Mixed-Effects LASSO	0.000	0	0	0.081	0	0	0	0	0
	M-E Elastic Net ($\alpha = 0.2$)	0.256	0.034	0	0.113	0.003	0	0	0	0.075
	M-E Elastic Net ($\alpha = 0.3$)	0.219	0	0	0.132	0.003	0	0.054	0.004	0.100
	M-E Elastic Net ($\alpha = 0.4$)	0.940	0.004	0	0.043	0.003	0	0	0	0.459
	M-E Elastic Net ($\alpha = 0.5$)	0.594	0.002	0	0.075	0.107	0	0.069	0.004	0.218
	M-E Elastic Net ($\alpha = 0.6$)	0.242	0	0	0.011	0.015	0	0.059	0.023	0
	M-E Elastic Net ($\alpha = 0.8$)	0.945	0	0.001	0.030	0.006	0	0.272	0.022	0
	Mixed-Effects Models	0	0.632	0.004	0.323	0.014	0	0.933	0	0.004
WAIC	Semi-Conjugate SABRE	0.002	0	0	0	0	0	0	0.017	0
	BM Conjugate SABRE	0.892	0.597	0.155	0.649	0.205	0.731	0.583	0.227	0.500

B.2 Foot-and-Mouth Disease Virus Data

This section gives a complete list of results for all the real datasets discussed in the main paper. Tables B.7, B.9 and B.11 give full lists of results for the original SAT1, extended SAT1 and SAT2 datasets based on taking the top $J\hat{\pi}$ variables from the model. Tables B.8, B.10 and B.12 give similar results for when only the branch variables are used. Finally Figure B.1 gives the complete phylogenetic tree for the extended SAT1 dataset when the $J\hat{\pi}$ variables with the highest predicted marginal probability of inclusion are used, as opposed to any variables with greater than 0.5 predicted marginal inclusion probability as shown in Figure 8b of the main paper.

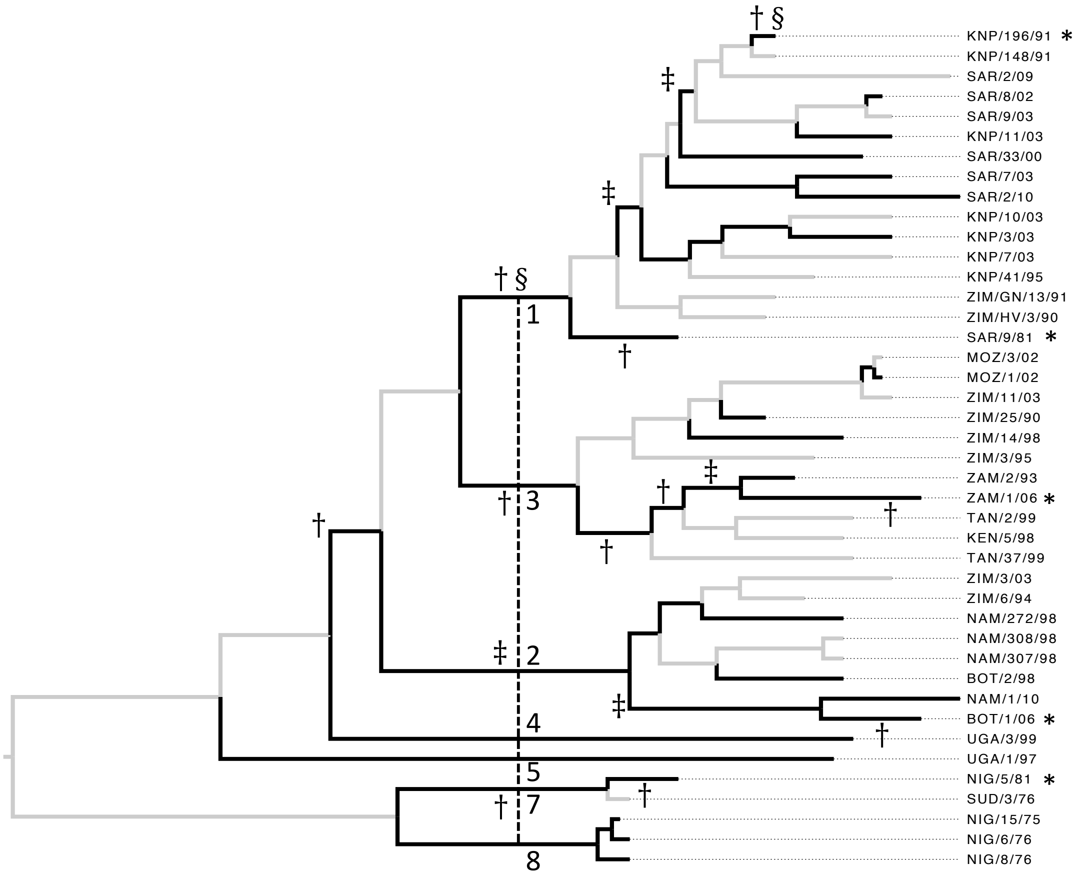


Figure B.1: **Phylogenetic tree indicating significant branches in the evolutionary history of the SAT1 serotype at a low threshold.** The phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*) and topotype defining branches (dashed vertical line). Branches inferred by the SABRE method are highlighted (black). Symbols indicate whether this was inferred to be a change in virus antigenicity (†), virus reactivity (‡) or virus immunogenicity (§). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance was taken to be the $J\hat{\pi}$ variables with the highest probability of inclusion given in Table B.10.

Table B.7: **Selected variables using the original SAT1 data with challenge strain and antiserum used as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Additionally the cut-off at 0.5 is marked by a horizontal line. Residues are given by their protein sequence alignment (Reeve et al., 2010), where for instance VP3 138 is position 138 on the VP3 protein. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3 and the labelled phylogenetic tree for this dataset is given in Figure 2.2.

Variable	Inclusion Prob.	Plausibility	Complete Correlations
VP2 74	0.87	Proven	-
VP3 74	0.51	Plausible	-
bran 1A	0.50	Plausible	-
VP1 143	0.49	Proven	-
VP1 189	0.48	Plausible	-
bran 2A	0.46	Proven	VP3 177; VP2 82; VP1 201; VP2 131; VP2 187; VP3 141
VP1 47	0.45	Plausible	-
bran 0014	0.45	Plausible	-
VP3 193	0.43	Plausible	-
VP1 150	0.43	Proven	-
VP1 62	0.41	Proven	bran 1C
VP3 67	0.38	Plausible	-
VP3 9	0.38	Implausible	-
VP2 198	0.37	Plausible	-
VP3 199	0.35	Plausible	bran 0002
VP1 149	0.35	Proven	-
react 3A	0.34	Plausible	-
anti 0013	0.32	Plausible	-
VP1 219	0.31	Proven	-
VP3 72	0.31	Proven	-
VP3 77	0.31	Proven	-
VP2 79	0.30	Proven	VP2 81; bran 0007
VP3 176	0.30	Plausible	-
bran 2C	0.29	Plausible	-
VP3 171	0.29	Plausible	-
bran 1F	0.29	Plausible	-
bran 0011	0.28	Plausible	-
VP1 144	0.28	Proven	-
VP1 216	0.28	Proven	-

Table B.8: **Selected variables using the original SAT1 branch data with challenge strain and antiserum used as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3. The labelled phylogenetic tree for this dataset is given in Figure 2.2 here and the inferred phylogenetic tree in Figure 8a of the main paper.

Variable	Inclusion Prob.	Complete Correlations
anti 0013	1	-
anti 0010	0.99	-
anti 0004	0.92	-
bran 2A	0.82	-
bran 0014	0.80	-
anti 1B	0.75	-
bran 1C	0.72	-
bran 1A	0.70	-
anti 4A	0.68	-
bran 0012	0.64	-
bran 0020	0.59	-
bran 0002	0.57	-
bran 1F	0.50	-
bran 0001	0.48	-
bran 0006	0.44	-
bran 0007	0.43	-
bran 3C	0.43	-
bran 0019	0.39	-

Table B.9: **Selected variables using the extended SAT1 data with challenge strain, date and antiserum used as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Additionally the cut-off at 0.5 is marked by a horizontal line Residues are given by their protein sequence alignment (Reeve et al., 2010), where for instance VP3 138 is position 138 on the VP3 protein. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3 and the labelled phylogenetic tree for this dataset is given in Figure 2.3.

Variable	Inclusion Prob.	Plausibility	Complete Correlations
VP1 149	1	Proven	-
VP2 72	0.99	Proven	-
VP3 138	0.97	Proven	-
VP1 209	0.81	Proven	-
anti 0031	0.69	Plausible	-
VP3 171	0.68	Plausible	-
VP3 72	0.66	Proven	-
VP1 144	0.65	Proven	-
VP1 147	0.63	Proven	-
react 4A	0.58	Proven	-
VP2 198	0.57	Plausible	-
VP1 116	0.54	Plausible	-
VP2 74	0.53	Proven	-
VP3 77	0.53	Proven	-
bran 1G	0.53	Plausible	-
immun 0018	0.52	Proven	immun 1H, 2D, 3C, 4B, 5A, 6A, 7A
VP1 148	0.51	Proven	-
VP1 163	0.51	Proven	-
VP3 223	0.51	Plausible	-
VP2 79	0.49	Proven	-
VP1 211	0.46	Proven	-
bran 0016	0.45	Plausible	-
VP1 150	0.45	Proven	-
VP1 207	0.45	Proven	-
immun 8A	0.45	Proven	-
VP1 86	0.44	Implausible	-

Table B.9 **Selected variables using the extended SAT1 data**

Variable	Inclusion Prob.	Plausibility	Complete Correlations
bran 2A	0.44	Proven	VP3 177; VP2 82; VP1 201; VP2 131; VP2 187; VP3 141
VP2 95	0.43	Plausible	-
react 1C	0.43	Plausible	-
bran 1A	0.43	Plausible	-
VP3 67	0.43	Plausible	-
anti 0029	0.43	Plausible	-
immun 9A	0.42	Plausible	-
VP1 218	0.41	Proven	-
react 6A	0.41	Plausible	-
bran 2F	0.41	Plausible	-
VP1 142	0.41	Proven	-
bran 1J	0.41	Plausible	-
VP3 58	0.4	Proven	-
bran 3B	0.4	Plausible	-
react 1K	0.4	Plausible	-
anti 2G	0.4	Plausible	-
react 0007	0.4	Plausible	-
VP3 61	0.4	Proven	-
bran 2B	0.39	Plausible	-
VP1 156	0.39	Proven	bran 0017
anti 1K	0.39	Plausible	-
bran 0002	0.38	Plausible	-
bran 0030	0.38	Plausible	-
VP1 143	0.38	Proven	-
bran 0038	0.38	Plausible	-
bran 0024	0.38	Plausible	-
bran 0027	0.38	Plausible	-
VP3 199	0.38	Plausible	-
anti 3E	0.38	Plausible	-
VP1 45	0.38	Plausible	-
VP3 182	0.38	Plausible	-
bran 0006	0.38	Plausible	-
VP3 76	0.38	Plausible	-

Table B.9 **Selected variables using the extended SAT1 data**

Variable	Inclusion Prob.	Plausibility	Complete Correlations
bran 3D	0.38	Plausible	-
bran 0001	0.37	Plausible	-
VP1 42	0.37	Plausible	bran 0013
VP3 69	0.37	Plausible	-
VP1 155	0.37	Proven	-
react 5A	0.37	Plausible	-
react 3A	0.36	Plausible	-
VP3 134	0.36	Plausible	-
VP1 164	0.36	Proven	-
VP3 178	0.36	Plausible	VP2 194, bran 0009
anti 0007	0.36	Plausible	-
VP2 192	0.36	Plausible	bran 0026
bran 2C	0.36	Plausible	-
bran 1D	0.36	Plausible	-
react 7A	0.36	Plausible	-
VP3 16	0.35	Implausible	bran 0010
bran 0023	0.35	Plausible	-

Table B.10: **Selected variables using the extended SAT1 branch data using challenge strain and antiserum as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Additionally the cut-off at 0.5 is marked by a horizontal line. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3. The labelled phylogenetic tree for this dataset is given in Figure 2.3 here. The inferred phylogenetic tree for a $J\hat{\pi}$ cut-off is given in Figure B.1 here and for the 0.5 cut-off in Figure 5.9.

Variable	Inclusion Prob.	Plausibility
anti 0007	1	-
anti 0029	1	-
anti 0031	1	-
anti 8A	1	-
bran 1G	0.91	-

Table B.10 **Selected variables using the extended SAT1 branch data**

Variable	Inclusion Prob.	Plausibility
anti 0018	0.85	-
anti 0004	0.80	-
bran 0016	0.73	-
anti 1B	0.71	-
react 4A	0.70	-
bran 2A	0.70	-
bran 0030	0.69	-
bran 1A	0.68	-
bran 0024	0.66	-
bran 0038	0.63	-
anti 6B	0.62	-
immun 0018	0.61	immun 1H, 2D, 3C, 4B, 5A, 6A, 7A
bran 0039	0.61	-
anti 2G	0.58	-
bran 1J	0.56	-
anti 3E	0.54	-
bran 0006	0.54	-
bran 0013	0.53	-
bran 0042	0.52	-
bran 0027	0.51	-
react 6A	0.50	-
bran 0002	0.50	-
react 1C	0.49	-
bran 3D	0.49	-
react 1K	0.48	-
bran 0035	0.48	-
bran 0017	0.48	-
bran 1M	0.47	-
bran 0023	0.46	-
bran 0001	0.45	-
anti 10A	0.43	-
bran 0021	0.41	-
bran 0008	0.40	-

Table B.10 **Selected variables using the extended SAT1 branch data**

Variable	Inclusion Prob.	Plausibility
bran 2F	0.40	-
immun 8A	0.39	-
bran 3B	0.39	-
react 3C	0.39	-
bran 0041	0.39	-
bran 0003	0.39	-
bran 2B	0.38	-

Table B.11: **Selected variables using the SAT2 data using challenge strain and antiserum as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Residues are given by their protein sequence alignment (Reeve et al., 2010), where for instance VP3 138 is position 138 on the VP3 protein. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3 and the labelled phylogenetic tree for this dataset is given in Figure 2.4.

Variable	Inclusion Prob.	Complete Correlations
VP1 88	0.91	-
VP1 48	0.77	VP1 66, anti 0013
VP2 71	0.73	VP2 72, VP1 180, VP1 208, anti 0003
VP1 103	0.65	-
VP1 210	0.60	-
VP1 166	0.41	-
VP2 101	0.39	-
VP1 209	0.38	-
immun 0003	0.36	immun 1A, 2A, 3A, 4A, 5A
VP2 134	0.36	-
VP3 69	0.35	-
immun 6A	0.35	immun 7A
VP1 102	0.34	-
VP3 199	0.33	-
VP2 132	0.33	-
VP2 193	0.32	-
VP1 178	0.29	-
VP1 211	0.29	-
VP1 144	0.28	-
VP1 54	0.28	-
react 8A	0.27	-
VP2 80	0.26	VP1 189
VP1 207	0.26	-
VP1 47	0.26	-
VP1 60	0.26	-
VP3 68	0.26	VP2 78, VP1 101, VP2 140, bran 0022
VP3 88	0.26	-
VP2 85	0.26	VP2 195, bran 0005

Table B.12: **Selected variables using the SAT2 branch data using challenge strain and antiserum as random effects factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Branches are given as to indicate: a reactivity effect associated with the challenge strain (react), an immunogenic effect of the protein strain (immun), an antigenic effect (anti) or an unknown effect which is either a reactivity or antigenic effect (bran). More details on the types of branches can be found in Section 2.1.3. The labelled phylogenetic tree for this dataset is given in Figure 2.4 here and the inferred phylogenetic tree in Figure 5.12.

Variable	Inclusion Prob.	Complete Correlations
anti 0003	1	-
anti 0013	1	-
anti 1G	0.98	-
anti 0016	0.91	-
bran 0015	0.46	-
bran 0018	0.46	-
immun 0003	0.45	immun 1A, 2A, 3A, 4A, 5A
bran 1B	0.43	-
bran 1H	0.35	-
immun 6A	0.34	immun 7A
bran 0022	0.34	-
bran 0009	0.34	-
bran 0014	0.32	-
bran 0005	0.31	-
immun 0020	0.3	immun 1G

Table B.13: **Antigenic SAT1 Residues Selected by Maree et al. (2015).** The table gives the results of Maree et al. (2015) that are equivalent to those reported in this paper. Due to Maree et al. (2015) having a different overall aim to this current paper, these results were not directly reported in their paper. The horizontal line indicates the cut-off based on the Holm-Bonferroni correction and the results are reported up until the first implausible residue is selected. Residues are given by their protein sequence alignment (Reeve et al., 2010), where for instance VP3 138 is position 138 on the VP3 protein. Selected branches are not stated.

Variable	Plausibility
VP2 72	Proven
VP1 149	Proven
VP1 144	Proven
VP3 138	Proven
VP3 72	Proven
VP3 171	Plausible
VP1 164	Proven
VP1 209	Proven
VP3 77	Proven
VP1 102	Implausible

B.3 Influenza Data

This section gives a complete list of results for H1N1 dataset discussed in the main paper. Table B.14 gives the full list of results for the H1N1 dataset described in Section 4.5 of the main paper based on taking the top $J\hat{\pi}$ variables from the model.

Table B.14: **Selected variables using the conjugate SABRE method on the reduced H1N1 dataset using challenge strain as a random effects factor.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Residues are given by their position of the H1 common alignment (Harvey et al., 2016). Selected branches are not stated except where they have have a correlation coefficient of 1 with a selected residue variable. In this case the branch is given simply as 'branch' as a phylogenetic tree is not given.

Variable	Inclusion Prob.	Plausibility	Complete Correlations
187	1	Proven	-
190	1	Proven	-
43	1	Implausible	-
141	1	Proven	-
252	0.73	Plausible	branch
142	0.68	Proven	branch
313	0.65	Implausible	branch
189	0.64	Proven	-
323	0.51	Implausible	-
66	0.50	Plausible	branch
310	0.45	Implausible	branch
130	0.42	Proven	-
146	0.38	Plausible	-
139	0.36	Proven	branch
153	0.35	Proven	-
74	0.34	Proven	-
327	0.33	Implausible	-
69	0.33	Proven	branch
72	0.28	Proven	branch

Table B.15: **Selected variables using the eSABRE method on the full H1N1 data using challenge strain and the date of the experiment as random effect factors.** The table gives a list of the variables selected using the eSABRE method with a cut-off of $J\hat{\pi}$. Residues are given by their position of the H1 common alignment (Harvey et al., 2016). Selected branches are not stated except where they have have a correlation coefficient of 1 with a selected residue variable. In this case the branch is given simply as 'branch' as a phylogenetic tree is not given.

Variable	Inclusion Prob.	Plausibility	Complete Correlations
187	1	Proven	-
43	1	Implausible	-
141	1	Proven	-
190	1	Proven	-
153	0.91	Plausible	-
142	0.86	Proven	branch
313	0.69	Implausible	branch
324	0.64	Implausible	325, 326
130	0.54	Proven	-
193	0.47	Plausible	54, 125, 127, branch
146	0.44	Plausible	-
72	0.43	Proven	branch
310	0.43	Implausible	branch
74	0.39	Proven	-
189	0.37	Proven	-
170	0.35	Proven	-
66	0.35	Plausible	134, branch
252	0.33	Plausible	branch
327	0.33	Implausible	-
69	0.31	Proven	branch

Table B.16: **Selected variables using the eSABRE method on the full H3N2 data using challenge strain, protective strain and the date of the experiment as random effect factors.** The table gives a list of the variables selected using the conjugate SABRE method with a cut-off of $J\hat{\pi}$. Residues are given by their position of the H1 common alignment (Harvey et al., 2016). Selected branches are not stated except where they have have a correlation coefficient of 1 with a selected residue variable. In this case the branch is given simply as 'branch' as a phylogenetic tree is not given. $*$ indicates that the residue was removed from the results due to the recorded genetic code being inaccurate.

Variable	Inclusion Prob.	Plausibility	Complete Correlations
135	1	Proven	-
138	1	Plausible	-
144	1	Proven	-
145	1	Proven	-
156	1	Proven	-
158	1	Proven	-
164	1	Proven	-
189	1	Proven	-
193	1	Proven	-
197	1	Proven	-
262	1	Proven	-
276	0.98	Proven	-
25	0.97	Plausible	75, branch
155	0.97	Proven	-
279	0.89	Plausible	-
183	0.87	Proven	-
212	0.64	Plausible	-
269	0.57	Implausible*	-
159	0.56	Proven	-
14	0.54	Implausible	43, branch
142	0.47	Proven	-
2	0.45	Implausible	-
190	0.41	Proven	-
207	0.40	Proven	-
194	0.37	Plausible	-
131	0.37	Proven	-
196	0.34	Proven	-
126	0.34	Proven	-

Table B.16 **Selected variables using the H3N2 data**

Variable	Inclusion Prob.	Plausibility	Complete Correlations
58	0.32	Implausible	-
140	0.30	Plausible	-
27	0.28	Implausible	-
57	0.26	Proven	-
318	0.25	Implausible	-
18	0.23	Implausible	-
3	0.23	Implausible	-
242	0.22	Proven	-
147	0.22	Implausible	-
216	0.22	Plausible	-
34	0.20	Implausible	-

References

Aderhold, A., Husmeier, D., and Grzegorczyk, M. (2014). Statistical inference of regulatory networks for circadian regulation. *Statistical Applications in Genetics and Molecular Biology*, 13(3):227–273. 64

Aktas, S. and Samuel, A. R. (2000). Identification of antigenic epitopes on the foot and mouth disease virus isolate O-1/Manisa/Turkey/69 using monoclonal antibodies. *Scientific and Technical Review of the Office International des Epizooties*, 19(3):744–753. 17, 74, 81

Andrieu, C. and Doucet, A. (1999). Joint bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, 47(10):2667–2676. 53, 91

Barbieri, L. and Berger, J. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897. 60

Barnett, P., Ouldridge, E., Rowlands, D., Brown, F., and Parry, N. (1989). Neutralizing epitopes of type O Foot-and-Mouth disease virus. I. Identification and characterization of three functionally independent, conformational sites. *The Journal of general virology*, 70 (Pt 6):1483–1491. 17

Barr, I. G., Russell, C., Besselaar, T. G., Cox, N. J., Daniels, R. S., Donis, R., Engelhardt, O. G., Grohmann, G., Itamura, S., Kelso, A., McCauley, J., Odagiri, T., Schultz-Cherry, S., Shu, Y., Smith, D., Tashiro, M., Wang, D., Webby, R., Xu, X., Ye, Z., and Zhang, W. (2014). WHO recommendations for the viruses used in the 2013-2014 Northern Hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from October 2012 to January 2013. *Vaccine*, 32(37):4713–25. 16

Bates, D., Maechler, M., and Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. 59

Baxt, B., Vakharia, V., Moore, D., Franke, A., and Morgan, D. (1989). Analysis of neutralizing antigenic sites on the surface of type A12 Foot-and-Mouth disease virus. *Journal of Virology*, 63(5):2143–2151. 17, 74

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418. 26

BBC (2016). When foot-and-mouth disease stopped the UK in its tracks. `http://www.bbc.co.uk/news/magazine-35581830`. BBC article author: Claire Bates. 1

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 22, 23, 24, 50, 115, 119

Bolwell, C., Brown, A., Barnett, P., Campbell, R., Clarke, B., Parry, N., Ouldridge, E., Brown, F., and Rowlands, D. (1989). Host cell selection of antigenic variants of Foot-and-Mouth disease virus. *The Journal of general virology*, 70 (Pt 1):45–57. 17, 74

Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. (1999). Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution*, 16(11):1457–1465. 19

Caton, A. J., Brownlee, G. G., Yewdell, J. W., and Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, 31(2 Pt 1):417–427. 18

Crowther, J., Farias, S., Carpenter, W., and Samuel, A. (1993a). Identification of a fifth neutralizable site on type O Foot-and-Mouth disease virus following characterization of single and quintuple monoclonal antibody escape mutants. *The Journal of general virology*, 74 (Pt 8):1547–1553. 17, 74

Crowther, J., Rowe, C., and Butcher, R. (1993b). Characterization of monoclonal antibodies against a type SAT 2 Foot-and-Mouth disease virus. *Epidemiology and Infection*, 111(2):391–406. 18, 74, 79, 80

Davies, V., Reeve, R., Harvey, W., Maree, F. F., and Husmeier, D. (2014). Sparse Bayesian variable selection for the identification of antigenic variability in the Foot-and-Mouth Disease Virus. *Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS)*, 33:149–158. iv, xii, xiii, 3, 13, 29, 37, 38, 42, 50, 57, 58, 60, 61, 63, 71, 72, 77, 79, 93, 108, 110, 111, 119

Davies, V., Reeve, R., Harvey, W., Maree, F. F., and Husmeier, D. (2016a). A sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution. *Computational Statistics (Under Revision)*. iv, 3, 11, 13, 21, 25, 29, 37, 42, 44, 46, 48, 57, 60, 62, 73, 74, 79, 85, 86, 89, 90, 91, 92, 93, 98, 99, 108, 109, 125, 128

Davies, V., Reeve, R., Harvey, W. T., and Husmeier, D. (2016b). Selecting random effect components in a sparse hierarchical Bayesian model for identifying antigenic variability. In Angelini, C., Rancoita, P. M. V., and Rovetta, S., editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 14–27. iv, 3, 13, 56, 58, 62, 70

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499. 23

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, 93:93–114. 113

Gelman, A. (2004). Parameterization and bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545. 48

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3). 38, 47, 48, 62, 90

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Ventari, A., and Rubin, D. B. (2013a). *Bayesian Data Analysis*. Chapman & Hall, third edition. 3, 28, 30, 34, 37, 44, 54

Gelman, A., Hwang, J., and Vehtari, A. (2013b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016. 35

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511. 28, 59, 98

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741. 27

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889. 29, 30, 31

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373. 29, 30, 31

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467):799–804. 29

Grazioli, S., Fallacara, F., and Brocchi., E. (2013). Mapping of antigenic sites of foot-and-mouth disease virus serotype Asia 1 and relationships with sites described in other serotypes. *The Journal of general virology*, 94(3):559–569. 17, 74, 75, 80

Grazioli, S., Moretti, M., Barbieri, I., Crosatti, M., and Brocchi, E. (2006). Use of monoclonal antibodies to identify and map new antigenic determinants involved in neutralisation on FMD viruses type SAT 1 and SAT 2. In *Report of the Session of the Research Group of the Standing Technical Committee of the European Commission for the Control of Foot-and-Mouth Disease*, pages 287–297. Appendix 43. 17, 18, 73, 74, 75, 76, 79, 80, 81, 83

Grzegorczyk, M. and Husmeier, D. (2013). Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Machine Learning*, 91:105–151. 28, 60, 79

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4). 114

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36. 33, 64

Harvey, W. T. (2016). *Quantifying the genetic basis of antigenic variation among human influenza A viruses*. PhD thesis, University of Glasgow. 23, 106, 113

Harvey, W. T., Benton, D. J., Gregory, V., Hall, J. P. J., Daniels, R. S., Bedford, T., Haydon, D. T., Hay, A. J., McCauley, J. W., and Reeve, R. (2016). Identification of low- and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A(H1N1) viruses. *PLoS Pathog*, 12(4):1–23. 10, 16, 23, 59, 81, 82, 105, 106, 113, 147, 148, 149

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. 23, 43, 89

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. 27

Heydari, J., Lawless, C., Lydall, D. A., and Wilkinson, D. J. (2016). Bayesian hierarchical modelling for inferring genetic interactions in yeast. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3):367–393. 55, 110

Hirst, G. K. (1942). The quantitative determination of influenza virus and antibodies by means of red cell agglutination. *The Journal of experimental medicine*, 75(1):49–64. 7

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and VandePol, S. (1982). Rapid evolution of RNA genomes. *Science*, 215:1577–1585. 6

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70. 22

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307. 25

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773. 30

Jow, H., Boys, R. J., and Wilkinson, D. J. (2014). Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data. *Statistical Applications in Genetics and Molecular Biology*, 13(5):531–551. 31, 68

Kitson, J., McCahon, D., and Belsham, G. (1990). Sequence analysis of monoclonal antibody resistant mutants of type O Foot and Mouth disease virus: evidence for the involvement of the three surface exposed capsid proteins in four antigenic sites. *Virology*, 179(1):26–34. 17, 74, 80

Knowles, N. and Samuel, A. (2003). Molecular epidemiology of Foot-and-Mouth disease virus. *Virus Res*, 91:65–80. 10

Lea, S., Hernandez, J., Blakemore, W., Brocchi, E., Curry, S., Domingo, E., Fry, E., Abu Ghazaleh, R., King, A., Newman, J., Stuart, D., and Mateu, M. (1994). The structure and antigenicity of a type C Foot-and-Mouth disease virus. *Structure*, 2(2):123–139. 17, 74, 76, 80, 81

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2015). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, pages 1–17. 85, 94, 111

Maree, F. F., Borley, D. W., Reeve, R., Upadhyaya, S., Lukhwareni, A., Mlingo, T., Esterhuysen, J. J., Harvey, W. T., Fry, E. E., Parida, S., Paton, D. J., and Mahapatra, M. (2015). Tracking the antigenic evolution of foot-and-mouth disease virus. *(In Submission)*. xi, 12, 13, 14, 22, 58, 59, 71, 76, 77, 83, 98, 112, 113, 146

Mateu, M. (1995). Antibody recognition of picornaviruses and escape from neutralization: a structural view. *Virus Research*, 38(1):1–24. 17, 74, 75, 76, 80

Mattion, N., König, G., Seki, C., Smitsaart, E., Maradei, E., Robiolo, B., Duffy, S., León, E., Piccone, M., Sadir, A., Bottini, R., Cosentino, B., Falczuk, A., Maresca, R., Periolo, O., Bellinzoni, R., Espinoza, A., Torre, J., and Palma, E. (2004). Reintroduction of Foot-and-Mouth disease in Argentina: characterisation of the isolates and development of tools for the control and eradication of the disease. *Vaccine*, 22:4149–4162. 2, 7

McDonald, N. J., Smith, C. B., and Cox, N. J. (2007). Antigenic drift in the evolution of H1N1 influenza A viruses resulting from deletion of a single amino acid in the haemagglutinin gene. *The Journal of General Virology*, 88(Pt 12):3209–3213. 18

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092. 27

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032. 3, 29, 30, 31, 40, 89

Mohamed, S., Heller, K., and Ghahramani, Z. (2012). Bayesian and l_1 approaches for sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 751–758. 3, 29, 30, 37, 40, 109

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.* MIT Press, Cambridge, MA. 29, 33, 55, 64, 68, 110, 124

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482). 29

Paton, D., Valarcher, J., Bergmann, I., Matlho, O., Zakharov, V., Palma, E., and Thomson, G. (2005). Selection of Foot and Mouth disease vaccine strains - a review. *Rev Sci Tech*, 24:981–993. 2, 7

Pinheiro, J. C. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer. 22

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11. 59

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. 25, 59

Reeve, R., Blignaut, B., Esterhuysen, J. J., Opperman, P., Matthews, L., Fry, E. E., de Beer, T. A. P., Theron, J., Rieder, E., Vosloo, W., O'Neill, H. G., Haydon, D. T., and Maree, F. F. (2010). Sequence-based prediction for vaccine strain selection and identification of antigenic variability in Foot-and-Mouth disease virus. *PLoS Comput Biol*, 6(12). 1, 3, 9, 10, 12, 13, 14, 15, 17, 21, 22, 23, 37, 58, 59, 71, 72, 74, 76, 77, 79, 83, 101, 111, 112, 113, 137, 139, 144, 146

Ripley, B. (1979). Algorithm AS 137: Simulating spatial patterns: Dependent samples from a multivariate density. *Journal of the Royal Statistical Society. Series C*, 28(1):109–112. 27

Ruyssinck, J., Huynh-Thu, V., Geurts, P., Dhaene, T., Demeester, P., and Saeys, Y. (2014). NIMEFI: Gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS ONE*, 9(3). 24, 64

Sabatti, C. and James, G. M. (2005). Bayesian sparse hidden components analysis for transcription networks. *Bioinformatics*, 22(6):739–746. 41, 50, 92, 119

Saiz, J. C., Gonzalez, M. J., Borca, M. V., Sobrino, F., and Moore, D. M. (1991). Identification of neutralizing antigenic sites on VP1 and VP2 of type A5 Foot-and-Mouth disease virus, defined by neutralization-resistant variants. *Journal of Virology*, 65(5):2518–2524. 17, 74, 80, 81

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using $\ell 1$-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214. 21, 24, 25, 59, 60, 61, 63, 71, 72, 109

Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S., and Li, W.-H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, 104(15):6283–6288. 19

Skehel, J. J. and Wiley, D. C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual review of biochemistry*, 69(1):531–569. 18

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. 35

Thomas, A., Woortmeijer, R., Barteling, S., and Meloen, R. (1988a). Evidence for more than one important, neutralizing site on Foot-and-Mouth disease virus. Brief report. *Archives of virology*, 99(3-4):237–242. 17

Thomas, A., Woortmeijer, R., Puijk, W., and Barteling, S. (1988b). Antigenic sites on Foot-and-Mouth disease virus type A10. *Journal of Virology*, 62(8):2782–2789. 17

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288. 23

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective (with comments). *Journal of the Royal Statistical Society: Series B*, 73(3):273–282. 23

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228. 93

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594. 34, 35, 56, 64, 85, 93, 95

WHO (2005). Ten things you need to know about pandemic influenza. `https://web.archive.org/web/20091008223707/http://www.who.int/csr/disease/influenza/pandemic10things/en/index.html`. 1

WHO (2009). WHO Influenza fact sheet. 1, 16

WHO (2011). Manual for the laboratory diagnosis and virological surveillance of influenza. `http://whqlibdoc.who.int/publications/2011/9789241548090_eng.pdf`. 7

Wiley, D. C. and Skehel, J. J. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry*, 56:365–394. 18

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320. 24