

Selecting Random Effect Components in a Sparse Hierarchical Bayesian Model for Identifying Antigenic Variability

Vinny Davies^(✉), Richard Reeve, William T. Harvey, and Dirk Husmeier

University of Glasgow, Glasgow, Scotland, UK
v.davies.1@research.gla.ac.uk

Abstract. In Foot-and-Mouth Disease Virus (FMDV), understanding how viruses offer protection against related emerging strains is vital for creating effective vaccines. With testing large numbers of vaccines being infeasible, the development of an *in silico* predictor of cross-protection between virus strains has been a vital area of recent research. The current paper reviews a recent contribution to this area, the SABRE method, a sparse hierarchical Bayesian model which uses spike and slab priors to identify key antigenic sites within FMDV serotypes. WAIC is then combined with the SABRE method and its ability to approximate Bayesian Cross Validation performance in terms of correctly selecting random effect components analysed. WAIC and the SABRE method have then been applied to two FMDV datasets and the results analysed.

Keywords: Model selection · Spike and slab prior · Foot-and-Mouth Disease Virus · Bayesian hierarchical models · WAIC · Cross Validation

1 Introduction

In Foot-and-Mouth Disease Virus (FMDV) where new virus strains continuously emerge, choosing effective vaccines is vital. However FMDV has high genetic variability due to changes in the virus proteins which affect recognition by the host immune system. With the high antigenic variability, FMDV vaccines are only effective against strains that are closely related genetically and antigenically similar to the vaccine strain. As a result it is important to estimate antigenic similarity between different strains and understand how one strain can confer protection against another. The South African Territories types 1 and 2 (SAT1 and SAT2) serotypes both show significant levels of antigenic variability and can be used to explore the relationship between antigenic variation and changes in the protein structure.

In order to understand the relationship between antigenic variation and changes in the protein structure, we need a measure of the antigenic similarity between any two virus strains. Virus Neutralisation (VN) titre is *in vitro* measure which approximates the extent one strain confers protection on another by examining how well one strain (the challenge strain) is able to neutralise a

second strain (the protective strain). Higher values of VN titre indicate that the protective strain offers a higher level of protection against the challenge strain and that the strains are more antigenically similar.

The antigenic differences between virus strains can be explained by changes in the protein structure on the surface of the virus shell. While many changes can occur, only some of these affect recognition by the host immune systems and result in a reduction in the observed VN titre. Identifying the individual areas of the surface exposed proteins, residues, that are considered to be key antigenic regions is critical to understanding the antigenic similarities between viruses. Similarly, understanding how antigenicity is affected by the evolutionary history of the virus strains is important and must be accounted for.

Predicting VN titre, the *in vitro* measure of antigenic similarity, based on the changes in the virus proteins and the shared evolutionary history of the virus strains is complicated by the presence of variation in the VN test, the test to determine VN titre. It is possible that certain virus strains will produce higher or lower VN titre measurements against all other virus strains due a reactivity or immunogenic effect caused by non-antigenic properties of the challenge or protective strains. Additionally the serum used as part of the VN test and the date of the experiment, a proxy for lab conditions, can affect the measured VN titre. Where available, see Sect. 5, the challenge strain, protective strain, serum and date are included as potential random effects and choosing which of these should be used in the analysis is an important problem as including irrelevant components will introduce unnecessary variation into the models.

To account for both the random and fixed effects, Reeve et al. (2010) used classical mixed-effects models, e.g. Pinheiro and Bates (2000), to predict the antigenic similarity between any two virus strains. The authors firstly selected the random effect components and then added terms to account for the evolutionary history of the virus using a forward inclusion algorithm. A univariate test for significance was then carried out on the residue variables with a p-value of less than 0.05 corresponding to an antigenically significant residue. Davies et al. (2014) then introduced a sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution (SABRE) method which was shown to outperform the method of Reeve et al. (2010) in terms variable selection. The first aim of the current work is to review the SABRE method of Davies et al. (2014), propose a slight methodological improvement and show how the SABRE method outperforms both classical mixed-effects models and the mixed-effects Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996; Schelldorfe et al. 2011) in terms of variable selection.

The SABRE method of Davies et al. (2014) combines a Bayesian hierarchical mixed-effects model with spike and slab priors. Hierarchical models allow for consistent inference of all parameters and hyper-parameters with the inferences borrowing strength from the sharing and combination of information; see Gelman et al. (2013). The introduction of spike and slab priors into the model allows for simultaneous model selection not offered by the classical mixed-effects models of Reeve et al. (2010) and improved variable selection over the ℓ_1 regularisation offered by the mixed-effects LASSO (Mohamed et al. 2012).

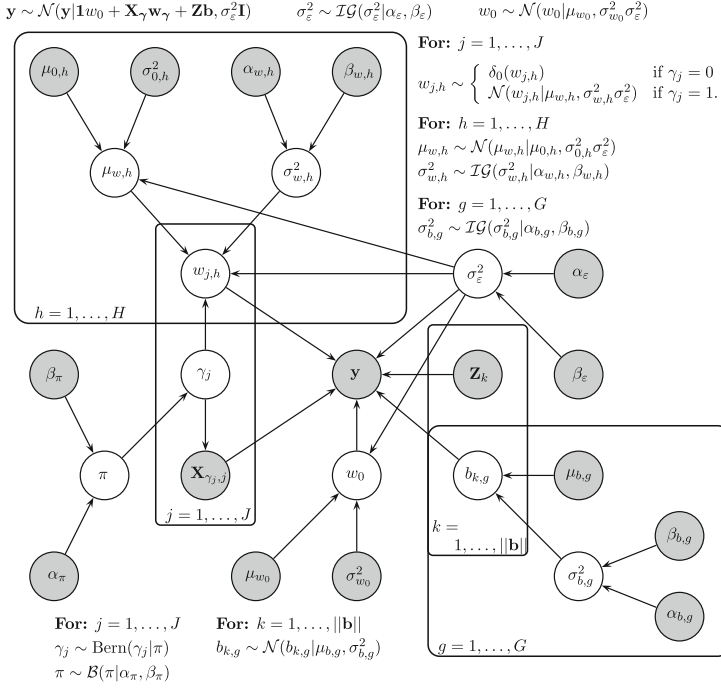


Fig. 1. Compact representation of the SABRE method as a PGM. The *grey* circles refer to the data and fixed (higher-order) hyperparameters, while the *white* circles refer to parameters and hyperparameters that are inferred.

The second and new contribution of the current work is to investigate how best to choose the random effect components that should be included in the SABRE method for each dataset. In larger datasets, where the SABRE method is computationally expensive, using Bayesian Cross Validation (CV) methods is computationally infeasible. The current work investigates whether the Widely Applicable Information Criterion (WAIC) (Watanabe 2010) can be used as a less computationally intensive alternative to Bayesian CV. While WAIC is asymptotically justified, it is unlikely to provide as accurate performance as Bayesian CV in terms of correctly including or excluding random effect components. The purpose of the current study is to understand the size of this reduction in accuracy and assess the suitability of WAIC to be used in larger more computationally demanding datasets, e.g. Harvey et al. (2015).

The final contribution of the current work is to give examples of the SABRE method and WAIC applied to two FMDV datasets. We apply the SABRE method with each possible combination of random effect components and then apply WAIC to find the best choice of model. The results are then analysed in terms of selecting relevant antigenic sites (residues).

2 SABRE Method

In this section we mathematically describe the SABRE method proposed in Davies et al. (2014) with the addition of a separate intercept parameter and increased conjugacy, where the Probabilistic Graphical Model (PGM) is shown in Fig. 1. The model parameters are sampled from the posterior distribution using Markov chain Monte Carlo (MCMC), using the distributions in Sect. 2.5.

2.1 Likelihood

The likelihood of the SABRE method is similar to that of classical mixed-effects models, e.g. Pinheiro and Bates (2000), where the response $\mathbf{y} = (y_1, \dots, y_N)^\top$ is taken as the log VN titre. In classical mixed-effects models, the response, \mathbf{y} , is modelled by a combination of the intercept, w_0 , the explanatory variables, \mathbf{X} , and corresponding regression coefficients, \mathbf{w} , as well as random effects, \mathbf{b} , and the design matrix, \mathbf{Z} . The SABRE method uses a similar structure as can be seen in Fig. 1, however it only includes the relevant explanatory variables, \mathbf{X}_γ , and regression coefficients, \mathbf{w}_γ :

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

The relevance of the j th column of \mathbf{X} is determined by $\gamma_j \in \{0, 1\}$, where feature j is said to be relevant if $\gamma_j = 1$, giving $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)^\top \in \{0, 1\}^J$. We then define \mathbf{X}_γ to be the matrix of relevant explanatory variables with $\|\boldsymbol{\gamma}\|$ columns and N rows, where $\|\boldsymbol{\gamma}\| = \sum_{j=1}^J \gamma_j$ is the number of non-zero elements of $\boldsymbol{\gamma}$. Similarly \mathbf{w}_γ is given as the column vector of regressors, where the inclusion of each parameter is dependent on $\boldsymbol{\gamma}$.

2.2 Noise and Intercept Priors

As with classical mixed-effects models, we assume iid Gaussian noise, σ_ε^2 , for the log VN titre, \mathbf{y} . σ_ε^2 is then given a conjugate Inverse-Gamma prior:

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

where α_ε and β_ε are fixed, as indicated by the grey nodes in Fig. 1.

In addition to being used in the likelihood, (1), σ_ε^2 is also included in the distributions for w_0 , \mathbf{w}_γ , $\boldsymbol{\mu}_\mathbf{w} = (\mu_{w,1}, \dots, \mu_{w,H})^\top$ in Sect. 2.3. These additional relationships, indicated by the edges in Fig. 1, increase information sharing and mean that the error variance in terms of model fit is reflected in the distribution of the regression coefficients. Including these relationships also makes the model conjugate rather than semi-conjugate, see Chap. 3 of Gelman et al. (2013), and allows the creation of an improved sampling strategy based on using collapsed Gibbs sampling, e.g. Andrieu and Doucet (1999).

We also require a distribution for the intercept, w_0 :

$$w_0 \sim \mathcal{N}(w_0 | \mu_{w_0}, \sigma_{w_0}^2 \sigma_\varepsilon^2). \quad (3)$$

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).

2.3 Spike and Slab Priors

Spike and slab priors are known to outperform ℓ_1 methods such as the LASSO both in terms of variable selection and out-of-sample performance (Mohamed et al. 2012). They have been used in a number of forms, but were originally proposed by Mitchell and Beauchamp (1988) as a mixture of a Gaussian distribution and a Dirac spike, as used for the SABRE method in (4). Alternatives to the specification of Mitchell and Beauchamp (1988) include the mixture of two Gaussian distributions proposed by George and McCulloch (1993) and the Binary mask model, e.g. Jow et al. (2014).

The spike and slab prior reflects the relevance of each variable $w_{j,h}$ based on the value of the corresponding latent indicator variable, γ_j . If $\gamma_j = 0$, i.e. the j th variable, \mathbf{X}_j , is irrelevant, then we expect that $w_{j,h} = 0$. Conversely if $\gamma_j = 1$, we think the j th variable is relevant and the corresponding regression coefficient should be non-zero, $w_{j,h} \neq 0$, and we specify a conjugate Gaussian prior. To increase generality we allow the models to have multiple groups of variables $h \in \{1, \dots, H\}$ which are defined by j , i.e. $w_{j,h}$ is shorthand for w_{j,h_j} , but only a single group is used for the results in Sects. 7 and 8.

$$p(w_{j,h}|\gamma_j, \mu_{w,h}, \sigma_{w,h}^2, \sigma_\varepsilon^2) = \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h}|\mu_{w,h}, \sigma_{w,h}^2, \sigma_\varepsilon^2) & \text{if } \gamma_j = 1 \end{cases} \quad (4)$$

for $j \in 1, \dots, J$ and where δ_0 is the delta function. Here we have a spike at 0 and as $\sigma_{w,h}^2 \sigma_\varepsilon^2 \rightarrow \infty$ the distribution, $p(w_{j,h}|\gamma_j = 1)$, approaches a uniform distribution, a slab of constant height.

We give the hyper-parameters of (4) conjugate priors, specifying $\sigma_{w,h}^2$ to have an Inverse-Gamma prior with fixed hyper-parameters $\alpha_{w,h}$ and $\beta_{w,h}$, and $\mu_{w,h}$ a Gaussian prior with fixed hyper-parameters $\mu_{0,h}$ and $\sigma_{0,h}^2$:

$$\sigma_{w,h}^2 \sim \mathcal{IG}(\sigma_{w,h}^2|\alpha_{w,h}, \beta_{w,h}) \quad \mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \sigma_{0,h}^2, \sigma_\varepsilon^2) \quad (5)$$

where σ_ε^2 is again included in the variance of $\mu_{w,h}$ for further conjugacy. We allow $\mu_{w,h}$ to vary in order to reflect our biological understanding of the problem. In the FMDV data we are likely to observe a comparatively large intercept, with negative regression coefficients, $w_{j,h}$, reflecting the fact that any mutational or evolutionary changes are likely to reduce the similarity between virus strains, therefore reducing the measured VN titre.

For convenience we define $\mathbf{w}_\gamma^* = (w_0, \mathbf{w}_\gamma^\top)^\top$ with the following distribution:

$$\mathbf{w}_\gamma^* \sim \mathcal{N}(\mathbf{w}_\gamma^*|\mathbf{m}_\gamma, \sigma_\varepsilon^2 \boldsymbol{\Sigma}_{\mathbf{w}_\gamma^*}) \quad (6)$$

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

The priors related to the latent inclusion parameters, γ , are given by:

$$p(\gamma|\pi) = \prod_{j=1}^J \text{Bern}(\gamma_j|\pi) \quad \pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi) \quad (7)$$

where we define π to be the probability of an individual variable being relevant. Given we do not a-priori know the value of π , it is given a conjugate Beta prior where α_π and β_π are fixed to represent our vague knowledge that only a small proportion of variables should be included in the model; see Sect. 6.

2.4 Random-Effects Priors

The random-effect coefficients are given as $\mathbf{b} = (\mathbf{b}_1^\top, \dots, \mathbf{b}_G^\top)^\top$, where each \mathbf{b}_g relates to a vector of coefficients related to different levels within a particular random effect component or group, $g \in \{1, \dots, G\}$, e.g. challenge strain. Each \mathbf{b}_g has $\|\mathbf{b}_g\|$ coefficients and follows a zero mean Gaussian distribution with a group dependent variance, $\mathbf{b}_g \sim \mathcal{N}(\mathbf{b}_g|\mathbf{0}, \sigma_{b,g}^2 \mathbf{I})$, where \mathbf{I} is the identity matrix. From this we then define all the random-effect coefficients to have a joint distribution $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \mathbf{\Sigma}_\mathbf{b})$, where we define $\mathbf{\Sigma}_\mathbf{b}$ to be a diagonal matrix with $(\sigma_{b,1}^2, \dots, \sigma_{b,1}^2, \sigma_{b,2}^2, \dots, \sigma_{b,G}^2)^\top$ on the diagonal with each $\sigma_{b,g}^2$ being repeated $\|\mathbf{b}_g\|$ times.

We give $b_{k,g}$, the k th coefficient of \mathbf{b} , a Gaussian distribution with a fixed zero mean, $\mu_{b,g} = 0$, and a group dependant variance parameter, $\sigma_{b,g}^2$, which is in turn given an Inverse-Gamma prior:

$$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g}, \sigma_{b,g}^2) \quad \sigma_{b,g}^2 \sim \mathcal{IG}(\sigma_{b,g}^2|\alpha_{b,g}, \beta_{b,g}). \quad (8)$$

The group g is defined by k , i.e. $b_{k,g}$ is shorthand for b_{k,g_k} and the hyper-parameters $\alpha_{b,g}$ and $\beta_{b,g}$ are fixed for each g .

2.5 Posterior Inference

To sample from the posterior distribution we have used an MCMC algorithm. As we have chosen mainly conjugate priors (see Sect. 2), we can use a Gibbs sampling scheme. The conditional dependence relations are shown in the graphical model of Fig. 1, and the detailed forms of the conditional distributions are available from Sect. 10.

Sampling γ 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. Here we have used collapsing methods to achieve faster mixing and convergence:

$$p(\gamma|\theta', \mathbf{X}_\gamma, \mathbf{Z}, \mathbf{y}) \propto \int p(\gamma, \pi, \sigma_\varepsilon^2, \mathbf{w}_\gamma^*, \boldsymbol{\mu}_\mathbf{w}|\theta', \mathbf{X}_\gamma, \mathbf{Z}, \mathbf{y}) d\boldsymbol{\mu}_\mathbf{w} d\mathbf{w}_\gamma^* d\pi d\sigma_\varepsilon^2 \quad (9)$$

where the fixed hyper-parameters (given as grey circles in Fig. 1) have been dropped to improve notational clarity. In the current work we update multiple γ_j

simultaneously via a Metropolis-Hastings step (Metropolis et al. 1953; Hastings 1970), which Davies et al. (2014) found to be more computationally efficient than the more established component-wise Gibbs sampler. In addition to being used within the conditional distribution of γ , collapsing steps are also used for \mathbf{w}_γ^* , $\boldsymbol{\mu}_\mathbf{w}$, σ_ϵ^2 and π . These steps are not detailed here, but using them leads to improved mixing and convergence, e.g. Andrieu and Doucet (1999).

3 Random Effect Selection Methods

3.1 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 estimates 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 computational 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 SABRE method to partial data, \mathbf{y}_{-k} , $\mathbf{X}_{\gamma,-k}$ and \mathbf{Z}_{-k} , and use thinned samples of the model parameters, $\boldsymbol{\theta}^\ell$, for $\ell \in \{1, \dots, I\}$, from $p(\boldsymbol{\theta}|\mathbf{y}_{-k}, \mathbf{X}_{\gamma,-k}, \mathbf{Z}_{-k})$, to estimate the performance on the remaining data, \mathbf{y}_k , $\mathbf{X}_{\gamma,k}$ and \mathbf{Z}_k , using (1). Doing this for each of the K groups gives the 10-fold Bayesian CV performance:

$$p_{CV} = \frac{1}{K} \sum_{k=1}^K \log \frac{1}{I} \sum_{\ell=1}^I p(\mathbf{y}_k | \boldsymbol{\theta}^\ell, \mathbf{X}_{\gamma,k}, \mathbf{Z}_k). \quad (10)$$

3.2 WAIC

WAIC, as proposed in Watanabe (2010) is a useful criterion for selecting the correct model when the underlying model is singular, e.g. the SABRE method. Additionally WAIC has the desirable property of averaging over the posterior distribution, as opposed to the Deviance Information Criterion (DIC) which uses a point estimate. Watanabe (2010) showed how WAIC is asymptotically equivalent to Bayesian LOO-CV and can be computed using the thinned parameter samples, $\boldsymbol{\theta}^\ell$, from the posterior distribution of the full dataset, $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}_\gamma, \mathbf{Z})$, meaning we only have to sample the model parameters once:

$$p_{WAIC} = -2 \sum_{i=1}^N \left(\log \left(\frac{1}{I} \sum_{\ell=1}^I p(y_i | \boldsymbol{\theta}^\ell, \mathbf{X}_{\gamma,i}, \mathbf{Z}_i) \right) - \text{Var}(\log(p(y_i | \boldsymbol{\theta}^\ell, \mathbf{X}_{\gamma,i}, \mathbf{Z}_i))) \right). \quad (11)$$

3.3 Multiple Parameter Spike and Slab Prior

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 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.

4 Simulated Data

To show that the SABRE method proposed in Davies et al. (2014), with the addition of a separate intercept parameter and increased conjugacy, still outperforms classical mixed-effects models and the mixed-effects LASSO in terms of variables selection, we generated 100 simulated datasets. Each of these datasets were given 40 possible variables, where the corresponding coefficients were set to be non zero with probability $\pi \sim \mathcal{U}(0.2, 0.4)$. Additionally 2 random effect components were added, each with 8 levels.

Additionally, to compare WAIC and 10-fold Bayesian CV, we generated 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 viruses the variables remain identical as in the real FMDV 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 FMDV Data

Davies et al. (2014) analysed a dataset from the SAT1 serotype of FMDV, which was originally used in Reeve et al. (2010). Since that analysis, additional data has been collected and been analysed using mixed-effects models in

(Maree et al. 2015). We call this the extended SAT1 dataset and it contains 2125 VN titre measurements with 5 protective and 42 challenge strains, and 221 variables related to the residues and phylogenetic structure. Possible random effects included the serum used to get the VN titre measurement, the challenge strain, the protective strain and the date of the experiment (a proxy to lab conditions).

Reeve et al. (2010) also used a dataset on the SAT2 serotype, although it was not analysed in Davies et al. (2014). The SAT2 dataset contains 320 VN titre measurements from 4 protective and 22 challenge strains. In total there are 148 variables when the residues and phylogenetic data are combined. Possible random effects include the serum used to get the VN titre measurement, the challenge strain and the protective strain.

6 Computational Inference

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

We ran MCMC simulations for 10,000 and 15,000 samples respectively for the simulated and real datasets removing an appropriate proportion for burn-in based on convergence diagnostics. Convergence was determined by computing the potential scale reduction factor (PSRF) (Gelman and Rubin 1992), where a $PSRF \leq 1.05$ for 95 % of the variables was taken as the threshold for convergence. The latent inclusion parameters were sampled using a block Metropolis-Hastings algorithm following Davies et al. (2014).

In general, the fixed hyper-parameters, shown as grey nodes in Fig. 1, 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 (7), which was set to be weakly informative such that $\alpha_\pi = 1$ and $\beta_\pi = 4$. This corresponds to prior knowledge that only a small number of residues or branches have a significant antigenic effect. The remaining hyper-parameters, shown as grey nodes in Fig. 1, 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 $\sigma_{0,h}^2 = 100$ for all h , $\mu_\xi = 0$, $\sigma_\xi^2 = 100$, $\mu_{w_0} = \max(\mathbf{y})$, $\sigma_{w_0}^2 = 100$ and $\alpha_\varepsilon = \beta_\varepsilon = 0.001$. The only informative choice is $\mu_{w_0} = \max(\mathbf{y})$ 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.

7 Simulation Study Results

To compare the variable selection accuracy of the SABRE method compared to the mixed-effects LASSO and classical mixed-effects models we produced receiver

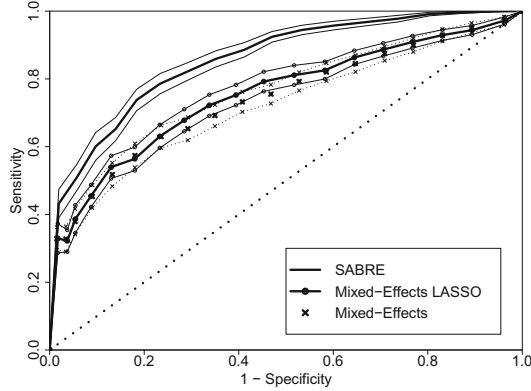


Fig. 2. ROC curves and 95 % confidence intervals for classical mixed-effects models (crosses), the mixed-effects LASSO (solid black and points), and the SABRE method (solid black) when applied to the simulated data in Sect. 4.

operating characteristic (ROC) curves for each of the methods by ordering the inclusion of variables. For the SABRE method we ordered the marginal posterior inclusion probabilities of each variable. For the mixed-effects LASSO, the model was run for different values of the penalty parameter λ and then the so-called LASSO path created (Hastie et al. 2009). Finally for the classical mixed-effects models we ran a forward inclusion algorithm with no stopping point, ranking the variables based on when they were included in the model. For each method the moving average (mean) and standard deviation of the ROC curves for each of the 100 simulated datasets was taken. The mean ROC curve for each method was then plotted with the corresponding 95 % confidence interval in Fig. 2. Using ROC curves to compare the methods gives a more general indication of performance than simply looking at the performance of a specific cut-off point.

Figure 2 shows that the SABRE method outperforms both the mixed-effects LASSO and classical mixed-effects models across all cut-offs. The improved performance is shown by the Area Under the ROC (AUROC) value, where the SABRE method achieves an AUROC value and 95 % confidence interval of 0.87 (0.86,0.88) compared to 0.76 (0.74,0.78) and 0.75 (0.73,0.77) for the mixed-effects LASSO and classical mixed-effects models. One-sided paired t-tests showed that the SABRE method performed better in terms of AUROC value than both the mixed-effects LASSO (p-value < 0.001) and standard mixed-effects models (p-value < 0.001). Similarly the mixed-effects LASSO performed better than standard mixed effects models (p-value = 0.035).

To analyse the performance of WAIC in comparison to 10-fold Bayesian CV, we looked at how accurate each method was at correctly selecting the random effect components used to generate the datasets simulated in Sect. 4. We applied both methods to each of the 16 possible models for each dataset and selected the best model in each case. We then analysed the ability of the best models to correctly include or exclude the random effect components that were used or

Table 1. Results comparing the model selection performance of WAIC compared to 10-fold Bayesian CV. The mean and 95 % confidence intervals are given in terms of correctly including or excluding random effect components in the simulated datasets described in Sect. 4.

	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)

not used to generate each of the datasets. Table 1 gives the results in terms of sensitivity, specificity, predictive accuracies and F-scores.

The results of Table 1 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 min, as opposed to 761 min for 10-fold Bayesian CV.

8 FMDV Results

Having tested the use of WAIC on simulated datasets in Sect. 7, we have then used WAIC to find the best choice of random effect components for two FMDV datasets. After applying WAIC to the extended SAT1 dataset, the best model was found to contain only the protective strain and the serum as random effect components. Choosing these random effect components is an interesting result as the work of Davies et al. (2014), on the original SAT1 dataset of Reeve et al. (2010), was based on using the challenge strain and the serum. The results suggest that it is important to effectively chose the random effect components rather than simply choosing them based on biological prior knowledge.

Choosing the most appropriate random effect components will have an affect on which of the fixed effects are selected by the SABRE method. Based on the model selected by WAIC and using $\hat{\pi} \times J$ as the cut-off, the SABRE method found a total of 9 proven and 24 plausible residues or branches. Classification is based on the residue being experimentally validated in the SAT1 serotype or validated in 4 or more FMDV serotypes (proven), being experimentally validated in 3 or less serotypes (plausible), or from a region not known to be antigenic in any of the FMDV serotypes (implausible). The proven residues come from the 4 known antigenic regions (Grazioli et al. 2006); VP1 C-terminus, VP1 G-H, VP2 B-C and VP3 B-C. Additionally, other residues which are not known to be antigenic were also found in these regions and should be experimentally investigated.

Applying WAIC to the SAT2 dataset resulted in all of the random effect components being included in the model; challenge strain, protective strain and serum. As less is known about the SAT2 serotype we do not classify the branches and residues into different categories, and instead treat the best model as a tool for hypothesis generation. While we do not discuss the results in detail here due to space restrictions and the lack of biological prior knowledge that could be used for assessment, it is worth noting that residues were selected from 3 out of the 4 regions identified in the SAT1 serotype above; VP1 C-terminus, VP1 G-H and VP3 B-C. These residues included a large number in close proximity to each other on the VP1 G-H loop and this area could be of experimental interest.

9 Discussion

In the current work we described and improved the SABRE method and shown how it outperforms established alternatives in terms of variable selection in Fig. 2. In addition we have compared the performance of WAIC and 10-fold Bayesian CV in the context of correctly selecting random effect components. The results, given in Table 1, show that in terms of model selection (concerning the random effects to be included) WAIC achieves a similar performance at a lower computational cost to 10-fold Bayesian CV. We have quantified both the difference in performance and the reduction in computational cost. Finally we have applied the SABRE method with WAIC to two FMDV datasets, identifying a number of antigenically important locations on the surface of the virus shell and a number of residues worthy of investigation.

Further work will develop the SABRE method to better take into account the structure of the data. For any given pair of virus strains tested, the fixed effects will remain the same. By introducing a latent structure into the model we can more precisely account for the data generation process. An additional computational advantage can also be gained for larger datasets, e.g. Harvey et al. (2015), as often the datasets will have far more VN titre measurements than tested virus pairs. A latent variable model could take advantage of the structure and reduce the computational complexity of the conditional distribution for γ .

10 Appendix

For the Gibbs sampling we sample the intercept and regression coefficients together and define $\mathbf{w}_\gamma^* = (w_0, \mathbf{w}_\gamma^\top)^\top$, $\mathbf{X}_\gamma^* = (\mathbf{1}, \mathbf{X}_\gamma)$, $\mathbf{m}_\gamma = (\mu_{w_0}, \mu_{w,1}, \dots, \mu_{w,1}, \mu_{w,2}, \dots, \mu_{w,H})^\top$ and $\Sigma_{\mathbf{w}_\gamma^*} = \text{diag}(\sigma_{\mathbf{w}^*}^2)$ with $\sigma_{\mathbf{w}^*}^2 = (\sigma_{w_0}^2, \sigma_{w,1}^2, \dots, \sigma_{w,1}^2, \sigma_{w,2}^2, \dots, \sigma_{w,H}^2)^\top$. Each $\mu_{w,h}$ and $\sigma_{w,h}^2$ is repeated with length $\|\mathbf{w}_{\gamma,h}\|$ dependent on γ . The Gibbs sampling distributions are then given as follows, with θ' used to

denote all the parameters not on the left of the conditioning bar:

$$\mathbf{w}_\gamma^* | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mathbf{w}_\gamma^* | \mathbf{V}_{\mathbf{w}_\gamma^*} \mathbf{X}_\gamma^{*\top} (\mathbf{y} - \mathbf{Z}\mathbf{b}) + \mathbf{V}_{\mathbf{w}_\gamma^*} \boldsymbol{\Sigma}_{\mathbf{w}_\gamma^*}^{-1} \mathbf{m}_\gamma, \sigma_\varepsilon^2 \mathbf{V}_{\mathbf{w}_\gamma^*}) \quad (12)$$

$$\mathbf{b} | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mathbf{b} | \frac{1}{\sigma_\varepsilon^2} \mathbf{V}_{\mathbf{b}} \mathbf{Z}^\top (\mathbf{y} - \mathbf{X}_\gamma^* \mathbf{w}_\gamma^*), \mathbf{V}_{\mathbf{b}}) \quad (13)$$

$$\sigma_{b,g}^2 | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma_{b,g}^2 | \|\mathbf{b}_g\|/2 + \alpha_{b,g}, \beta_{b,g} + \frac{1}{2} \mathbf{b}_g^\top \mathbf{b}_g) \quad (14)$$

$$\mu_{w,h} | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}(\mu_{w,h} | V_{\mu_{w,h}}^{-1} (\sum (\mathbf{w}_{\gamma,h}) / \sigma_{w,h}^2 + \mu_{0,h} / \sigma_{0,h}^2), \sigma_\varepsilon^2 V_{\mu_{w,h}}) \quad (15)$$

$$\sigma_{w,h}^2 | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \quad (16)$$

$$\mathcal{IG}(\sigma_{w,h}^2 | \|\mathbf{w}_{\gamma,h}\|/2 + \alpha_{w,h}, \beta_{w,h} + \frac{1}{2\sigma_\varepsilon^2} (\mathbf{w}_{\gamma,h} - \mathbf{1}\mu_{w,h})^\top (\mathbf{w}_{\gamma,h} - \mathbf{1}\mu_{w,h}))$$

$$\sigma_\varepsilon^2 | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}(\sigma_\varepsilon^2 | (N + \|\mathbf{w}_\gamma^*\| + H)/2 + \alpha_\varepsilon, \beta_\varepsilon + \frac{1}{2} R_{\sigma_\varepsilon^2}) \quad (17)$$

$$\pi | \theta', \mathbf{X}_\gamma^*, \mathbf{Z}, \mathbf{y} \sim \mathcal{B}(\pi | \alpha_\pi + \|\gamma\|, \beta_\pi + J - \|\gamma\|) \quad (18)$$

where we sample $\sigma_{b,g}^2$, $\mu_{w,h}$ and $\sigma_{w,h}^2$ for each g and h respectively. We also define $\mathbf{V}_{\mathbf{w}_\gamma^*} = (\mathbf{X}_\gamma^{*\top} \mathbf{X}_\gamma^* + \boldsymbol{\Sigma}_{\mathbf{w}_\gamma^*}^{-1})^{-1}$, $\mathbf{V}_{\mathbf{b}} = (\frac{1}{\sigma_\varepsilon^2} \mathbf{Z}^\top \mathbf{Z} + \boldsymbol{\Sigma}_{\mathbf{b}}^{-1})^{-1}$, $V_{\mu_{w,h}} = ((\|\mathbf{w}_{\gamma,h}\|/\sigma_{w,h}^2)^{-1} + (\sigma_{0,h}^2)^{-1})^{-1}$ and $R_{\sigma_\varepsilon^2} = (\mathbf{y} - \mathbf{X}_\gamma^* \mathbf{w}_\gamma^* - \mathbf{Z}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}_\gamma^* \mathbf{w}_\gamma^* - \mathbf{Z}\mathbf{b}) + (\mathbf{w}_\gamma^* - \mathbf{m}_\gamma)^\top \boldsymbol{\Sigma}_{\mathbf{w}_\gamma^*}^{-1} (\mathbf{w}_\gamma^* - \mathbf{m}_\gamma) + \sum_{h=1}^H (\mu_{w,h} - \mu_{0,h})^2 / \sigma_{0,h}^2$ for notational simplicity.

References

- Andrieu, C., Doucet, A.: Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Sig. Process.* **47**(10), 2667–2676 (1999)
- Bates, D., Maechler, M., Bolker, B.: *lme4: linear mixed-effects models using Eigen and Eigenpack* (2013)
- Davies, V., Reeve, R., Harvey, W., Maree, F., Husmeier, D.: Sparse Bayesian variable selection for the identification of antigenic variability in the Foot-and-Mouth Disease Virus. *J. Mach. Learn. Res. Workshop Conf. Proc. (AISTATS)* **33**, 149–158 (2014)
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Ventari, A., Rubin, D.B.: *Bayesian Data Analysis*, 3rd edn. Chapman & Hall, London (2013)
- Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992)
- George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**(423), 881–889 (1993)
- Grazioli, S., Moretti, M., Barbieri, I., Crosatti, M., Brocchi, E.: 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, pp. 287–297, Appendix 43 (2006)
- Harvey, W.T., Gregory, V., Benton, D.J., Hall, J.P., Daniels, R.S., Bedford, T., Haydon, D.T., Hay, A.J., McCauley, J.W., Reeve, R.: Identifying the genetic basis of antigenic change in influenza A (H1N1). *arXiv preprint arXiv:1404.4197* (2015)
- Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer, New York (2009)
- Hastings, W.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970)

- Jow, H., Boys, R.J., Wilkinson, D.J.: Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data. *Stat. Appl. Genet. Mol. Biol.* **13**(5), 531–551 (2014)
- 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., Mahapatra, M.: Tracking the antigenic evolution of Foot-and-Mouth Disease Virus (2015, in submission)
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087–1092 (1953)
- Mitchell, T., Beauchamp, J.: Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **83**(404), 1023–1032 (1988)
- Mohamed, S., Heller, K., Ghahramani, Z.: Bayesian and l_1 approaches for sparse unsupervised learning. In: *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 751–758 (2012)
- Pinheiro, J.C., Bates, D.: *Mixed-Effects Models in S and S-PLUS*. Springer, New York (2000)
- 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., Maree, F.F.: Sequence-based prediction for vaccine strain selection and identification of antigenic variability in Foot-and-Mouth Disease Virus. *PLoS Comput. Biol.* **6**(12), e1001027 (2010)
- Schelldorfer, J., Bühlmann, P., van de Geer, S.: Estimation for high-dimensional linear mixed-effects models using ℓ_1 -penalization. *Scand. J. Stat.* **38**(2), 197–214 (2011)
- Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* **58**, 267–288 (1996)
- Watanabe, S.: Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010)

Computational Intelligence Methods for Bioinformatics
and Biostatistics

12th International Meeting, CIBB 2015, Naples, Italy,

September 10-12, 2015, Revised Selected Papers

Angelini, C.; Rancoita, P.M.; Rovetta, S. (Eds.)

2016, XII, 286 p. 89 illus., Softcover

ISBN: 978-3-319-44331-7