Sparse Bayesian Variable Selection for the Identification of Antigenic Variability in the Foot-and-Mouth Disease Virus

Vinny Davies
School of Mathematics and Statistics,
University of Glasgow, UK
v.davies.1@research.gla.ac.uk

Richard Reeve **William Harvey**
Boyd Orr Centre for Population and Ecosystem Health, and
Institute of Biodiversity, Animal Health and Comparative Medicine,
University of Glasgow, UK

Francois F. Maree
Transboundary Animal Diseases Programme,
Onderstepoort Veterinary Institute, SA

Dirk Husmeier
School of Mathematics and Statistics,
University of Glasgow, UK

Abstract

Vaccines created from closely related viruses are vital for offering protection against newly emerging strains. For Foot-and-Mouth disease virus (FMDV), where multiple serotypes co-circulate, testing 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. Here we describe a novel sparse Bayesian variable selection model using spike and slab priors which is able to predict antigenic variability and identify sites which are important for the neutralisation of the virus. We are able to identify multiple residues which are known to be key indicators of antigenic variability. Many of these were not identified previously using Frequentist mixed-effects models and still cannot be found when an ℓ_1 penalty is used. We further explore how the Markov chain Monte Carlo (MCMC) proposal method for the inclusion of variables can offer significant reductions in computational requirements, both for spike and slab priors in general, and our hierarchical Bayesian model in particular.

1 INTRODUCTION

With the continual emergence of new virus strains, the need to produce effective vaccines has become ever more vital. Predicting where past exposure to a closely related virus strain can offer protection is an important field of research, as testing large numbers of vaccines can be time consuming and expensive. In particular for Foot-and-Mouth Disease Virus (FMDV), where a variety of virus strains co-circulate, 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.

Reeve et al. (2010) used mixed-effects models to account for the variation in virus neutralisation (VN) titre, an *in vitro* measure of antigenic variability; the extent to which one strain confers protection on the other. They identified a specific residue at which substitutions had caused a drop in antigenic variability. Their results have been backed up experimentally in Grazioli et al. (2006).

To achieve this, the authors corrected for the fact that viruses with similar evolutionary paths are likely to be more closely related. They did this by accounting for the shared evolutionary history of the virus strains. Through including the branches of the phylogenetic tree as explanatory variables in the model, they were able to account for the similarities within individual topotypes, groups of genetically similar viruses that have been isolated for long periods, within the tree.

One of the main weaknesses of the models of Reeve et al. (2010) was their reliance on stepwise regression techniques. The method used, forward inclusion using the Holm-Bonferroni correction (Holm 1979), does not explore all variable configurations and can result in a non-optimal solution. A potential extension to this is the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996), which uses an ℓ_1 penalty for simultaneous variable selection. This work has recently been extended to mixed-effects models by

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

Schelldorfer et al. (2011).

A drawback of the classical LASSO, and its close cousin, the Elastic Net (Zou & Hastie 2005), is the way regularisation parameters are selected. Information criteria, such as the Akaike information criterion (AIC), corrected AIC (AICc) (Hurvich & Tsai 1989) and Bayesian information criterion (BIC), are asymptotically justified, but can have poor small-sample performance. Alternative methods include hold-out cross-validation, which reduces the training set size, and k-fold or leave-one-out cross validation, which are potentially biased (Bengio & Grandvalet 2004, Rao et al. 2008). The sub-optimality of these methods compared to a Bayesian approach has been reported before (Dalton & Dougherty 2012).

A more serious drawback is the ℓ_1 regularisation term itself, which in a Bayesian context corresponds to a Laplace prior (Park & Casella 2008). This choice is computationally efficient, leading to a convex optimisation problem for penalised maximum likelihood or Bayesian maximum a posteriori (MAP) inference. However, ℓ_1 regularisation combines the problems of insufficient sparsity of selection with increased bias caused by shrinkage, as discussed in detail in Chapter 13 of Murphy (2012). A preferred alternative, which improves variable selection and avoids excessive shrinkage, is the spike and slab prior proposed in Mitchell & Beauchamp (1988), which, however, leads to a non-convex optimisation problem in a penalised likelihood context. In the present work, we integrate the spike and slab prior into the context of hierarchical Bayesian models, whose advantages have been discussed on various occasions elsewhere; see e.g. Gelman et al. (2004). In particular, Bayesian hierarchical models allow consistent inference of all parameters and hyperparameters, and inference borrows strength by the systematic sharing and combination of information.

The model we propose is designed specifically to deal with the various aspects of the FMDV data. The understanding of antigenic variability requires a focus on selecting variables, while still taking into account the experimental condition under which the data was gathered. To do this we propose a novel Bayesian hierarchical model into which we integrate the prior originally proposed in Mitchell & Beauchamp (1988). In particular we allow for confounding experimental variation through the specification of random effects. We also specify an additional layer in the hierarchical model in order to allow the mean of the coefficients to vary. This comes from biological knowledge of the problem, where we expect a high intercept and a negative impact on the responses from the inclusion of additional variables.

In this paper we evaluate the advantages of our model over the use of both the standard and ℓ_1 penalised mixed-effects model. We use simulated data with random effects designed to mimic variation in experimental conditions to objectively assess prediction and variable selection performance. Finally the model is used on the real life data set of Reeve et al. (2010) in order to assess its capability in identifying surface exposed residues at which substitutions are known to cause a significant drop in antigenic variability. This is done by accounting for the evolutionary history of a strain through including branches of the phylogenetic tree that divide antigenically distinct groups within the virus.

2 CLASSICAL METHODS

Before we describe the novel Bayesian model, we review established classical methods, which are more commonly used within the biological community.

2.1 Classical Mixed-Effects Model

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 coefficient, w_j, to control its influence on the response.

We further set $\mathbf{Z}$ as the matrix of indicators with N rows and $k \in \{1, \ldots, ||\mathbf{b}||\}$ columns. $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_G^\top)^\top$ represents a vector of parameters related to each of the groups $g \in \{1, \ldots, G\}$, where each $\mathbf{b}_g$ has length $||\mathbf{b}_g||$ and $||\mathbf{b}|| = \sum_{g=1}^{G} ||\mathbf{b}_g||$. For more details on mixed-effects models see Pinheiro & Bates (2000).

The model is therefore defined as:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{1}$$

where we assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}})$, where we define $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma_\varepsilon^2 \mathbf{I}$ and $\boldsymbol{\Sigma}_{\mathbf{b}} = diag(([\boldsymbol{\sigma}_{b,1}^2]^\top, \ldots, [\boldsymbol{\sigma}_{b,G}^2]^\top)^\top)$ for notational simplicity and $\mathbf{I}$ is the identity matrix. Integrating over $\mathbf{b}$ gives us the likelihood:

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

In classical mixed-effects models, model comparison techniques must be used to choose which variables are included within the model. To get a sparse model, Reeve et al. (2010) used forward inclusion, making an adjustment for multiple testing using the Holm-Bonferroni correction. They firstly included the variables correcting for the shared evolutionary paths using prior biological knowledge, before adding the residue data.

2.2 LASSO

A classical alternative to forward variable selection is the LASSO of Tibshirani (1996, 2011), which allows for simultaneous variable selection. In the simplest case of linear regression, this gives the following parameter estimates:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \lambda \sum_{j=1}^{J} |w_j| \right\}. \quad (3)$$

This is a convex optimisation problem, for which a variety of fast and effective algorithms exist (e.g. Hastie et al. (2009)). The effect of eq. 3 is to simultaneously shrink and prune parameters $\mathbf{w}$, thereby promoting a sparse model. The degree of sparsity depends on the regularization parameter λ, which can be optimised with cross-validation or information criteria, e.g. BIC.

A recent 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 variance σ_b^2 and the variance of the noise σ_ε^2 as:

$$(\hat{\mathbf{w}},\hat{\sigma}_b^2,\hat{\sigma}_\varepsilon^2) = \underset{\mathbf{w},\sigma_b^2>0,\sigma_\varepsilon^2>0}{\operatorname{argmin}} \left\{ \tfrac{1}{2}\log|\mathbf{V}| + \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{w}) + \lambda \sum_{j=1}^{J} |w_j| \right\} \quad (4)$$

where $\mathbf{V} = \mathbf{Z}\mathbf{\Sigma_b}\mathbf{Z}^\top + \sigma_\varepsilon \mathbf{I}$. In the package of Schelldorfer et al. (2011), this is minimised using a block coordinate gradient descent scheme. To select the value of λ we test the use of BIC, as recommended in Schelldorfer et al. (2011), and AICc (Hurvich & Tsai 1989).

We point out that the mixed effects Lasso of Schelldorfer et al. (2011) has only been developed for a single random effect. To deal with multiple random effects, the Cartesian product of several random effects has to be mapped onto a single random effect, which can lead to excessive model complexity. We also note that to the best of our knowledge, a mixed-effects model version of the Elastic net (Zou & Hastie 2005) has not yet been developed.

3 NOVEL BAYESIAN METHOD

To perform variable selection within Bayesian statistics, we must firstly define the model that is used. This is usually done by constructing the posterior distribution using Bayes' rule:

$$p(\boldsymbol{\gamma}|\mathcal{D},\boldsymbol{\theta}') \propto p(\boldsymbol{\gamma})p(\mathcal{D},\boldsymbol{\theta}'|\boldsymbol{\gamma}). \quad (5)$$

We then sample the parameters using Markov chain Monte Carlo (MCMC), where we are interested in $\boldsymbol{\gamma}$, a vector of latent indicators of whether a variable is included in the regression model. Each parameter is sampled subject to the data, $\mathcal{D}$, and the other model parameters, $\boldsymbol{\theta}'$.

3.1 Likelihood

The likelihood for our Bayesian variable selection model is similar to the classical mixed-effects model described in Section 2.1. However instead of including all the variables, $\mathbf{X}$, and their corresponding regression coefficient, we now only include relevant variables, $\mathbf{X}_{\boldsymbol{\gamma}}$, and regressors, $\mathbf{w}_{\boldsymbol{\gamma}}$:

$$p(\mathbf{y}|\mathbf{w}_{\boldsymbol{\gamma}},\mathbf{b},\sigma_\varepsilon^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \mathbf{\Sigma}_\varepsilon) \quad (6)$$

The relevance of variable j is determined by $\gamma_j \in \{0,1\}$, where feature j is said to be relevant if $\gamma_j = 1$. This gives $\boldsymbol{\gamma} = (\gamma_0,\gamma_1,\ldots,\gamma_J)^\top \in \{0,1\}^J$ 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 $||\boldsymbol{\gamma}||$ columns and N rows, where $||\boldsymbol{\gamma}|| = \sum_{j=0}^{J}\gamma_j$, the number of non-zero elements of $\boldsymbol{\gamma}$. Similarly $\mathbf{w}_{\boldsymbol{\gamma}}$ is given as the column vector of regressors, where the inclusion of each parameter is again dependent on $\boldsymbol{\gamma}$.

3.2 Priors

For computational convenience, conjugate priors have been chosen where possible. In this manner, as in classical mixed-effects models, we choose each $b_{k,g}$ to have group dependent Gaussian priors:

$$b_{k,g} \sim \mathcal{N}(b_{k,g}|\mu_{b,g},\sigma_{b,g}^2). \quad (7)$$

We define this to have a fixed mean $\mu_{b,g} = 0$ and a common variance parameter for each random effect group g. Further to this, we put a conjugate Inverse-Gamma prior on each $\sigma_{b,g}^2$:

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

where $\alpha_{b,g}$ and $\beta_{b,g}$ are fixed hyper-parameters for each g. This gives the model the flexibility to learn each $\sigma_{b,g}^2$ instead of predefining their values.

The prior for $\mathbf{w}_{\boldsymbol{\gamma}}$ is set in the manner proposed in Mitchell & Beauchamp (1988) such that it reflects whether a feature is relevant. 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 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

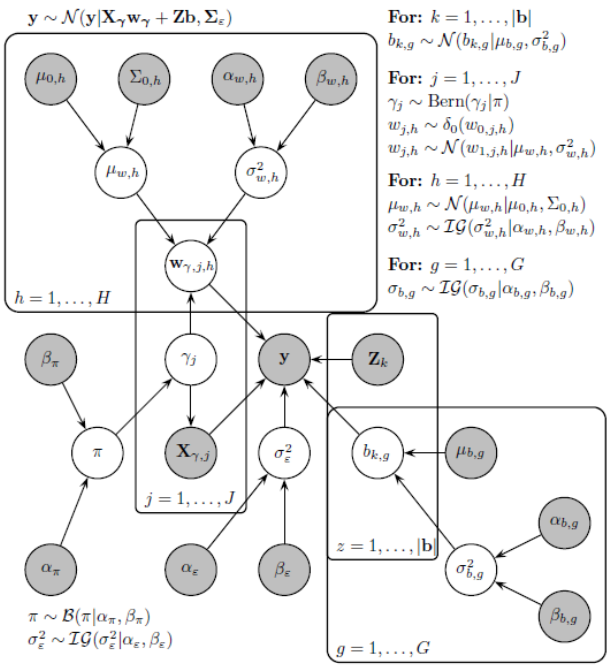


Figure 1: Compact representation of the complex spike and slab model as a Directed Acyclic Graph (DAG). The *grey* circles refer to data and hyper-parameters which are fixed, while the *white* circles refer to parameters that are inferred with MCMC.

chosen when the feature is relevant:

$$p(w_{j,h}|\gamma_{j,h},\mu_{w,h},\sigma^2_{w,h}) = \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_{j,h}|\mu_{w,h},\sigma^2_{w,h}) & \text{if } \gamma_j = 1. \end{cases} \quad (9)$$

where δ_0 is the delta function. Here we have a spike at the mean, $\mu_{w,h}$, and as $\sigma^2_{w,h} \to \infty$ 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. Similar non-singular versions of this prior are also possible (George & McCulloch 1993, 1997).

Through giving each group h a separate hyper-parameter $\sigma^2_{w,h}$ in eq. 9, we leave the model open to penalising the groups of variables to different degrees through the priors:

$$\sigma^2_{w,h} \sim \mathcal{IG}(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h}). \quad (10)$$

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:

$$\mu_{w,h} \sim \mathcal{N}(\mu_{w,h}|\mu_{0,h}, \Sigma_{0,h}) \quad (11)$$

where the hyper-parameters $\mu_{0,h}$ and $\Sigma_{0,h}$ are fixed. This specification comes from the expected biological values of each regression coefficients $w_{j,h}$. In the FMDV 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. Similarly, traversing a significant branch of the phylogenetic tree is likely to cause differences between the strains.

As with the classical mixed-effects model in Section 2.1, we assume the errors are independent and identically distributed. Again specifying a conjugate prior gives us an Inverse-Gamma distribution:

$$\sigma^2_\varepsilon \sim \mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon). \quad (12)$$

where the hyper-parameters α_ε and β_ε are fixed.

A prior must also be given for $\boldsymbol{\gamma}_{1:J}$, the parameters which determine the relevance of the variables:

$$p(\boldsymbol{\gamma}_{1:J}|\pi) = \prod_{j=1}^{J} \text{Bern}(\gamma_j|\pi) \quad (13)$$

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 & 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:

$$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi). \quad (14)$$

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 $\alpha_\pi\beta_\pi/((\alpha_\pi + \beta_\pi)^2(\alpha_\pi + \beta_\pi + 1)) \to 0$.

3.3 Posterior

Using Bayes' theorem we can construct the posterior distribution for the inferred parameters $\boldsymbol{\theta}_1 = (\boldsymbol{\gamma}, \mathbf{w}_{\boldsymbol{\gamma}}, \mathbf{b}, \boldsymbol{\sigma}_{\mathbf{b}}, \boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\sigma}_{\mathbf{w}}, \sigma^2_\varepsilon, \pi)$ given the fixed parameters $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}_{\mathbf{b}}, \boldsymbol{\beta}_{\mathbf{b}}, \boldsymbol{\alpha}_{\mathbf{w}}, \boldsymbol{\beta}_{\mathbf{w}}, \boldsymbol{\mu}_{\mathbf{0}}, \boldsymbol{\Sigma}_{\mathbf{0}}, \alpha_\varepsilon, \beta_\varepsilon, \alpha_\pi, \beta_\pi)$, where we define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. This combines the likelihood and priors specified, as shown in Figure 1:

$$\begin{aligned} p(\boldsymbol{\theta}_1|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}_2) \propto\ & \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Zb}, \boldsymbol{\Sigma}_\varepsilon)\mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}})\mathcal{IG}(\sigma^2_\varepsilon|\alpha_\varepsilon, \beta_\varepsilon) \\ & \times \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\mu_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}) \prod_{g=1}^{G} \{\mathcal{IG}(\sigma^2_{b,g}|\alpha_{b,g}, \beta_{b,g})\} \\ & \times \prod_{h=1}^{H} \{\mathcal{IG}(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h})\mathcal{N}(\mu_{w,h}|\mu_{0,h}, \Sigma_{0,h})\} \\ & \times \prod_{i=1}^{J} \{\text{Bern}(\boldsymbol{\gamma}_j|\pi)\}\mathcal{B}(\pi|\alpha_\pi, \beta_\pi) \end{aligned} \quad (15)$$

where $\boldsymbol{\Sigma}_{\mathbf{b}} = diag(([\boldsymbol{\sigma}^2_{b,1}]^\top, \ldots, [\boldsymbol{\sigma}^2_{b,G}]^\top)^\top)$ and each $\boldsymbol{\sigma}^2_{b,g}$ has length $||\boldsymbol{\sigma}^2_{b,g}||$. We similarly set $\boldsymbol{\Sigma}_{\mathbf{w}_{\boldsymbol{\gamma}}} = diag(([\boldsymbol{\sigma}^2_{w,1}]^\top, \ldots, [\boldsymbol{\sigma}^2_{w,H}]^\top)^\top)$ and $\boldsymbol{\mu}_{\mathbf{w}} = (\boldsymbol{\mu}^\top_{w,1}, \ldots, \boldsymbol{\mu}^\top_{w,H})^\top$ where each $\boldsymbol{\sigma}^2_{w,h}$ and $\boldsymbol{\mu}_{w,h}$ have length $||\boldsymbol{\sigma}^2_{w,h}|| = ||\boldsymbol{\mu}_{w,h}||$.

3.4 Posterior Inference

In order to explore the posterior distribution of the parameters we use an MCMC algorithm. Having chosen conjugate priors where possible means we can run a Gibbs sampler for the majority of parameters (Ripley 1979, Geman & Geman 1984). The only exception is $\boldsymbol{\gamma}$, although it is possible to use component-wise Gibbs sampling with a small adaptation; see Section 3.5.1. The conditional distributions for those parameters amenable to standard Gibbs sampling are:

$$\mathbf{w}_{\boldsymbol{\gamma}}|\boldsymbol{\theta}' \sim \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\mathbf{X}^\top_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}^{-1}_{\varepsilon}(\mathbf{y}-\mathbf{Z}\mathbf{b})+ \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}_{\boldsymbol{\gamma}}}\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}_{\boldsymbol{\gamma}}}) \quad (16)$$

$$\mathbf{b}|\boldsymbol{\theta}' \sim \mathcal{N}(\mathbf{b}|\mathbf{V}_{\mathbf{b}}\mathbf{Z}^\top\boldsymbol{\Sigma}^{-1}_{\varepsilon}(\mathbf{y}-\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}}), \mathbf{V}_{\mathbf{b}}) \quad (17)$$

$$\sigma^2_{b,g}|\boldsymbol{\theta}' \sim \mathcal{IG}(\sigma^2_{b,g}|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}^\top_g\mathbf{b}_g) \quad (18)$$

$$\mu_{w,h}|\boldsymbol{\theta}' \sim \mathcal{N}(\mu_{w,h}| \boldsymbol{\Sigma}^{-1}_{\mu}(\boldsymbol{\Sigma}^{-1}_{\mathbf{w}}\mathbf{w}_{\boldsymbol{\gamma},h} + \boldsymbol{\Sigma}^{-1}_0\mu_{0,h}), \boldsymbol{\Sigma}_{\mu}) \quad (19)$$

$$\sigma^2_{w,h}|\boldsymbol{\theta}' \sim \mathcal{IG}(\sigma^2_{w,h}|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h}+ \tfrac{1}{2}(\mathbf{w}_{\boldsymbol{\gamma},h} - \boldsymbol{\mu}_{\boldsymbol{\gamma},h})^\top(\mathbf{w}_{\boldsymbol{\gamma},h} - \boldsymbol{\mu}_{\boldsymbol{\gamma},h})) \quad (20)$$

$$\sigma^2_{\varepsilon}|\boldsymbol{\theta}' \sim \mathcal{IG}(\sigma^2_{\varepsilon}|N/2 + \alpha_{\varepsilon}, \beta_{\varepsilon} + \tfrac{1}{2}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b})) \quad (21)$$

$$\pi|\boldsymbol{\theta}' \sim \mathcal{B}(\pi|\alpha_{\pi} + ||\boldsymbol{\gamma}_{1:N}||, \beta_{\pi} + J - ||\boldsymbol{\gamma}_{1:N}||) \quad (22)$$

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

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 3.5, however in order to do this we need a conditional distribution:

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}') \propto p(\boldsymbol{\gamma}_{1:J}|\pi)\int \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}_{\varepsilon}) \mathcal{N}(\mathbf{w}_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})d\mathbf{w}_{\boldsymbol{\gamma}} \quad (23)$$

$$\propto \pi^{||\boldsymbol{\gamma}_{1:N}||}(1-\pi)^{J-||\boldsymbol{\gamma}_{1:N}||} \mathcal{N}(\mathbf{y}|\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\mathbf{w}} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}_{\varepsilon} + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{X}^\top_{\boldsymbol{\gamma}}) \quad (24)$$

where there are J variables. Here we have used a collapsing step as in Sabatti & James (2005), integrating out $\mathbf{w}_{\boldsymbol{\gamma}}$ through the application of standard Gaussian integrals (Bishop 2006) to reduce the computational requirements. The normalisation constant is not required in eq. 24 as it cancels out in all of the methods discussed in Section 3.5: eq. 26 and eq. 29.

3.5 Sampling the Latent Indicators

Multiple methods have been proposed for sampling the latent variables, $\boldsymbol{\gamma}$. In this paper we look at two of these in particular; the component-wise Gibbs sampling approach of George & McCulloch (1993) and through a Metropolis-Hastings step where we can propose changes to multiple parameters simultaneously for a computational improvement (Metropolis et al. 1953, Hastings 1970).

3.5.1 Component-wise Gibbs Sampling

Following George & McCulloch (1993) we can use a component-wise Gibbs sampler to consecutively sample each γ_j from $\boldsymbol{\gamma}$ in a random order. To do this we first define a conditional distribution for γ^i_j, the value of the ith iteration of γ_j, from eq. 24:

$$\gamma^i_j \sim p(\gamma^i_j|\pi, \mathbf{b}, \sigma^2_{\varepsilon}, \boldsymbol{\mu}_{\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\gamma}^i_{-j}) \quad (25)$$

where $\boldsymbol{\gamma}^i_{-j} = (\gamma^i_1, \ldots, \gamma^i_{j-1}, \gamma^{i-1}_{j+1}, \ldots, \gamma^{i-1}_J)$ in the case of ordered inclusion parameters. Each distribution can then be given a Bernoulli distribution with probability:

$$P(\gamma^i_j = 1|\pi, \mathbf{b}, \sigma^2_{\varepsilon}, \boldsymbol{\mu}_{\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\gamma}^i_{-j}) = \frac{a}{a+b} \quad (26)$$

where we define:

$$a = p(\gamma^i_j = 1|\pi, \mathbf{b}, \sigma^2_{\varepsilon}, \boldsymbol{\mu}_{\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\gamma}^i_{-j}) \quad (27)$$

$$b = p(\gamma^i_j = 0|\pi, \mathbf{b}, \sigma^2_{\varepsilon}, \boldsymbol{\mu}_{\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\gamma}^i_{-j}). \quad (28)$$

3.5.2 Metropolis-Hastings Sampling

Unlike with the Gibbs sampling approach, sampling via a Metropolis-Hastings step leads to some proposals being rejected. However an advantage can be gained through proposing multiple variables simultaneously. We define the acceptance rate for the Metropolis-Hastings step in this case as:

$$\alpha(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}^{i-1}) := \min\left\{\frac{q(\boldsymbol{\gamma}^{i-1}|\boldsymbol{\gamma}^*)p(\boldsymbol{\gamma}^*|\boldsymbol{\theta}')}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{i-1})p(\boldsymbol{\gamma}^{i-1}|\boldsymbol{\theta}')}, 1\right\} \quad (29)$$

where $q(.)$ is a proposal density, which we set to be: $q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{i-1}) = q(\boldsymbol{\gamma}^*) = \text{Bern}(\boldsymbol{\gamma}^*|\pi)$. Proposed moves for groups of randomly ordered inclusion parameters, $\boldsymbol{\gamma}*$, are then accepted if $\alpha(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}^{i-1})$ is greater than a random variable $u \sim \mathcal{U}[0,1]$.

Tuning the number of simultaneous proposals can give significant computational improvements to the algorithm. Changing only a few variables in $\boldsymbol{\gamma}$ gives a high acceptance rate, but takes a long time to cycle through the variables leading to poor mixing. Conversely if we propose too many simultaneously we reject too many proposals to be efficient. In this paper we investigate how best to tune the proposals and compare this against the component-wise Gibbs sampler.

4 DATA

4.1 Simulated Data

20 sets of data were simulated to reflect the structure of the real FMDV data such that there were two groups of variables. For each response, 20 count variables were simulated from a Poisson distribution and then an additional set of 10 binary variables was generated. These were both simulated such that there was a basic correlation within the groups in order to reflect some of the correlations found in the real data. Additionally 10 data sets were given one group of random effects, with the remaining sets given two groups, in order to mimic random variation in the real experimental data.

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.

4.2 Real Life Example

The FMDV data analysed in Reeve et al. (2010) comes from sub-Saharan Africa, where the virus is endemic. The authors evaluated data on two different serotypes, but in this paper we only consider the South African Territories (SAT) type 1 serotype. This contains 246 measures of VN titre, an *in vitro* measure of whether the sites that contribute to the neutralization of the virus remain sufficiently similar to cross-react. The VN titre measures have been shown to be log-normally distributed (Reeve et al. 2010), so we take the log of the data as our response.

The fixed effects used in our model are divided into two groups. The first contains 137 variables which count amino acid mutations (substitutions) at each residue in the proteins that form the virus shell. The second group consists of 38 variables which indicate whether a specific branch in the phylogenetic tree was traversed between the protective and challenge strains, where the protective strain is the strain the animal has been vaccinated with and the challenge is the strain it is tested against. Additionally, the model requires the inclusion of two groups of random effects. The first is the challenge strain, used to account for the variability in reactivity of the viruses. The second accounts for the sera taken from the individual animals exposed to a strain.

5 SIMULATIONS

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.

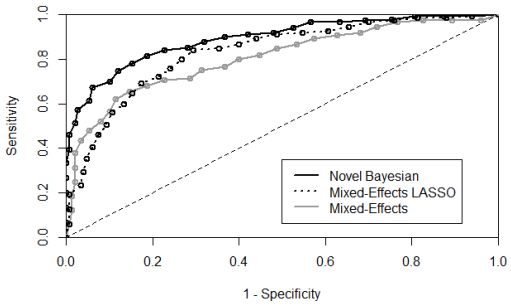
For our MCMC chains we sampled 10,000 and 50,000 iterations respectively for the simulated and real data. The fixed parameters, $\boldsymbol{\theta}_2$, were all set to represent vague priors: $\boldsymbol{\alpha_b} = \boldsymbol{\beta_b} = \boldsymbol{\alpha_w} = \boldsymbol{\beta_w} = \mathbf{0.001}$, $\boldsymbol{\mu_0} = \mathbf{0}$, $\boldsymbol{\Sigma_0} = \mathbf{100}$, $\alpha_\varepsilon = \beta_\varepsilon = 0.001$, $\alpha_\pi = \beta_\pi = 1$. The only exception to this is the shape parameter for the variance of the intercept, which is given as $\alpha_{w,1} = 1.501$ 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 test the convergence of the parameters, 4 chains were ran for each model and a potential scale reduction factor (PSRF) (Gelman & Rubin 1992) was computed from the within-chain and between-chain variances using the *R* package *coda* (Plummer et al. 2006). We take a PSRF ≤ 1.1 as a threshold for convergence and terminate the burn-in when this is satisfied for 95% of the variables.

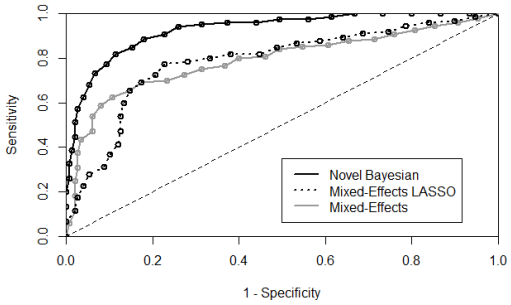
To analyse the best proposal method we tested the component-wise Gibbs sampler and several specifications of the Metropolis-Hastings sampler on the FMDV data set. For the Metropolis-Hastings sampler, we proposed the inclusion or exclusion of the variables in groups of 4, 8, 16, 32 and 64. We analysed convergence by monitoring the percentage of variables with a PSRF ≤ 1.1 as in Grzegorczyk & Husmeier (2013).

6 RESULTS

Our results compare the accuracy of predicting significant variables, as well as out-of-sample predictive performance for all methods on the simulated data. When calculating the out-of-sample predictive performance, we defined a 0.5 marginal probability for the selection criteria of the novel Bayesian model, allowing a small amount of deviation to account for the variation in chains due to the finite effective sample size. For the mixed-effects models, forward inclusion was used with the Holm-Bonferroni correction at a fixed significance threshold of 0.05. Both AICc and BIC were used for selecting the regularisation parameter of the mixed-effects LASSO. These selection criteria were then used to compute out-of-sample likelihood. The same processes were used to evaluate the model performance on the FMDV data and compare the convergence speed of the different proposal methods.



(a) One Random-Effect Group



(b) Two Random-Effect Groups

Figure 2: ROC curves for classical mixed-effects (grey), mixed-effects LASSO (black dotted), and the novel Bayesian (black) models when applied to the simulated data. The simulated data was generated with (a) one and (b) two random effect groups; see section 4.1.

To investigate the accuracy of the selection of significant variables we produced receiver operating characteristic (ROC) curves for each of the methods by ordering the inclusion of variables. This can be achieved for the novel Bayesian method by ordering of the variables using their predicted marginal posterior probabilities. For the standard mixed-effects models this is done by removing the significance threshold and ranking the edges by order of inclusion. Finally for the mixed-effects LASSO we predict models for a variety of different penalty parameters, λ, to create the so-called LASSO path (Hastie et al. 2009). This defines a ranking of the variable from which again a ROC curve can be obtained.

6.1 Simulation Study

Figure 2 shows ROC curves for the classical mixed-effects, mixed-effects LASSO and novel sparse Bayesian selection model. ROC curves show model classification performance under different levels of sensitivity and specificity. This is more general than evaluating performance at a specific cut-off point, determined using model selection criteria. ROC curves also provide a convenient comparison measure in the form of the area under curve (AUC) value.

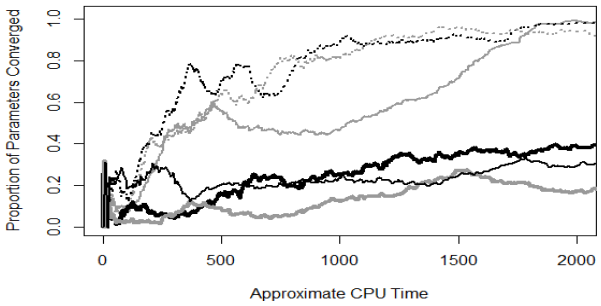


Figure 3: Convergence diagnostics. The lines show the proportion of parameters that have converged (PSRF $\leq$ 1.1) when using component-wise Gibbs sampling (black) and Metropolis-Hastings sampling proposing 4 (grey), 8 (black dashed), 16 (grey dashed), 32 (black thick) and 64 (grey thick) inclusion parameters simultaneously.

For two random effects groups (Figure 2b), the proposed Bayesian model, AUC = 0.93, consistently outperforms the mixed effects LASSO, AUC = 0.79, and standard mixed effects model, AUC = 0.79. This is presumably a consequence of the fact that the mixed-effects LASSO, as developed by 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. For data with a single random effect (Figure 2a), the novel Bayesian method still achieves a greater AUC value, 0.89, than the LASSO, 0.83, and standard mixed effects model, 0.81. Note that realistic data need to be modeled with more than one random effect, though. To our knowledge, a mixed effects LASSO for such data has not been developed.

In addition to the comparison of AUC values, we also looked at the predictive performance. For the data with 2 groups of random effects the novel Bayesian method got a mean out-of-sample log-likelihood of -113.8, outperforming the mixed effect LASSO 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.

6.2 Foot-and-Mouth Disease Virus Data

Figure 3 shows that proposing a larger proportion of 8 or 16 binary selection hyperparameters, γ, simul-

taneously in a 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 Gibbs sampling should not always be the default method of choice, and that further improvements may be obtained by including posterior correlations in the proposal moves (see Section 8).

With respect to the evaluation of the prediction, we need to point out that the proposed novel Bayesian 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 'relevant' residues (as informed by the 'gold-standard'; see below) and the branches of the phylogenetic tree as potential explanatory variables. We then iteratively excluded strongly correlated 'non-relevant' variables 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, where mutations potentially allow the virus to escape the host immune response. For evaluation, we used a list of known 'true positives' from Grazioli et al. (2006) - these are residues in exposed regions of the virus protein known to be targeted by the immune system. We also used a list of 'true negatives', which are areas not found in any study of FMDV antibody targets (see Reeve et al. (2010) and references therein) - these typically lie in buried regions of the virus protein that are inaccessible to the immune system. The predictions are shown in Figure 4. It can be seen that the novel Bayesian hierarchical model finds no 'false positive', while also showing an increased number of 'true positives'. In combination with the fully automated inference procedure this can be seen as a method improvement.

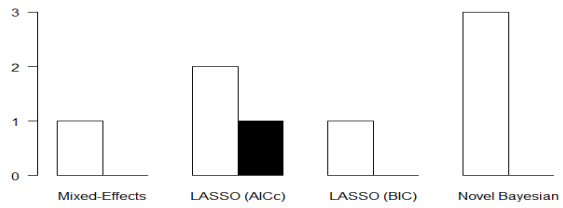


Figure 4: Bar plot showing 'true positives' (white) and 'false positive' (black) for the mixed-effects model results of Reeve et al. (2010), the mixed-effects LASSO using AICc and BIC (Schelldorfer et al. 2011) and the novel Bayesian variable selection model.

7 CONCLUSION

We have addressed the problem of identifying residues responsible for changes in antigenic variability within FMDV. We have proposed a novel sparse Bayesian variable selection scheme based on spike and slab priors, which outperforms competing methods (Figure 2 and 4). In the process we have identified three key residues that are known to be critical to understanding cross-protection between virus.

Further to this we have investigated the sampling of the latent inclusion variables, $\boldsymbol{\gamma}$. We have shown that proposing multiple moves simultaneously through Metropolis-Hastings sampling can give a significant computational improvement over the more widely used component-wise Gibbs sampler (Figure 3).

8 FURTHER WORK

Further work on this paper comes in several forms. Firstly, there may be room for improvement by replacing the Inverse-Gamma prior for the variances of the random effects ($\sigma^2_{b,g}$ in Figure 1) by a half-Cauchy distribution. As discussed in Gelman (2006), this may result in a reduced dependence of the results on the prior as well as an improvement in the convergence of the MCMC simulations. Secondly, the model can be extended to include a spike and slab prior for the selection of random effects (i.e. the $b_{k,g}$'s in Figure 1). Thirdly, improved proposal distributions that account for the estimated posterior correlation in the binary inclusion hyperparameters, $\boldsymbol{\gamma}$, could potentially improve convergence, in the same way as for the continuous case (Haario et al. 2006). A method to generate correlated binary variables has been proposed (Leisch et al. 2012). However, to apply this method in the context of MCMC, a proper proposal distribution has to be developed, which has to enter the Metropolis-Hastings ratio. Finally we would like to extend the model to further serotypes and diseases. In particular combining all available data for FMDV serotypes would lead to a larger, more complete data set, which would give us the best chance of identifying all the key residues associated with changes in antigenic variability.

References

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

Bengio, Y. & Grandvalet, Y. (2004), 'No unbiased estimator of the variance of k-fold cross-validation', *The Journal of Machine Learning Research* **5**, 1089–1105.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer, Singapore.

Dalton, L. & Dougherty, E. (2012), 'Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error - part II: Consistency and performance analysis', *Signal Processing, IEEE Transactions on* **60**(5), 2588–2603.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', *Bayesian Analysis* **1**(3).

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall.

Gelman, A. & Rubin, D. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**, 457–511.

Geman, S. & 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.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**, 339–373.

Grazioli, S., Moretti, M., Barbieri, I., Crosatti, M. & 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', pp. 287–297. Appendix 43.

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

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

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*, Springer.

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

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

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

Leisch, F., Weingessel, A. & Hornik, K. (2012), *bindata: Generation of Artificial Binary Data*.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21**(6), 1087–1092.

Mitchell, T. & Beauchamp, J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association* **83**(404), 1023–1032.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT Press, Cambridge, MA.

Park, T. & Casella, G. (2008), 'The Bayesian lasso', *Journal of the American Statistical Association* **103**(482).

Pinheiro, J. C. & Bates, D. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', *R News* **6**(1), 7–11.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rao, R. B., Fung, G. & Rosales, R. (2008), 'On the dangers of cross-validation. an experimental evaluation', *SIAM Data Mining* .

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. (2010), 'Sequence-based prediction for vaccine strain selection and identification of antigenic variability in Foot-and-Mouth disease virus', *PLoS Comput Biol* **6**(12).

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.

Sabatti, C. & James, G. M. (2005), 'Bayesian sparse hidden components analysis for transcription networks', *Bioinformatics* **22**(6), 739–746.

Schelldorfer, J., Bühlmann, P. & 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.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

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.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B* **67**(2), 301–320.