HIGH DIMENSIONAL REGRESSION TECHNIQUES FOR COMPLEX DATA

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Statistics

by
Chase Joyner
December 2019

Accepted by:
Dr. Christopher McMahan, Committee Chair
Dr. Andrew Brown
Dr. Brook Russell
Dr. Xiaoqian Sun

Abstract

This dissertation focuses on developing mixed effects models for large scale and complex data. Our motivating applications involve areas where this data is common, including epidemiological studies, environmental sciences, and genetics. Two key attributes for most of the modeling techniques discussed in this dissertation are that they scale easily to large data and that they achieve full variable selection, which is often a desirable trait in mixed effects models. These attributes are primarily handled in two ways. The first is with carefully constructed latent variables that we introduce to make the posterior distributions more tractable. This allows a Markov chain Monte Carlo (MCMC) sampler to be carried out with Gibbs steps, which results in efficient computation of posterior estimates, especially in large data scenarios. The second is through a decomposition of the covariance matrix associated with the random effects and with the use of spike and slab priors, we can achieve full variable selection in not only the fixed effects, but also the random effects. The finite sample performance of our techniques are assessed through extensive simulations and are used to analyze motivating data sets, which includes data from group testing procedures, human disease surveillance studies, and genetics.

Acknowledgments

Many thanks to my dissertation advisor, Dr. Christopher McMahan, without whom none of this work would have been possible. I would also like to thank the other members of my dissertation committee, Drs. Andrew Brown, Brook Russell, and Xiaoqian Sun for their feedback and advice throughout my dissertation. Much of this work was completed in collaboration with other co-authors. In addition to my committee members, I would like to thank the National Institute of Health, Biorealm Principal Investigator, and the National Science Foundation for funding individual pieces of this work.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern advancements in technology have allowed experimenters to gather and store larger amounts of data. While this advancement is favorable, it comes at the expense of more complicated modeling for statisticians. For example, special modeling techniques are required for the large p small n problems, or efficient modeling is necessary when either p or n are excessively large. These situations are practically ubiquitous in high dimensional analyses. High dimensional data is common in a variety of applications, including epidemiological studies, environmental sciences, and genetics. While more data can be a good thing, typical issues with high dimensional data arising from experimental designs in these areas are the number of predictors being considered (i.e., large p) or the complexity of the data. For example, group testing data arises in epidemiological studies, where the observed variables are often not on an individual level and also are error-contaminated observations due to imperfect testing. The latter issue is problematic as traditional binary data models are no longer valid, and the former issue further exacerbates the complexity of the required modeling framework. The Bayesian paradigm has proven most beneficial in these complicated models, and as such is used throughout this dissertation.

This dissertation primarily handles these complicated data structures in two ways. The first is that we utilize clever data augmentation strategies that make posterior distributions more tractable. For example, in binary regression, the addition of a carefully constructed latent variable allows the logistic or probit link function to be written into a normal distribution. The structure of these latent variables is inherently tied to the link function being used; e.g., follow Polson et al. [2013] and Albert and Chib [1993] for the logistic link and probit link, respectively. This allows model fitting to be carried out with primarily Gibbs steps, which provides for an efficient modeling framework that scales well to larger data. The second method used often in this dissertation is variable selection. In complicated models, such as in mixed effects models, this becomes a major task as the dimensions of the parameter space grow quickly due to the inclusion of random effects. Chapter 2 develops a Bayesian mixed effects model that achieves full variable selection (i.e., in both fixed effects and random effects). Motivated by Chen and Dunson [2003], the variable selection within the random effects is achieved by placing spike and slabs priors on diagonal elements of a matrix that is the result of a Cholesky decomposition of the covariance matrix of the random effects.

In high dimensions, dimension reduction either before or during model fitting enables the analysis of large amounts of data. For example, Chapter 5 develops a two-phase methodology where the first phase prescreens the predictors to reduce it to a more promising set of candidates that are then jointly analyzed in the second phase. However, in some situations, it is more beneficial to jointly analyze all predictors simultaneously due to possible correlations and interactions. Chapter 4 develops a computationally efficient expectation-maximization that is motivated by Armagan et al. [2013a] and jointly analyzes large amounts of predictors by performing dimension reduction during the model fitting process. Due to the penalty structure of the prior used on the regression coefficients, once a regression coefficient is dropped from the model (i.e., is set to zero), it cannot return.

The remainder of this dissertation is organized as follows. Chapter 2 develops a Bayesian mixed effects logistic regression model for group testing data, where individuals of a population are screened for an infectious disease. Modern group testing procedures are moving towards *multiplex* testing assays, which have the ability to test for multiple diseases simultaneously. To account for this, Chapter 3 extends this model to a multivariate setting, which incorporates possible correlations between diseases. Chapter 4 develops a Bayesian linear mixed effects model that relates single-nucleotide polymorphisms (SNPs) of rice plants to the amount of yield produced. Chapter 5 involves developing a Bayesian logistic regression model to associate human SNPs and covariate information to colorectal cancer, where the number of predictors in the model is much larger than the sample size. We conclude with Chapter 6, a brief discussion of this dissertation.

Chapter 2

From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data

2.1 Introduction

Group testing involves taking specimens (e.g., blood, urine, swabs, etc.) from different individuals and forming a pooled specimen that is then tested for disease. In most group testing protocols, if a pooled specimen tests negatively, then all individuals are declared to be disease free at the expense of a single diagnostic test. In contrast, if a pooled specimen tests positively, the pool is resolved algorithmically to determine which individuals are positive. Dorfman [1943] is credited with conceptualizing the group testing idea during World War II to screen military recruits for syphilis. Since then, group testing, or “pooling”, has become a mainstream approach to screen large populations for multiple diseases. The primary reason for pooling is to save money. For example, the State Hygienic Laboratory (SHL) at the University of Iowa has reported savings of approximately $3.1 million during a recent 5-year period after adopting a variant of Dorfman’s protocol to screen Iowa residents for chlamydia and gonorrhea; see Tebbs et al. [2013] and McMahan et al. [2017]. Pooling biospecimens through group testing arises in other applications, including testing for HIV and HCV [Sarov et al., 2007, Krajden et al., 2014], environmental testing [Heffernan et al., 2014], and drug discovery [Hughes-Oliver, 2006].

While testing pools can be far more cost effective than performing individual tests, it also leads to a more complicated data structure. This is true because specimens are pooled and hence individual-level responses may never be observed. Recent statistical research has focused on developing regression methods to model the probability of disease for individuals based on pooled outcomes; e.g., see Vansteelandt et al. [2000], Bilder and Tebbs [2009], Huang [2009], Delaigle and Meister [2011], and Delaigle et al. [2014]. All of the aforementioned regression methods are designed to analyze test results arising from assaying the initially formed (master) pools; i.e., pools formed by assigning each individual to exactly one initial pool for testing. As a consequence, these methods cannot incorporate retesting information that becomes available when positive pools are resolved [Kim et al., 2007] or when quality control steps are implemented [Gastwirth and Johnson, 1994, Johnson and Gastwirth, 2000]. To incorporate retesting information, Xie [2001] developed an expectation-maximization algorithm to estimate the individual-level probability of disease for general regression models. Wang et al. [2014] developed a semiparametric framework to estimate single-index models. Most recently, McMahan et al. [2017] proposed a Bayesian approach to estimate generalized linear models while incorporating historical information on disease prevalence and uncertainty in assay performance.

At most public health laboratories like the SHL, individual specimens arrive at the lab from different locations throughout a particular geographic region. For example, in Iowa, specimens are collected at different types of clinics (e.g., family planning clinics, STD clinics, etc.) in multiple locations from all over the state and are then shipped to the SHL for testing. Given the vast differences among clinic types and the additional differences between rural and metropolitan areas, it is natural to suspect that heterogeneity may exist from location to location. However, when individual specimens are pooled together, it becomes a significant challenge to account for this source of variability while also estimating covariate effects like age, gender, race, and sexual history. In fact, most previous regression methods for group testing data, such as those outlined above, are not able to incorporate the effects due to observing data from different locations−especially when individual specimens from different locations are pooled together.

In this article, we develop a Bayesian generalized linear mixed model approach for group testing data which uses fixed effects to describe the population-level mean structure and random effects to account for differential variability among population subgroups. Our work generalizes the random effects modeling techniques for group testing data proposed by Chen et al. [2009] and simultaneously offers a far more flexible approach for data analysis. First, by taking a Bayesian point of view, we can incorporate historical information about disease prevalence, and our approach allows assay accuracy probabilities to be estimated from the observed data. Second, a limitation of Chen et al. [2009] is that it can incorporate only master pool responses; i.e., it does not allow one to include additional retests that will be performed for disease classification purposes. On the other hand, our estimation framework is flexible and, as in McMahan et al. [2017], it can accommodate data from any group testing protocol as well as quality control screening procedures [Gastwirth and Johnson, 1994, Johnson and Gastwirth, 2000]. Third, and perhaps most limiting, the methods in Chen et al. [2009] allow only for pools to consist of individuals from within the same location. In practice, this can be markedly prohibitive because individual specimens are often pooled sequentially based on their arrival date for testing. Furthermore, for those locations performing a small number of tests, it may be impractical to wait and to pool within location. Our approach removes this limitation and includes “pooling within location” as a special case. Finally, given the complexity of the considered mixed effects model, we use spike and slab priors to perform variable selection−both within the fixed and random effects components. In particular, three of the most common spike and slab priors are considered with details of implementation under each being provided. No existing group testing regression procedure has considered such an automated variable selection technique; i.e., for both fixed and random effects. For implementation purposes, a computationally efficient Markov chain Monte Carlo (MCMC) sampling algorithm is developed which can estimate the proposed model.

Subsequent sections of this article are organized as follows. Section 2 provides preliminary information regarding the proposed mixed effects model, the modeling assumptions, and the derivation of the observed data likelihood. Section 3 presents the specifics of the approach, including prior model specifications and data augmentation steps used to construct an efficient posterior sampling algorithm. Section 4 outlines the development of the full conditional distributions. Section 5 reports the results of an extensive numerical study conducted to assess the performance of the proposed approach. Section 6 presents an analysis of chlamydia testing data collected by the SHL in Iowa. Section 7 concludes with a summary discussion. Additional technical details and additional simulation results are provided in Appendix A.

2.2 Notation and preliminaries

Consider a setting in which N individuals are screened for an infectious agent by a group testing protocol. As a part of this process, each of the N individuals visit one of K distinct clinics, where a specimen (e.g., blood, urine, saliva, etc.) is collected. Testing is then performed either at the clinic site or at a regional laboratory; e.g., the SHL in Iowa. Note the former scenario would mandate pooling of individuals within clinic sites while the latter allows for pooling across sites, with our methodology being applicable in either case. Let $\widetilde{Y}_i$ denote the true infection status of the ith individual, for $i = 1, ..., N$, with $\widetilde{Y}_i = 1$ indicating that the individual is truly positive and $\widetilde{Y}_i = 0$ otherwise. Furthermore, let $\mathbf{x}_i = (1, x_{i1}, ..., x_{i,q_1-1})'$ and $\mathbf{t}_i = (1, t_{i1}, ..., t_{i,q_2-1})'$ denote vectors of covariate values taken on the ith individual which correspond to fixed and random effects, respectively, where $\mathbf{t}_i$ is a subvector $\mathbf{x}_i$. We assume throughout that individuals' infection statuses are conditionally independent given the covariate information and the random effects. The individuals' true infection statuses (i.e., the $\widetilde{Y}_i$) are never observed due to the effect of imperfect testing, while the covariate information for each individual is observed. For ease of exposition, we aggregate the individuals' infection statuses as $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, ..., \widetilde{Y}_N)'$ and denote $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$ and $\mathbf{T} = (\mathbf{t}_1, ..., \mathbf{t}_N)'$ as the design matrices.

The goal of this work is to relate the individuals' latent infection statuses to their covariate values through the following generalized linear mixed model

$$g^{-1}\{P(\widetilde{Y}_i = 1 \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{k(i)})\} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\gamma}_{k(i)}, \tag{2.1}$$

where $g^{-1}(\cdot)$ is a known link function, $\boldsymbol{\beta}$ is a q_1-dimensional vector of fixed effects, $\boldsymbol{\gamma}_{k(i)} := \boldsymbol{\gamma}_k$ if the ith individual presented at the kth clinic, and $\boldsymbol{\gamma}_k$ is a q_2-dimensional vector of clinic-specific random effects, for $k = 1, ..., K$. It is assumed that the $\boldsymbol{\gamma}_k$ are independent and identically distributed and follow a mean zero multivariate Gaussian distribution with covariance matrix $\mathbf{D}$; i.e., $\boldsymbol{\gamma}_k \overset{iid}{\sim} N(\mathbf{0}, \mathbf{D})$. Note, to track clinic membership, herein we adopt the functional notation $k(\cdot)$ and specify that $k(i) = k$ if the ith individual presented at the kth clinic.

A typical challenge that arises in mixed modeling involves the selection of both the fixed and random effects components, which is tantamount to selecting the proper subsets of the available covariates to be retained in the final model. To accomplish this task, we adopt spike and slab priors [George and McCulloch, 1993, 1997, Kuo and Mallick, 1998]. These specifications proceed as usual for the fixed effects and follow the proposal of Chen and Dunson [2003] for the random effects, which requires a reparameterization of the proposed model. The reparameterized model is

$$g^{-1}\{P(\widetilde{Y}_i = 1 \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)})\} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)}, \tag{2.2}$$

where $\mathbf{b}_{k(i)} := \mathbf{b}_k$ if the ith individual presented at the kth clinic, $\mathbf{b}_k \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I})$, $\mathbf{I}$ is the identity matrix, $\boldsymbol{\Lambda}$ is a non-negative diagonal $q_2 \times q_2$ matrix, and $\mathbf{A}$ is a $q_2 \times q_2$ lower triangular matrix with unit main diagonal elements and free elements given by a_{ml} for $l = 1, ..., q_2 - 1; m = l + 1, ..., q_2$. For ease of exposition, we introduce $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{q_2})'$ such that $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ and $\mathbf{a}$ which denotes the vector of free elements of the matrix $\mathbf{A}$; i.e., $\mathbf{a} = (a_{ml} \colon l = 1, ..., q_2 - 1; m = l + 1, ..., q_2)'$. Note that the matrices $\boldsymbol{\Lambda}$ and $\mathbf{A}$ are obtained via a modified Cholesky decomposition and satisfy $\mathbf{D} = \boldsymbol{\Lambda}\mathbf{A}\mathbf{A}'\boldsymbol{\Lambda}$. Under this reparameterization, if λ_l (the lth diagonal element of $\boldsymbol{\Lambda}$) is zero, then so is the lth diagonal element of $\mathbf{D}$. That is, if $\lambda_l = 0$, then the variance of the lth random effect is zero, which is equivalent to dropping the lth random effect from the model. Thus, to perform variable selection for the random effects, the proposed methodology places spike and slab priors on each λ_l. Our approach also models the a_{ml} values, which allows for the estimation of $\mathbf{D}$ without imposing any prior form or structure.

The observed data that arises from implementing a group testing protocol can be quite complex. First of all, there are many protocols available for use [e.g., see Dorfman, 1943, Phatarfod and Sudbury, 1994, Kim et al., 2007, Kim and Hudgens, 2009]. Secondly, in an effort to reduce testing cost, a given protocol often requires that individuals be tested in multiple (possibly overlapping) pools and may even mandate confirmatory testing [Gastwirth and Johnson, 1994, Johnson and Gastwirth, 2000]. Thus, to provide a general framework which can incorporate and account for the complexity of data observed from implementing any group testing protocol, we define the index set $\mathcal{P}_j \subset \{1, ..., N\}$ which identifies the individuals contributing to the jth pool, for $j = 1, ..., J$. Let $\widetilde{Z}_j$ denote the true status of the jth pool, under the convention that the pool is positive ($\widetilde{Z}_j = 1$) if it contains at least one infected individual and negative otherwise ($\widetilde{Z}_j = 0$); i.e., $\widetilde{Z}_j = I\big(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0\big)$. Like the individuals' true statuses, the $\widetilde{Z}_j$'s are unobserved due to the effect of imperfect testing. Instead, we observe the diagnosed status Z_j which can be viewed as an error-contaminated version of $\widetilde{Z}_j$, with $Z_j = 1$ indicating that the jth pool tested positively and $Z_j = 0$ otherwise. To quantify the effect of imperfect testing, let $S_{ej} = P\big(Z_j = 1 \mid \widetilde{Z}_j = 1\big)$ and $S_{pj} = P\big(Z_j = 0 \mid \widetilde{Z}_j = 0\big)$ denote the sensitivity and specificity, respectively, of the assay for the jth pool. We allow S_{ej} and S_{pj} to be pool specific, thus allowing for the potential use of different types of assays and/or the potential effect that pool size (i.e., the cardinality of $\mathcal{P}_j$) may have on an assay's performance.

To relate the individual-level model in (2.2) to the observed testing outcomes $\mathbf{Z} = (Z_1, ..., Z_J)'$, it is assumed that the testing responses in $\mathbf{Z}$ are conditionally independent given $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, ..., \widetilde{Z}_J)'$ and that the conditional distribution $\mathbf{Z} \mid \widetilde{\mathbf{Z}}$ does not depend on the covariates. Under these assumptions, the conditional distribution of $\mathbf{Z}$ can be written as

$$\pi(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) = \sum_{\widetilde{\mathbf{Y}} \in \{0,1\}^N} \left[\prod_{j=1}^{J} \left\{ S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \right.$$
$$\left. \times \prod_{i=1}^{N} g(\eta_i)^{\widetilde{Y}_i} \{1 - g(\eta_i)\}^{1-\widetilde{Y}_i} \right], \tag{2.3}$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)}$ and $\mathbf{b} = (\mathbf{b}_1, ..., \mathbf{b}_K)'$. Note, on the right hand side of (2.3) we are marginalizing the joint conditional distribution of the observed testing responses and the latent statuses of the individuals, denoted by $\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})$, over $\widetilde{\mathbf{Y}}$; i.e., $\pi(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) = \sum_{\widetilde{\mathbf{Y}} \in \{0,1\}^N} \pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})$. Unfortunately, (2.3) involves a very high dimensional sum effectively rendering direct numerical evaluation infeasible. To circumvent this issue, a two-stage data augmentation procedure in Section 3.2 is proposed which leads to an efficient posterior sampling algorithm.

2.3 Data augmentation and prior specification

The full hierarchy of the proposed model is

$$\begin{array}{rcll}
\widetilde{Y}_i \mid \eta_i & \sim & \text{Bernoulli}\{g(\eta_i)\}, & \eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)} \\
\beta_q \mid v_q & \sim & (1 - v_q)\pi_{\text{spike}}(\beta_q) + v_q \pi_{\text{slab}}(\beta_q), & q = 1, ..., q_1 \\
\lambda_l \mid w_l & \sim & (1 - w_l)\pi_{\text{spike}}(\lambda_l) + w_l \pi_{\text{slab}}(\lambda_l), & l = 1, ..., q_2 \\
\mathbf{a} & \sim & N(\mathbf{m}_0, \mathbf{C}_0), & \\
\mathbf{b}_k & \sim & N(\mathbf{0}, \mathbf{I}), & k = 1, ..., K \\
v_q \mid \tau_{v_q} & \sim & \text{Bernoulli}(\tau_{v_q}), & q = 1, ..., q_1 \\
w_l \mid \tau_{w_l} & \sim & \text{Bernoulli}(\tau_{w_l}), & l = 1, ..., q_2 \\
\tau_{v_q} & \sim & \text{Beta}(a_v, b_v), & q = 1, ..., q_1 \\
\tau_{w_l} & \sim & \text{Beta}(a_w, b_w), & l = 1, ..., q_2,
\end{array}$$

where $\pi_{\text{spike}}(\cdot)$ and $\pi_{\text{slab}}(\cdot)$ denote the "spike" and "slab" components, respectively, of our spike and slab prior (for further details see Section 3.1) and $\mathbf{m}_0$, $\mathbf{C}_0$, a_v, a_w, b_v, and b_w are hyperparameters. In specifying these hyperparameters, the prior on $\mathbf{a}$ should be made to be informative (e.g., specified with $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{C}_0 = 0.5\mathbf{I}$) to avoid imposing a strong *a priori* correlation between any two random effects; for further details, see Chen and Dunson [2003]. Further, we also assume the *a priori* independence of the β_q's and the

λ_l's. Proceeding in this fashion greatly simplifies the calculations necessary for posterior sampling and is common in the literature; e.g., see George and McCulloch [1993], George and McCulloch [1997], Kuo and Mallick [1998], and Chen and Dunson [2003].

2.3.1 Spike and slab prior

The model hierarchy presented thus far provides a general representation of the spike and slab prior. To ground the description of our approach and to illustrate our methodology, we discuss three commonly used spike and slab priors: the stochastic search variable selection (SSVS), the normal mixture inverse gamma (NMIG), and the Dirac spike, see George and McCulloch [1993], George and McCulloch [1997], and Kuo and Mallick [1998], respectively.

Following the work of George and McCulloch [1993], the SSVS approach used herein makes use of spike and slab priors of the following form:

$$\beta_q \mid v_q \sim N(0, r(v_q)\phi_q^2) \tag{2.4}$$

$$\lambda_l \mid w_l \sim TN\big(0, r(w_l)\psi_l^2, (0, \infty)\big), \tag{2.5}$$

where $r(\cdot)$ is a function serving as a binary switch (i.e., $r(0) = r$ and $r(1) = 1$) that transitions the prior between the spike and the slab, ϕ_q^2 and ψ_l^2 are specified variance components, and $TN(\mu, \psi^2, (a, b))$ denotes the usual truncated normal distribution which arises from restricting the support of a $N(\mu, \psi^2)$ distribution to the interval (a, b). In the specification of (2.4) and (2.5), one should provide large values of ϕ_q^2 and ψ_l^2 and a small value for r. In particular, these specifications should be made such that r^{-1} is sufficiently larger than the variance components; i.e., $r^{-1} >> \phi_q^2$ and $r^{-1} >> \psi_l^2$; for further discussion, see Wagner and Duller [2012]. Proceeding in this fashion leads to a flat slab and a spike that is concentrated around zero. It is important to note that specifying appropriate values of the variance components can, in some instances, be challenging and moreover has the potential to greatly influence the analysis.

To avoid specifying the variance components, one could instead use the NMIG prior specification outlined in George and McCulloch [1997] and Ishwaran and Rao [2003]. This approach proceeds identically to that of SSVS with the exception that the variance components are viewed as unknown quantities and an inverse gamma prior is specified for them. That is, $\phi_q^2 \sim$ Inv-Gamma(a_ϕ, b_ϕ) and $\psi_l^2 \sim$ Inv-Gamma(a_ψ, b_ψ). This addition to the hierarchy removes the need to specify these nuisance parameters and allows one to estimate them through data driven means. One still must specify the value of r; i.e., the proportional difference between the variance components of the spike and slab densities. Experience suggests that the selection of r tends to impact the spike distribution far more than the slab, with the model selection process being too liberal when r is chosen too large and vice versa.

To avoid specifying r, a Dirac delta function could be used for the spike; see Kuo and Mallick [1998] and Wagner and Duller [2012]. This can be viewed as a limiting case of SSVS where the variance of the continuous spike distribution is driven to zero; i.e., $r \to 0$. In this situation, $v_q = 0$ if and only if $\beta_q = 0$ and similarly for w_l and λ_l. Although this seems favorable, it also introduces an absorbing state in the Markov chain. To handle this issue, rather than sampling the binary variables from their full conditional distributions, $\boldsymbol{\beta}$ is integrated out when updating $\boldsymbol{v} = (v_1, ..., v_{q_1})'$ and $\boldsymbol{\lambda}$ is integrated out when updating $\boldsymbol{w} = (w_1, ..., w_{q_2})'$. Thus, to develop a computationally efficient posterior sampling algorithm, one must be able to analytically marginalize the posterior distribution over both $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. The ability to do so is inherently tied to the link function being used. Fortunately, this can be accomplished under both the probit and logistic link functions after a series of data augmentation steps; this process is outlined in Section 3.2. The distribution for the slab can take on any diffuse continuous distribution. To closely mimic the slab priors in SSVS and NMIG, we take *a priori* the β_q's to be independent with slab component $N(0, \phi_q^2)$ and the λ_l's to be independent with slab component $TN\big(0, \psi_l^2, (0, \infty)\big)$, where ϕ_q^2 and ψ_l^2 are again specified to be large.

2.3.2 Data augmentation

To facilitate the development of an efficient posterior sampling algorithm, a two-stage data augmentation procedure is proposed which focuses on implementation under both the probit and logistic link functions. In the first stage, we introduce the individuals' true statuses $\widetilde{\mathbf{Y}}$ as latent random variables and consider the joint conditional distribution of the observed testing responses and the latent statuses of the individuals, which is

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) = \prod_{j=1}^{J} \left\{ S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j} \right\}^{1-\widetilde{Z}_j}$$

$$\times \prod_{i=1}^{N} g(\eta_i)^{\widetilde{Y}_i} \big\{1 - g(\eta_i)\big\}^{1-\widetilde{Y}_i}.$$

In the second stage, a carefully constructed latent random variable, ω_i, is introduced for each of the individuals. Under the probit and logistic link functions, these random variables obey specifically structured normal and Pólya-Gamma distributions, respectively; for further details, see Albert and Chib [1993] and Polson et al.

[2013]. In either case, this stage yields the following joint conditional distribution

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) \propto \prod_{j=1}^{J} \left\{ S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \times \exp\left\{ -\frac{1}{2} (\mathbf{h} - \boldsymbol{\eta})' \boldsymbol{\Omega} (\mathbf{h} - \boldsymbol{\eta}) \right\} \prod_{i=1}^{N} \xi(\omega_i), \quad (2.6)$$

where $\boldsymbol{\omega} = (\omega_1, ..., \omega_N)'$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_N)'$. Under the probit link, $\mathbf{h} = (\omega_1, ..., \omega_N)'$, $\boldsymbol{\Omega} = \mathbf{I}$, and $\xi(\omega_i) = I(\omega_i \geq 0, \widetilde{Y}_i = 1) + I(\omega_i < 0, \widetilde{Y}_i = 0)$. Under the probit link, $\xi(\omega_i)$ acts to control the support of ω_i such that given $\widetilde{Y}_i = 0$ or 1 results in ω_i being constrained to $(-\infty, 0)$ or $(0, \infty)$, respectively. Under the logistic link, $\mathbf{h} = (\kappa_1/\omega_1, ..., \kappa_N/\omega_N)', \kappa_i = \widetilde{Y}_i - 1/2$, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$, and $\xi(\omega_i) = f(\omega_i \mid 1, 0) \exp\{\kappa_i^2/(2\omega_i)\}$, where $f(\omega_i \mid a, b)$ denotes the Pólya-Gamma density with parameters (a, b); see Polson et al. [2013].

2.4 Posterior computation and inference

To facilitate estimation and inference, a posterior sampling algorithm consisting solely of Gibbs steps is constructed. In what follows, the necessary full conditional distributions used in this algorithm are provided. A symbolic representation of the entire posterior sampling algorithm is provided in Appendix A.1.

Attention is first turned to the latent random variables introduced through the data augmentation procedure. The full conditional distribution of the individuals' latent statuses is given by $\widetilde{Y}_i \mid \widetilde{\mathbf{Y}}_{-i}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)} \sim$ Bernoulli$\{p_{i1}^{\star}/(p_{i0}^{\star} + p_{i1}^{\star})\}$, where

$$p_{i1}^{\star} = g(\eta_i) \prod_{j \in \mathcal{I}_i} S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j}$$
$$p_{i0}^{\star} = \{1 - g(\eta_i)\} \prod_{j \in \mathcal{I}_i} \left\{ S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j} \right\}^{I(s_{ij} > 0)} \left\{ (1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j} \right\}^{I(s_{ij} = 0)},$$

$s_{ij} = \sum_{i' \in \mathcal{P}_j : i' \neq i} \widetilde{Y}_{i'}$, and the index set $\mathcal{I}_i = \{j : i \in \mathcal{P}_j\}$ keeps track of the indices of the pools to which the ith individual contributed. We also adopt the convention that $\mathbf{V}_{-i}$ represents the vector $\mathbf{V}$ after removing the ith component. The full conditional distribution of ω_i is link function dependent and is given by

$$\omega_i \mid \widetilde{Y}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)} \sim \begin{cases} TN\{\eta_i, 1, (0, \infty)\}, & \text{if } \widetilde{Y}_i = 1, \\ TN\{\eta_i, 1, (-\infty, 0)\}, & \text{if } \widetilde{Y}_i = 0, \end{cases}$$

or

$$\omega_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)} \sim \text{PG}(1, \eta_i),$$

under the probit and logistic link, respectively, where PG(a, b) denotes the Pólya-Gamma distribution with parameters (a, b); see Polson et al. [2013].

We now describe how to sample the fixed and random effects. Focusing on the quadratic form in the exponential in (2.6), we have that

$$\begin{aligned}(\mathbf{h} - \boldsymbol{\eta})'\boldsymbol{\Omega}(\mathbf{h} - \boldsymbol{\eta}) &= \sum_{i=1}^{N}(h_i - \eta_i)^2\boldsymbol{\Omega}_{ii} \\ &= \sum_{i=1}^{N}(h_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)})^2\boldsymbol{\Omega}_{ii} \\ &= \sum_{i=1}^{N}(h_{\boldsymbol{\beta}i} - \mathbf{x}_i'\boldsymbol{\beta})^2\boldsymbol{\Omega}_{ii} = (\mathbf{h}_{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{h}_{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}),\end{aligned}$$

where $\boldsymbol{\Omega}_{ii}$ is the ith diagonal element of $\boldsymbol{\Omega}$, $h_{\boldsymbol{\beta}i} = h_i - \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)}$, and $\mathbf{h}_{\boldsymbol{\beta}} = (h_{\boldsymbol{\beta}1}, ..., h_{\boldsymbol{\beta}N})'$. Thus, it is easy to see that under the SSVS and NMIG spike and slab priors, the full conditional distribution of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v} \sim N\left\{\left(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X} + \boldsymbol{\Phi}^{-1}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{h}_{\boldsymbol{\beta}}, \left(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X} + \boldsymbol{\Phi}^{-1}\right)^{-1}\right\},$$

where $\boldsymbol{\Phi} = \text{diag}\big(r(v_1)\phi_1^2, ..., r(v_{q_1})\phi_{q_1}^2\big)$. Under the Dirac spike, the full conditional distribution of β_q is degenerate at 0 if $v_q = 0$, while the non-zero elements of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{\boldsymbol{v}}$, have the following normal full conditional

$$\boldsymbol{\beta}_{\boldsymbol{v}} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v} \sim N\left\{\left(\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}\right)^{-1}\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}\mathbf{h}_{\boldsymbol{\beta}}, \left(\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}\right)^{-1}\right\},$$

where $\mathbf{X}_{\boldsymbol{v}}$ is the design matrix consisting of those columns of $\mathbf{X}$ corresponding to non-zero elements of $\boldsymbol{v}$ and $\boldsymbol{\Phi}_{\boldsymbol{v}}$ is the diagonal matrix formed by retaining the diagonal elements of $\boldsymbol{\Phi} = \text{diag}(\phi_1^2, ..., \phi_{q_1}^2)$ corresponding to the non-zero elements of $\boldsymbol{v}$. Due to the data augmentation steps described above, one can also obtain the following full conditionals

$$\begin{gathered}\lambda_l \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l \sim TN\{\mu_{\lambda_l}(w_l), \sigma^2_{\lambda_l}(w_l), (0,\infty)\} \\ \mathbf{a} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}) \\ \mathbf{b}_k \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}),\end{gathered}$$

where the specific forms of these distributions are provided in Appendix A.1. Sampling these parameters is equivalent to sampling the random effects as well as the covariance matrix of the distribution of the random effects.

For completion, the full conditional distributions of v_q and w_l are Bernoulli, with the success probability depending on the specified spike and slab prior; see Appendix A.1. The full conditional distribution for the mixing weights τ_{v_q} and τ_{w_l} are conveniently $\tau_{v_q} \mid v_q \sim \text{Beta}(a_v + v_q, 1 - v_q + b_v)$ and $\tau_{w_l} \mid w_l \sim \text{Beta}(a_w + w_l, 1 - w_l + b_w)$, respectively. Finally, under the NMIG prior, the full conditionals of the variance parameters are $\phi^2_q \mid \beta_q, v_q \sim \text{Inv-Gamma}\big(a_\phi + 1/2, b_\phi + \beta^2_q/\{2r(v_q)\}\big)$ and $\psi^2_l \mid \lambda_l, w_l \sim \text{Inv-Gamma}\big(a_\psi + 1/2, b_\psi + \lambda^2_l/\{2r(w_l)\}\big)$.

Up until this point, the assay accuracies (i.e., S_{ej} and S_{pj}) have been assumed to be known. When these quantities are unknown, we may estimate them along with the rest of the model parameters following the approach outlined in McMahan et al. [2017]. Briefly, this approach allows for different assays to be used throughout the testing process (e.g., screening and confirmatory testing) and/or can account for the effect of pool size on the accuracy of the assay; i.e., sensitivity and specificity might change with the pool size. Define the index set $\mathcal{M}_m$ which identifies the indices of the pools which were tested by the mth assay, for $m = 1, ..., M$. Further, let $S_{e(m)}$ and $S_{p(m)}$ denote the sensitivity and specificity of the mth assay such that $S_{ej} = S_{e(m)}$ and $S_{pj} = S_{p(m)}$ for all $j \in \mathcal{M}_m$. Under these conventions, (2.6) can be written as

$$\begin{aligned}\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{S}_e, \mathbf{S}_p) \propto \prod_{m=1}^{M} \prod_{j\in\mathcal{M}_m} & \left\{S^{Z_j}_{e(m)}(1 - S_{e(m)})^{1-Z_j}\right\}^{\widetilde{Z}_j} \left\{(1 - S_{p(m)})^{Z_j} S^{1-Z_j}_{p(m)}\right\}^{1-\widetilde{Z}_j} \\ & \times \exp\left\{-\frac{1}{2}(\mathbf{h} - \boldsymbol{\eta})'\boldsymbol{\Omega}(\mathbf{h} - \boldsymbol{\eta})\right\} \prod_{i=1}^{N} \pi(\omega_i),\end{aligned}$$

where $\mathbf{S}_e = (S_{e(1)}, ..., S_{e(M)})'$ and $\mathbf{S}_p = (S_{p(1)}, ..., S_{p(M)})'$. Given the form of the conditional distribution above, independent beta priors are a natural choice; i.e., $S_{e(m)} \sim \text{Beta}(a_{e(m)}, b_{e(m)})$ and $S_{p(m)} \sim$

Beta$(a_{p(m)}, b_{p(m)})$. These specifications lead to the following full conditionals

$$S_{e(m)} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{Beta}\big(a^{\star}_{e(m)}, b^{\star}_{e(m)}\big)$$
$$S_{p(m)} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{Beta}\big(a^{\star}_{p(m)}, b^{\star}_{p(m)}\big),$$

where $a^{\star}_{e(m)} = a_{e(m)} + \sum_{j \in \mathcal{M}_m} Z_j \widetilde{Z}_j$, $b^{\star}_{e(m)} = b_{e(m)} + \sum_{j \in \mathcal{M}_m} (1 - Z_j)\widetilde{Z}_j$, $a^{\star}_{p(m)} = a_{p(m)} + \sum_{j \in \mathcal{M}_m} (1 - Z_j)(1 - \widetilde{Z}_j)$, and $b^{\star}_{p(m)} = b_{p(m)} + \sum_{j \in \mathcal{M}_m} Z_j(1 - \widetilde{Z}_j)$. The other posterior distributions are left unchanged up to acknowledging dependence on the testing accuracies and accounting for the slight change in notation.

2.5 Simulation

To investigate the performance of our regression and variable selection methods, we designed a simulation study which emulates the primary features of our Iowa data application in Section 6. To this end, $K = 50$ clinic sites were conceptualized and the infection statuses for 100 individuals within each of these sites were generated; i.e., $N = 5000$. This sample size is roughly a third of the sample size available in our data application. The individuals' true statuses were generated according to the following model

$$g^{-1}\{P(\widetilde{Y}_i = 1 \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)})\} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)}, \text{ for } i = 1, ..., N,$$

where $g^{-1}(\cdot)$ denotes the probit link, $\boldsymbol{\beta} = (-3, -1.5, 0.5, 0.25, 0, 0)'$, $\boldsymbol{\lambda} = (1, 0.75, 0.25, 0, 0, 0)'$, $\mathbf{a} = (1, 0.5, 0.7, 0, ..., 0)'$, $\mathbf{b}_{k(i)} = \mathbf{b}_k$ if individual i presented at clinic site k, and $\mathbf{b}_k \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I})$. The covariate vectors $\mathbf{x}_i$ and $\mathbf{t}_i$ are taken to be equal and are standardized versions of $\mathbf{x}^*_i = (1, x^*_{i1}, x^*_{i2}, x^*_{i3}, x^*_{i4}, x^*_{i5})'$, where $x^*_{i1}, x^*_{i5} \sim N(0, 1)$ and $x^*_{i2}, x^*_{i3}, x^*_{i4} \sim$ Bernoulli(0.5). Under these specifications, the generating model consists of four non-zero fixed effects (one intercept and three slopes) as well as three non-zero random effects (one intercept and two slopes). The parameter configurations above provide for an overall prevalence of approximately 9%, which is in keeping with the motivating data application. This model was used to generate 1000 independent data sets.

To generate testing outcomes, we implement three group testing protocols; namely, master pool testing (MPT), Dorfman testing (DT), and array testing (AT). Briefly, under MPT, each individual is assigned to exactly one master pool which is tested and no further testing is performed regardless of the outcome. DT completes the decoding process initiated by MPT by retesting all individuals in positive master pools.

Similarly, AT completes decoding in two-stages, but unlike DT it starts by assigning individuals to an array. In the first stage, AT tests pools formed by combining individuals who share a common row or column. The second stage retests individuals identified to be likely positives; e.g., individuals residing at the intersection of positive rows and columns. For the specific retesting protocol adopted for AT, see Kim et al. [2007]. Following the pooling practices used in the motivating example, we consider implementing MPT and DT using master pools of size 4 and AT using 4×4 arrays. For comparative purposes, individual testing (IT) was also implemented.

For each of the 1000 individual-level data sets, we simulate IT, MPT, DT and AT. To implement the group testing protocols, individuals were randomly assigned to pools (arrays), so that individuals would be pooled across sites rather than within sites. Proceeding in this fashion poses the most difficult estimation configuration; that is, individuals within the same pool have different random effects. Moreover, this mirrors large-scale surveillance studies such as the Iowa chlamydia application in Section 6. Under all testing protocols, the testing response for the jth pool was simulated as $Z_j \mid \widetilde{Z}_j \sim$ Bernoulli$\{S_{ej}\widetilde{Z}_j + (1 - S_{pj})(1 - \widetilde{Z}_j)\}$, where $\widetilde{Z}_j = I\left(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0\right)$. Two different simulation settings are considered regarding the testing accuracies. In the first setting, sensitivity and specificity are assumed to be known and are set to be $S_{ej} = 0.95$ and $S_{pj} = 0.98$ for all $j = 1, ..., J$. The second setting considers two assays, where the first ($m = 1$) is used to test pools and the second ($m = 2$) is used for individual-level testing with $\mathbf{S}_e = (S_{e(1)}, S_{e(2)})' = (0.95, 0.98)'$ and $\mathbf{S}_p = (S_{p(1)}, S_{p(2)})' = (0.98, 0.99)'$. Under this setting, we assume that these accuracies are unknown and have to be estimated along with the other model parameters. In the second setting we only consider DT and AT since both protocols mandate both pool and individual-level testing.

We assess the performance under all three spike and slab priors described in Section 3; we set $\mathbf{m}_0 = \mathbf{0}, \mathbf{C}_0 = 0.5\mathbf{I}$, and used flat priors for all mixing weights and all testing assay accuracies; i.e., Beta$(1, 1)$. As mentioned previously, we specify a slightly informative prior on $\mathbf{a}$ to avoid a strongly informative prior distribution on the prior correlation between any two random effects [Chen and Dunson, 2003]. We chose $r = 0.00025$ for both SSVS and NMIG, and $a_\phi = a_\psi = 5$ and $b_\phi = b_\psi = 50$ when using NMIG, closely resembling the values chosen in Scheipl [2011]. To provide a fair comparison, the prior mean for the variance component under NMIG was used as the variance component in SSVS and the Dirac spike; i.e., $\phi_q^2 = \psi_l^2 = 50/4$. To perform posterior estimation and inference, our MCMC algorithm was used to draw 100000 iterates, with every 50th being retained after a burn-in of 50000; i.e., we draw a posterior sample consisting of 1000 iterates. Point estimates of the model parameters were obtained as the empirical means of the posterior distributions. To assess the performance of the variable selection techniques, estimates of the posterior inclusion probabilities were also computed, where the posterior inclusion probability refers to the probability that $v_q(w_l) = 1$. These estimates were taken to be the sample mean of the posterior draws of v_q and w_l. To assess out of sample classification accuracy, we conducted a receiver operating characteristic curve (ROC) analysis. In particular, for each model fit, we simulated 10000 new individuals (i.e., statuses and covariates) and used our model fits to predict their infection probabilities. This gives 1000 ROC curve estimates which are summarized as the average area under the curve (AUC). For purposes of comparison, we also fit the competing model discussed in McMahan et al. [2017]. This comparison is aimed at demonstrating the gains in classification accuracy that are possible via including cite specific random effects and using variable selection to guide model selection.

Table 6.1 summarizes the results under the Dirac spike when $S_{ej} = 0.95$ and $S_{pj} = 0.98$ are known. The summary includes the empirical bias (Bias), sample standard deviation of the point estimates (SSD), and the average estimated probability of inclusion (PI). Tables 6.12 and 6.13 provide the analogous results under SSVS and NMIG, respectively. These results illustrate that our approach reliably estimates the fixed and random effects; i.e., the empirical bias and the variability of the estimators are small relative to the true value of the corresponding parameter. These results also indicate that the proposed methodology is adept at identifying non-zero fixed and random effects. That is, covariates with strong (no) effects almost always have posterior inclusion probabilities being near 1 (0) in all data sets. Table 6.14 summarizes the results of our ROC analysis. These results show that the average AUC for our model was markedly higher than the competing procedure, across all configurations. This indicates that our approach provides better classification than this existing technique. Further, when comparing Table 6.1 to Tables 6.12 and 6.13, one will note that the Dirac spike tends to outperform both SSVS and NMIG in terms of variable selection.

Among all testing protocols, MPT tends to perform the worst; i.e., the estimates obtained from analyzing MPT data exhibit more bias and variability. This is expected because MPT does not complete classification like the other considered testing protocols; i.e., data collected via MPT consists of less information about the individuals' latent statuses when compared to DT, AT, and IT. In contrast, the estimation performance under the two classification protocols (DT and AT) is as good if not better than the performance under IT. Keep in mind, these estimates are obtained at nearly half the testing cost on average. Specifically, to complete IT, 5000 tests are used, while DT and AT require on average 2747 and 3258 tests, respectively. These results illustrate the "get more for less" phenomenon that has previously been reported with group testing regression [Zhang et al., 2013, McMahan et al., 2017].

Table 6.2 summarizes the setting when the assay accuracy probabilities are unknown under the

Dirac spike. Under this configuration, the proposed methodology is tasked with estimating four additional parameters through analyzing DT and AT data. Tables 6.15 and 6.16 provide the analogous results under SSVS and NMIG, respectively. These results indicate that the proposed approach can accurately estimate the unknown assay accuracies; i.e., these estimates exhibit little (if any) average bias and the variability in the estimates is small relative to the true value of the testing accuracies. Moreover, there are no appreciable differences between the estimates displayed in Tables 6.1 and 6.2 for DT and AT; i.e., the estimation of the fixed and random effects are not unduly impacted by the additional task of estimating the assay accuracies. It is important to note that flat priors are specified for the testing accuracies in this application to provide for the most challenging case; i.e., we have no prior information about the testing accuracies. In other settings it might be desirable to set informative priors; e.g., if one believes that sensitivity or specificity are around 0.95 an informative prior could be specified as Beta$(19c, c)$, where large(small) values of c would reflect strong(weak) prior belief. Informative priors could also be designed based on validation trials as we demonstrate in Section 6. In either case, it is reasonable to assume that the proposed methodology would perform as well if not better when informative priors are specified for the testing accuracies.

In addition to the studies described above, we have also performed a complementary study aimed at assessing the robustness of our approach to severe violations of the conditional independence assumption; i.e., the assumption that the testing responses in $\mathbf{Z}$ are conditionally independent given $\widetilde{\mathbf{Z}}$. Appendix A.2 provides the specific details on how this study was conducted along with a summary discussion of the results. Briefly, this study reveals that the performance of our proposed regression method is not degraded even under severe violations of this assumption.

[Table 1 about here.]

[Table 2 about here.]

2.6 Chlamydia testing application

As Iowa's public health and environmental laboratory, the SHL serves all of the state's counties through infectious disease detection and surveillance. This includes annually screening thousands of residents for the two most common sexually transmitted diseases (STDs): chlamydia and gonorrhea. This process begins with specimens (e.g., urine, swab, etc.) being collected from residents at different clinics (e.g., STD screening clinics, family planning, etc.) throughout the state. These specimens are then transported to the SHL for testing. Current SHL screening protocols mandate that all male specimens and female urine specimens be tested individually while a variant of Dorfman testing is used to classify female swab specimens; for further discussion, see Tebbs et al. [2013]. The SHL uses the Aptima Combo 2 Assay to test both pooled and individual specimens.

Our analysis focuses on the chlamydia data collected on female patients during the 2014 calendar year. During this time period, 64 different clinics submitted specimens to the SHL for testing. The available data consist of results collected on 4316 individual urine specimens, 416 individual swab specimens, and 2286 swab master pools (1 of size 2, 12 of size 3, and 2273 of size 4), as well as the test results required to resolve the positive master pools. In addition to the test data, several covariates were collected on each individual: age (in years, denoted by x_1^*), a race indicator ($x_2^* = 1$ if Caucasian and $x_2^* = 0$ otherwise), an indicator denoting whether the patient reported a new sexual partner in the last 90 days ($x_3^* = 1$ if affirmative and $x_3^* = 0$ otherwise), an indicator denoting whether the patient reported having multiple sexual partners in the last 90 days ($x_4^* = 1$ if affirmative and $x_4^* = 0$ otherwise), an indicator denoting whether the patient reported sexual contact with an STD-positive partner in the previous year ($x_5^* = 1$ if affirmative and $x_5^* = 0$ otherwise), and an indicator denoting whether the patient presented with symptoms ($x_6^* = 1$ if affirmative and $x_6^* = 0$ otherwise). To relate the individuals' disease statuses to the available covariate information, we assume that $g^{-1}\{P(\widetilde{Y}_i = 1 \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{k(i)})\} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)}$, where $g^{-1}(\cdot)$ denotes the probit link. The covariate vectors $\mathbf{x}_i$ and $\mathbf{t}_i$ are taken to be equal and are standardized versions of $\mathbf{x}_i^* = (1, x_{i1}^*, x_{i2}^*, x_{i3}^*, x_{i4}^*, x_{i5}^*, x_{i6}^*)'$. This was done so that the spike and slab distributions will have the same impact on the regression coefficients across all covariates. In this analysis, a random effect vector $\mathbf{b}_k$ is specified for each of the 64 clinics, with the convention that $\mathbf{b}_{k(i)} = \mathbf{b}_k$ if the ith individual presented at the kth clinic site.

Given the results of the numerical studies presented in Section 5, in this analysis we chose to implement the proposed approach under the Dirac spike only. All other prior specifications were made in the exact same fashion as was described in Section 5. The only difference being that three sets of testing accuracies were conceptualized to account for the SHL's screening protocol: $S_{e(1)}$ and $S_{p(1)}$ for swab specimens tested individually, $S_{e(2)}$ and $S_{p(2)}$ for urine specimens tested individually, and $S_{e(3)}$ and $S_{p(3)}$ for swab specimens tested in pools. Flat priors were again specified for these parameters; i.e., $S_{e(m)}, S_{p(m)} \sim \text{Beta}(1, 1)$, for $m = 1, 2, 3$.

Table 6.3 provides the estimated posterior mean and standard deviation for all model parameters, along with estimated posterior probabilities of inclusion. To aid interpretability, we have unstandardized these estimates. Further, Table 6.18 provides the six models with the highest posterior probabilities; see

Kuo and Mallick [1998]. The direction (sign) of the point estimates of the fixed effects are congruous with the findings from other similar epidemiological studies involving chlamydia infection. That is, the risk of contracting chlamydia tends to decrease with age, and Caucasians appear to be associated with a lower risk when compared to other races. In contrast, having a new sexual partner, multiple partners, and contact with an STD are all associated with an increased risk. Lastly, our modeling framework identified the random intercept parameter to be strongly significant; i.e., there exists strong evidence of heterogeneity across the various clinics throughout the state. It is worthwhile to point out that these random intercepts act as crude proxies for clinic level unmeasured confounders such as an areas socioeconomic status, rural verses urban areas, etc. By examining these estimated random effects alongside other predictor variables (e.g., census data) one could reveal new covariates that are related to chlamydia prevalence.

Shifting attention to the estimates of the testing assay accuracies, one may notice the lower estimates of $S_{e(2)}$ and $S_{e(3)}$, suggesting potential underestimation of these parameters. To examine this further, we performed the analysis again using informative priors which were set based on the product literature and validation trials available on the Aptima Combo 2 Assay; see McMahan et al. [2017]. A summary of the results is provided in Table 6.17. From these results, one will note that there are no appreciable differences in the regression parameter estimates. Therefore, even if the testing accuracies are slightly underestimated, this does not appear to unduly affect the estimation of the regression parameters, which are likely of primary interest. When comparing Table 6.3 to Table 6.17, we find evidence that all testing accuracies are identifiable under the Iowa testing protocol, with $S_{e(2)}$ and $S_{e(3)}$ being weaker "learners" than $S_{e(1)}$. This feature is likely attributable to the specified testing protocol; e.g., testing all female urine specimens individually provides very little confirmatory and/or counter-factual information that can be used to estimate $S_{e(2)}$.

Lastly, based on our posterior sampling strategy we are able to estimate a subject specific posterior infection probability for each individual given all of the observed data. This is accomplished by averaging over the sampled latent statuses for the ith individual; i.e., we compute the subject specific posterior infection probability as $G^{-1}\sum_{g=1}^{G}\widetilde{Y}_i^{(g)}$, where $\widetilde{Y}_i^{(g)}$ is the gth posterior draw of $\widetilde{Y}_i$, for $g = 1, ..., G$. Figure 6.6 displays these posterior probabilities for the individuals in this study, stratified by diagnosed status. These probabilities can be used as a measure of diagnostic certainty; e.g., probabilities near 1(0) indicate that the individual is (not) infected. Further, we believe that these probabilities could also be used to guide back-end confirmatory screening via informative group testing procedures; e.g., see McMahan et al. [2012] and Bilder et al. [2019].

[Table 3 about here.]

2.7 Discussion

This work has developed a Bayesian generalized linear mixed model that can be used to analyze data arising from any group testing protocol. To further disseminate our work, R programs which implement the proposed approach have been developed and are available at www.chrisbilder.com/grouptesting and a description of the main functions can be found in Appendix A.3.

Given the performance of the proposed approach, several modeling extensions could be of interest. For example, many large-scale screening laboratories are now adopting *multiplex* assays, which test specimens for multiple diseases simultaneously. That is, a multiplex assay generates a multivariate outcome consisting of correlated binary data. Extending the proposed modeling framework to account for this type of data would be of interest. Further, the proposed methodology could also be generalized to an additive model framework with variable selection being applied to nonlinear functionals of the continuous predictors.

Chapter 3

Mixed effects Bayesian regression for multivariate group testing data

3.1 Introduction

Group testing has received a considerable amount of attention in recent years; rightfully so as it has the potential to save practitioners a considerable amount of money in testing costs. In its essence, a group testing protocol reduces the total number of tests required to screen a population for an infectious disease. This phenomenon was first well established by Robert Dorfman in 1943 when he proposed his Dorfman group testing algorithm to test United States soldiers for syphilis during World War II. Individuals were placed into groups where their specimens (e.g., blood, urine, swabs, etc) were physically combined into a pool. That pool is then tested for an infectious agent at the expense of a single diagnostic assay. Group testing protocols proceed in the following manner: If the pooled specimen tests negatively, then all individuals are declared disease free, and contrastly, if a pool tests positively, contributing members are then retested algorithmicially to determine which individuals are positive. Today, the State Hygienic Laboratory (SHL) in Iowa currently employs a group testing protocol and has reported a $3.1 million savings over a course of 5 years after adopting the Dorfman group testing algorithm as their primary testing protocol. Pooling biospecimens through group testing arises in many applications, including testing for HIV, HBV, and HCV [Kleinman et al., 2005, Sarov et al., 2007, Krajden et al., 2014], testing animal and insect populations [Dhand et al., 2010, Speybroeck et al., 2012], environmental testing [Heffernan et al., 2014], and drug discovery

[Hughes-Oliver, 2006].

The group testing literature has grown vastly in recent years due to the dramatic cost effectiveness that group testing offers over individual level testing. To name a few, some of the more recent notable works include Vansteelandt et al. [2000], Bilder and Tebbs [2009], Delaigle and Meister [2011], and McMahan et al. [2017]. However, these methods analyze data from a group testing algorithm through a univariate analysis; i.e., to model data from a group testing protocol that used multiplex testing assays, they would have to perform multiple univariate analyses. However, this practice may miss important features in the data, such as correlation between diseases. That is, an individual who has disease one may be more likely to have disease two. However, to date, there is a lack of existing methodology in the group testing literature to perform a multivariate analysis of group testing data where a multiplex testing assay was used.

To further the model complexity, most public health laboratories like the SHL receive individual specimens from different clinics. Given the different types of clinics (e.g., family planning clinics, STD testing clinics, etc.) and different locations of clinics (e.g., urban area versus rural), it is reasonable to believe that heterogeneity may exist from clinic to clinic. However, like the SHL, most laboratories pool individual specimens as they arrive. That is, contributing members may share different clinic effects. This provides for a challenging modeling framework.

In this paper, we develop a general multivariate Bayesian generalized linear mixed effects model to analyze data arising from any group testing protocol. That is, individuals may be pooled together and restested in any fashion, and may be tested for more than one disease simultaneously. The novelty of this methodology is the multivariate analysis of such data, and furthermore, we achieve full model selection via spike and slab priors.

The remainder of this paper is organized as follows. Section 2 introduces the notation seen throughout this manuscript and lays out the developed methodology. Section 3 assesses the models capability of estimating the unknown parameters for a given group testing data set. Section 4 analyzes the motivating group testing data set, provided by the State Hygienic Laboratory (SHL) in Iowa. Finally, Section 5 concludes the manuscript with a brief discussion.

3.2 Methodology

3.2.1 Notation and preliminaries

Suppose that there are N individuals who are tested for the presence of any of D diseases through a group testing procedure that used a multiplex testing assay. These individuals each visit one of K distinct clinics, where an individual's specimen (e.g., blood, urine, swabs, etc) is extracted for evaluation. These specimens are either tested in house at the clinic site or sent to a central hub for testing; e.g., the State Hygienic Laboratory in Iowa. In either case, the testing is done by forming J total groups (pools), where the jth pool involves c_j individuals; i.e., $N = \sum_{j=1}^{J} c_j$. For the methodology to incorporate data from any group testing algorithm, define $\mathcal{P}_j$ to be the set of individuals involved in the jth pool. Regardless of the group testing algorithm, keeping track of pool membership will suffice for model fitting. Let the true status of individual $i = 1, 2, ..., N$ be $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, ..., \widetilde{Y}_{iD})'$, a D dimensional binary vector, and the true status of pool $j = 1, 2, ..., J$ be $\widetilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, ..., \widetilde{Z}_{jD})'$, also a D dimensional binary vector. Here, $\widetilde{Y}_{id} = 1$ ($\widetilde{Y}_{id} = 0$) if individual i is truly positive (negative) for the dth disease. Group testing protocols mandate that a pool is positive if at least one member is truly positive; i.e., $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$. Unfortunately, these binary vectors are never truly known due to testing assays of any kind being subject to error; i.e., false positives or false negatives. To this end, denote the observed diagnoses as $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iD})'$ and $\mathbf{Z}_j = (Z_{j1}, ..., Z_{jD})'$. Moreover, to acknowledge these imperfect testing assays, each pool receives a sensitivity $\mathbf{S}_{e_j} = (S_{e_j:1}, ..., S_{e_j:D})'$ and specificity $\mathbf{S}_{p_j} = (S_{p_j:1}, ..., S_{p_j:D})'$, where $S_{e_j:d} = P(Z_{jd} = 1 \mid \widetilde{Z}_{jd} = 1)$ and $S_{p_j:d} = P(Z_{jd} = 0 \mid \widetilde{Z}_{jd} = 0)$. The testing assay accuracies may be known constants and provided to the model, or the model can simultaneously estimate them during model fitting; more details in Section 3.2.4.

To address the possible heterogenity that may exist across the clinic sites, a mixed effects model is used to relate an individual's covariate information to their infectious probability. For each of the K clinics, the kth site gets assigned a random effect vector $\boldsymbol{\gamma}_{kd}$, which is assumed to be a multivariate normal random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}_d$, i.e., $\boldsymbol{\gamma}_{kd} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_d)$. Define for individual i and disease d, $\mathbf{x}_{id}$ as the p_d dimensional covariate vector associated with the fixed effects and $\mathbf{t}_{id}$ as the q_d dimensional covariate vector associated with the random effects. Let $\boldsymbol{\beta}_d$ and $\boldsymbol{\gamma}_{kd}$ be the unknown fixed and random effects vectors for the dth disease, respectively. It is assumed that the random effects are pairwise independent across all sites and all diseases, i.e., $\boldsymbol{\gamma}_{kd}$ is independent of $\boldsymbol{\gamma}_{k'd'}$ for all $k, d \neq k', d'$. For individual i, we set $\boldsymbol{\gamma}_{(i)d} = \boldsymbol{\gamma}_{kd}$ if and only if individual i was a patient at the kth clinic site. For ease of notation, aggregate $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_D')'$, a $p = \sum_{d=1}^{D} p_d$ dimensional vector, $\boldsymbol{\gamma}_{(i)} = (\boldsymbol{\gamma}_{(i)1}', ..., \boldsymbol{\gamma}_{(i)D}')'$, a $q = \sum_{d=1}^{D} q_d$

dimensional vector, $\mathbf{X}_i = \text{diag}(\mathbf{x}'_{i1}, ..., \mathbf{x}'_{iD})$, a $D \times p$ covariate matrix associated with the fixed effects, and $\mathbf{T}_i = \text{diag}(\mathbf{t}'_{i1}, ..., \mathbf{t}'_{iD})$, a $D \times q$ dimensional covariate matrix associated with the random effects. Then, the multivariate infectious probability of the ith individual is related to their covariate information through the following multivariate generalized linear mixed model

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\theta}) = g(\boldsymbol{\eta}_i; \boldsymbol{\theta}) \tag{3.1}$$

where g is a known multivariate link function, $\boldsymbol{\theta}$ is a collection of link dependent parameters (e.g., a correlation matrix in the multivariate probit link) and $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{iD})'$, where $\eta_{id} = \mathbf{x}'_{id}\boldsymbol{\beta}_d + \mathbf{t}'_{id}\gamma_{(i)d}$.

Mixed effects models are important for clustered observations (clinics), but the dimension of free parameters quickly becomes an issue and furthermore, the prior specification of each covariance matrix $\boldsymbol{\Sigma}_d$ may not be clear. To overcome these challenges, Chen and Dunson [2003] propose reparameterizatizing the covariance matrices as $\boldsymbol{\Sigma}_d = \boldsymbol{\Lambda}_d\mathbf{A}'_d\mathbf{A}_d\boldsymbol{\Lambda}_d$ via a modified Cholesky decomposition. Here, $\boldsymbol{\Lambda}_d$ is a nonnegative q_d dimensional diagonal matrix with entries $\boldsymbol{\lambda}_d$ and $\mathbf{A}_d$ is a $q_d \times q_d$ lower triangular matrix with entries $\mathbf{a}_d = (a_{mld} \colon l = 1, ..., q_d - 1; m = l + 1, ..., q_d)'$ and unit main diagonal. Aggregating $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_1, ..., \boldsymbol{\lambda}'_D)'$ and $\mathbf{a} = (\mathbf{a}'_1, ..., \mathbf{a}'_D)'$, the reparameterized model is

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \boldsymbol{\theta}) = g(\boldsymbol{\eta}_i; \boldsymbol{\theta}) \tag{3.2}$$

where now $\eta_{id} = \mathbf{x}'_{id}\boldsymbol{\beta}_d + \mathbf{t}'_{id}\boldsymbol{\Lambda}_d\mathbf{A}_d\mathbf{b}_{(i)d}$, and $\mathbf{b}_{(i)d}$ is the standardized clinic specific random effect associated with the dth disease. Specifically, $\mathbf{b}_{kd} \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if and only if the ith individual's specimen was extracted at the kth clinic. Here, the standardized random effects adopt the same assumptions as the original random effects. The benefit of model (3.2) over the unparameterized model is twofold. First, it is no longer necessary to specify, or posit prior structure on, the covariance matrices $\boldsymbol{\Sigma}_d$, $d = 1, ..., D$; instead they are estimated through $\boldsymbol{\Lambda}_d$ and $\mathbf{A}_d$. Second, by setting a diagonal element of $\boldsymbol{\Lambda}_d$ to zero effectively zeros out the corresponding row and column of $\boldsymbol{\Sigma}_d$, rendering that random effect insignificant. To this end, a spike and slab prior distribution is utilized to exploit this feature, facilitating variable selection in both the fixed effects and random effects.

The following conditional distribution of the observed testing outcomes $\mathbf{Z} = (\mathbf{Z}'_1, ..., \mathbf{Z}'_J)'$ shows the relationship of the group testing data to the individual level model expressed in (3.2), which is given by

$$\pi(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}) = \sum \left\{ \prod_{d=1}^{D} \prod_{j=1}^{J} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p_j:d}^{1-Z_{jd}} (1 - S_{p_j:d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}} \right.$$
$$\left. \times \prod_{i=1}^{N} P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \boldsymbol{\theta}) \right\}, \tag{3.3}$$

where $\mathbf{b} = (\mathbf{b}_1, ..., \mathbf{b}_K)'$. Note that in expressing this conditional distribution, a few mild assumptions have been made. First, it is assumed that the testing outcomes for each disease are conditionally independent given the true pool statuses $\widetilde{\mathbf{Z}}$ (i.e., $Z_{jd} \mid \widetilde{\mathbf{Z}}$ is independent of $Z_{j'd'} \mid \widetilde{\mathbf{Z}}$) and that the conditional distribution $\mathbf{Z} \mid \widetilde{\mathbf{Z}}$ does not depend on the individuals' covariates. Second, the individuals' true statuses $\widetilde{\mathbf{Y}}_i$ are conditionally independent given the covariates and the random effects. To proceed, note that the summation in (3.3) is over all possible D dimensional binary true statuses for all N individuals, rendering direct evaluation infeasible. To overcome this, we utilize a data augmentation strategy used in error prone group testing literature; see McMahan et al. [2017]. By introducing the true latent statuses $\widetilde{\mathbf{Y}}_i$ as random variables, we instead consider the joint conditional distribution

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}) = \prod_{d=1}^{D} \prod_{j=1}^{J} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p_j:d}^{1-Z_{jd}} (1 - S_{p_j:d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}}$$
$$\times \prod_{i=1}^{N} P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \boldsymbol{\theta}). \tag{3.4}$$

where $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1', ..., \widetilde{\mathbf{Y}}_N')'$. It should be noted that this data augmentation strategy will require sampling the latent random variables $\widetilde{\mathbf{Y}}$; however, this poses no issue as the full conditional distribution is obtainable; see

Section 3.2.2. The prior specifications of the proposed model is

$$\begin{array}{rcll}
\beta_{rd} \mid v_{rd} & \sim & (1-v_{rd}) \cdot \delta_0(\beta_{rd}) + v_{rd} \cdot N(0, \phi^2_{rd}), & r = 1, ..., p_d \\
\lambda_{ld} \mid w_{ld} & \sim & (1-w_{ld}) \cdot \delta_0(\lambda_{ld}) + w_{ld} \cdot TN\{0, \psi^2_{ld}, (0, \infty)\}, & l = 1, ..., q_d \\
\mathbf{a}_d & \sim & N(\mathbf{m}_d, \mathbf{C}_d) & \\
\mathbf{b}_{kd} & \sim & N(\mathbf{0}, \mathbf{I}), & k = 1, ..., K \\
v_{rd} \mid \tau_{v_{rd}} & \sim & \text{Bernoulli}(\tau_{v_{rd}}), & r = 1, ..., p_d \\
w_{ld} \mid \tau_{w_{ld}} & \sim & \text{Bernoulli}(\tau_{w_{ld}}), & l = 1, ..., q_d \\
\tau_{v_{rd}} & \sim & \text{Beta}(a_v, b_v), & r = 1, ..., p_d \\
\tau_{w_{ld}} & \sim & \text{Beta}(a_w, b_w), & l = 1, ..., q_d,
\end{array}$$

where $\mathbf{m}_d$, $\mathbf{C}_d$, a_v, a_w, b_v, and b_w are all hyperparameters, along with the variance components ϕ^2_{rd} and ψ^2_{ld}, which are specified to be large to impose a diffuse slab distribution. The prior distributions stated above for the fixed effects and random effects is the Dirac spike prior distribution, where $\delta_0(x)$ is a degenerate distribution for x with mass at zero and $TN(a, b, c)$ is a truncated normal distribution that arises by restricting a normal distribution with mean a and variance b to the interval c; for further details about the Dirac spike prior, see Wagner and Duller [2012]. It is worth noting that $\mathbf{m}_d$ and $\mathbf{C}_d$ should be specified in an informative fashion (e.g., $\mathbf{m}_d = \mathbf{0}$ and $\mathbf{C}_d = 0.5\mathbf{I}$). This is noted in Chen and Dunson [2003], where failing to do so results in a strong *a priori* correlation between any two random effects within the dth disease.

3.2.2 Posterior computation via probit link

To ground our methodology, we illustrate posterior computation through the multivariate probit link function. Recall that the multivariate probit link function relates an individual's covariate information to their probability of infection through the following relationship:

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = g(\boldsymbol{\eta}_i; \boldsymbol{\theta}) = \int_{I_{i1}} \int_{I_{i2}} \cdots \int_{I_{iD}} \phi_D(\mathbf{t}; \mathbf{0}, \mathbf{R}) d\mathbf{t} \qquad (3.5)$$

where $\phi_D(\cdot \mid \mathbf{0}, \mathbf{R})$ is a D dimensional multivariate normal density with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}$. For each integral, the region of integration is given by $I_{id} = (-\infty, \eta_{id})$ if $\widetilde{Y}_{id} = 1$ and $I_{id} = [\eta_{id}, \infty)$ otherwise. To avoid identifiability issues, $\mathbf{R}$ must be a correlation matrix; see Chib and Greenberg [1998]. Equation 3.4, in conjunction with equation 3.5, shows that the full conditional distribution of the individuals' latent statuses is Bernoulli for each disease. That is, $\widetilde{Y}_{id} \mid \widetilde{\mathbf{Y}}_{-id}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R} \sim \text{Bernoulli}\{p^\star_{id1}/(p^\star_{id0}+$

$p^{\star}_{id1})\}$, where $\widetilde{\mathbf{Y}}_{-id}$ is the vector $\widetilde{\mathbf{Y}}_i$ with the dth element removed and

$$\begin{aligned} p^{\star}_{id1} &= g_{id1}(\boldsymbol{\eta}_i; \mathbf{R}) \prod_{j \in \mathcal{I}_i} S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \\ p^{\star}_{id0} &= g_{id0}(\boldsymbol{\eta}_i; \mathbf{R}) \prod_{j \in \mathcal{I}_i} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \right\}^{I(s_{ijd}>0)} \left\{ (1 - S_{p_j:d})^{Z_{jd}} S_{p_j:d}^{1-Z_{jd}} \right\}^{I(s_{ijd}=0)}. \end{aligned}$$

In the above expressions, the index set $\mathcal{I}_i = \{j : i \in \mathcal{P}_j\}$ keeps track of which pools the ith individual was a member of, $s_{ijd} = \sum_{i' \in \mathcal{P}_j :\, i' \neq i} \widetilde{Y}_{i'd}$, and $g_{id1}(g_{id0})$ is the integral in equation 3.5 when $\widetilde{Y}_{id} = 1(0)$.

Unfortunately, the formulation of (3.5) inside of (3.4) is not very tractable in regards to the regression parameters and the random effects. However, the seminal work of Albert and Chib [1993] can be readily extended to higher dimensions. For each individual, introduce the latent random vector $\boldsymbol{\omega}_i = (\omega_{i1}, ..., \omega_{iD})'$, which denotes a D dimensional normal random vector with mean $\boldsymbol{\eta}_i$ and correlation matrix $\mathbf{R}$; i.e., $\boldsymbol{\omega}_i \sim N(\boldsymbol{\eta}_i, \mathbf{R})$. These latent vectors necessarily follow that $Y_{id} = 1$ if $\omega_{id} \geq 0$ and $Y_{id} = 0$ if $\omega_{id} < 0$. Then, the augmented likelihood function of (3.4) to be considered is

$$\begin{aligned} \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}) \propto & \prod_{d=1}^{D} \prod_{j=1}^{J} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p_j:d}^{1-Z_{jd}} (1 - S_{p_j:d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}} \\ & \times \prod_{i=1}^{N} |\mathbf{R}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i)' \mathbf{R}^{-1} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i) \right\} \prod_{i=1}^{N} f(\boldsymbol{\omega}_i), \end{aligned} \tag{3.6}$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_N)$ and $f(\boldsymbol{\omega}_i) = \prod_{d=1}^{D} I(\omega_{id} \geq 0, \widetilde{Y}_{id} = 1) + I(\omega_{id} < 0, \widetilde{Y}_{id} = 0)$. This framework, in conjunction with the posited prior specifications, allows posterior inference of the regression parameters to be carried out with a full Gibbs sampling algorithm. To elucidate, under the Dirac spike, β_{rd} is set to zero if $v_{rd} = 0$, and the nonzero elements of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{\boldsymbol{v}}$, has the full conditional distribution $\boldsymbol{\beta}_{\boldsymbol{v}} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{v} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_D)'$ and $\boldsymbol{v}_d = (v_{1d}, ..., v_{p_d d})'$. The specific form of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, along with all other full conditionals needed to carry out the Gibbs sampler, are outlined in Appendix B.

3.2.3 Sampling correlation matrix

Up to this point, $\mathbf{R}$ has been treated as a known matrix of correlations between diseases. One may view this as an impractical assumption and to that end, a posterior sampling algorithm that can be used to simultaneously estimate these correlations is provided. Note that the sampling of this matrix, say with a Wishart distribution, is not trivial due to the correlation constraint, i.e., bounded off-diagonal and unit main-diagonal elements. Following the work of Zhang et al. [2006], a parameter-extended Metropolis-Hastings (PX-MH) algorithm for the sampling of correlation matrices can be utilized. Here, parameter-extended refers to the introduction of an extra variance parameter matrix $\mathbf{D}$, a D dimensional diagonal matrix with dth element denoted as $\mathbf{D}_{dd}$. Then $\mathbf{R}$ can be sampled by generating a pair of $\mathbf{R}$ and $\mathbf{D}$ together in the following manner. Assume that $\mathbf{W}$ follows a Wishart distribution with m_0 degrees of freedom and scale matrix $\mathbf{S}$; i.e., $\mathbf{W} \sim \text{Wishart}(m_0, \mathbf{S})$. To relate the extra variance parameter matrix $\mathbf{D}$ and $\mathbf{W}$ to the desired matrix of correlations $\mathbf{R}$, we force $\mathbf{W} = \mathbf{D}^{\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}$. Then PX-MH is carried out in the following manner:

PX-MH Algorithm

1. Initialize $(\mathbf{R}^{(0)}, \mathbf{D}^{(0)})$ so that $\mathbf{W}^{(0)} = \sqrt{\mathbf{D}^{(0)}}\mathbf{R}^{(0)}\sqrt{\mathbf{D}^{(0)}}$ is a covariance matrix. Set $t = 1$.

2. Sample $\mathbf{W}^{\star} = \sqrt{\mathbf{D}^{\star}}\mathbf{R}^{\star}\sqrt{\mathbf{D}^{\star}}$ from Wishart$(m, \mathbf{W}^{(t)}/m)$, where m is a tuning parameter.

3. Propose $(\mathbf{R}^{\star}, \mathbf{D}^{\star})$ as $\mathbf{R}^{\star} = \mathbf{D}^{\star^{-\frac{1}{2}}}\mathbf{W}^{\star}\mathbf{D}^{\star^{-\frac{1}{2}}}$ and $\mathbf{D}^{\star}_{dd} = \mathbf{W}^{\star}_{dd}$.

4. Generate $(\mathbf{R}^{(t+1)}, \mathbf{D}^{(t+1)})$ according to

$$(\mathbf{R}^{(t+1)}, \mathbf{D}^{(t+1)}) = \begin{cases} (\mathbf{R}^{\star}, \mathbf{D}^{\star}) & \text{with probability } \alpha \\ (\mathbf{R}^{(t)}, \mathbf{D}^{(t)}) & \text{otherwise.} \end{cases}$$

5. Increment t and return to step 2.

The acceptance probability in step 4 is given by

$$\alpha = \min\left\{1, \frac{p(\mathbf{R}^{\star}, \mathbf{D}^{\star} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})}{p(\mathbf{R}^{(t)}, \mathbf{D}^{(t)} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})} \frac{f(\mathbf{W}^{(t)} \mid m^{-1}\mathbf{W}^{\star})}{f(\mathbf{W}^{\star} \mid m^{-1}\mathbf{W}^{(t)})}\right\},$$

where $p(\mathbf{R}, \mathbf{D} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})$ is the joint posterior density of $(\mathbf{R}, \mathbf{D})$, which is up to a constant proportional to

$$|\mathbf{R}|^{\frac{m-D-1}{2}}|\mathbf{D}|^{\frac{m}{2}-1}\exp\left\{-\text{tr}(\mathbf{S}^{-1}\mathbf{D}^{\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}})/2\right\} \times \prod_{i=1}^{N}\phi_D(\boldsymbol{\omega}_i; \boldsymbol{\eta}_i, \mathbf{R}).$$

Furthermore, $f(\cdot \mid m^{-1}\mathbf{W}^{(t)})$ in α is the product of the Jacobian $\prod_{d=1}^{D}\mathbf{D}_{dd}^{\frac{D-1}{2}}$ and a Wishart distribution with m degrees of freedom and scale matrix $m^{-1}\mathbf{W}^{(t)}$. Note that the degrees of freedom m is a tuning parameter that can adjust the acceptance rates in the fourth step.

3.2.4 Sampling assay accuracies

The methodology thus far has assumed the assay accuracies $S_{e_j:d}$ and $S_{p_j:d}$ are known constants. In some settings, if not most, this assumption is not appropriate and so simultaneous estimation of the assay accuracies is required. To do so, we follow the work of McMahan et al. [2017] and extend it to D dimensions. Aggregate the indices of pools $j = 1, ..., J$ that were tested with the mth multiplex testing assay and denote as $\mathcal{I}_m$. Accordingly, notate $S_{e(m):d}$ and $S_{p(m):d}$ as the sensitivity and specificity of the mth assay for the dth disease, $m = 1, ..., M$; i.e., $S_{e_j:d} = S_{e(m):d}$ and $S_{p_j:d} = S_{p(m):d}$ for all $j \in \mathcal{I}_m$. With this convention, equation (3.4) can be rewritten as

$$\prod_{d=1}^{D}\prod_{m=1}^{M}\prod_{j\in\mathcal{I}_m}\left\{S_{e(m):d}^{Z_{jd}}(1-S_{e(m):d})^{1-Z_{jd}}\right\}^{\widetilde{Z}_{jd}}\left\{S_{p(m):d}^{1-Z_{jd}}(1-S_{p(m):d})^{Z_{jd}}\right\}^{1-\widetilde{Z}_{jd}}$$
$$\times\prod_{i=1}^{N}P(\widetilde{\mathbf{Y}}_i=\widetilde{\mathbf{y}}_i\mid\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{a},\mathbf{b}_{(i)},\boldsymbol{\theta}).$$

Given the form of this conditional distribution, natural prior choices for the testing assay accuracies are $S_{e(m):d} \sim \text{Beta}(a_{e(m):d}, b_{e(m):d})$ and $S_{p(m):d} \sim \text{Beta}(a_{p(m):d}, b_{p(m):d})$. These specifications lead to the following full conditionals

$$S_{e(m):d}\mid\mathbf{Z},\widetilde{\mathbf{Y}}\sim\text{Beta}(a^{\star}_{e(m):d},b^{\star}_{e(m):d})$$
$$S_{p(m):d}\mid\mathbf{Z},\widetilde{\mathbf{Y}}\sim\text{Beta}(a^{\star}_{p(m):d},b^{\star}_{p(m):d}),$$

where $a^{\star}_{e(m):d} = a_{e(m):d} + \sum_{j\in\mathcal{I}_m} Z_{jd}\widetilde{Z}_{jd}$, $b^{\star}_{e(m):d} = b_{e(m):d} + \sum_{j\in\mathcal{I}_m}(1 - Z_{jd})\widetilde{Z}_{jd}$, $a^{\star}_{p(m):d} = a_{p(m):d} + \sum_{j\in\mathcal{I}_m}(1 - Z_{jd})(1 - \widetilde{Z}_{jd})$, and $b^{\star}_{p(m):d} = b_{p(m):d} + \sum_{j\in\mathcal{I}_m} Z_{jd}(1 - \widetilde{Z}_{jd})$. The posterior distributions derived in Section 3.2.2 are left unchanged up to acknowledging dependence on the testing accuracies and accounting for the slight change in notation.

3.3 Simulations

To assess the performance of our estimation and variable selection methods, we simulated group testing data in the following manner. To closely mimic the features of the motivating Iowa data, $K = 50$ clinic sites were conceptualized with 100 individuals at each clinic; i.e., $N = 5000$. This sample size provides for a stress test of our model's capability of recovering the true covariates, as compared to the amount of data available in the motivating data set. The infection status for each individual was generated as a $D = 2$ dimensional binary vector $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, \widetilde{Y}_{i2})'$ according to $\widetilde{Y}_{id} = 1$ if $\omega_{id} \geq 0$ and 0 otherwise, where $\boldsymbol{\omega}_i = (\omega_{1d}, \omega_{2d})'$ is a random draw from a multivariate normal distribution with mean $\boldsymbol{\eta}_i$ and correlation matrix $\mathbf{R}$. The mean vectors $\boldsymbol{\eta}_i$ are constructed with the following values: $\boldsymbol{\beta}_1 = (-4, -1.5, 0.5, 0.25, 0)'$, $\boldsymbol{\beta}_2 = (-2.5, 1, -0.75, 0.3, 0, 0)'$, $\boldsymbol{\lambda}_1 = (1, 0.75, 0.25, 0, 0)'$, $\boldsymbol{\lambda}_2 = (0.8, 0.3, 0.15, 0, 0, 0)'$, $\mathbf{a}_1 = (0.9, 0.5, 0.7, \mathbf{0}_7)'$, $\mathbf{a}_2 = (0.75, -0.5, 0.8, 1.2, 0.5, 0.3, \mathbf{0}_9)'$, $\mathbf{0}_\ell$ is an ℓth dimensional zero vector, $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if individual i presented at clinic site k, $\mathbf{b}_{kd} \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I})$, and $\mathbf{R}$'s off-diagonal elements are set to $\rho = 0.99$. We set the covariate vectors associated with the fixed effects to standardized versions of $\mathbf{x}^\star_{i1} = (1, N(0,1), \mathrm{B}(0.5), \mathrm{B}(0.5), N(0,1))'$ and $\mathbf{x}^\star_{i2} = (1, N(0,1), \mathrm{B}(0.5), \mathrm{B}(0.5), N(0,1), \mathrm{B}(0.5))'$, where $\mathrm{B}(0.5)$ represents a Bernoulli(0.5) random variable. The random effects covariates for each disease are taken to be equal to the fixed effects covariates. These covariate and parameter configurations provide for an overall disease prevalence rate of about 3% and 9%, which is in keeping with the observed prevalence rate of gonorrhea and chlamydia, respectively, from the motivating data. This process was used to generate 500 individual level data sets.

Group testing outcomes were generated with these individual level data sets in the following manner. Each individual was randomly placed into a group of size $c_j = 4$ individuals. This poses the most difficult estimation configuration; that is, individuals are pooled across sites rather than within sites and thus each pool contains multiple random effects. To proceed, one of three group testing algorithms was implemented to test each of these master pools; namely, master pool testing (MPT), Dorfman testing (DT), and array testing (AT). Briefly, under MPT, each master pool is tested with no further retesting, regardless of the outcome. DT proceeds by individually retesting each contributing member of a master pool that tested positively. Similarly, AT views the master pools as rows and columns of an array and will retest individuals identified to be likely positives; e.g., individuals residing at the intersection of positive rows and columns. For the specific AT protocol we adopted, see Kim et al. [2007]. Under each protocol, the testing response for the jth pool was simulated as $Z_{jd} \mid \widetilde{Z}_{jd} \sim$ Bernoulli$\{S_{e_j:d}\widetilde{Z}_{jd} + (1 - S_{p_j:d})(1 - \widetilde{Z}_{jd})\}$, where $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$. For comparative purposes, individual testing (IT) was also implemented. Regarding the testing assay accuracies, two different simulation configurations were considered. The first setting assumes that the sensitivity and specificity for each pool are known; i.e., $S_{e_j:1} = S_{e_j:2} = 0.95$ and $S_{p_j:1} = S_{p_j:2} = 0.98$ for all $j = 1, ..., J$. In the second setting, two different multiplex testing assays are considered unknown and estimated; the first assay $(m = 1)$ is used to test pools and the second assay $(m = 2)$ is used to retest individuals. More specifically, we set $S_{e(1):d} = 0.95, S_{e(2):d} = 0.98, S_{p(1):d} = 0.98$, and $S_{p(2):d} = 0.99$ for $d = 1, 2$. In this setting, we only consider DT and AT since both protocols mandate both pool and individual level testing.

To perform the assessment of the proposed model, we set $\mathbf{m}_0 = \mathbf{0}, \mathbf{C}_0 = 0.5\mathbf{I}$, and used flat priors for all mixing weights. Recall that the prior parameters for $\mathbf{a}$ should be chosen in a somewhat informative fashion to avoid a strong *a priori* correlation between any two random effects; see Chen and Dunson [2003]. To impose a diffuse prior variance on the slab distribution of the fixed and random effects, we took $\phi^2_{rd} = \psi^2_{ld} = 100$. The prior degrees of freedom and scale matrix for $\mathbf{W}$ was set as $m_0 = D + 1 = 3$ and $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is a $D \times D$ identity matrix, and the proposal degrees of freedom was set to $m = 300$; these values are discussed in Zhang et al. [2006]. To perform posterior estimation and inference, the last half of 50000 iterations of our MCMC algorithm was used for summary statistics. Point estimates of the model parameters were obtained as the empirical means of the posterior distributions.

Table 6.4 summarizes estimation performance in the first simulation setting; that is, when $S_{ej:d}$ and $S_{pj:d}$ are known. Overall, these results illustrate our approach provides reliable inference for the fixed and random effects; i.e., the empirical bias and the variability in the estimates are small relative to the true value of the corresponding parameter. These results also indicate the proposed methodology is adept at identifying non-zero fixed and random effects. That is, covariates with strong (no) effects almost always have posterior inclusion probabilities near 1 (0) in all data sets.

Among the group testing protocols presented in Table 6.4, MPT generally performs the worst in terms of estimation i.e., estimates obtained from analyzing MPT data exhibit the most bias and variability. This is expected because MPT does not complete the classification process like the other protocols and therefore results in less information about the individuals' latent statuses. In contrast, the estimation performance under the two classification protocols (DT and AT) is as good if not better than the performance under IT; furthermore, DT/AT estimates are obtained at about 60% of the testing cost on average when compared to IT. Specifically, 5000 tests are used to complete IT, while DT and AT require 2747 and 3258 tests on average, respectively. These results illustrate the “get more for less” phenomenon that has previously been reported with group testing regression [Zhang et al., 2013, McMahan et al., 2017].

Table 6.5 summarizes estimation performance when the assay accuracy probabilities are unknown. In this second setting, the proposed methodology is tasked with estimating 8 additional parameters; i.e., the testing assay accuracies. These results indicate we can accurately estimate these parameters and the variability in the estimates is small relative to the true values. Moreover, there are no appreciable differences between the estimates in Tables 6.4 and 6.5 for DT and AT; i.e., inference for the fixed and random effects is not impacted by having to estimate these additional parameters.

[Table 4 about here.]

[Table 5 about here.]

3.4 Data application

The State Hygienic Laboratory (SHL) in Iowa annually screens thousands of residents for two of the most common sexually transmitted diseases (STDs): gonorrhea and chlamydia. In an effort to more accurately detect the infectious agents, while also saving money in testing costs, the SHL employs group testing and tests for these two diseases simultaneously with the use of multiplex testing assays. The process begins with drawing specimens (e.g., urine, swab, etc.) from individuals at different clinics (e.g., family planning clinics, STD testing clinics, etc.) throughout the state which are then transported to the SHL for testing. Current SHL screening protocols mandate that all male specimens and female urine specimens be tested individually while all female swab specimens are processed through a variant of Dorfman testing (DT); for further discussion, see Tebbs et al. [2013]. The SHL uses the Aptima Combo 2 Assay to test both pooled and individual specimens.

During the 2014 calendar year, 64 different clinics submitted specimens to the SHL for testing. The available data consist of results collected on 4316 individual urine specimens, 416 individual swab specimens, and 2286 swab master pools (1 of size 2, 12 of size 3, and 2273 of size 4), as well as the test results required to resolve the positive master pools. That is, a master pool that tests positive for either disease is resolved by retesting all individuals for both diseases simulatenously. In addition to the test data, several covariates were collected on each individual: age (in years, denoted by x_1^*), a race indicator ($x_2^* = 1$ if Caucasian and $x_2^* = 0$ otherwise), an indicator denoting whether the patient reported a new sexual partner in the last 90 days ($x_3^* = 1$ if affirmative and $x_3^* = 0$ otherwise), an indicator denoting whether the patient reported having multiple sexual partners in the last 90 days ($x_4^* = 1$ if affirmative and $x_4^* = 0$ otherwise), an indicator denoting whether the patient reported sexual contact with an STD-positive partner in the previous year ($x_5^* = 1$ if affirmative and $x_5^* = 0$ otherwise), and an indicator denoting whether the patient presented with symptoms ($x_6^* = 1$ if affirmative and $x_6^* = 0$ otherwise). We relate the individuals' disease statuses to the available covariate information via the multivariate probit link function in Equation (3.5), where the covariate vectors $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{t}_{i1}$, and $\mathbf{t}_{i2}$ are taken to be equal and are standardized versions of $(1, x_{i1}^*, x_{i2}^*, x_{i3}^*, x_{i4}^*, x_{i5}^*, x_{i6}^*)'$. Standardization was used so that the spike and slab distributions have the same impact on the regression coefficients across all covariates. For each of the 64 clinics, a random effect vector $\mathbf{b}_{kd}$ is conceptualized for each disease, with the convention that $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if the ith individual was a patient at the kth clinic site.

To complete model specification, all prior specifications were made in the exact same fashion as was described in Section 3.3 with the exception of the testing assay accuracies. For this analysis, three sets of multiplex testing accuracies were conceptualized to account for the SHL's testing protocol: $S_{e(1):d}$ and $S_{p(1):d}$ for individually testing swab specimens, $S_{e(2):d}$ and $S_{p(2):d}$ for individually tested urine specimens, and $S_{e(3):d}$ and $S_{p(3):d}$ for swab specimens tested in pools. In total, this gives 12 testing assay accuracies to be estimated. Informative priors were used on the testing assay accuracies. Specifically, for gonorrhea, we specified $a^{\star}_{e(m):1}\big(a^{\star}_{p(m):1}\big)$ to be 126(1335), 116(1347), and 126(1335) for $m = 1, 2, 3$, respectively, and took $b^{\star}_{e(m):1}\big(b^{\star}_{p(m):1}\big)$ to be 1(17), 11(10), and 1(17) for $m = 1, 2, 3$, respectively. For chlamydia, we set $a^{\star}_{e(m):2}\big(a^{\star}_{p(m):2}\big)$ to be 195(1154), 197(1170), and 195(1154) for $m = 1, 2, 3$, respectively, and $b^{\star}_{e(m):2}\big(b^{\star}_{p(m):2}\big)$ to be 12(28), 11(13), and 12(28) for $m = 1, 2, 3$, respectively. These values are well vetted in extensive pilot studies and have been used in previous literature; e.g., see McMahan et al. [2017].

Table 6.6 displays estimates of the posterior mean and standard deviation for all model parameters and estimates of the posterior probabilities of inclusion for the fixed and random effects. The direction of the estimates of the fixed effects are expected in light of known epidemiological patterns of gonorrhea and chlamydia infections. That is, the risk of infection tends to decrease with age and Caucasian females are associated with a lower risk when compared to females of other races. In contrast, having a new sexual partner, multiple partners, and contact with STDs are all associated with an increased risk. Our analysis also identifies the random intercept parameter for both diseases, and potentially the random effect for new sexual partner associated with chlamydia, to be strongly significant indicating clear evidence of heterogeneity across the clinics throughout the state.

[Table 6 about here.]

3.5 Discussion

We have proposed a Bayesian approach to estimate multivariate generalized linear mixed models with data arising from a group testing protocol. The novelty of the proposed methodology is the multivariate nature of the model, allowing practitioners to analyze group testing data arising from multiplex testing assays. Moreover, when compared to existing regression techniques for group testing data, the appeal of our methodology is twofold. First, including random effects allows one to account for heterogeneity that may exist across subgroups of the population; i.e., clinic sites. Second, our approach employs automatic variable selection for both the fixed and random effects by using spike and slab priors. Through a series of data augmentation steps, we illustrate how our regression methods can be used with the multivariate probit link function.

Chapter 4

A Bayesian Hierarchical Model for Identifying Significant Polygenic Effects while Controlling for Confounding and Repeated Measures

4.1 Introduction

Oryza sativa, or Asian rice, is a staple food in Asian countries, and its continual production is essential to food security. As the fourth most populous country in the world, Indonesia is also one of the biggest producers and consumers of rice. With a current annual population growth rate of 1.2% [Bank, 2012], the Indonesian population is predicted to reach 337 million in 2050 [Facts, 2012]. With the current rate of rice consumption at 139 kg per capita per year [Shean, 2012], Indonesia must reach an annual rice production of 47 million tons by 2050 to meet population needs. These needs have spurred research aimed at increasing crop yield by better understanding which rice varieties respond favorably/unfavorably to certain growing conditions. For example, the Indonesian Center for Rice Research (ICRR) is continuously evaluating new rice varieties from breeding programs. The practice of cross breeding plants to create new varieties with desirable characteristics dates back to the origins of agriculture. To aid this endeavor, this paper develops statistically sound methods that can identify genetic factors related to specific phenotypes of interest, while controlling for confounding variables, genetic similarities, and allowing for repeated measurements. Our methods provide agro-scientists with a new tool that can be used to predict the potential of new plant varieties, without requiring expensive field testing.

Several key concerns arise when new variety accessions are evaluated. For example, it is hypothesized that climate change will affect rice production through a rise in average temperatures and increasingly frequent and prolonged floods and droughts in Southeast Asia [Singh et al., 2014]. For every degree Celsius increase in temperature, rice yields are estimated to decline by 7% [Matthews et al., 1997]. Further, drought is the largest constraint to rice production in the rainfed agricultural systems of Asia [Pandey and Bhandari, 2009]. To address such issues, researchers seek to identify/develop varieties of rice that are resilient to adverse climate conditions and have desirable production qualities. The proposed methods, by controlling for covariate effects, have two beneficial characteristics. First, they allow for a more accurate assessment of genetic effects that could influence variety development. Second, they allow one to predict how a particular phenotype of interest will perform in conditions where data are not taken.

New plant development based on genetic variation has, of course, been extensively considered elsewhere. For example, marker-assisted selection (MAS) uses DNA markers to identify and develop plants with desirable traits, including disease resistance and yield improvements. This process involves linking variations in the genome, particularly single-nucleotide polymorphisms (SNPs), to important characteristics and then using those genetic variants to select seeds for planting or breeding. MAS programs have had limited success when multiple genetic and environmental factors are involved [Kilian et al., 2012, Schielzeth and Husby, 2014, Sun and Wu, 2015]. On the statistical side, rudimentary analyses often fail to appropriately control for environmental variables. A single genetic variant typically has a small effect on rice yield; however, their combined effects can be significant. On the other hand, field factors such as seasonal time of planting, duration in the field, intensity of stress, and overall climatic conditions strongly influence rice yield. Thus, by not appropriately accounting for the latter, evaluation of the former is a difficult task.

From a statistical point of view, this study seeks to identify and assess the joint effect of genetic markers while controlling for confounding covariates, a task tantamount to model selection in a high dimensional regression framework. Many techniques exist for such problems; e.g., the least absolute shrinkage and selection operator (LASSO) of Tibshirani [1996a], smoothly clipped absolute deviations regression of Fan and Li [2001a], the elastic net of Zou and Hastie [2005a], and the adaptive LASSO Zou [2006a], etc. These techniques treat the observed phenotypic responses as statistically independent, which is unrealistic since the rice varieties in question are genetically similar to each other. To account for this issue, Zhou et al. [2013] proposed a Bayesian sparse linear mixed model, which uses a "spike and slab" prior to induce sparsity. This innovative approach is not directly applicable here as it does not allow for repeated measurements taken on the same rice variety, which is needed to evaluate environmental factors. Another notable contribution in this area is that of Yazdani and Dunson [2015a], which proposed a two-stage approach that is a hybrid of a Bayesian single and simultaneous analysis; i.e., the first stage screens markers independently to develop a candidate set, the candidate set of markers is then jointly modeled in the second stage. In both of the aforementioned methods, joint estimation and inference is completed through standard Markov Chain Monte Carlo (MCMC) techniques, which can be computationally burdensome when the number of genetic markers is large. Thus, a general sparse regression methodology is developed here for variable selection in a high dimensional context in the presence of confounding and genetic variables. The proposed approach explicitly accounts for genetic similarities and allows for repeated measures (e.g., across fields, seasons, etc.). From the hierarchical representation of the proposed model, a computationally efficient expectation-maximization (EM) algorithm is developed for parameter estimation, providing almost instantaneous estimates of all model parameters for studies similar in size to the motivating application.

The remainder of this article is organized as follows. Section 2 introduces a sparse regression model and describes an EM algorithm for parameter estimation. Section 3 studies the finite sample properties of the proposed estimator through simulation. Section 4 applies the proposed methodology to yield data for 467 rice varieties planted in three fields in Indonesia. Section 5 concludes with comments about the limitations and extensions of the model and study design.

4.2 Model

To assess environmental and genetic effects while accounting for genetic similarities, the regression model

$$Y_i = \beta_0 + \mathbf{F}_i'\boldsymbol{\beta}_1 + \mathbf{S}_i'\boldsymbol{\beta}_2 + \mathbf{G}_i'\boldsymbol{\gamma} + \epsilon_i, \quad i \in \{1, \ldots, n\}, \tag{4.1}$$

is posited. Here, Y_i is a response variable representing a phenotype of interest measured on the ith observation (e.g., crop yield), $\mathbf{F}_i = (F_{i1}, \ldots, F_{ir})'$ is an r-dimensional vector of covariates (e.g., field identifiers, temperature, humidity, etc.), $\mathbf{S}_i = (S_{i1}, \ldots, S_{iq})'$ is a q-dimensional vector of single-nucleotype polymorphism (SNP) genotypes, $\mathbf{G}_i$ is a k-dimensional binary vector indicating the plant variety of the ith observation, and

ϵ_i is the error term. The regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are covariate and genetic marker effects, respectively, with β_0 denoting the usual intercept and $\boldsymbol{\gamma}$ being a k-dimensional vector of variety specific random effects. For modeling purposes, it is assumed that the error terms are independent and follow a normal distribution with zero mean and common variance σ^2; i.e., $\boldsymbol{\epsilon}|\sigma^2 \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ and $\mathbf{I}$ is an $n \times n$ identity matrix. Through the specification of $\mathbf{G}_i$, one can handle multiple observations (i.e., repeated measurements) from the same plant variety by allowing them to share a common random effect.

In this model, genetic similarities between distinct varieties are quantified through random effects. In particular, as in Zhou et al. [2013] and Zhou [2016], we assume that

$$\boldsymbol{\gamma}|\sigma^2 \quad \sim \quad N(\mathbf{0}, \sigma^2\mathbf{C}),$$

where $\mathbf{C}$ is a known $k \times k$ "relatedness matrix" that describes the genetic similarities between the k different plant varieties. Several forms of $\mathbf{C}$ have been proposed; for further discussion see Dodds et al. [2015] and the references therein. Most forms of $\mathbf{C}$ are based on measured genotypes, which are unique to the k varieties under consideration. The metric implemented by the genome-wide efficient mixed model association (GEMMA) algorithm is used here; for further details and discussion, see Zhou et al. [2013] and Zhou [2016]. In particular, $\mathbf{C} = q^{-1}\mathbf{S}_u\mathbf{S}_u'$, where $\mathbf{S}_u$ is a $k \times q$ matrix whose ℓth row consists of the genotypes for the ℓth plant variety, for $\ell = 1, \ldots, k$. Other relatedness matrices, such as those discussed in Dodds et al. [2015], are easily incorporated into our approach.

For ease of exposition, make the aggregations $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)'$, $\mathbf{F} = (\mathbf{F}_1, \ldots, , \mathbf{F}_n)'$, $\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_n)'$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' = (\beta_0, \beta_1, ..., \beta_p)'$, $p = r + q$, and $\mathbf{X} = (\mathbf{1}, \mathbf{F}, \mathbf{S})$, where $\mathbf{1}$ is an n-dimensional vector of ones. Then (4.1) is succinctly expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{4.2}$$

where the ith row of the design matrix $\mathbf{X}$ is $\mathbf{X}_i = (1, F_{i1}, \ldots, F_{ir}, S_{i1}, \ldots, S_{iq})$. It is worthwhile to point out that the proposed approach can also be used to evaluate SNP-SNP interactions and/or SNP-covariate interactions, by including the necessary and usual terms in the design matrix $\mathbf{X}$. Although, due to the combinatorial explosion in the potential number of such interactions, it is generally advisable that these interactions be chosen judiciously. To complete the Bayesian model formulation, the following prior distributions for β_0,

β_j, and σ^2 are specified:

$$\begin{aligned}\beta_0|\sigma^2 &\sim N(0, \sigma^2 T_0),\\ \beta_j|\sigma^2, \alpha, \eta &\sim \text{GDP}(\psi = \sigma\eta/\alpha, \alpha),\ \text{for } j = 1, \ldots, p,\\ \sigma^2 &\sim \pi(\sigma^2) \propto 1/\sigma^2.\end{aligned}$$

Here $\theta \sim \text{GDP}(\psi, \alpha)$ indicates that the random variable θ has a generalized double Pareto distribution whose probability density function is

$$f(\theta|\psi, \alpha) = \frac{1}{2\psi}\left(1 + \frac{|\theta|}{\alpha\psi}\right)^{-(\alpha+1)}, \quad -\infty < \theta < \infty,$$

where $\psi > 0$ and $\alpha > 0$ are scale and shape parameters, respectively. These prior specifications put a vague independent normal prior on β_0 when T_0 is large, and independent generalized double Pareto shrinkage priors on the other regression coefficients. For further details, see Armagan et al. [2013b]. As such, our approach is referred to as the genetic generalized double Pareto (GGDP) regression model. Through the shrinkage prior, our method can handle the scenario where $p > n$, which are ubiquitous in genomic association studies such as the one considered herein.

The hyperparameters α and η play a crucial role in the shrinkage prior. Larger values of α correspond to a more peaked prior density with lighter tails, thus imposing stronger shrinkage on the regression parameters. In contrast, larger η provide a flatter density with less shrinkage. As suggested in Armagan et al. [2013b], a suitable default choice for these hyperparameters is $\alpha = \eta = 1$, which leads to priors with Cauchy-like tails. To circumvent specification of these hyperparameters, we use the following hyper-priors:

$$\begin{aligned}\alpha &\sim \text{Uniform}(\tau_{1\alpha}, \tau_{2\alpha}),\ \tau_{2\alpha} > \tau_{1\alpha} > 0,\\ \eta &\sim \text{Uniform}(\tau_{1\eta}, \tau_{2\eta}),\ \tau_{2\eta} > \tau_{1\eta} > 0.\end{aligned}$$

This makes the data inform us about the values of α and η, and serves as an attempt to prevent over/under shrinking the regression coefficients.

A key feature of the generalized double Pareto shrinkage prior is that it can be represented as a scale mixture of normal distributions, see Proposition 1 in Armagan et al. [2013b]. Thus, for the regression coefficients, the following hierarchical representation provides for the same prior specifications as those above:

$$\begin{aligned}\boldsymbol{\beta}|\sigma^2, \mathbf{T} &\sim N(\mathbf{0}, \sigma^2\mathbf{T}),\\ T_j|\lambda_j &\sim \text{Exponential}(\lambda_j^2/2),\ \text{for } j = 1, ..., p,\\ \lambda_j|\alpha, \eta &\sim \text{Gamma}(\alpha, \eta),\ \text{for } j = 1, ..., p,\end{aligned}$$

where $\mathbf{T} = \text{diag}(T_0, ..., T_p)$ and T_0 is again a specified constant. This hierarchical representation uses the rate parameterization of both the exponential and gamma distributions; for example, the mean of an Exponential variate with parameter λ is λ^{-1}.

Under the above hierarchy, an efficient Markov Chain Monte Carlo (MCMC) sampling algorithm can be constructed through a sequence of Gibbs and Metropolis Hastings steps. Unfortunately, inference via standard MCMC techniques will not provide a sparse estimate of $\boldsymbol{\beta}$, despite the specified shrinkage prior. Obtaining a sparse estimator allows one to estimate the unknown parameters in the model while simultaneously identifying variables that are significantly related to the response. To this end, an EM algorithm is developed to obtain a sparse Bayesian maximum a posteriori probability (MAP) estimator of $\boldsymbol{\beta}$. Essentially, this blends standard frequentist and Bayesian methods, as motivated by Armagan et al. [2013b]. The use of a shrinkage prior and our non-standard estimator are the primary improvements over standard GEMMA implementations. In particular, GEMMA specifies a "spike and slab" prior for the regression coefficients and completes model fitting through MCMC techniques, which can be computationally burdensome and does not yield a sparse estimator.

4.2.1 Sparse Estimation for Variable Selection

The key problem addressed here is to identify which covariates influence the response in (4.1). Motivated by the GDP prior framework and its hierarchical formulation, an EM algorithm will now be developed to facilitate both model fitting and parameter selection via a MAP estimator. The EM algorithm developed for the model in (4.1) is similar to that in Armagan et al. [2013b], with a few differences. Specifically, our formulation allows one to account for genetic similarities between plant varieties through the random effects $\boldsymbol{\gamma}$, and the parameters that control the shrinkage/regularization (i.e., α and η) are estimated along with the other model parameters.

The EM algorithm is developed by viewing the posterior distribution, resulting from the hierarchical representation of the GDP prior, as a complete data likelihood in which T_j and λ_j are regarded as missing (i.e., latent) data, after integrating over the distribution of $\boldsymbol{\gamma}$. After integrating over the distribution of the random effects, one obtains

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{Q}), \tag{4.3}$$

where $\mathbf{Q} = \mathbf{I} + \mathbf{G}\mathbf{C}\mathbf{G}'$. The parameters updated at the maximization (M) step of the algorithm are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \alpha, \eta)$. The derivation of the EM algorithm begins by computing the expectation of the logarithm of the complete data likelihood (i.e., the logarithm of the posterior distribution) with respect to the missing data, conditional on the observed data $\mathcal{D} = \{\mathbf{Y}, \mathbf{X}, \mathbf{G}\}$ and current parameter estimates $\boldsymbol{\theta}^{(d)} = (\boldsymbol{\beta}^{(d)}, \sigma^{2^{(d)}}, \alpha^{(d)}, \eta^{(d)})$ (where d indicates the iteration level in the algorithm). This yields $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = Q_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) + Q_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) + Q_3(\boldsymbol{\theta}^{(d)})$, where

$$\begin{aligned}
Q_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) =& -\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \beta_0^2 T_0^{-1} + \sum_{j=1}^p \beta_j^2 E(T_j^{-1})}{2\sigma^2} \\
& - \frac{n+p+3}{2}\log(\sigma^2), \\
Q_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) =& \sum_{j=1}^p \alpha\log(\eta) - \log\{\Gamma(\alpha)\} + (\alpha - 1)E\{\log(\lambda_j)\} - E(\lambda_j)\eta,
\end{aligned}$$

and $Q_3(\boldsymbol{\theta}^{(d)})$ is a function of $\boldsymbol{\theta}^{(d)}$, but is free of $\boldsymbol{\theta}$. Here and elsewhere, the conditioned variables in expectations is suppressed for notational brevity; i.e., $E(\cdot) = E(\cdot|\mathcal{D}, \boldsymbol{\theta}^{(d)})$. Using the model's hierarchical formulation, it is possible to express all needed expectations in closed form:

$$\begin{aligned}
E(T_j^{-1}) &= (\alpha^{(d)} + 1)\sigma^{2^{(d)}}/\{|\beta_j^{(d)}|(|\beta_j^{(d)}| + \eta^{(d)}\sigma^{(d)})\}, \\
E\{\log(\lambda_j)\} &= \Psi(\alpha^{(d)} + 1) - \log(|\beta_j^{(d)}|/\sigma^{(d)} + \eta^{(d)}), \\
E(\lambda_j) &= (\alpha^{(d)} + 1)/(|\beta_j^{(d)}|/\sigma^{(d)} + \eta^{(d)}),
\end{aligned}$$

where $\Psi(x) = \Gamma'(x)/\Gamma(x)$; i.e., $\Psi(\cdot)$ is the digamma function.

The M step of the EM algorithm has $\boldsymbol{\theta}^{(d+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$. Maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ over $\boldsymbol{\beta}$ and σ^2 yields the closed form updates

$$\begin{aligned}
\boldsymbol{\beta}^{(d+1)} &= (\mathbf{X}'\mathbf{Q}^{-1}\mathbf{X} + \mathbf{D}^{(d)})^{-1}\mathbf{X}'\mathbf{Q}^{-1}\mathbf{Y}, \\
\sigma^{2^{(d+1)}} &= \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(d+1)})'\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(d+1)}) + \boldsymbol{\beta}^{(d+1)'}\mathbf{D}^{(d)}\boldsymbol{\beta}^{(d+1)}}{n+p+3},
\end{aligned}$$

where $\mathbf{D}^{(d)} = \text{diag}\{T_0^{-1}, E(T_1^{-1}), \ldots, E(T_p^{-1})\}$. The updates $\alpha^{(d+1)}$ and $\eta^{(d+1)}$ are the maximizers of

$Q_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ and are computed via standard numerical optimization techniques. Note, the uniform priors for α and η dictate that the updates of $\alpha^{(d+1)}$ and $\eta^{(d+1)}$ be computed over the intervals $(\tau_{1\alpha}, \tau_{2\alpha})$ and $(\tau_{1\eta}, \tau_{2\eta})$, respectively.

The EM algorithm can now be succinctly stated:

1. Initialize $\boldsymbol{\theta}^{(0)}$ and set $d = 0$.

2. Compute $\boldsymbol{\beta}^{(d+1)}$ and $\sigma^{2^{(d+1)}}$ via the aforementioned expressions.

3. Obtain $\alpha^{(d+1)}$ and $\eta^{(d+1)}$ as the maximizers of

$$\sum_{j=1}^{p} \alpha \log(\eta) - \log\{\Gamma(\alpha)\} + (\alpha - 1)E\{\log(\lambda_j)\} - E(\lambda_j)\eta.$$

4. Set $d = d + 1$, and return to Step 2.

Steps 2-4 are iterated until convergence, at which point a sparse estimator of the regression coefficients is obtained. Due to the penalty form in the GDP prior, once a regression coefficient is dropped from the model (i.e., is set to zero), it can not return. Thus, the computational burden lessens as the algorithm iterates through steps 2-4.

Note, when $p >> n$ the computationally expensive aspect of the proposed EM algorithm involves the inversion of a $(p+1) \times (p+1)$ dense matrix in order to compute the update of the regression coefficients. This computational burden can easily be avoided by exploiting the Sherman-Morrison-Woodbury formula, so that one has that

$$(\mathbf{X}'\mathbf{Q}^{-1}\mathbf{X} + \mathbf{D}^{(d)})^{-1} = \mathbf{D}^{(d)^{-1}} - \mathbf{D}^{(d)^{-1}}\mathbf{X}'(\mathbf{Q} + \mathbf{X}\mathbf{D}^{(d)^{-1}}\mathbf{X}')^{-1}\mathbf{X}\mathbf{D}^{(d)^{-1}},$$

where the inversion of $\mathbf{D}^{(d)}$ is trivial since it is diagonal and the other matrix inversion step on the right-hand side involves only an $n \times n$ matrix. Utilizing this inversion formula, the proposed approach can be used when p is on the order of 10^5, which is a situation which is commonly encountered in genome-wide association studies.

We point out that if $\mathbf{Q} = \mathbf{I}$, a model that ignores genetic similarities is fitted; i.e., the model reduces to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. We refer to this model as the generalized double Pareto (GDP) regression model. Further, by setting $\alpha = \tau_{1\alpha} = \tau_{2\alpha}$, $\eta = \tau_{1\eta} = \tau_{2\eta}$, and $\mathbf{Q} = \mathbf{I}$, the proposed approach reduces to that in Armagan et al. [2013b].

4.3 Numerical Studies

A simulation study was conducted to evaluate the finite sample performance of our approach. The characteristics assessed include the method's ability to 1) identify significant covariates under various signal to noise ratios, 2) accurately estimate the effect size of significant covariates, 3) classify covariates not related to the response as such, and 4) capably handle the complex data structures that are ubiquitous in genomic association studies. To accomplish this, data were simulated to mimic the design of our ensuing application: $k = 430$ unique rice varieties, each of which are planted in three distinct fields. This results in $n = 1290$ observations. For this study, the 430 unique SNP vectors available in our application were used; thus, $q = 1232$. This setup allows us to include the complex SNP relationship that naturally exists between rice varieties that would be difficult to otherwise simulate. Yields were generated from the model

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{G}_i'\boldsymbol{\gamma} + \epsilon_i,$$

where $\mathbf{X}_i = (1, F_{i1}, F_{i2}, \mathbf{S}_i')'$, F_{ij}, for $j = 1, 2$, is a field indicator, $\epsilon_i \sim N(0, \sigma^2)$, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2\mathbf{C})$, $\mathbf{C} = q^{-1}\mathbf{S}_u\mathbf{S}_u'$, $\mathbf{S}_u$ is a $k \times q$ dimensional matrix whose rows contain the 430 unique SNP vectors, and $\mathbf{G}_i$ is a k-dimensional binary variety identification vector for the ith observation. The study considers $\sigma \in \{0.5, 1.0, 2.0\}$.

To generate yields, we posit nine covariates as non-zero. In particular, the intercept and the two field effects were taken as $\beta_0 = 3.00$, $\beta_1 = 3.50$, and $\beta_2 = 1.00$. The six significant SNPs were selected at random, without replacement, from the set of common SNPs with minor allele frequency greater than 0.1, and the corresponding effects, after reordering the SNP values for notational convenience, were set to $\beta_3 = 0.25$, $\beta_4 = 0.50$, $\beta_5 = 0.75$, $\beta_6 = 1.00$, $\beta_7 = 1.50$, and $\beta_8 = 2.00$. All other regression coefficients are zero. The process of randomly selecting SNP values was repeated three times; for each replication, 500 independent data sets were constructed for each σ. Overall, the generating process produced 4500 independent data sets. Our algorithm used $T_0 = 1000$ and the tuning parameters $\tau_{1\alpha} = \tau_{1\eta} = 0.001$ and $\tau_{2\alpha} = \tau_{2\eta} = 5$. Other choices for the tuning parameters were investigated (results not shown) and produced no appreciable differences from those reported below.

Our results are compared to a standard marginal analysis, which is a staple in genome association studies. In particular, q models of the form

$$Y_i = \beta_0 + \beta_1 F_{i1} + \beta_2 F_{i2} + \beta_3 S_{il} + \epsilon_i, \tag{4.4}$$

were fit to each data set and the estimate of β_3 along with its p-value was calculated. To assess the importance of including variety specific random effects (i.e., $\boldsymbol{\gamma}$) in (4.1), the GDP regression model

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i,$$

was also fitted to each data set using the same parameter configurations as above. In order to provide a comparison between the proposed approach and existing methods, we also analyzed each data set using the methodology outlined in Armagan et al. [2013b], hereafter referred to as ADL. In this implementation we utilized the suggested default choice for the hyperparameters; i.e., we set $\alpha = \eta = 1$.

Table 6.7 summarizes simulation results obtained from the GGDP, GDP, and ADL regression models for the first set of randomly selected SNPs for all considered σ. This summary includes the empirical bias, empirical mean-squared error, and sample standard deviation of the parameter estimates that were estimated as non-zero, as well as the empirical percentage of runs where a regression coefficient was identified as being non-zero From these results, all three methods seem to perform well across most of the simulation configurations. In particular, for all considered σ, the three techniques identified the significant regression coefficients nearly 100% of the time, with accuracy decreasing with larger σ and smaller effect sizes. Moreover, the estimators obtained from these techniques exhibit little evidence of bias in most configurations. It is worthwhile to point out that of the three approaches the GGDP model in general provided the smallest mean-squared errors.

Table 6.7 also provides the empirical false discovery rate (the number of insignificant covariates identified as being significant divided by the total number of insignificant variables) in all simulation configurations. While neither the GGDP or GDP methods perform poorly, some distinctions are apparent. In particular, the GGDP regression model, which makes use of the genetic similarity matrix, actually reduces the number of false discoveries, on average, by more than a factor of four. To clarify, in this study, a false discovery rate of 0.3% was obtained by the proposed approach, while a false discovery rate of 1.35% was obtained for the GDP regression model. Hence, the GGDP regression model, when compared to its counterpart that ignores genetic similarities, helps reduce false discoveries. In contrast, the false discovery rate for ADL was 26%, which was far worse than the other two procedures. Table 6.7 also provides the average number of iterations and computational time required to fit the three models. The time trials were run on a Dell Optiplex 790, with a 2.9GHz Intel Core i7-2600 CPU. From these results one can see that the proposed approach is far more computationally efficient than the ADL method; i.e., the GGDP and GDP methods complete in far fewer iterations and in a shorter time frame when compared to the ADL method.

[Table 7 about here.]

A few concluding remarks follow. From additional studies (results not shown), it was ascertained that the proposed EM algorithm, for both the GGDP and GDP models, is robust to initialization; i.e., in these studies multiple initial values were specified resulting in the same point of convergence. Results from the other two sets of randomly selected SNPs were almost identical to those in Table 6.7 and are therefore omitted. Marginal analyses again yielded higher false discovery rates (not shown here). Section 5 provides a more detailed discussion on the appropriateness and pitfalls of marginal analysis in these settings.

To complement the studies described here, an additional simulation study was conducted to examine the performance of the proposed methodology in higher-dimensional settings. In particular, this study considered values of $q \in \{10^4, 10^5\}$. Briefly, the findings from this additional study reinforces all of the findings discussed above. That is, these studies tend to indicate that the proposed methodology can be used to efficiently analyze genetics data sets consisting of a large number of SNPs. Moreover, this analysis can be completed in a relatively short period of time; e.g., when $q = 10^4$ and $q = 10^5$ the average model fitting time in this study was approximately 3.5 and 40 minutes respectively.

4.4 Application

The developed methods were used in a genetic association study of rice varieties in Indonesia. The purpose of the study was to investigate genetic diversity and identify SNPs linked to crop properties, with the ultimate goal of improving rice varieties and ensuring food security.

A diverse Indonesian rice germplasm collection of 467 accessions, including 136 local varieties, 162 improved lines, 11 wild species, 34 near-isogenic-lines, 29 released varieties, and 95 newly introduced varieties were used in this study. The land rice accessions were selected to represent the diverse geographic and climatic range of the many Indonesian islands. The other accessions were chosen to build upon several previous studies and related breeding programs.

The rice collection was extensively phenotyped for complex traits, including times to flowering and harvest, panicle number and length, total and productive tiller, plant height, grain numbers and weight, and yield. Our analysis herein focuses on the yield measurements, which were extrapolated to tons per hectare. Phenotyping was conducted in three fields representing different agro-ecosystems across two planting seasons. The three fields were located in Kuningan (rainy season 2010-2011), Subang (rainy season 2011-2012), and Citayam (rainy season 2012-2013). Regrettably, the available environmental variables (e.g., rainfall, temperature, humidity, etc.) purported to influence yield were practically identical at these three sites. As a consequence, this analysis only considers a field effect to account for the unmeasured confounders at the three sites.

The rice genome is approximately 389 megabases and consists of 12 chromosomes. Genotyping was performed on the 467 accessions using a custom Illumina high-throughput genotyping array (GoldenGate assay) [Pardamean et al., 2018]. The 1536 markers measured by this array were selected from several bioinformatics resources, including the Rice-SNP-Seek Database [Alexandrov et al., 2014], an existing rice genotyping array [Zhao et al., 2010], and the rice diversity project (`www.ricediversity.org`).

Genotypes were called using Illumina's GenomeStudio software. SNPs and samples were excluded when missing rates exceeded 25%. For the remaining 430 samples, dosages of the reference allele were imputed using BIMBAM [Servin and Stephens, 2007] for missing genotypes. Monomorphic SNPs were excluded, leaving 1232 SNPs. The correlation among these remaining SNPs vary in strength and are shown in Figure 6.1. Overall, 697 yield measurements were available for joint analysis. The genetic relatedness matrix $\mathbf{C}$ needed in the GGDP is graphically depicted in Figure 6.2 and was computed as described in Section 3. The GGDP and GDP models were both fit to the data with $T_0 = 1000$, $\tau_{1\alpha} = \tau_{1\eta} = 0.001$, and $\tau_{2\alpha} = \tau_{2\eta} = 5$. Other tuning parameter choices were considered but did not produce appreciable differences. The EM algorithm described in Section 2 was run on a Dell Optiplex 790, with a Intel Core i7-2600 CPU 2.9GHz, and completed model fitting in approximately twenty seconds for both the GGDP and GDP regression models. Standard techniques were employed to assess model adequacy, with no major violations being observed; e.g., normal quantile plots indicate that the residuals from both models are near to normally distributed (Figure 6.3).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The GGDP model identified that the two field variables and seven of the SNPs jointly influence rice yield. Not surprisingly, the field effects were strong ($\beta_1 = 3.30$, $\beta_2 = 3.59$), suggesting that yield is highly dependent on field factor conditions (see Table 6.8). The SNP effects were modest compared to the field effects. Five SNPs were associated with lower yields and two were judged to induce higher yields (Table 6.8). In contrast, the GDP model suggests that eight SNPs and all field variables jointly influence rice yield. There was much concordance between the estimates obtained from the two models. For differences, the GGDP model identified S_{941} as influencing yield but the GDP model did not, while the GDP model identified S_{664} and S_{1118} as influencing yield but the GGDP model did not. Table 6.8 also provides a 5-fold cross validation statistic for each model, as a means to evaluate their predictive performance. From this measure it appears that the GGDP model performs slightly better than the GDP model. Based on this finding and per the discussion in Section 3, we are more confident in the GGDP model's conclusions. It is worthwhile to note that the effect sizes for SNPs in both models sometimes differed, suggesting that genetic relatedness can indeed confound yield.

[Table 8 about here.]

These findings were annotated by the Rice Annotation Project [Sakai et al., 2013] and UniProt [Consortium, 2014]. One of the SNPs, S_{64}, is within a gene that encodes for a mitochondrial processing peptidase (MEROP M41 family) that has been associated with cellular activities pertinent to rice growth and development [Huang et al., 2013, Teixeira and Glaser, 2013]. The protein product of another SNP, S_{768}, is a mitogen-activated kinase, whose pathway plays a role in rice plant disease resistance and pathogenic defense [Sheikh et al., 2013, Yang et al., 2015]. While not directly related to rice, the S_{941} SNP was found in a gene that encodes a protein product thought to be related to the salt tolerance protein 3 in sugar beets [Trivedi et al., 2012]. The S_{1014} SNP was found within a gene encoding for a pentatricopeptide repeat protein, which is a part of a family of proteins with a wide range of roles from selection diversification [Geddy and Brown, 2007] to stress and developmental response [Sharma and Pandey, 2016] in a variety of plants, including rice.

4.5 Discussion

The introduced methods improve existing approaches for polygenic modeling of agriculture traits by allowing for important confounding factors and repeated measurements in the model. The proposed approach completes model selection and estimation via a Bayesian MAP estimator under the generalized double Pareto shrinkage prior. From the hierarchical representation of our model, a computationally efficient EM algorithm was developed for identifying the MAP estimator. The proposed methods were evaluated through an extensive simulation study and were used to analyze data collected during a genomic association study conducted by the Indonesian Center for Rice Research.

A standard analysis in genomic association studies is a marginal scan, i.e., the SNPs are analyzed one at a time. As such, a marginal analysis for each of our Section 3 simulated data sets was also conducted based on the model in (4.4). Through this analysis, several key findings arose; first, the regression parameter estimates were often severely biased, and second, the false discovery percentages were egregiously high, even after applying standard multiple testing corrections in an effort to control the family-wise error rate. Further investigations attribute this to the strong correlations between the individual SNPs considered in our application, which are quantified in Figure 6.2. For these reasons, these results were omitted from the manuscript; however, it is worthwhile noting that both the GGDP and GDP approaches were practically immune to the high correlation issues that were so detrimental to the marginal approach. Given the amount of correlation that exists, future work could be aimed at extending the proposed methodology to allow for the penalization of groups of highly correlated variables. This could be accomplished by following the development of the group lasso [Yuan and Lin, 2006] and/or sparse-group lasso [Simon et al., 2013].

To further disseminate this work, code written in R has been developed and is available upon request. This code could benefit plant researchers studying large genomic and crop data sets. While the data analyzed here had limited environmental information, data collection and analysis of rice varieties is ongoing in Indonesia. Future data will include historic and new field factors (e.g., soil, weather, etc.), crop outcomes over seasons and locations, and genomic information on the rice varieties planted. A large database should produce yield prediction models and drive experimental designs to validate them. Ultimately, these models could advise farmers on optimal rice varieties for given or predicted field and climatic conditions.

Chapter 5

A two-phase Bayesian methodology for the analysis of binary phenotypes in genome-wide association studies

5.1 Introduction

In genetics, a genome-wide association study (GWAS) is an observational study of a genome-wide set of genetic markers across individuals with the intent of identifying one or more markers that are associated with a trait of interest. For example, recent GWAS have led to the identification of common genetic variants which are predictive of a subject's predisposition towards colorectal cancer [Peters et al., 2015]. Regretfully, the field of complex disease genetics has been plagued by irreproducibility with respect to marker identification and low predictive fidelity; for further discussion see Zeggini and Ioannidis [2009]. There remains a gap between the estimated genetic component of most complex diseases and the associated genetic variants discovered so far [Manolio et al., 2009]. This "missing heritibility" problem cannot be completely solved by association scans on increasing sample sizes. Methods are needed that acknowledge the inherent complexity of both the genome and these diseases. While new approaches have emerged that attempt to aggregate results based on linkage disequilibrium patterns [Bulik-Sullivan et al., 2015] or that use biological knowledge to focus on relevant regions of the genome [Baurley and Conti, 2013], comprehensive genome-wide analytic approaches are still lacking.

In general, GWAS focuses on measuring and analyzing single-nucleotide polymorphisms (SNPs) across the genome. Historically, researchers have primarily focused on marginal screening methods (i.e., one at a time analyses of the available SNPs) for the purpose of detecting associations, while appropriately adjusting for false discoveries. This approach tends to be conservative and has the propensity to miss important joint behaviour. As a solution, the current research paradigm is shifting to SNP assessment via joint models. This new direction also poses significant challenges; i.e., given the advances in sequencing and genotyping technologies, modern GWAS considers millions of SNPs. From a statistical point of view, this is the classic large p small n problem (i.e., $p >> n$) encountered in high-dimensional regression. In general, high-dimensional regression techniques leverage the bias-variance trade-off by imposing penalties on the regression coefficients. For a continuous outcome, through specifying an L_1 penalty, Tibshirani [1996b] proposed the LASSO which is able to identify a sparse estimator of the regression coefficients, thus completing model fitting and variable selection simultaneously. Following the seminal work of Tibshirani [1996b], many other proposals have been developed under other penalization schemes; e.g., see Fan and Li [2001b], Zou and Hastie [2005b], Zou [2006b], and Candes and Tao [2007]. Extensions of penalized regression methods have been made to generalized linear models; e.g., Wu et al. [2009] and Friedman et al. [2010] incorporated the LASSO and elastic net penalties, respectively, when fitting the logistic regression model. Interestingly, many of these frequentist based techniques have Bayesian analogs which make use of shrinkage priors; e.g., the Bayesian LASSO [Park and Casella, 2008]. In many instances, analytic and computational tractability are aided by the fact that shrinkage priors can be represented as scale mixtures of normals; e.g., see Park and Casella [2008] and Armagan et al. [2013a]. Though theoretically justified in the case of high-dimensional data, the aforementioned techniques are known to struggle and provide inaccurate results when p is large relative to n, which is unarguably the norm in GWAS. To pointedly address this feature, Yazdani and Dunson [2015b] proposed a hybrid Bayesian approach for quantitative traits which combined the marginal scan and joint modeling paradigms.

Motivated by the work of Yazdani and Dunson [2015b] and a recent colorectal cancer study, herein we develop a two-phase Bayesian methodology that can be used to identify significant polygenic effects in genome-wide association studies of binary traits. Like Yazdani and Dunson [2015b], we advocate for the use of a preliminary scan, via Bayes factors, of the available SNPs in an effort to form a reduced set of promising markers. These markers are then analyzed by a joint model along with other confounding variables. The generalized double Pareto shrinkage prior of Armagan et al. [2013a] is specified for the regression coefficients in the joint model and a sparse estimator of these quantities is obtained via a novel maximum a posteriori

(MAP) estimation technique. For finding the MAP estimator, an expectation-maximization (EM) algorithm is derived by introducing carefully constructed latent variables. In particular, through the introduction of these latent variables both the data model and shrinkage prior are decomposed into a convenient hierarchical form. The proposed methodology is thoroughly vetted through an extensive numerical study, and is further illustrated through an analysis of a genome-wide association study of colorectal cancer in Indonesia.

The remainder of this article is organized as follows. Section 2 provides the details of the proposed methodology to include the data augmentation steps and EM algorithm development. Section 3 provides the results of an extensive numerical study conducted to assess the performance of the proposed methodology. Section 4 presents the results of the analysis of the motivating colorectal cancer data. Section 5 concludes with a summary discussion.

5.2 Methodology

In the context of the motivating example, we wish to relate a binary trait (e.g., presence/absence of colorectal cancer) to genetic markers. Let Y_i encode the binary trait for the ith individual, for $i = 1, ..., n$, with the event $Y_i = 1$ denoting that the individual is a case and $Y_i = 0$ otherwise. Similarly, we let E_{iq}, for $q = 1, ..., q_1$, denote the qth confounding variable (e.g., age, BMI, smoking status, etc.) measured on the ith individual. For notational ease, we aggregate these variables as $\mathbf{E}_i = (E_{i1}, \ldots, E_{iq_1})'$. Finally, let S^*_{iq}, for $q = 1, ..., q^*_2$, denote the qth SNP genotype of the ith individual. In order to evaluate both the confounding variables and genetic markers, we propose the following two-phase methodology.

5.2.1 Phase 1

In Phase 1 of our approach, the genetic markers undergo a preliminary scan to identify a promising set of possible significant genotypes, while controlling for confounding variables. More specifically, in this phase, we seek to rank order each of the SNPs via Bayes factors. Briefly, a Bayes factor is a summary of the evidence provided by the data for a model relative to another model. This evidence is computed as

$$B_{q0} = \int_{\boldsymbol{\Theta}_q} p_q(\mathbf{Y} \mid \boldsymbol{\theta}_q)\pi_q(\boldsymbol{\theta}_q)d\boldsymbol{\theta}_q \left\{ \int_{\boldsymbol{\Theta}_0} p_0(\mathbf{Y} \mid \boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0 \right\}^{-1}, \text{ for } q = 1, ..., q^*_2, \tag{5.1}$$

where p_0 and p_q are binary data models (e.g., logisitic or probit regression models) for the observed data $\mathbf{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_q$ denote collections of regression coefficients, and π_0 and π_q are prior distributions. Here, the baseline model (p_0) makes use of a linear predictor consisting of only linear effects in the confounding variables, while p_q considers the same and adds a linear effect associated with S^*_{iq}, for $q = 1, ..., q^*_2$. If B_{q0} is large then there exists strong evidence in favor of p_q when compared to p_0; e.g., $B_{q0} > 20$ and $B_{q0} > 150$ offer strong and very strong evidence, respectively. In addition to comparing various models to the baseline model, one may rank order models without the need to recompute Bayes factors. For example, the event $B_{q'0} > B_{q0}$ suggests that $p_{q'}$ is favorable when compared to p_q, given the available data. In order to avoid prior influence, it is standard to specify non-informative or vague priors which are often improper. It is well known that Bayes factors should not be computed using improper priors [Wasserman, 2000], and thus we suggest the use of vague independent normal priors for the regression coefficients; i.e., $\boldsymbol{\theta}_0 \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{q_1+1})$ and $\boldsymbol{\theta}_q \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{q_1+2})$, where $\mathbf{I}_q$ denotes a $q \times q$ identity matrix.

The multi-dimensional integrals depicted in the numerator and denominator of (5.1) are analytically intractable and therefore have to be approximated. Many techniques for approximating such integrals have been proposed; e.g., see Raftery [1996]. Herein, we proceed to approximate the necessary integrals through the following Laplacian approximation:

$$\widehat{p}_q(\mathbf{Y}) = p_q(\mathbf{Y} \mid \widetilde{\boldsymbol{\theta}}_q)\pi_q(\widetilde{\boldsymbol{\theta}}_q)|\mathbf{C}|^{1/2}(2\pi)^{\dim(\widetilde{\boldsymbol{\theta}})/2} \approx \int_{\boldsymbol{\Theta}_q} p_q(\mathbf{Y} \mid \theta_q)\pi_q(\boldsymbol{\theta}_q)d\boldsymbol{\theta}_q, \text{ for } q = 0, ..., q^*_2 \quad (5.2)$$

where $\widetilde{\boldsymbol{\theta}}_q$ is the minimizer of $h(\boldsymbol{\theta}_q) = -\log\{p_q(\mathbf{Y} \mid \theta_q)\pi_q(\boldsymbol{\theta}_q)\}$, $\mathbf{C}$ is the inverse of the hessian of $h(\cdot)$ evaluated at $\widetilde{\boldsymbol{\theta}}_q$, and the function dim$(\cdot)$ provides the dimension of the vector argument. Thus, an approximation to B_{q0} can be constructed as $\widehat{B}_{q0} = \widehat{p}_q(\mathbf{Y})/\widehat{p}_0(\mathbf{Y})$. After computing this approximate Bayes factor for each of the genetic markers, Phase 1 of our methodology concludes by rank ordering the SNPs based on $\widehat{B}_{q0}$ and retaining the top q_2 as promising markers. Let the q_2-dimensional vector $\mathbf{S}_i = (S_{i1}, \ldots, S_{iq_2})'$ aggregate the SNP genotypes that were identified as promising markers. In Section 5.3 we discuss a pragmatic approach that can be used to choose the value of q_2.

5.2.2 Phase 2

In this phase, we build a joint model which relates the confounding variables and all SNPs selected in Phase 1 to the binary trait. To this end, we proceed under the following generalized linear model (GLM):

$$g^{-1}\{P(Y_i = 1 \mid \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\} = \beta_0 + \mathbf{E}'_i\boldsymbol{\beta}_1 + \mathbf{S}'_i\boldsymbol{\beta}_2, \quad (5.3)$$

where $g(\cdot)$ is the link function. For the purposes of this work, we allow $g(\cdot)$ to take on two forms (i.e., logisitic and probit) and provide details of implementation under each. The regression coefficients $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1q_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2q_2})'$ are covariate and genetic marker effects, respectively, with β_0 denoting the usual intercept. Throughout, it is assumed that the independent variables (i.e., $\mathbf{E}_i$ and $\mathbf{S}_i$) have been standardized.

To complete the proposed Bayesian GLM and to induce sparsity into the estimation of the effects (i.e., β_{lq}), we impose a vague independent normal prior on β_0 and independent shrinkage priors on the other regression coefficients through the following specifications:

$$\begin{aligned} \beta_0 \mid T_0 &\sim N(0, T_0), \\ \beta_{lq} \mid \alpha, \eta &\sim \text{GDP}(\psi = \eta/\alpha, \alpha), \text{ for } q = 1, \ldots, q_l \text{ and } l = 1, 2, \end{aligned}$$

where GDP(ψ, α) refers to the generalized double Pareto distribution outlined in Armagan et al. [2013a]. Under these prior choices, setting T_0 to be large provides a vague prior on β_0, while the hyper-parameters $\alpha > 0$ and $\eta > 0$ govern the amount of shrinkage which is imparted on the regression coefficients. In particular, the density of the generalized double Pareto distribution becomes more peaked with lighter tails as α is increased, while larger values of η provide for less shrinkage through a flatter density. Armagan et al. [2013a] suggest a default setting of $\alpha = \eta = 1$, leading to a prior density similar to that of a Cauchy distribution. However, given the computationally efficient nature of our approach, one may explore multiple settings for these hyper-parameters and make use of model selection criteria (e.g., AIC, BIC, cross-validation, etc.) to choose the "optimal" configuration.

In order to avoid the computational burden of Markov chain Monte Carlo in high dimensions and to identify a sparse estimator of the regression coefficients, we develop a computationally efficient EM algorithm that can be used to compute the MAP estimator. To develop this algorithm, we introduce two different sets of latent variables which allow us to decompose both the proposed data model and shrinkage priors into a convenient hierarchical representation. In particular, a hierarchical representation of the proposed data model is formed by introducing latent random variables ω_i, for $i = 1, ..., n$. The specific structure of these random variables is inherently tied to the chosen link function, with the distribution of ω_i being normal or Pólya gamma if one proceeds under the probit or logistic link, respectively; for further details see Albert and Chib [1993] and Polson et al. [2013]. Under either specification, the joint density of the observed and latent data is given by

$$\pi(\mathbf{Y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})\right\}\prod_{i=1}^{n}\xi(\omega_i), \tag{5.4}$$

where $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)'$, and $\mathbf{X}_i = (1, \mathbf{E}_i', \mathbf{S}_i')'$. Under the probit link, $\mathbf{h} = (\omega_1, ..., \omega_n)'$, $\boldsymbol{\Omega} = \mathbf{I}$, and $\xi(\omega_i) = I(\omega_i \geq 0, Y_i = 1) + I(\omega_i < 0, Y_i = 0)$, where $I(\cdot)$ denotes the usual indicator function. In contrast, under the logistic link, $\mathbf{h} = (\kappa_1/\omega_1, ..., \kappa_n/\omega_n)', \kappa_i = Y_i - 1/2$, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$, and $\xi(\omega_i) = f(\omega_i \mid 1, 0)\exp\{\kappa_i^2/(2\omega_i)\}$, where $f(\omega_i \mid a, b)$ denotes the Pólya-Gamma density with parameters (a, b); see Polson et al. [2013].

Attention is now turned to constructing a hierarchical representation of the joint prior distribution. As noted by Proposition 1 in Armagan et al. [2013a], the generalized double Pareto shrinkage prior can be represented as a scale mixture of normal distributions. Thus, for the regression coefficients, the following hierarchical representation provides for the same prior specifications as those given above:

$$\begin{aligned}
\boldsymbol{\beta} \mid \mathbf{T} &\sim N(\mathbf{0}, \mathbf{T}),\\
T_{lq} \mid \lambda_{lq} &\sim \text{Exponential}(\lambda_{lq}^2/2), \text{ for } q = 1, \ldots, q_l \text{ and } l = 1, 2,\\
\lambda_{lq} \mid \alpha, \eta &\sim \text{Gamma}(\alpha, \eta), \text{ for } q = 1, \ldots, q_l \text{ and } l = 1, 2,
\end{aligned}$$

where $\mathbf{T} = \text{diag}(T_0, \mathbf{T}_1', \mathbf{T}_2')$ and $\mathbf{T}_l = (T_{l1}, ..., T_{lq_l})'$. Here the rate parametrization of both the exponential and gamma distributions are utilized.

Given these hierarchical representations, our proposed EM algorithm can be derived viewing $\boldsymbol{\omega}$, $\mathbf{T}$, and λ_{lq}, for $q = 1, \ldots, q_l$ and $l = 1, 2$, as missing data. The E-step of our algorithm identifies the function $Q(\cdot, \cdot)$ as the conditional expectation of the natural logarithm of the posterior distribution, given the observed data (denoted as $\mathcal{D}$) and the current set of parameter estimates (denoted as $\boldsymbol{\beta}^{(d)}$). This yields

$$\begin{aligned}
Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}) = &-\frac{1}{2}E\{(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta}) \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}\}\\
&-\frac{1}{2}\beta_0^2 T_0^{-1} - \frac{1}{2}\sum_{l=1}^{2}\sum_{q=1}^{q_l}\beta_{lq}^2 E(T_{lq}^{-1} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}) + Q_r(\boldsymbol{\beta}^{(d)}),
\end{aligned} \tag{5.5}$$

where $Q_r(\boldsymbol{\beta}^{(d)})$ is a function which is free of $\boldsymbol{\beta}$. The M-step of the algorithm then updates the set of unknown parameters as the maximizer of $Q(\cdot, \cdot)$. Given the form of (5.5), the maximizer obtained in the M-step of the algorithm is given by

$$\boldsymbol{\beta}^{(d+1)} = (\mathbf{X}'\boldsymbol{\Omega}^{(d)}\mathbf{X} + \mathbf{D}^{(d)})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{(d)}\mathbf{h}^{(d)} = \text{argmax}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}), \tag{5.6}$$

where $\mathbf{D}^{(d)} = E(\mathbf{T}^{-1} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)})$ and $E(T_{lq}^{-1} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}) = (\alpha + 1)/\{|\beta_{lq}^{(d)}|(|\beta_{lq}^{(d)}| + \eta)\}$. The form of $\boldsymbol{\Omega}^{(d)}$ and $\mathbf{h}^{(d)}$ in (5.6) are link function dependent. In particular, under the probit link $\boldsymbol{\Omega}^{(d)} = \mathbf{I}$ and $\mathbf{h}^{(d)} = E(\boldsymbol{\omega} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)})$, where

$$\begin{aligned} E(\omega_i \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}) = \mathbf{X}_i'\boldsymbol{\beta}^{(d)} &+ Y_i\phi(\mathbf{X}_i'\boldsymbol{\beta}^{(d)})\{\Phi(\mathbf{X}_i'\boldsymbol{\beta}^{(d)})\}^{-1} \\ &- (1 - Y_i)\phi(\mathbf{X}_i'\boldsymbol{\beta}^{(d)})\{1 - \Phi(\mathbf{X}_i'\boldsymbol{\beta}^{(d)})\}^{-1}, \end{aligned}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ denoting the density and cumulative distribution functions of the standard normal distribution, respectively. Under the logistic link $\boldsymbol{\Omega}^{(d)} = E(\boldsymbol{\Omega} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)})$ and $\mathbf{h}^{(d)} = (\boldsymbol{\Omega}^{(d)})^{-1}\boldsymbol{\kappa}$, where $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_n)'$ and

$$E(\omega_i \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}) = \{P(Y_i = 1 \mid \boldsymbol{\beta}^{(d)}) - 0.5\}(\mathbf{X}_i'\boldsymbol{\beta}^{(d)})^{-1}.$$

Thus, the proposed EM algorithm continues to update $\boldsymbol{\beta}^{(d)}$ via these two steps until convergence is attained; see Abbi et al. [2008] for a discussion on convergence criterion. At the point of convergence, the final update of $\boldsymbol{\beta}^{(d)}$ is our sparse MAP estimator. For computational reasons, it is important to note that due to the carefully constructed hierarchical representations provided above, we are able to identify closed form expressions for all of the necessary expectations in (5.5) as well as to compute closed form updates of the regression coefficients in the M-step given in (5.6).

From a computational perspective, the proposed approach has a few key attributes which are worth outlining. First, due to the nature of the penalty arising from the GDP prior, once a regression coefficient is dropped from the model (i.e., is set to zero), it cannot return. This fact can be exploited to reduce the number of computational steps required to compute $\boldsymbol{\beta}^{(d)}$, thus alleviating a computational bottle neck. Second, in scenarios where $p >> n$, with $p = 1 + q_1 + q_2$, which are common among GWAS, the computationally expensive aspect of the proposed EM algorithm involves the inversion of a $p \times p$ dense matrix in order to compute $\boldsymbol{\beta}^{(d)}$. This computational burden can be avoided by exploiting the Sherman-Morrison-Woodbury formula, which allows one to effectively compute the inversion of the $p \times p$ matrix at the same computational expense as inverting a $n \times n$ matrix. Specifically, we may compute the necessary inversion in (5.6) as

$$(\mathbf{X}'\mathbf{\Omega}^{(d)}\mathbf{X} + \mathbf{D}^{(d)})^{-1} = \mathbf{D}^{(d)^{-1}} - \mathbf{D}^{(d)^{-1}}\mathbf{X}'(\mathbf{\Omega}^{(d)^{-1}} + \mathbf{X}\mathbf{D}^{(d)^{-1}}\mathbf{X}')^{-1}\mathbf{X}\mathbf{D}^{(d)^{-1}},$$

where the inversion of $\mathbf{D}^{(d)}$ and $\mathbf{\Omega}^{(d)}$ are trivial since they are diagonal matrices and the other matrix inversion step on the right-hand side involves only an $n \times n$ matrix. Lastly, the proposed EM algorithm can easily, through the point of initialization, accommodate warm starts [Koh et al., 2007] when fitting models for multiple specifications of the hyper-parameters α and η.

5.3 Numerical studies

In order to evaluate the finite sample performance of the proposed approach, the following simulation study was conducted. Given that Bayes factors are a common tool and have been well vetted, this study focuses on assessing the performance of the MAP estimator developed in Section 5.2.2. The assessed characteristics include the method's ability to 1) identify significant covariates under various signal strengths, 2) accurately estimate the effect size of significant covariates, 3) classify covariates not related to the response as such, and 4) capably handle the complex data structures that are ubiquitous in GWAS. To accomplish this, datasets were simulated to mimic our motivating application; i.e., we consider simulating data for n individuals, where $n \in \{200, 500\}$. For each individual, we simulate the collection of confounding variables $\mathbf{E}_i = (E_{i1}, E_{i2})$, where E_{i1} and E_{i2} are standardized draws that were sampled independently from a $N(0,1)$ and Bernoulli(0.5) distribution, respectively. For this study, we consider SNP vectors of various lengths for the different sample sizes; specifically, we consider $q_2 \in \{100, 200, 500\}$. Rather than randomly generating these variables, we make use of the SNP data from our motivating example. Proceeding in this fashion allows us to capture the complex SNP relationships that naturally exist and would be hard to simulate. To have adequate representation with respect to minor allele frequency, SNPs were first classified according to their minor allele frequency into one of two categories: low (0.20-0.35) and high (0.35-0.50). Then, at random, the q_2 SNPs used in this study were selected from the two categories, with equal representation being taken from each. Let $\mathbf{S}_i$ denote the vector of selected SNPs for subject i, after standardization. The individuals' statuses were then simulated according to the following model:

$$g^{-1}\{P(Y_i = 1 \mid \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\} = \beta_0 + \mathbf{E}_i'\boldsymbol{\beta}_1 + \mathbf{S}_i'\boldsymbol{\beta}_2,$$

where $\beta_0 = -1$, $\boldsymbol{\beta}_1 = (1,1)'$, $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}'_{q_2-12})'$, $\boldsymbol{\beta}^* = (0.25, 0.25, 0.5, 0.5, 1.0, 1.0)$, $\mathbf{0}_q$ is a q-dimensional vector of zeros, and $g(\cdot)$ is the logistic link. This data generating process was used to create 500 independent data sets.

A few comments on the design of this study are warranted. First, the SNPs S_{i1} through S_{i6} were selected from the low minor allele frequency category and SNPs S_{i7} through S_{i12} were selected from the high frequency category. This allows us to examine the ability of the proposed approach to identify small (0.25), medium (0.50), and large (1.00) effects across these different allelic frequencies. Second, this study focuses on the logistic link. Complementary studies were performed under the probit link and resulted in a practically identical conclusion and are therefore omitted for purposes of brevity.

The proposed methodology was used to analyze each of the generated data sets. In this implementation, a vague prior was placed on the intercept by specifying $T_0 = 1000$ and we considered different values of the penalty parameters; i.e., $\alpha \in \{0.1, 0.2, ..., 1.0\}$ and $\eta \in \{0.1, 0.2, 0.3\}$. These choices were made based on prior experience which showed that η should be set to a small value and that values of $\alpha \in (0.1, 1)$ perform well for binary outcomes. It is important to note that a MAP estimator is computed under each of these hyper-parameter configurations. Thus, to choose the "best" from among them we make use of the Bayesian information criterion [Neath and Cavanaugh, 2012]. The computational expense associated with identifying all of the MAP estimators under the various configuration of (α, η) was minimal and scalable.

Table 6.9 summarizes the MAP estimators that were obtained from analyzing the 500 data sets when $n = 200$. This summary includes the empirical bias and standard deviation of the MAP estimators of the truly nonzero coefficients, as well as the percentage of the time that they were identified to be nonzero; i.e., the percentage of time that they were found to be related to the response. We also summarize the false discovery proportion which we define to be the proportion of coefficients which are truly zero but are identified to be nonzero by the MAP estimator. Table 6.10 provides an analogous summary when $n = 500$. From these results, one can see that the proposed approach can be used to reliably identify important explanatory variables as well as estimate their effects. In general, the observed bias is small and is on the same scale as the bias resulting from the oracle model (results not shown); i.e., the model which is provided the correct set of covariates. Moreover, the bias tends to fade as the sample size increases and more importantly does not tend to grow rapidly in the number of considered variables; i.e., in q_2. With respect to selection accuracy, for smaller sample sizes (e.g., $n = 200$) the proposed approach can aptly and reliably detect moderate and strong signals, across different allelic frequencies and values of q_2. The ability to detect smaller signals improves, as one would imagine, when a larger sample size is available. Further, the small false discovery proportions convey that the proposed approach is capable of identifying unrelated coefficients as being such. Finally, Tables 6.9 and 6.10 also report the average time required to compute the MAP estimator that minimizes BIC over the considered (α, η) combinations. From these results, one can see that the proposed approach is both computationally efficient and scalable. In summary, this study has demonstrated the strengths of the proposed MAP estimator with regard to identifying coefficients that are truly related to a binary response. These results also serve to indicate that Phase 1 of our methodology should be used to create a set of candidate SNPs which are on the same order as the available sample size.

[Table 9 about here.]

[Table 10 about here.]

5.4 Colorectal cancer data

Colorectal cancer is one of the most common forms of cancer and is a leading cause of cancer related deaths [Jemal et al., 2011]. Genetic association studies have previously identified markers associated with colorectal cancer risk, but have predominantly focused on subjects from European ancestory. Given the potential differences between South East Asia and European ancestry, a recent study conducted in South Sulawesi, Indonesia was aimed at investigating the genetic and environmental risk factors of colorectal cancer within this South East Asian population. To aid in the discovery of genetic and environmental risk factors, the analysis presented herein focuses on data arising from this seminal study.

The data available for this analysis consists of 173 observations which were taken on 84 cases and 89 controls. These participants were recruited from throughout Makassar, Indonesia between the years of 2014 and 2016. Environmental risk factor information was collected via voluntary questionnaires and medical records. This information includes, but is not limited to, demographics, family history, smoking behavior, alcohol use, and dietary history. To collect genetic information, each participant provided a blood sample for genotyping. DNA was extracted from these samples at Mochtar Riady Institute for Nanotechnology Laboratory in Tangerang, Indonesia. After extraction, the DNA was sent to RUCDR Infinite Biologics for genotyping (Piscataway, NJ, USA). Genotyping was completed using the Smokescreen Genotyping Array (BioRealm LLC). Analysis of the raw data was performed using Affymetrix Power tools (APT) v-1.16 according to the Affymetrix best practices workflow. Additional quality control steps were performed using SNPolisher to identify and select best performing probe sets and high quality SNPs for analysis. After QC filtering, 495,532 SNPs remained for analysis.

To reduce the number of candidate SNPs, Phase 1 of our methodology was used to conduct a preliminary scan of the SNP data, while accounting for environmental risk factors. In this analysis, we control for gender (1=male, 0=female), age (in years), body mass index (BMI), and smoking status (1=Yes, 0=No). In the specification of the Bayes factors, the prior variance (i.e., σ^2) was set to be 100 to provide a vague, yet proper, prior on the regression coefficients. Figure 6.4 provides a histogram of the Bayes factors associated with the 495,532 SNPs and Figure 6.5 provides a plot of the same across chromosomes. From this initial phase, and the results obtained in Section 5.3, we decided to focus attention on the top 200 SNPs; i.e., the SNPs with largest associated Bayes factors. This set of candidate SNPs are denoted as triangles in Figure 6.5. In Phase 2, we fit the following first order model to the data:

$$\text{logit}\{P(Y_i = 1 \mid \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\} = \beta_0 + \mathbf{E}_i'\boldsymbol{\beta}_1 + \mathbf{S}_i'\boldsymbol{\beta}_2,$$

where $\mathbf{E}_i$ is the vector of environmental risk factors, and $\mathbf{S}_i$ is the vector of top SNPs identified in Phase 1 for the ith participant. Note that all variables in $\mathbf{E}_i$ and $\mathbf{S}_i$ were standardized. Here, $\mathbf{E}_i = (E_{i1}, ..., E_{i4})'$, where E_{i1} denotes standardized gender, E_{i2} denotes standardized age, E_{i3} denotes standardized BMI, and E_{i4} denotes standardized smoking status. The proposed EM algorithm was used to fit this model and identify the hyper-parameter dependent MAP estimator for each considered configuration of (α, η), where $\alpha \in \{0.1, 0.2, ..., 1.0\}$ and $\eta \in \{0.1, 0.2, 0.3\}$ with $T_0 = 1000$. Final model selection, as in Section 5.3, was guided by the Bayesian information criterion.

[Figure 4 about here.]

[Figure 5 about here.]

Table 6.11 presents the results of this analysis. These results include the chromosome number, coordinate, reference allele, minor allele frequency, and estimated effect for all SNPs identified by the proposed MAP estimator to be the related to colorectal cancer. Also included are effect estimates for the considered environmental risk factors. First, the interpretation of the results pertaining to the environmental risk factors should be made cautiously. That is, by design, the study at enrollment frequency matched cases and controls based on age, sex, and ethnicity. Thus, the interpretation of the findings associated with the various environmental risk factors is limited but important to take account of when assessing genetic risk factors. Second, this analysis identified 10 SNPs which appear to have a relatively strong association (i.e., large effect size) with the risk of developing colorectal cancer. Four of these SNPs lie in intergenic regions; four lie in introns of *ARHGEF3*, *PLCG2*, *RGMB*, and *CTC-340A15.2*; one is a deletion in *PIGN*; and one is an insertion in *SHISA9*. *ARHGEF3* has been implicated in promoting nasopharyngeal carcinoma in Asians Liu et al. [2016]. *RGMB* has been shown to promote colorectal cancer growth Shi et al. [2015].

[Table 11 about here.]

5.5 Discussion

Motivated by a recent study aimed at assessing environmental and genetic risk factors associated with colorectal cancer, we have proposed a Bayesian two-phase methodology for the analysis of binary phenotypes in GWAS. Phase 1 of our methodology makes use of a preliminary scan, via Bayes factors, of the available SNPs. The primary goal of this phase is to render a reduced set of promising markers. These markers are then analyzed via a joint model along with other confounding variables in Phase 2. Through utilizing the generalized double Pareto shrinkage prior and constructing a novel EM algorithm, we are able to develop a computationally efficient approach to identifying a sparse MAP estimator. The performance of the proposed methodology has been illustrated thorough an extensive numerical study, and was used to analyze the motivating cancer data. Through this application, 10 SNPs were identified to be associated with colorecetal cancer via the proposed approach. To further disseminate this work, scripts written in R which implement all aspects of these techniques have been developed and are available in the supporting information accompanying this work, while the motivating colorectal cancer data is available either from the corresponding author upon request or from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO).

Given statistical limitations with respect to the classic large p small n problem and recent advances in sequencing and genotyping technologies, it is natural to believe that two-phase methodologies such as the one proposed here will become standard in GWAS. For this reason, future work could be aimed at examining different marginal analysis techniques that could be used to identify a reduced set of promising SNPs. This could be accomplished by using sparse estimation techniques (e.g., LASSO, elastic net, etc.) or through adopting ideas from the recent advances in polygenic risk scores Dudbridge [2013]. Though prescan techniques, such as Phase 1 of the proposed approach, are common [e.g., see Wang et al., 2018], it is important to note that they in fact limit the set of candidate variables that can be considered in the joint model; i.e., once a set of candidate SNPs have been identified additions in Phase 2 are not considered. For this reason, it could be of interest to merge the goals of Phase 1 and 2 into a more flexible formulation that would allow one to consider all available SNPs in the joint model. With that being said, an approach of this nature would likely pose many challenges from both a methodological and a computational perspective.

Chapter 6

Discussion

This dissertation provides a thorough set of Bayesian methodologies for high dimensional and complicated data. Chapters 2 and 3 outline how to analyze complicated group testing data with a mixed effects model while simultaneously achieving full variable selection in the fixed effects and random effects. These models are then used to analyze the motivating data set provided by the State Hygienic Laboratory in Iowa, where Iowa citizens are screened for chlamydia and gonorrhea. Furthermore, Chapter 2 reveals that when the random effects are ignored, the result is lower classification accuracy. Chapter 4 outlines a Bayesian linear mixed effects model that relates single-nucleotide polymorphisms (SNPs) of rice plants to the amount of yield produced. This data was provided by fields in Indonesia in a local breeding effort to produce more rice. Chapter 5 concludes this dissertation by developing a Bayesian logistic regression model to associate human SNPs and covariate information to individual probabilities of having colorectal cancer. In analyzing the colorectal cancer data with this model, 10 SNPs were identified to be significant with relatively large magnitude. This provides grounds for possible exploration from practitioners in an effort to better detect colorectal cancer in patients.

Appendices

Appendix A Supplementary Material for Chapter 2

A.1 Full conditional distributions and posterior sampling

Herein we provide the specific form of the full conditional distributions required to complete our posterior sampling algorithm. We then outline the step-by-step implementation of the posterior sampling algorithm under the three considered spike and slab priors.

Full conditional of λ_l: Define the $q_2 \times 1$ vectors $\mathbf{e}_i, i = 1, ..., N$, with lth entry given by $e_{il} = t_{il} b_{k(i)l} + t_{il} \sum_{m=1}^{l-1} b_{k(i)m} a_{lm}$ so that the linear predictor can be expressed as

$$\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{e}_i'\boldsymbol{\lambda}.$$

Then the full conditional distribution of λ_l is given by

$$\lambda_l \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l \sim TN\{\mu_{\lambda_l}(w_l), \sigma^2_{\lambda_l}(w_l), (0, \infty)\},$$

where the mean and variance are

$$\begin{aligned}\mu_{\lambda_l}(w_l) &= (\mathbf{E}_l'\boldsymbol{\Omega}\mathbf{E}_l + 1/\{r(w_l)\psi_l^2\})^{-1}\mathbf{E}_l'\boldsymbol{\Omega}\mathbf{h}_{\lambda_l} \\ \sigma^2_{\lambda_l}(w_l) &= (\mathbf{E}_l'\boldsymbol{\Omega}\mathbf{E}_l + 1/\{r(w_l)\psi_l^2\})^{-1},\end{aligned}$$

and $\mathbf{E}$ is the $N \times q_2$ matrix with ith row $\mathbf{e}_i$, $\mathbf{E}_l$ is the lth column of $\mathbf{E}$, $\mathbf{E}_{-l}$ is all columns except for the lth column, and $\mathbf{h}_{\lambda_l} = \mathbf{h} - \mathbf{X}\boldsymbol{\beta} - \mathbf{E}_{-l}\boldsymbol{\lambda}_{-l}$.

Full conditional of $\mathbf{a}$: For each individual $i = 1, ..., N$, define the $q_2(q_2-1)/2 \times 1$ vector $\mathbf{u}_i = (b_{k(i)l}\lambda_m t_{im} : l = 1, ..., q_2 - 1; m = l+1, ..., q_2)'$ so that the linear predictor becomes

$$\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{A}\mathbf{b}_{k(i)} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{b}_{k(i)} + \mathbf{u}_i'\mathbf{a}.$$

The full conditional distribution of $\mathbf{a}$ is then given by

$$\mathbf{a} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}),$$

where the mean and covariance matrix is

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{a}} &= (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U} + \mathbf{C}_0^{-1})^{-1}(\mathbf{U}'\boldsymbol{\Omega}\mathbf{h}_{\mathbf{a}} + \mathbf{C}_0^{-1}\mathbf{m}_0)\\ \boldsymbol{\Sigma}_{\mathbf{a}} &= (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U} + \mathbf{C}_0^{-1})^{-1},\end{aligned}$$

and $\mathbf{U}$ is a matrix with rows $\mathbf{u}_i$ and $\mathbf{h}_{\mathbf{a}} = (h_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{t}_i'\boldsymbol{\Lambda}\mathbf{b}_{k(i)} : i = 1, ..., N)'$.

Full conditional of $\mathbf{b}_k$: Note that only individuals tested at site k contribute to the posterior distribution of $\mathbf{b}_k$. Thus define the index set $\mathcal{S}_k = \{i : \mathbf{b}_{k(i)} = \mathbf{b}_k\}$; i.e., the index set $\mathcal{S}_k$ identifies the individuals tested at site k. Define $\mathbf{M}(\mathcal{S})$ to be the matrix formed by retaining the rows of the matrix $\mathbf{M}$ identified by the index set $\mathcal{S}$; with the analogous extension for vectors. Thus, the full conditional distribution of $\mathbf{b}_k$ is given by

$$\mathbf{b}_k \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}),$$

where the mean and covariance matrix is

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{b}_k} &= (\mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}_k'\boldsymbol{\Omega}_k\mathbf{T}_k\boldsymbol{\Lambda}\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}_k'\boldsymbol{\Omega}_k(\mathbf{h}_k - \mathbf{X}_k\boldsymbol{\beta})\\ \boldsymbol{\Sigma}_{\mathbf{b}_k} &= (\mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}_k'\boldsymbol{\Omega}_k\mathbf{T}_k\boldsymbol{\Lambda}\mathbf{A} + \mathbf{I})^{-1},\end{aligned}$$

and $\mathbf{T}_k = \mathbf{T}(\mathcal{S}_k)$, $\mathbf{h}_k = \mathbf{h}(\mathcal{S}_k)$, $\mathbf{X}_k = \mathbf{X}(\mathcal{S}_k)$, and $\boldsymbol{\Omega}_k = \boldsymbol{\Omega}(\mathcal{S}_k)$.

Full conditional of v_q: For SSVS and NMIG, the full conditional distribution for v_q is Bernoulli with success probability p_{v_q}; i.e., $v_q \mid \beta_q, \tau_{v_q} \sim$ Bernoulli(p_{v_q}), where

$$p_{v_q} = \frac{\pi_{\text{slab}}(\beta_q)\tau_{v_q}}{\pi_{\text{spike}}(\beta_q)(1 - \tau_{v_q}) + \pi_{\text{slab}}(\beta_q)\tau_{v_q}}.$$

Under the Dirac spike, we draw $\boldsymbol{v}$ from its marginal posterior, which is obtained after integrating over $\boldsymbol{\beta}$; i.e.,

$$\begin{aligned}\pi(\boldsymbol{v} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) &\propto \pi(\boldsymbol{v}) \int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})\pi(\boldsymbol{\beta} \mid \boldsymbol{v})d\boldsymbol{\beta}\\ &\propto \pi(\boldsymbol{v})\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}),\end{aligned}$$

where

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}) \propto |\boldsymbol{\Phi}_{\boldsymbol{v}}|^{-1/2} |\boldsymbol{\Sigma}_{\boldsymbol{v}}|^{-1/2} \exp\left\{-\frac{1}{2}\Big[\mathbf{h}_{\boldsymbol{\beta}}'\boldsymbol{\Omega}\mathbf{h}_{\boldsymbol{\beta}} - \boldsymbol{\mu}_{\boldsymbol{v}}'\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{\mu}_{\boldsymbol{v}}\Big]\right\},$$

and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{v}} = \boldsymbol{\Sigma}_{\boldsymbol{v}}^{-1}\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}\mathbf{h}_{\boldsymbol{\beta}}$. It is worth noting that if $\boldsymbol{v} = \mathbf{0}$, then this marginalized likelihood reduces to $\exp\left\{-\frac{1}{2}\mathbf{h}_{\boldsymbol{\beta}}'\boldsymbol{\Omega}\mathbf{h}_{\boldsymbol{\beta}}\right\}$. Thus, it is easy to see that the full conditional distribution of v_q, after marginalizing over $\boldsymbol{\beta}$, is Bernoulli, with success probability p_{v_q}; i.e., $v_q \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-q}, \tau_{v_q} \sim$ Bernoulli(p_{v_q}), where

$$p_{v_q} = \frac{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-q}, v_q = 1)\tau_{v_q}}{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-q}, v_q = 0)(1-\tau_{v_q}) + \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-q}, v_q = 1)\tau_{v_q}}.$$

Full conditional of w_l: For SSVS and NMIG, the full conditional distribution for w_l is Bernoulli, with success probability p_{w_l}; i.e., $w_l \mid \lambda_l, \tau_{w_l} \sim$ Bernoulli(p_{w_l}), where

$$p_{w_l} = \frac{\pi_{\text{slab}}(\lambda_l)\tau_{w_l}}{\pi_{\text{spike}}(\lambda_l)(1-\tau_{w_l}) + \pi_{\text{slab}}(\lambda_l)\tau_{w_l}}.$$

Under the Dirac spike, we draw w_l from its marginal posterior, which is obtained after integrating over λ_l; i.e.,

$$\begin{aligned}\pi(w_l \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}) &\propto \pi(w_l)\int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})\pi(\lambda_l \mid w_l)d\lambda_l \\ &\propto \pi(w_l)\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l),\end{aligned}$$

where

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l) \propto 2\psi_l^{-1}\sigma_l\{1 - \Phi(-\mu_l/\sigma_l)\}\exp\left\{-\frac{1}{2}\Big[\mathbf{h}_{\lambda_l}'\boldsymbol{\Omega}\mathbf{h}_{\lambda_l} - \mu_l^2/\sigma_l^2\Big]\right\},$$

and $\sigma_l^2 = (\mathbf{E}_l'\boldsymbol{\Omega}\mathbf{E}_l + 1/\psi_l^2)^{-1}$, $\mu_l = \sigma_l^2\mathbf{E}_l'\mathbf{h}_{\lambda_l}$, and $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. It is worth noting that if $w_l = 0$, then this marginalized likelihood reduces to $\exp\left\{-\frac{1}{2}\mathbf{h}_{\lambda_l}'\boldsymbol{\Omega}\mathbf{h}_{\lambda_l}\right\}$. Thus, it is easy to see that the full conditional distribution of w_l, after marginalizing over λ_l, is Bernoulli with success probability p_{w_l}; i.e., $w_l \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, \tau_{w_l} \sim$ Bernoulli(p_{w_l}), where

$$p_{w_l} = \frac{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l = 1)\tau_{w_l}}{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l = 0)(1-\tau_{w_l}) + \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-l}, \mathbf{a}, \mathbf{b}, w_l = 1)\tau_{w_l}}.$$

Posterior sampling algorithm under SSVS:

1. Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\mathbf{a}^{(0)}$, $\boldsymbol{v}^{(0)}$, $\boldsymbol{w}^{(0)}$, $\widetilde{\mathbf{Y}}^{(0)}$, $\mathbf{b}_k^{(0)}$, $\tau_{v_q}^{(0)}$, and $\tau_{w_l}^{(0)}$ for all $k = 1, ..., K$, $q = 1, ..., q_1$, and $l = 1, ..., q_2$. If estimating assay accuracies, then initialize $\mathbf{S}_e^{(0)}$ and $\mathbf{S}_p^{(0)}$; otherwise set $\mathbf{S}_e^{(t)} = \mathbf{S}_e$ and $\mathbf{S}_p^{(t)} = \mathbf{S}_p$ for all t, where $\mathbf{S}_e$ and $\mathbf{S}_p$ are known. Set $t = 1$.

2. For $i = 1, .., N$, do one of the following:

   a. If using probit link, sample $h_i^{(t)} = \omega_i^{(t)} \sim \begin{cases} TN\{\eta_i, 1, (0, \infty)\}, & \text{if } \widetilde{Y}_i^{(t-1)} = 1 \\ TN\{\eta_i, 1, (-\infty, 0)\}, & \text{if } \widetilde{Y}_i^{(t-1)} = 0 \end{cases}$

   b. If using logistic link, sample $\omega_i^{(t)} \sim \text{PG}(1, \eta_i)$, and set $h_i^{(t)} = (\widetilde{Y}_i^{(t-1)} - 1/2)/\omega_i^{(t)}$

   where η_i is evaluated at $\boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$. Aggregate $\boldsymbol{\omega}^{(t)} = (\omega_1^{(t)}, ..., \omega_N^{(t)})'$ and $\mathbf{h}^{(t)} = (h_1^{(t)}, ..., h_N^{(t)})'$. Set $\boldsymbol{\Omega}^{(t)} = \mathbf{I}$ under probit or $\boldsymbol{\Omega}^{(t)} = \text{diag}(\boldsymbol{\omega}^{(t)})$ under logistic.

3. Sample $\boldsymbol{\beta}^{(t)} \sim N\left\{\left(\mathbf{X}'\boldsymbol{\Omega}^{(t)}\mathbf{X} + \boldsymbol{\Phi}^{-1}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{(t)}\mathbf{h}_{\boldsymbol{\beta}}, \left(\mathbf{X}'\boldsymbol{\Omega}^{(t)}\mathbf{X} + \boldsymbol{\Phi}^{-1}\right)^{-1}\right\}$, where $\mathbf{h}_{\boldsymbol{\beta}}$ is evaluated at $\mathbf{h}^{(t)}$, $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$, and $\boldsymbol{\Phi}$ is evaluated at $\boldsymbol{v}^{(t-1)}$.

4. For $l = 1, ..., q_2$, sample $\lambda_l^{(t)} \sim TN\{\mu_{\lambda_l}(w_l^{(t-1)}), \sigma_{\lambda_l}^2(w_l^{(t-1)}), (0, \infty)\}$, where the mean and variance are evaluated at $\mathbf{a}^{(t-1)}$, $\boldsymbol{\Omega}^{(t)}$, $\mathbf{h}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, $\lambda_1^{(t)}, ..., \lambda_{l-1}^{(t)}, \lambda_{l+1}^{(t-1)}, ..., \lambda_{q_2}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$. Aggregate $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, ..., \lambda_{q_2}^{(t)})'$.

5. Sample $\mathbf{a}^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}})$, where the mean and covariance matrix are evaluated at $\boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\Omega}^{(t)}$, $\mathbf{h}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$.

6. For $k = 1, ..., K$, sample $\mathbf{b}_k^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k})$, where the mean and covariance matrix are evaluated at $\mathbf{a}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\Omega}_k^{(t)}, \mathbf{h}^{(t)}$, and $\boldsymbol{\beta}^{(t)}$.

7. For $q = 1, ..., q_1$, sample $v_q^{(t)} \sim \text{Bernoulli}(p_{v_q})$, where p_{v_q} is evaluated at $\beta_q^{(t)}$ and $\tau_{v_q}^{(t-1)}$.
   Aggregate $\boldsymbol{v}^{(t)} = (v_1^{(t)}, ..., v_{q_1}^{(t)})'$.

8. For $l = 1, ..., q_2$, sample $w_l^{(t)} \sim \text{Bernoulli}(p_{w_l})$, where p_{w_l} is evaluated at $\lambda_l^{(t)}$ and $\tau_{w_l}^{(t-1)}$.
   Aggregate $\boldsymbol{w}^{(t)} = (w_1^{(t)}, ..., w_{q_2}^{(t)})'$.

9. For $q = 1, ..., q_1$ and $l = 1, ..., q_2$, sample $\tau_{v_q}^{(t)} \sim \text{Beta}(a_v + v_q^{(t)}, 1 - v_q^{(t)} + b_v)$ and
   $\tau_{w_l}^{(t)} \sim \text{Beta}(a_w + w_l^{(t)}, 1 - w_l^{(t)} + b_w)$.

10. If estimating testing assay accuracies, then sample $S_{e(m)}^{(t)} \sim \text{Beta}(a^\star_{e(m)}, b^\star_{e(m)})$ and $S_{p(m)}^{(t)} \sim \text{Beta}(a^\star_{p(m)}, b^\star_{p(m)})$ for $m = 1, ..., M$, where $a^\star_{e(m)}, b^\star_{e(m)}, a^\star_{p(m)}$, and $b^\star_{p(m)}$ are evaluated at $\widetilde{\mathbf{Y}}^{(t-1)}$. Aggregate $\mathbf{S}_e^{(t)} = (S_{e(1)}^{(t)}, ..., S_{e(M)}^{(t)})'$ and $\mathbf{S}_p^{(t)} = (S_{p(1)}^{(t)}, ..., S_{p(M)}^{(t)})'$.

11. For $i = 1, ..., N$, sample $\widetilde{Y}_i^{(t)} \sim$ Bernoulli$\{p^\star_{i1}/(p^\star_{i0} + p^\star_{i1})\}$, where $p^\star_{i0}$ and $p^\star_{i1}$ are evaluated at $\widetilde{Y}_1^{(t)}, ..., \widetilde{Y}_{i-1}^{(t)}, \widetilde{Y}_{i+1}^{(t-1)}, ..., \widetilde{Y}_N^{(t-1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\lambda}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}_{k(i)}^{(t)}, \mathbf{S}_e^{(t)}$, and $\mathbf{S}_p^{(t)}$. Aggregate $\widetilde{\mathbf{Y}}^{(t)} = (\widetilde{Y}_1^{(t)}, ..., \widetilde{Y}_N^{(t)})'$.

12. Increment t and return to step 2.

Posterior sampling algorithm under NMIG:

Refer to the SSVS sampling algorithm. During step 1, also initialize $\phi_q^{2(0)}$ and $\psi_l^{2(0)}$ for $q = 1, ..., q_1$ and $l = 1, ..., q_2$. Complete steps 2 through 11 using $\phi_q^{2(t-1)}$ and $\psi_l^{2(t-1)}$ where necessary. Then sample $\phi_q^{2(t)} \sim$ Inv-Gamma$(a_\phi + 1/2, b_\phi + \beta_q^{(t)2}/\{2r(v_q^{(t)})\})$ and $\psi_l^{2(t)} \sim$ Inv-Gamma$(a_\psi + 1/2, b_\psi + \lambda_l^{(t)2}/\{2r(w_l^{(t)})\})$ for all q and all l.

Posterior sampling algorithm under Dirac:

1. Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\mathbf{a}^{(0)}$, $\boldsymbol{v}^{(0)}$, $\boldsymbol{w}^{(0)}$, $\widetilde{\mathbf{Y}}^{(0)}$, $\mathbf{b}_k^{(0)}$, $\tau_{v_q}^{(0)}$, and $\tau_{w_l}^{(0)}$ for all $k = 1, ..., K$, $q = 1, ..., q_1$, and $l = 1, ..., q_2$. If estimating assay accuracies, then initialize $\mathbf{S}_e^{(0)}$ and $\mathbf{S}_p^{(0)}$; otherwise set $\mathbf{S}_e^{(t)} = \mathbf{S}_e$ and $\mathbf{S}_p^{(t)} = \mathbf{S}_p$ for all t, where $\mathbf{S}_e$ and $\mathbf{S}_p$ are known. Set $t = 1$.

2. For $i = 1, .., N$, do one of the following:

   a. If using probit link, sample $h_i^{(t)} = \omega_i^{(t)} \sim \begin{cases} TN\{\eta_i, 1, (0, \infty)\}, & \text{if } \widetilde{Y}_i^{(t-1)} = 1 \\ TN\{\eta_i, 1, (-\infty, 0)\}, & \text{if } \widetilde{Y}_i^{(t-1)} = 0 \end{cases}$

   b. If using logistic link, sample $\omega_i^{(t)} \sim \text{PG}(1, \eta_i)$, and set $h_i^{(t)} = (\widetilde{Y}_i^{(t-1)} - 1/2)/\omega_i^{(t)}$

   where η_i is evaluated at $\boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$. Aggregate $\boldsymbol{\omega}^{(t)} = (\omega_1^{(t)}, ..., \omega_N^{(t)})'$ and $\mathbf{h}^{(t)} = (h_1^{(t)}, ..., h_N^{(t)})'$. Set $\boldsymbol{\Omega}^{(t)} = \mathbf{I}$ under probit or $\boldsymbol{\Omega}^{(t)} = \text{diag}(\boldsymbol{\omega}^{(t)})$ under logistic.

3. Set $\beta_q^{(t)} = 0$ if $v_q^{(t-1)} = 0$. Sample the remaining nonzero β_q's from $\boldsymbol{\beta}_{\boldsymbol{v}}^{(t)} \sim N\left\{\left(\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}^{(t)}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}\right)^{-1}\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}^{(t)}\mathbf{h}_{\boldsymbol{\beta}}, \left(\mathbf{X}_{\boldsymbol{v}}'\boldsymbol{\Omega}^{(t)}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}\right)^{-1}\right\}$, where $\mathbf{h}_{\boldsymbol{\beta}}$ is evaluated at $\mathbf{h}^{(t)}$, $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$.

4. For $l = 1, ..., q_2$, set $\lambda_l^{(t)} = 0$ if $w_l^{(t-1)} = 0$; otherwise sample $\lambda_l^{(t)} \sim TN\{\mu_{\lambda_l}(w_l^{(t-1)}), \sigma^2_{\lambda_l}(w_l^{(t-1)})\}$, where the mean and variance are evaluated at $\mathbf{a}^{(t-1)}$, $\boldsymbol{\Omega}^{(t)}$, $\mathbf{h}^{(t)}, \boldsymbol{\beta}^{(t)}, \lambda_1^{(t)}, ..., \lambda_{l-1}^{(t)}, \lambda_{l+1}^{(t-1)}, ..., \lambda_{q_2}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$. Aggregate $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, ..., \lambda_{q_2}^{(t)})'$.

5. Sample $\mathbf{a}^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}})$, where the mean and covariance matrix are evaluated at $\boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\Omega}^{(t)}$, $\mathbf{h}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, and $\mathbf{b}_{k(i)}^{(t-1)}$ for $i = 1, ..., N$.

6. For $k = 1, ..., K$, sample $\mathbf{b}_k^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k})$, where the mean and covariance matrix are evaluated at $\mathbf{a}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\Omega}_k^{(t)}, \mathbf{h}^{(t)}$, and $\boldsymbol{\beta}^{(t)}$.

7. For $q = 1, ..., q_1$, sample $v_q^{(t)} \sim$ Bernoulli(p_{v_q}), where p_{v_q} is evaluated at $v_1^{(t)}$, ..., $v_{q-1}^{(t)}$, $v_{q+1}^{(t-1)}$, ..., $v_{q_1}^{(t-1)}$, $\boldsymbol{\lambda}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{h}^{(t)}$, $\boldsymbol{\Omega}^{(t)}$, $\tau_{v_q}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t)}$ for $i = 1, ..., N$. Aggregate $\boldsymbol{v}^{(t)} = (v_1^{(t)}, ..., v_{q_1}^{(t)})'$.

8. For $l = 1, ..., q_2$, sample $w_l^{(t)} \sim$ Bernoulli(p_{w_l}), where p_{w_l} is evaluated at $w_1^{(t)}$, ..., $w_{l-1}^{(t)}$, $w_{l+1}^{(t-1)}$, ..., $w_{q_2}^{(t-1)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}_{-l}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{h}^{(t)}$, $\boldsymbol{\Omega}^{(t)}$, $\tau_{w_l}^{(t-1)}$, and $\mathbf{b}_{k(i)}^{(t)}$ for $i = 1, ..., N$. Aggregate $\boldsymbol{w}^{(t)} = (w_1^{(t)}, ..., w_{q_2}^{(t)})'$.

9. For $q = 1, ..., q_1$ and $l = 1, ..., q_2$, sample $\tau_{v_q}^{(t)} \sim$ Beta$(a_v + v_q^{(t)}, 1 - v_q^{(t)} + b_v)$ and
$\tau_{w_l}^{(t)} \sim$ Beta$(a_w + w_l^{(t)}, 1 - w_l^{(t)} + b_w)$.

10. If estimating testing assay accuracies, then sample $S_{e(m)}^{(t)} \sim$ Beta$(a_{e(m)}^{\star}, b_{e(m)}^{\star})$ and $S_{p(m)}^{(t)} \sim$ Beta$(a_{p(m)}^{\star}, b_{p(m)}^{\star})$ for $m = 1, ..., M$, where $a_{e(m)}^{\star}, b_{e(m)}^{\star}, a_{p(m)}^{\star}$, and $b_{p(m)}^{\star}$ are evaluated at $\widetilde{\mathbf{Y}}^{(t-1)}$. Aggregate $\mathbf{S}_e^{(t)} = (S_{e(1)}^{(t)}, ..., S_{e(M)}^{(t)})'$ and $\mathbf{S}_p^{(t)} = (S_{p(1)}^{(t)}, ..., S_{p(M)}^{(t)})'$.

11. For $i = 1, ..., N$, sample $\widetilde{Y}_i^{(t)} \sim$ Bernoulli$\{p_{i1}^{\star}/(p_{i0}^{\star} + p_{i1}^{\star})\}$, where $p_{i0}^{\star}$ and $p_{i1}^{\star}$ are evaluated at $\widetilde{Y}_1^{(t)}, ..., \widetilde{Y}_{i-1}^{(t)}, \widetilde{Y}_{i+1}^{(t-1)}, ..., \widetilde{Y}_N^{(t-1)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{b}_{k(i)}^{(t)}$, $\mathbf{S}_e^{(t)}$, and $\mathbf{S}_p^{(t)}$. Aggregate $\widetilde{\mathbf{Y}}^{(t)} = (\widetilde{Y}_1^{(t)}, ..., \widetilde{Y}_N^{(t)})'$.

12. Increment t and return to step 2.

A.2 Robustness to conditional independence assumption

Equation (2.3) in the corresponding manuscript is developed under the assumption that the testing outcomes are conditionally independent given the true status of the pools. This assumption could be questionable in situations where retesting is performed (e.g., DT and AT). To examine the performance of our approach under violations of this assumption, we simulated group testing data with highly dependent testing outcomes. That is, we first generate $N = 5000$ true individual statuses $\widetilde{Y}_i$ as in Section 5 of the manuscript. To generate the testing outcomes Z_j, we then generate an additional variable C_i for each individual based on their true status; specifically,

$$C_i = C_i^{+}\widetilde{Y}_i + C_i^{-}(1 - \widetilde{Y}_i),$$

where $C_i^{+} \sim N(1, 0.2)$ and $C_i^{-} \sim N(0.1, 0.02)$. These two normal distributions can be regarded as underlying assay biomarker distributions for truly positive and truly negative individuals, respectively. The testing outcomes were then defined by $Z_j = I(|\mathcal{P}_j|^{-1}\sum_{i\in\mathcal{P}_j} C_i > t_0)$, where $|\mathcal{P}_j|^{-1}\sum_{i\in\mathcal{P}_j} C_i$ denotes the average biomarker concentration of the jth pool and $t_0 = 0.2$ is a diagnostic threshold. We used the same threshold value for both DT and AT. Proceeding in this manner leads to a clear violation of the conditional independence assumption. As in the studies outlined in Section 5, two sets of unknown assay accuracies were considered: $S_{e(1)}$ and $S_{p(1)}$ for master pools and $S_{e(2)}$ and $S_{p(2)}$ for individuals. This process was used to simulate 1000 data sets, each of which was analyzed using the proposed approach in the exact same manner as was outlined in Section 5 of the corresponding manuscript. Given its performance, these studies solely focused on the performance of the proposed approach under the Dirac spike.

Table 6.19 summarizes the results of the robustness study. When comparing Table 6.19 to Table 6.2 of the manuscript, one will note that our estimation methods are not unduly impacted even when faced with extreme violations of the conditional independence assumption. That is, the bias of the estimated coefficients is still small on average, the values of SSD remain on par with that of Table 6.2, and variable selection is practically identical between studies.

A.3 R code for posterior sampling

Our group testing research website `www.chrisbilder.com/grouptesting` contains R programs that implement the methods in this paper. One can reproduce, up to MCMC error, our simulation results in Section 5 for SSVS, NMIG, and the Dirac spike. The provided files include the source files (`BayesLogit_0.6.tar`) to install the `Bayeslogit` package, six C++ routines that are intergal parts of the model fitting process, and three user friendly main functions that implement the proposed methods, one each for the three considered spike and slab configurations. The inputs for each main function are identical, and the inputs and outputs are described below.

Inputs: The following inputs are supplied to this function:

`G`: Matrix of testing responses whose jth row is of the form $(Z_j, c_j, m, \mathcal{P}_j, \mathbf{O})$, where Z_j is the observed testing response for the jth pool, c_j is the number of individuals in the jth pool, m refers to which assay is used to test the jth pool, and $\mathcal{P}_j$ is a vector of indices that identifies which individuals are within the jth pool. Note that $\mathbf{O}$ is a vector of unused values required to complete the dimensions of the matrix.

`Y`: Matrix of individual statuses whose ith row is of the form $(\widetilde{Y}_i^{(0)}, |\mathcal{I}_i|, k(i), \mathcal{I}_i, \mathbf{O})$, where $\widetilde{Y}_i^{(0)}$ is the initial value for the ith individual's status (if the group testing algorithm results in a diagnosed status, one could specify $\widetilde{Y}_i^{(0)}$ to be the diagnosed status), $\mathcal{I}_i$ is a vector of indices identifying the pool tests that the ith individual is a part of, $k(i)$ is the clinic that the ith individual visited, $|\mathcal{I}_i|$ is the cardinality of $\mathcal{I}_i$, and $\mathbf{O}$ is a vector of unused values to complete the dimensions of the matrix.

`X`: Matrix of covariates for the fixed effects, defined in Section 2. The ith row contains the covariates for the ith individual.

`Z`: Matrix of covariates for the random effects, defined in Section 2. The ith row contains the covariates for the ith individual.

`Se`: Vector of test sensitivities of the form $(S_{e(1)}, S_{e(2)}, ..., S_{e(M)})$. These are used as the true values if the assay accuracies are known and are used as initial values if assay accuracies are not known and are to be estimated.

`Sp`: Vector of test specificities of the form $(S_{p(1)}, S_{p(2)}, ..., S_{p(M)})$. These are used as the true values if the assay accuracies are known and are used as initial values if assay accuracies are not known and are to be estimated.

`link`: Specifies which link function is used; can be probit link or logistic link, with probit link as default.

`thin`: Specifies the thinning value of the chain; default is 50.

`iters`: Number of MCMC samples.

`est.error`: Logical; default `TRUE` indicates to estimate assay accuracies, `FALSE` otherwise.

`verbose`: Logical; `TRUE` indicates that a figure depicting posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is displayed during the MCMC, default `FALSE` otherwise.

Outputs for Dirac: The output from this function is a list object that contains the following:

`beta`: Matrix of samples of the fixed effects.

`lambda`: Matrix of samples of $\boldsymbol{\lambda}$.

`a`: Matrix of samples of $\mathbf{a}$.

`v`: Matrix of samples of $\mathbf{v}$.

`w`: Matrix of samples of $\mathbf{w}$.

`b`: Array of samples of $\mathbf{b}$.

`se`: Matrix of samples of $\mathbf{S}_e$.

`sp`: Matrix of samples of $\mathbf{S}_p$.

`D`: Array of samples of $\mathbf{D}$.

Outputs for SSVS: The output from this function is a list object that, in addition to Dirac's output, contains the following:

`tau_v`: Matrix of samples of mixing weights $(\tau_{v_1}, \tau_{v_2}, ..., \tau_{v_{q_1}})$.

`tau_w`: Matrix of samples of mixing weights $(\tau_{w_1}, \tau_{w_2}, ..., \tau_{w_{q_2}})$.

Outputs for NMIG: The output from this function is a list object that, in addition to Dirac's and SSVS's output, contains the following:

`phisq`: Matrix of samples of variance components $(\phi_1^2, \phi_2^2, ..., \phi_{q_1}^2)$.

`psisq`: Matrix of samples of variance components $(\psi_1^2, \psi_2^2, ..., \psi_{q_2}^2)$.

[Table 12 about here.]

[Table 13 about here.]

[Table 14 about here.]

[Table 15 about here.]

[Table 16 about here.]

[Table 17 about here.]

[Figure 6 about here.]

[Table 18 about here.]

[Table 19 about here.]

Appendix B Supplementary Material for Chapter 3

B.1 Full conditional distributions and posterior sampling

Herein we provide the specific form of each full conditional distribution when the probit link is utilized. We then outline the step-by-step implementation of the posterior sampling algorithm under the Dirac spike. Throughout, we make use of the following notation:

$$\begin{aligned}
\mathbf{X}_i &= \text{diag}(\mathbf{x}'_{i1}, ..., \mathbf{x}'_{iD}) \\
\mathbf{T}_i &= \text{diag}(\mathbf{t}'_{i1}, ..., \mathbf{t}'_{iD}) \\
\mathbf{V} &= \text{diag}(\mathbf{V}_1, ..., \mathbf{V}_D) \\
\mathbf{A} &= \text{diag}(\mathbf{A}_1, ..., \mathbf{A}_D) \\
\mathbf{b}_{(i)} &= (\mathbf{b}'_{(i)1}, ..., \mathbf{b}'_{(i)D})'.
\end{aligned}$$

Full conditional of $\boldsymbol{\omega}_i$: By inspection of the fully augmented likelihood function

$$\begin{aligned}
\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}) = &\prod_{d=1}^{D} \prod_{j=1}^{J} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p_j:d}^{1-Z_{jd}} (1 - S_{p_j:d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}} \\
&\times \prod_{i=1}^{N} |\mathbf{R}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i)' \mathbf{R}^{-1} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i) \right\} \prod_{i=1}^{N} f(\boldsymbol{\omega}_i),
\end{aligned}$$

where $f(\boldsymbol{\omega}_i) = \prod_{d=1}^{D} I(\omega_{id} \geq 0, \widetilde{Y}_{id} = 1) + I(\omega_{id} < 0, \widetilde{Y}_{id} = 0)$, it can be seen that the full conditional distribution of $\boldsymbol{\omega}_i$ is multivariate truncated normal with mean $\boldsymbol{\omega}_i$ and covariance matrix $\mathbf{R}$ on the feasible region $\mathcal{R}$ given by the hypercube $\mathbb{R}^D$ whose dth dimension is truncated below(above) by 0 if $\widetilde{Y}_{id} = 1(0)$; i.e.,

$$\boldsymbol{\omega}_i \mid \widetilde{\mathbf{Y}}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)} \sim TN(\boldsymbol{\eta}_i, \mathbf{R}, \mathcal{R}).$$

Full conditional of $\boldsymbol{\beta}$: The full conditional distribution of β_{rd} is degenerate at 0 if $v_{rd} = 0$, while the nonzero elements of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{\boldsymbol{v}}$, have the following normal full conditional distribution

$$\boldsymbol{\beta}_{\boldsymbol{v}} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where the mean and covariance matrix are

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\beta}} &= \left(\boldsymbol{\Phi}(\boldsymbol{v})^{-1} + \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})'\mathbf{R}^{-1}\mathbf{X}_i(\boldsymbol{v})\right)^{-1} \times \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})'\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta}i} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \left(\boldsymbol{\Phi}(\boldsymbol{v})^{-1} + \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})'\mathbf{R}^{-1}\mathbf{X}_i(\boldsymbol{v})\right)^{-1},\end{aligned}$$

and $\boldsymbol{\Phi} = \text{diag}(\phi^2_{rd}; r = 1, ..., p_d, d = 1, ..., D)$, $\boldsymbol{\Phi}(\boldsymbol{v})$ is the matrix of rows and columns of $\boldsymbol{\Phi}$ corresponding to non-zero elements of $\boldsymbol{v}$, $\mathbf{X}_i(\boldsymbol{v})$ is the columns of $\mathbf{X}_i$ corresponding to non-zero elements of $\boldsymbol{v}$, and $\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta}i} = \boldsymbol{\omega}_i - \mathbf{T}_i\mathbf{V}\mathbf{A}\mathbf{b}_{(i)}$.

Full conditional of $\boldsymbol{\lambda}$: Define for individual i a $q_d \times 1$ vector $\mathbf{e}_{id}$ whose lth element is $t_{idl}b_{(i)dl} + t_{idl}\sum_{m=1}^{l-1} b_{(i)dm}a_{dlm}$, where t_{idl} is the lth entry of $\mathbf{t}_{id}$, $b_{(i)dl}$ is the lth entry of $\mathbf{b}_{(i)d}$, and a_{dlm} is the (l, m)th entry of $\mathbf{A}_d$. Construct $\mathbf{E}_i = \text{diag}(\mathbf{e}'_{i1}, ..., \mathbf{e}'_{iD})'$. Then the full conditional distribution of λ_{ld} is degenerate at 0 if $w_{ld} = 0$, while the ℓth nonzero element of $\boldsymbol{\lambda}$, say λ_ℓ, has full conditional distribution

$$\lambda_\ell \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-\ell}, \mathbf{a}, \mathbf{b} \sim TN\{\mu_{\lambda_\ell}, \sigma^2_{\lambda_\ell}, (0, \infty)\},$$

where the mean and variance are given by

$$\begin{aligned}\mu_{\lambda_\ell} &= \left(1/\boldsymbol{\Psi}_{\ell\ell} + \sum_{i=1}^{N} \mathbf{E}_i^{\ell'}\mathbf{R}^{-1}\mathbf{E}_i^{\ell}\right)^{-1} \times \sum_{i=1}^{N} \mathbf{E}_i^{\ell'}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\lambda_\ell i} \\ \sigma^2_{\lambda_\ell} &= \left(1/\boldsymbol{\Psi}_{\ell\ell} + \sum_{i=1}^{N} \mathbf{E}_i^{\ell'}\mathbf{R}^{-1}\mathbf{E}_i^{\ell}\right)^{-1},\end{aligned}$$

and $\mathbf{E}_i^{\ell}$ is the ℓth column of $\mathbf{E}_i$, $\boldsymbol{\Psi}_{\ell\ell}$ is the ℓth diagonal element of $\boldsymbol{\Psi} = \text{diag}(\psi^2_{ld}; l = 1, ..., q_d, d = 1, ..., D)$, $\boldsymbol{\omega}^{\star}_{\lambda_\ell i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{E}_i^{-\ell}\boldsymbol{\lambda}_{-\ell}$, and $\mathbf{E}_i^{-\ell}$ is the matrix $\mathbf{E}_i$ after removing the ℓth column.

Full conditional of a: Define the $q_d \times (q_d - 1)/2$ vector $\mathbf{u}_{id} = (b_{(i)dl}\lambda_{dm}t_{id,m}; l = 1, ..., q_d - 1, m = l + 1, ..., q_d)'$, where $b_{(i)dl}$ is the lth element of $\mathbf{b}_{(i)d}$, λ_{dm} is the mth element of $\boldsymbol{\lambda}_d$, and $t_{id,m}$ is the mth element of $\mathbf{t}_{id}$. Then, the linear predictor η_{id} becomes

$$\mathbf{x}'_{id}\boldsymbol{\beta}_d + \mathbf{t}'_{id}\boldsymbol{\Lambda}_d\mathbf{A}_d\mathbf{b}_{(i)d} = \mathbf{x}'_{id}\boldsymbol{\beta} + \mathbf{t}'_{id}\boldsymbol{\Lambda}_d\mathbf{b}_{(i)d} + \mathbf{u}'_{id}\mathbf{a}_d.$$

Construct $\mathbf{U}_i = \text{diag}(\mathbf{u}'_{i1}, ..., \mathbf{u}'_{iD})'$. Then the full conditional distribution of $\mathbf{a}$ is given by

$$\mathbf{a} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}})$$

where the mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{a}} &= \left(\mathbf{C}^{-1} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \mathbf{U}_i\right)^{-1} \times \left(\mathbf{C}^{-1}\mathbf{m} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \boldsymbol{\omega}^{\star}_{\mathbf{a}i}\right) \\ \boldsymbol{\Sigma}_{\mathbf{a}} &= \left(\mathbf{C}^{-1} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \mathbf{U}_i\right)^{-1}, \end{aligned}$$

and $\boldsymbol{\omega}^{\star}_{\mathbf{a}i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{T}_i\boldsymbol{\Lambda}\mathbf{b}_{(i)}$, $\mathbf{C} = \text{diag}(\mathbf{C}_1, ..., \mathbf{C}_D)$, and $\mathbf{m} = (\mathbf{m}_1, ..., \mathbf{m}_D)'$.

Full conditional of $\mathbf{b}_k$: Define the index set $\mathcal{S}_k = \{i : \mathbf{b}_{(i)} = \mathbf{b}_k\}$; i.e., the index set of individuals who visited site k. Then the full conditional distribution of $\mathbf{b}_k$ is given by

$$\mathbf{b}_k \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}),$$

where the mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{b}_k} &= \left(\mathbf{I} + \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i \mathbf{R}^{-1} \mathbf{T}_i \boldsymbol{\Lambda}\mathbf{A}\right)^{-1} \times \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i \mathbf{R}^{-1} \boldsymbol{\omega}^{\star}_{\mathbf{b}_k i} \\ \boldsymbol{\Sigma}_{\mathbf{b}_k} &= \left(\mathbf{I} + \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i \mathbf{R}^{-1} \mathbf{T}_i \boldsymbol{\Lambda}\mathbf{A}\right)^{-1}, \end{aligned}$$

and $\boldsymbol{\omega}^{\star}_{\mathbf{b}_k i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta}$.

Full conditional of v_{rd}: Under the Dirac spike, $\boldsymbol{v}$ should be sampled from its marginal posterior, which is obtained after integrating over $\boldsymbol{\beta}$; i.e.,

$$\begin{aligned} \pi(\boldsymbol{v} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) &\propto \pi(\boldsymbol{v}) \int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}) \pi(\boldsymbol{\beta} \mid \boldsymbol{v}) d\boldsymbol{\beta} \\ &\propto \pi(\boldsymbol{v}) \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}), \end{aligned}$$

where

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}) \propto |\boldsymbol{\Phi}(\boldsymbol{v})|^{-1/2} |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{1/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N} {\boldsymbol{\omega}^{\star}}'_{\boldsymbol{\beta}i} \mathbf{R}^{-1} \boldsymbol{\omega}^{\star}_{\boldsymbol{\beta}i} - \boldsymbol{\mu}'_{\boldsymbol{\beta}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}} \boldsymbol{\mu}_{\boldsymbol{\beta}}\right]\right\}.$$

Here, $\boldsymbol{\Phi}(\boldsymbol{v})$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, and $\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta}i}$ are defined in the full conditional of $\boldsymbol{\beta}$ outlined above. It is worth noting that if $\boldsymbol{v} = \mathbf{0}$, then this marginalized likelihood reduces to $\exp\left\{-\frac{1}{2}\sum_{i=1}^{N} {\boldsymbol{\omega}^{\star}}'_{\boldsymbol{\beta}i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta}i}\right\}$. Thus, it is easy to see that the full conditional distribution of v_{rd}, after marginalizing over $\boldsymbol{\beta}$, is Bernoulli, with success probability $p_{v_{rd}}$; i.e., $v_{rd} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-rd}, \tau_{v_{rd}} \sim$ Bernoulli($p_{v_{rd}}$), where $\boldsymbol{v}_{-rd}$ is the vector $\boldsymbol{v}$ after removing the rth element of $\boldsymbol{v}_d$ and

$$p_{v_{rd}} = \frac{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-rd}, v_{rd} = 1)\tau_{v_{rd}}}{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-rd}, v_{rd} = 0)(1 - \tau_{v_{rd}}) + \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \boldsymbol{v}_{-rd}, v_{rd} = 1)\tau_{v_{rd}}}.$$

Full conditional of w_{ld}: Assume w_{ld} is the ℓth entry of $\boldsymbol{w}$, call it w_ℓ. Under the Dirac spike, w_ℓ should be sampled from its marginal posterior, which is obtained after integrating over $\lambda_\ell := \lambda_{ld}$; that is, sample from

$$\begin{aligned}\pi(w_\ell \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}) &\propto \pi(w_\ell)\int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b})\pi(\lambda_\ell \mid w_\ell)d\lambda_\ell \\ &\propto \pi(w_\ell)\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, w_\ell),\end{aligned}$$

where $\boldsymbol{\lambda}_{-ld}$ is the vector $\boldsymbol{\lambda}$ with the lth element of $\boldsymbol{\lambda}_d$ removed and

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, w_\ell) \propto \frac{\sigma_{\lambda_\ell}(1 - \Phi(-\mu_{\lambda_\ell}/\sigma_{\lambda_\ell}))}{\psi_{ld}(1 - 1/2)} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N} {\boldsymbol{\omega}^{\star}_{\lambda_\ell i}}'\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\lambda_\ell i} - \mu^2_{\lambda_\ell}/\sigma^2_{\lambda_\ell}\right]\right\}.$$

Here, all notation has been defined in the full conditional distribution of $\boldsymbol{\lambda}$. Note that when $w_{ld} = 0$, then this marginalized likelihood reduces to $\exp\left\{-\frac{1}{2}\sum_{i=1}^{N} {\boldsymbol{\omega}^{\star}}'_{\lambda_\ell i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\lambda_\ell i}\right\}$. Thus, it is easy to see that the full conditional distribution of w_{ld}, after marginalizing over λ_{ld}, is Bernoulli, with probability $p_{w_{ld}}$; i.e., $w_{ld} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}} \sim$ Bernoulli($p_{w_{ld}}$), where

$$p_{w_{ld}} = \frac{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, w_\ell = 1)\tau_{w_{ld}}}{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, w_\ell = 0)(1 - \tau_{w_{ld}}) + \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{-ld}, \mathbf{a}, \mathbf{b}, w_\ell = 1)\tau_{w_{ld}}}.$$

Complete posterior sampling algorithm for multivariate probit link:

1. Initialize $\boldsymbol{\beta}^{(0)}, \boldsymbol{\lambda}^{(0)}, \mathbf{a}^{(0)}, \mathbf{b}^{(0)}_{kd}, \boldsymbol{v}^{(0)}, \boldsymbol{w}^{(0)}, \widetilde{\mathbf{Y}}^{(0)}, \tau^{(0)}_{v_{rd}}$, and $\tau^{(0)}_{w_{ld}}$. If estimating $\mathbf{R}$, initialize $\mathbf{W}^{(0)}, \mathbf{D}^{(0)}$, and $\mathbf{R}^{(0)}$. If estimating testing assay accuracies, initialize $S_{e(m):d}$ and $S_{p(m):d}$. Set $t = 1$.

2. For $i = 1, ..., N$, sample $\boldsymbol{\omega}^{(t)}_i \sim TN(\boldsymbol{\eta}_i, \mathbf{R}^{(t-1)}, \mathcal{R})$, where $\boldsymbol{\eta}_i$ is evaluated at $\boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, and $\mathbf{b}^{(t-1)}_{(i)}$, and $\mathcal{R}$ is evaluated at $\widetilde{\mathbf{Y}}^{(t-1)}_i$. Aggregate $\boldsymbol{\omega}^{(t)} = (\boldsymbol{\omega}^{(t)}_1, ..., \boldsymbol{\omega}^{(t)}_N)'$.

3. Set $\beta^{(t)}_{rd} = 0$ if $v^{(t-1)}_{rd} = 0$. Sample nonzero β's from $\boldsymbol{\beta}^{(t)}_{\boldsymbol{v}} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where the mean and

covariance matrix are evaluated at $\boldsymbol{\lambda}^{(t-1)}$, $\mathbf{a}^{(t-1)}$, $\mathbf{b}^{(t-1)}$, $\boldsymbol{v}^{(t-1)}$, $\mathbf{R}^{(t-1)}$, and $\boldsymbol{\omega}^{(t)}$.

4. Set $\lambda_{ld}^{(t)} = 0$ if $w_{ld}^{(t-1)} = 0$. Sample the ℓth nonzero λ from $\lambda_{\ell}^{(t)} \sim TN\{\mu_{\lambda_\ell}, \sigma^2_{\lambda_\ell}, (0, \infty)\}$, where the mean and variance are evaluated at $\boldsymbol{\beta}^{(t)}$, $\mathbf{a}^{(t-1)}$, $\mathbf{b}^{(t-1)}$, $\mathbf{R}^{(t-1)}$, $\boldsymbol{\omega}^{(t)}$, and $\boldsymbol{\lambda}_{-\ell}^{(t)} = (\lambda_1^{(t)}, ..., \lambda_{\ell-1}^{(t)}, \lambda_{\ell+1}^{(t-1)}, ..., \lambda_L^{(t-1)})'$; i.e., there are a total of L nonzero λ's being sampled. Aggregate $\boldsymbol{\lambda}^{(t)} = (\boldsymbol{\lambda}_1^{(t)}, ..., \boldsymbol{\lambda}_D^{(t)})'$.

5. Sample $\mathbf{a}^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}})$, where the mean and covariance matrix are evaluated at $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}^{(t)}$, $\mathbf{b}^{(t-1)}$, $\mathbf{R}^{(t-1)}$, and $\boldsymbol{\omega}^{(t)}$.

6. For $k = 1, ..., K$, sample $\mathbf{b}_k^{(t)} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k})$, where the mean and covariance matrix are evaluated at $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}^{(t)}$, $\mathbf{a}^{(t)}$, and $\boldsymbol{\omega}^{(t)}$. Aggregate $\mathbf{b}^{(t)} = (\mathbf{b}_1^{(t)}, ..., \mathbf{b}_K^{(t)})'$.

7. For $r = 1, ..., p_d$ and $d = 1, ..., D$, sample $v_{rd}^{(t)} \sim$ Bernoulli$(p_{v_{rd}})$, where $p_{v_{rd}}$ is evaluated at $\boldsymbol{\lambda}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{b}^{(t)}$, $\boldsymbol{v}_{-rd}^{(t)}$, $\mathbf{R}^{(t-1)}$, $\boldsymbol{\omega}^{(t)}$, and $\tau_{v_{rd}}^{(t-1)}$. Here, $\boldsymbol{v}_{-rd}^{(t)}$ uses the tth iteration values of $v_{r_0 d_0}$ if $r_0 < r$ and $d_0 < d$, and $(t-1)$ otherwise. Aggregrate $\boldsymbol{v}^{(t)} = (\boldsymbol{v}_1^{(t)}, ..., \boldsymbol{v}_D^{(t)})'$.

8. For $l = 1, ..., q_d$ and $d = 1, ..., D$, sample $w_{ld}^{(t)} \sim$ Bernoulli$(p_{w_{ld}})$, where $p_{w_{ld}}$ is evaluated at $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}_{-ld}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{b}^{(t)}$, $\mathbf{R}^{(t-1)}$, $\boldsymbol{\omega}^{(t)}$, and $\tau_{w_{ld}}^{(t-1)}$. Here, $\boldsymbol{\lambda}_{-ld}^{(t)}$ uses the tth iteration of $\lambda_{l_0 d_0}$ if $l_0 < l$ and $d_0 < d$, and $(t-1)$ otherwise.

9. For $r = 1, ..., p_d$, $l = 1, ..., q_d$, and $d = 1, ..., D$, sample $\tau_{v_{rd}}^{(t)} \sim$ Beta$(a_v + v_{rd}^{(t)}, 1 - v_{rd}^{(t)} + b_v)$ and $\tau_{w_{ld}}^{(t)} \sim$ Beta$(a_w + w_{ld}^{(t)}, 1 - w_{ld}^{(t)} + b_w)$.

10. If estimating testing assay accuracies, then sample $S_{e(m):d}^{(t)} \sim$ Beta$(a^{\star}_{e(m):d}, b^{\star}_{e(m):d})$ and $S_{p(m):d}^{(t)} \sim$ Beta$(a^{\star}_{p(m):d}, b^{\star}_{p(m):d})$ for $m = 1, ..., M$, where $a^{\star}_{e(m):d}$, $b^{\star}_{e(m):d}$, $a^{\star}_{p(m):d}$, and $b^{\star}_{p(m):d}$ are evaluated at $\widetilde{\mathbf{Y}}^{(t-1)}$. Aggregate $\mathbf{S}_e^{(t)} = (\mathbf{S}_{e1}^{(t)}, ..., \mathbf{S}_{eD}^{(t)})'$ and $\mathbf{S}_p^{(t)} = (\mathbf{S}_{p1}^{(t)}, ..., \mathbf{S}_{pD}^{(t)})'$, where $\mathbf{S}_{ed}^{(t)} = (S_{e(1):d}^{(t)}, ..., S_{e(M):d}^{(t)})'$ and $\mathbf{S}_{pd}^{(t)} = (S_{p(1):d}^{(t)}, ..., S_{p(M):d}^{(t)})'$.

11. For $i = 1, ..., N$ and $d = 1, ..., D$, sample $\widetilde{Y}_{id}^{(t)} \sim$ Bernoulli$\{p^{\star}_{id1}/(p^{\star}_{id0} + p^{\star}_{id1})\}$, where $p^{\star}_{id0}$ and $p^{\star}_{id1}$ are evaluated at $\widetilde{Y}_{i1}^{(t)}, ..., \widetilde{Y}_{i,d-1}^{(t)}, \widetilde{Y}_{i,d+1}^{(t-1)}, ..., \widetilde{Y}_{iD}^{(t-1)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\lambda}^{(t)}$, $\mathbf{a}^{(t)}$, $\mathbf{b}_{(i)}^{(t)}$, $\mathbf{S}_e^{(t)}$, and $\mathbf{S}_p^{(t)}$. Aggregate $\widetilde{\mathbf{Y}}^{(t)} = (\widetilde{\mathbf{Y}}_1^{(t)}, ..., \widetilde{\mathbf{Y}}_N^{(t)})'$.

12. Increment t and return to step 2.

Bibliography

Revlin Abbi, Elia El-Darzi, Christos Vasilakis, and Peter Millard. Analysis of stopping criteria for the em algorithm in the context of patient grouping according to length of stay. In *2008 4th International IEEE Conference Intelligent Systems*, volume 1, pages 3–9. IEEE, 2008.

James Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Nickolai Alexandrov, Shuaishuai Tai, Wensheng Wang, Locedie Mansueto, Kevin Palis, Roven Rommel Fuentes, Victor Jun Ulat, Dmytro Chebotarov, Gengyun Zhang, Zhikang Li, et al. Snp-seek database of snps derived from 3000 rice genomes. *Nucleic acids research*, 43(D1):D1023–D1027, 2014.

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013a.

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013b.

World Bank. Indonesia population growth rate. 2012. URL `http://data.worldbank.org/country/indonesia.html`. Last accessed, September 29, 2016.

James W Baurley and David V Conti. A scalable, knowledge-based analysis framework for genetic association studies. *BMC Bioinformatics*, 14:312, October 2013.

Christopher Bilder and Joshua Tebbs. Bias, efficiency, and agreement for group-testing regression models. *Journal of Statistical Computation and Simulation*, 79:67–80, 2009.

Christopher R Bilder, Joshua M Tebbs, and Christopher S McMahan. Informative group testing for multiplex assays. *Biometrics*, 75(1):278–288, 2019.

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, March 2015.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Peng Chen, Joshua Tebbs, and Christopher Bilder. Group testing regression models with fixed and random effects. *Biometrics*, 65(4):1270–1278, 2009.

Zhen Chen and David Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003.

Siddhartha Chib and Edward Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2014.

A Delaigle, P Hall, and J Wishart. New approaches to non-and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*, 101:567–585, 2014.

Aurore Delaigle and Alexander Meister. Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*, 106(494):640–650, 2011.

Navneet Dhand, Wesley Johnson, and Jenny Toribio. A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(4):452–473, 2010.

Ken G Dodds, John C McEwan, Rudiger Brauning, Rayna M Anderson, Tracey C van Stijn, Theodor Kristjánsson, and Shannon M Clarke. Construction of relatedness matrices using genotyping-by-sequencing data. *BMC genomics*, 16(1):1047, 2015.

Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14(4):436–440, 12 1943. doi: 10.1214/aoms/1177731363. URL `http://dx.doi.org/10.1214/aoms/1177731363`.

Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

People Facts. Population growth. 2012. URL `http://os-connect.com/pop/p2ai.html`. Last accessed, September 29, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001a.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001b.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–20, 2010.

Joseph L Gastwirth and Wesley O Johnson. Screening with cost-effective quality control: potential applications to hiv and drug testing. *Journal of the American Statistical Association*, 89(427):972–981, 1994.

Rachel Geddy and Gregory G Brown. Genes encoding pentatricopeptide repeat (ppr) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC genomics*, 8(1): 130, 2007.

Edward George and Robert McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

Amy Heffernan, Lesa Aylward, Leisa Toms, Peter Sly, Matthew Macleod, and Jochen Mueller. Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology*, 24(3):225–232, 2014.

Shaobai Huang, Rachel N Shingaki-Wells, Nicolas L Taylor, and Harvey Millar. The rice mitochondria proteome and its response during development and to the environment. *Frontiers in plant science*, 4:16, 2013.

Xianzheng Huang. An improved test of latent-variable model misspecification in structural measurement error models for group testing data. *Statistics in Medicine*, 28(26):3316–3327, 2009.

Jacqueline M Hughes-Oliver. Pooling experiments for blood screening and drug discovery. In *Screening*, pages 48–68. Springer, New York, NY, 2006.

Hemant Ishwaran and J Rao. Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, 98(462):438–455, 2003.

Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2):69–90, 2011.

Wesley O Johnson and Joseph L Gastwirth. Dual group screening. *Journal of Statistical Planning and Inference*, 83(2):449–473, 2000.

Joachim Kilian, Florian Peschke, Kenneth W Berendzen, Klaus Harter, and Dierk Wanke. Prerequisites, performance and profits of transcriptional profiling the abiotic stress response. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(2):166–175, 2012.

Hae Kim and Michael Hudgens. Three-dimensional array-based group testing algorithms. *Biometrics*, 65(3): 903–910, 2009.

Hae Kim, Michael Hudgens, Jonathan Dreyfuss, Daniel Westreich, and Christopher Pilcher. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, 63(4):1152–1163, 2007.

S Kleinman, D Strong, G Tegtmeier, P Holland, J Gorlin, C Cousins, R Chiacchierini, and L Pietrelli. Hepatitis B virus (HBV) DNA screening of blood donations in minipools with the COBAS ampliscreen HBV test. *Transfusion*, 45(8):1247–1257, 2005.

Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8(Jul):1519–1555, 2007.

Mel Krajden, Darrel Cook, Annie Mak, Ken Chu, Navdeep Chahil, Malcolm Steinberg, Michael Rekart, and Mark Gilbert. Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *Journal of Clinical Virology*, 61(1):132–137, 2014.

Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

Tian-Hao Liu, Fang Zheng, Mu-Yan Cai, Lin Guo, Huan-Xin Lin, Jie-Wei Chen, Yi-Ji Liao, Hsiang-Fu Kung, Yi-Xin Zeng, and Dan Xie. The putative tumor activator ARHGEF3 promotes nasopharyngeal carcinoma cell pathogenesis by inhibiting cellular apoptosis. *Oncotarget*, 7(18):25836–25848, May 2016.

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

RB Matthews, MJ Kropff, T Horie, and D Bachelet. Simulating the impact of climate change on rice production in asia and evaluating options for adaptation. *Agricultural systems*, 54(3):399–425, 1997.

Christopher McMahan, Joshua Tebbs, Timothy Hanson, and Christopher Bilder. Bayesian regression for group testing data. *Biometrics*, 73(4):1443–1452, 2017.

Christopher S McMahan, Joshua M Tebbs, and Christopher R Bilder. Informative dorfman screening. *Biometrics*, 68(1):287–296, 2012.

Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

Sushil Pandey and Humnath Bhandari. Drought, coping mechanisms and poverty. *IFAD Occa-sional Papers*, 2009.

Bens Pardamean, James W Baurley, Anzaludin S Perbangsa, Dwinita Utami, Habib Rijzaani, and Dani Satyawan. Information technology infrastructure for agriculture genotyping studies. *Journal of Information Processing Systems*, 14(3), 2018.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008.

Ulrike Peters, Stephanie Bien, and Niha Zubair. Genetic architecture of colorectal cancer. *Gut*, 64(10): 1623–1636, 2015.

RM Phatarfod and Aidan Sudbury. The use of a square array scheme in blood testing. *Statistics in Medicine*, 13(22):2337–2343, 1994.

Nicholas Polson, James Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Adrian E Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996.

Hiroaki Sakai, Sung Shin Lee, Tsuyoshi Tanaka, Hisataka Numa, Jungsok Kim, Yoshihiro Kawahara, Hironobu Wakimoto, Ching-chia Yang, Masao Iwamoto, Takashi Abe, et al. Rice annotation project database (rap-db): an integrative and interactive database for rice genomics. *Plant and Cell Physiology*, 54(2): e6–e6, 2013.

B Sarov, L Novack, N Beer, J Safi, H Soliman, JS Pliskin, E Litvak, A Yaari, and E Shinar. Feasibility and cost–benefit of implementing pooled screening for HCVAg in small blood bank settings. *Transfusion Medicine*, 17(6):479–487, 2007.

Fabian Scheipl. spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *arXiv preprint arXiv:1105.5253*, 2011.

Holger Schielzeth and Arild Husby. Challenges and prospects in genome wide qtl mapping of standing genetic variation in natural populations. 2014.

Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114, 2007.

Manisha Sharma and Girdhar K Pandey. Expansion and function of repeat domain proteins during stress and development in plants. *Frontiers in plant science*, 6:1218, 2016.

M. Shean. Indonesia: Stagnating rice production ensures continued need for imports. 2012. URL `http://www.pecad.fas.usda.gov/highlights/2012/03/Indonesia_rice_Mar2012`. Last accessed, September 29, 2016.

Arsheed H Sheikh, Badmi Raghuram, Siddhi K Jalmi, Dhammaprakash P Wankhede, Pallavi Singh, and Alok K Sinha. Interaction between two rice mitogen activated protein kinases and its possible role in plant defense. *BMC plant biology*, 13(1):121, 2013.

Ying Shi, Guo-Bin Chen, Xiao-Xiao Huang, Chuan-Xing Xiao, Huan-Huan Wang, Ye-Sen Li, Jin-Fang Zhang, Shao Li, Yin Xia, Jian-Lin Ren, and Bayasi Guleng. Dragon (repulsive guidance molecule b, RGMb) is a novel gene that promotes colorectal cancer growth. *Oncotarget*, 6(24):20540–20554, August 2015.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Deepti Singh, Michael Tsiang, Bala Rajaratnam, and Noah S Diffenbaugh. Observed changes in extreme wet and dry spells during the south asian summer monsoon season. *Nature Climate Change*, 4(6):456, 2014.

Niko Speybroeck, Christopher Williams, Kora Lafia, Brecht Devleesschauwer, and Dirk Berkvens. Estimating the prevalence of infections in vector populations using pools of samples. *Medical and Veterinary Entomology*, 26(4):361–371, 2012.

Lidan Sun and Rongling Wu. Mapping complex traits as a dynamic system. *Physics of life reviews*, 13: 155–185, 2015.

Joshua Tebbs, Christopher McMahan, and Christopher Bilder. Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics*, 69(4):1064–1073, 2013.

Pedro Filipe Teixeira and Elzbieta Glaser. Processing peptidases in mitochondria and chloroplasts. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(2):360–370, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996a.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996b.

Dipesh Kumar Trivedi, Sandep Yadav, Neha Vaid, and Narendra Tuteja. Genome wide analysis of cyclophilin gene family from rice and arabidopsis and its comparison with yeast. *Plant signaling & behavior*, 7(12): 1653–1666, 2012.

Stijn Vansteelandt, Els Goetghebeur, and T Verstraeten. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, 56(4):1126–1133, 2000.

Helga Wagner and Christine Duller. Bayesian model selection for logistic regression models with random intercept. *Computational Statistics & Data Analysis*, 56(5):1256–1274, 2012.

D Wang, CS McMahan, CM Gallagher, and KB Kulasekera. Semiparametric group testing regression models. *Biometrika*, 101(3):587–598, 2014.

Xulong Wang, Vivek M Philip, Guruprasad Ananda, Charles C White, Ankit Malhotra, Paul J Michalski, Krishna R Murthy Karuturi, Sumana R Chintalapudi, Casey Acklin, Michael Sasner, et al. A bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset alzheimers disease. *Genetics*, 209(1):51–64, 2018.

Larry Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44 (1):92–107, 2000.

Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

Minge Xie. Regression analysis of group testing samples. *Statistics in Medicine*, 20(13):1957–1969, 2001.

Zeyu Yang, Haigang Ma, Hanming Hong, Wen Yao, Weibo Xie, Jinghua Xiao, Xianghua Li, and Shiping Wang. Transcriptome-based analysis of mitogen-activated protein kinase cascades in the rice response to xanthomonas oryzae infection. *Rice*, 8(1):4, 2015.

A Yazdani and David B Dunson. A hybrid bayesian approach for genome-wide association studies on related individuals. *Bioinformatics*, 31(24):3890–3896, 2015a.

A Yazdani and David B Dunson. A hybrid bayesian approach for genome-wide association studies on related individuals. *Bioinformatics*, 31(24):3890–3896, 2015b.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Eleftheria Zeggini and John PA Ioannidis. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201, 2009.

Boan Zhang, Christopher Bilder, and Joshua Tebbs. Group testing regression model estimation when case identification is a goal. *Biometrical Journal*, 55(2):173–189, 2013.

Xiao Zhang, W John Boscardin, and Thomas R Belin. Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896, 2006.

Keyan Zhao, Mark Wright, Jennifer Kimball, Georgia Eizenga, Anna McClung, Michael Kovach, Wricha Tyagi, Md Liakat Ali, Chih-Wei Tung, Andy Reynolds, et al. Genomic diversity and introgression in o. sativa reveal the impact of domestication and breeding on the rice genome. *PloS one*, 5(5):e10780, 2010.

X. Zhou. Gemma user manual. 2016. URL `http://www.xzlab.org/software.html`. Last accessed, September 29, 2016.

Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006a.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476):1418–1429, 2006b.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005a.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005b.

Figure 6.1: Graphical display of the correlation matrix between the 1232 SNPs in the data application. Darker shades indicate stronger correlation among genetic variants.

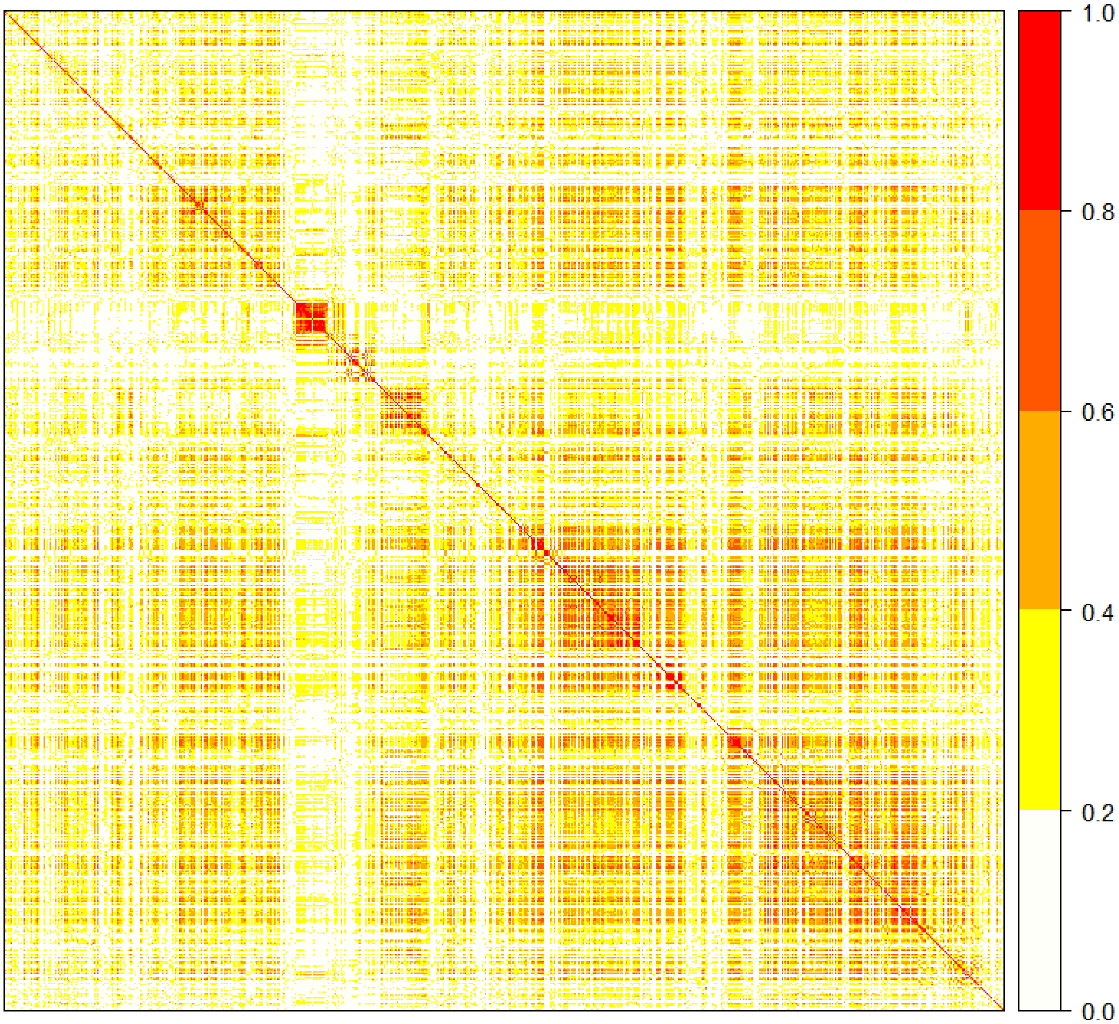


Figure 6.2: Graphical display of the genetic relatedness matrix **C** for the 430 rice varieties. Darker shades indicate varieties that are more genetically similar.

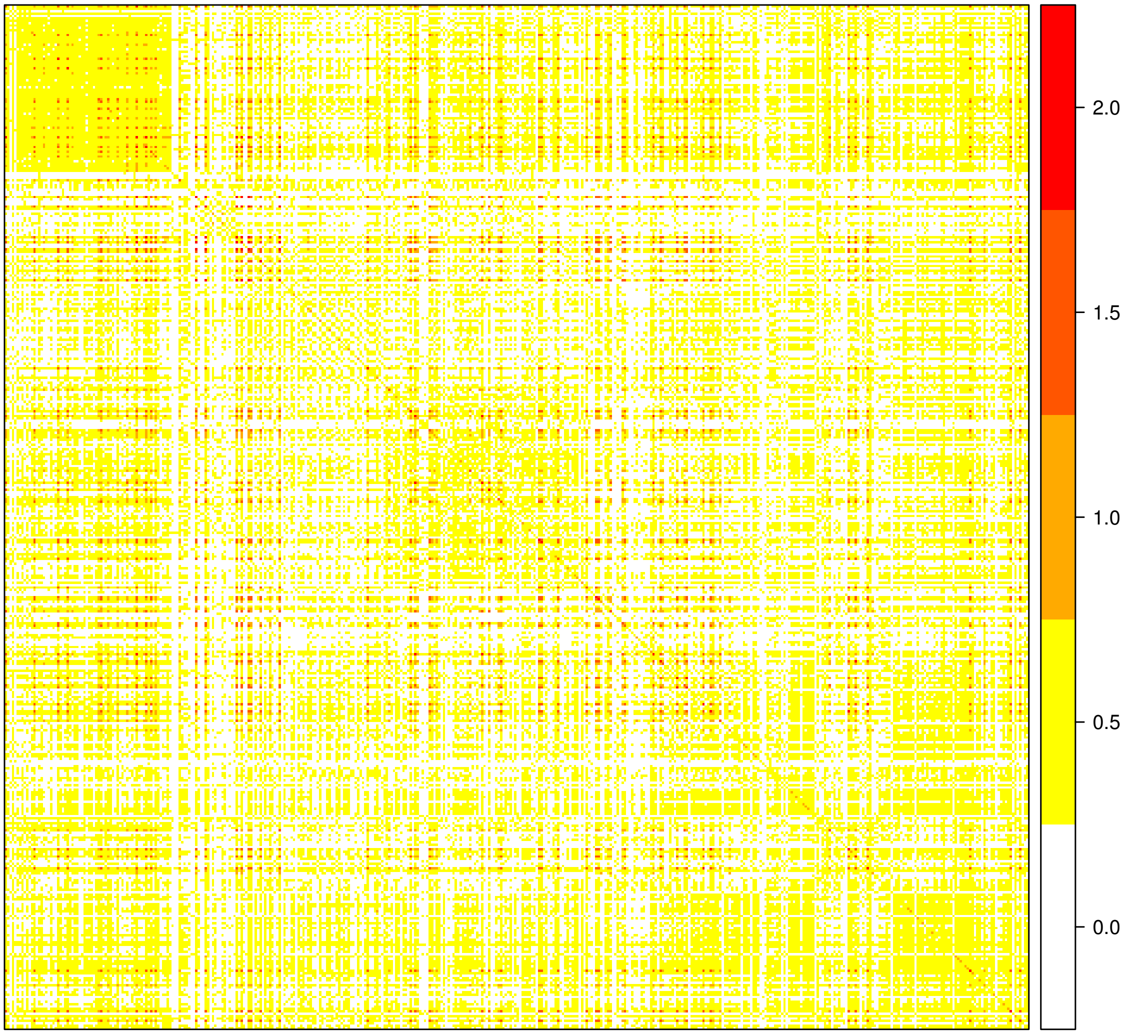
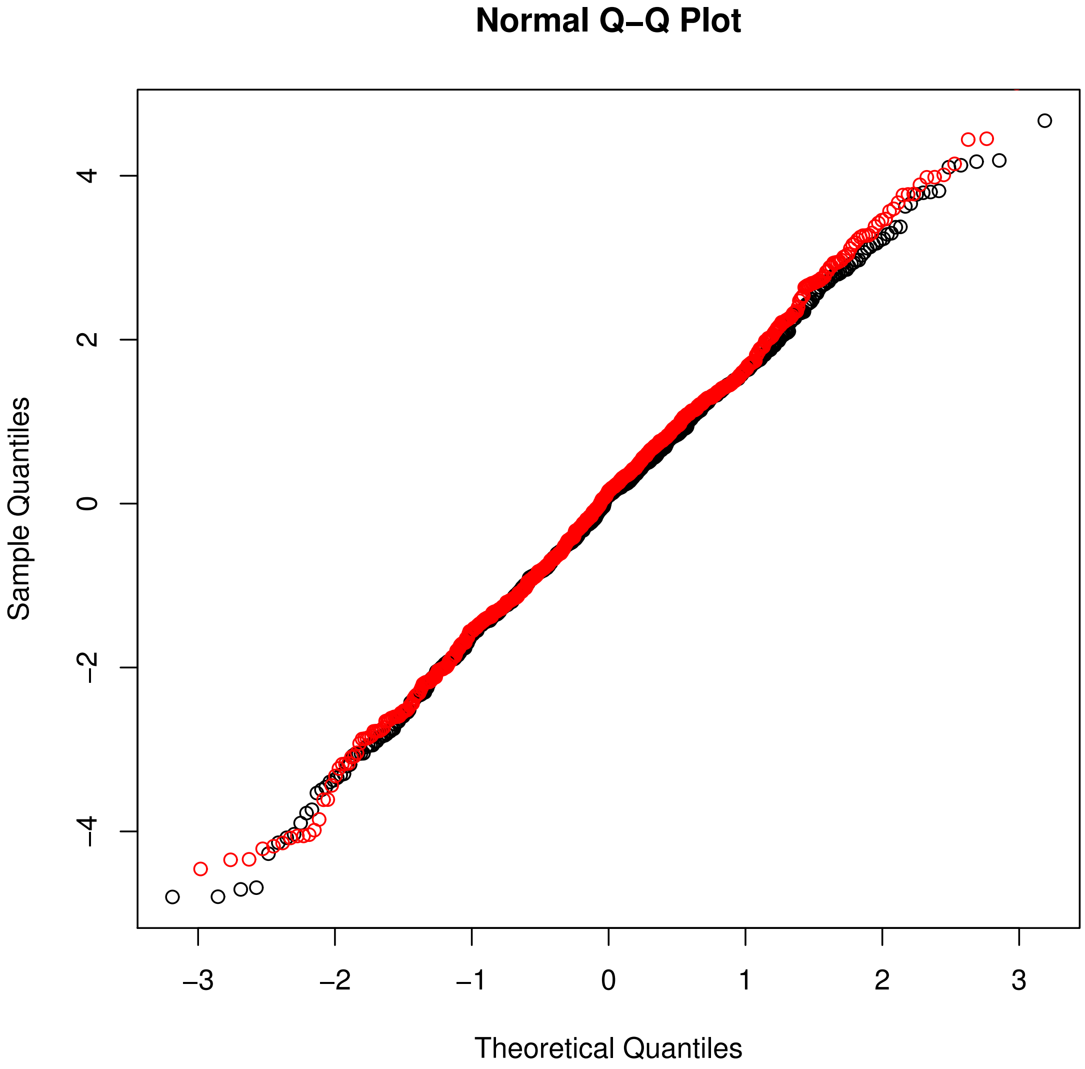


Figure 6.3: Normal QQ-plot of the residuals from the GGDP model (red line) and the GDP model (black line) for the rice data.



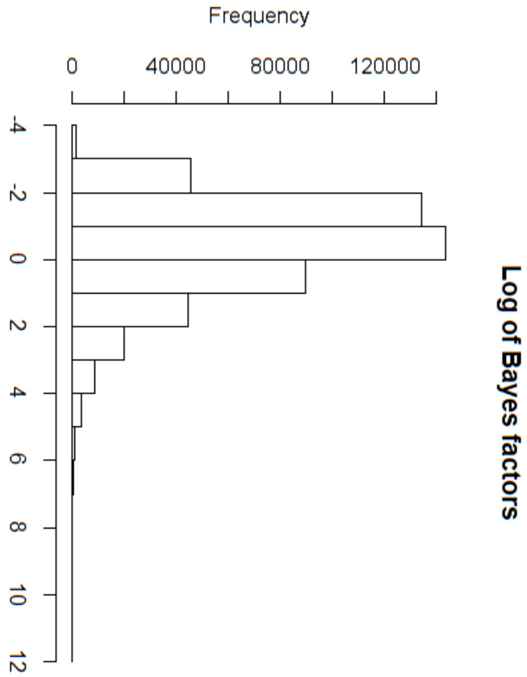


Figure 6.4: Histogram depicting the natural logarithm of the Bayes factors which were computed for each of 495,532 SNPs available in the CRC data.

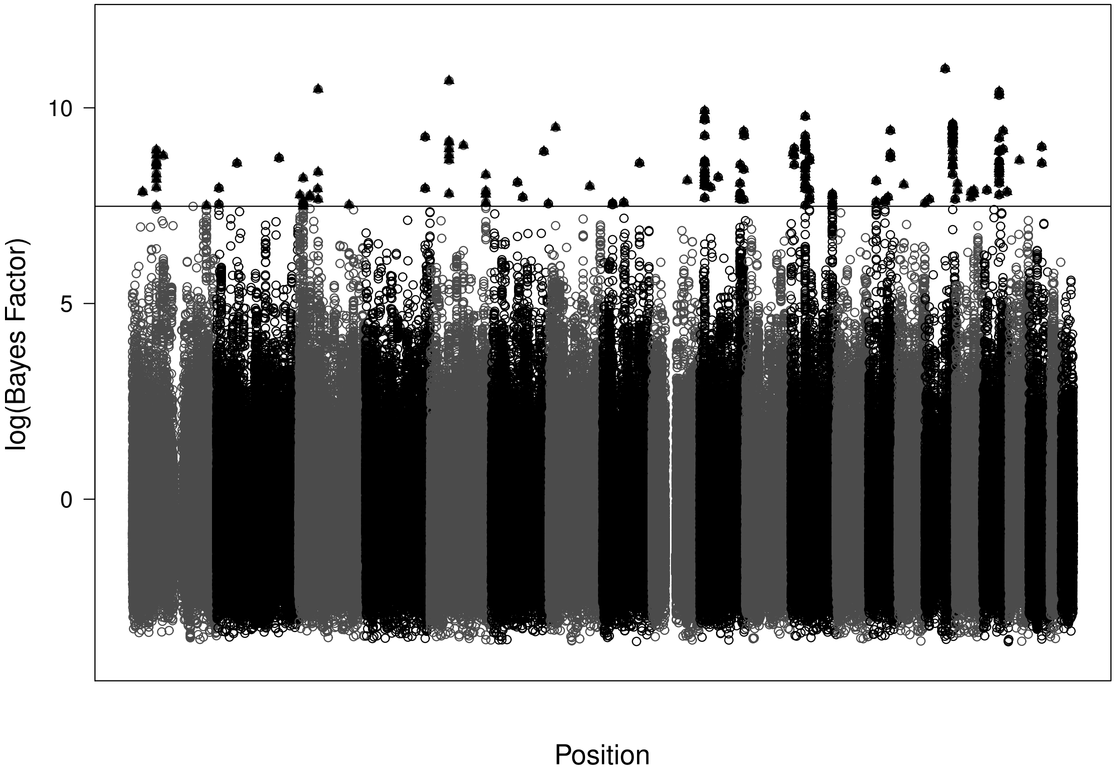


Figure 6.5: Plot of the natural logarithm of the Bayes factors which were computed for each of 495,532 SNPs verses their position in the genome. Each shade change represents the transition to a new chromosome and the black triangles above the horizontal line depict the 200 SNPs with the largest Bayes factors.

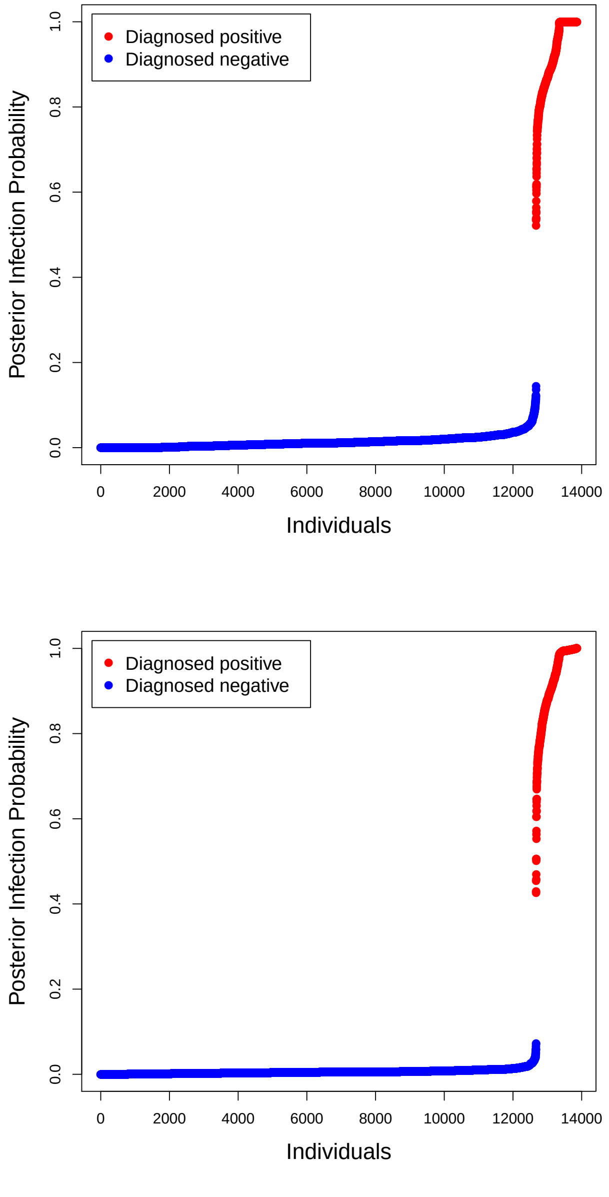


Figure 6.6: The left (right) figure depicts the posterior infection probabilities for all individuals from the Iowa chlamydia data corresponding to the analysis from Table 6.3 (Table 6.17) in the dissertation. The blue (red) points depict individuals who were diagnosed to be negative (positive).

Table 6.1: Simulation results with known assay accuracies ($S_{ej} = 0.95$ and $S_{pj} = 0.98$) under the Dirac spike. This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	IT			MPT			DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.06	0.27	1.00	0.10	0.40	1.00	-0.03	0.23	1.00	-0.02	0.22	1.00
β_1 = -1.5	-0.01	0.20	1.00	0.08	0.27	1.00	-0.01	0.19	1.00	-0.01	0.19	1.00
β_2 = 0.5	0.04	0.12	0.99	0.00	0.16	0.99	0.02	0.09	0.99	0.01	0.08	0.99
β_3 = 0.25	0.00	0.06	0.99	-0.05	0.11	0.77	0.00	0.05	0.99	0.00	0.05	0.99
β_4 = 0	0.00	0.01	0.03	0.00	0.02	0.04	0.00	0.01	0.03	0.00	0.01	0.03
β_5 = 0	0.00	0.01	0.03	0.00	0.01	0.04	0.00	0.01	0.03	0.00	0.01	0.03
λ_1 = 1	-0.01	0.32	0.93	-0.10	0.40	0.88	0.03	0.21	0.98	0.03	0.18	0.99
λ_2 = 0.75	0.01	0.15	0.98	-0.13	0.32	0.80	0.04	0.11	0.99	0.02	0.10	1.00
λ_3 = 0.25	-0.07	0.17	0.56	-0.16	0.18	0.21	-0.06	0.14	0.66	-0.06	0.13	0.69
λ_4 = 0	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
λ_5 = 0	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
λ_6 = 0	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
d_{11} = 1	0.13	0.52	–	0.03	0.67	–	0.14	0.41	–	0.13	0.38	–
d_{22} = 1.125	0.07	0.40	–	-0.13	0.63	–	0.10	0.34	–	0.06	0.31	–
d_{33} = 0.109	0.03	0.18	–	-0.01	0.21	–	0.01	0.11	–	0.00	0.08	–
d_{21} = 0.75	0.01	0.37	–	-0.12	0.52	–	0.04	0.30	–	0.03	0.27	–
d_{31} = 0.125	-0.07	0.08	–	-0.11	0.06	–	-0.05	0.08	–	-0.05	0.08	–
d_{32} = 0.225	-0.07	0.16	–	-0.14	0.17	–	-0.06	0.13	–	-0.06	0.12	–

Table 6.2: Simulation results with unknown assay accuracies under the Dirac spike. This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.05	0.23	1.00	-0.04	0.21	1.00
β_1 = -1.5	-0.04	0.19	1.00	-0.03	0.18	1.00
β_2 = 0.5	0.03	0.09	0.99	0.01	0.09	0.99
β_3 = 0.25	0.01	0.05	0.99	0.00	0.04	0.99
β_4 = 0	0.00	0.01	0.03	0.00	0.01	0.03
β_5 = 0	0.00	0.01	0.03	0.00	0.01	0.03
λ_1 = 1	0.04	0.23	0.98	0.03	0.19	0.99
λ_2 = 0.75	0.04	0.13	0.99	0.03	0.11	0.99
λ_3 = 0.25	-0.05	0.14	0.68	-0.05	0.13	0.71
λ_4 = 0	0.00	0.00	0.01	0.00	0.00	0.01
λ_5 = 0	0.00	0.00	0.01	0.00	0.00	0.01
λ_6 = 0	0.00	0.00	0.01	0.00	0.00	0.01
d_{11} = 1	0.17	0.44	–	0.14	0.38	–
d_{22} = 1.125	0.13	0.37	–	0.07	0.33	–
d_{33} = 0.109	0.02	0.13	–	0.01	0.09	–
d_{21} = 0.75	0.05	0.33	–	0.03	0.28	–
d_{31} = 0.125	-0.05	0.08	–	-0.06	0.08	–
d_{32} = 0.225	-0.05	0.14	–	-0.06	0.12	–
$S_{e(1)}$ = 0.95	-0.02	0.03	–	0.00	0.01	–
$S_{e(2)}$ = 0.98	-0.01	0.01	–	0.00	0.01	–
$S_{p(1)}$ = 0.98	0.00	0.01	–	0.00	0.00	–
$S_{p(2)}$ = 0.99	0.00	0.01	–	0.00	0.01	–

Table 6.3: Analysis of the Iowa chlamydia data. This summary includes the posterior mean estimate (Estimate), posterior standard deviation estimate (ESD), and the posterior probability of inclusion (PI). The unstandardized effect estimate (β^*) is reported.

Parameter	Description	Estimate	ESD	PI
$\beta_0^\star$	Intercept	-0.508	0.109	1.00
$\beta_1^\star$	Age	-0.037	0.004	1.00
$\beta_2^\star$	Race	-0.164	0.061	0.93
$\beta_3^\star$	New partner	0.145	0.050	0.94
$\beta_4^\star$	Multiple partners	0.137	0.093	0.74
$\beta_5^\star$	Contact with STD	0.732	0.067	1.00
$\beta_6^\star$	Symptoms	0.029	0.055	0.24
λ_1	Intercept	0.174	0.036	1.00
λ_2	Age	0.000	0.001	0.01
λ_3	Race	0.000	0.001	0.00
λ_4	New partner	0.003	0.017	0.03
λ_5	Multiple partners	0.000	0.002	0.01
λ_6	Contact with STD	0.000	0.000	0.00
λ_7	Symptoms	0.000	0.000	0.01
$S_{e(1)}$	Swab individual	0.998	0.002	–
$S_{e(2)}$	Urine individual	0.792	0.090	–
$S_{e(3)}$	Swab pool	0.909	0.062	–
$S_{p(1)}$	Swab individual	0.979	0.007	–
$S_{p(2)}$	Urine individual	0.987	0.007	–
$S_{p(3)}$	Swab pool	0.999	0.001	–

Table 6.4: Simulation results with known assay accuracies ($S_{e_j:1} = S_{e_j:2} = 0.95$ and $S_{p_j:1} = S_{p_j:2} = 0.98$). This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The average prevalence rate p for each disease is reiterated.

		IT			MPT			DT			AT		
	Parameter	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI
Disease one ($p = 0.03$)	β_{11} = -4	-0.51	0.87	1.00	-0.48	0.77	1.00	-0.23	0.46	1.00	-0.13	0.39	1.00
	β_{21} = -1.5	-0.20	0.38	0.99	-0.17	0.37	0.99	-0.07	0.25	1.00	-0.02	0.23	1.00
	β_{31} = 0.5	0.03	0.20	0.90	0.03	0.23	0.88	0.06	0.15	0.97	0.05	0.23	0.98
	β_{41} = 0.25	-0.07	0.15	0.56	-0.12	0.15	0.39	-0.04	0.12	0.72	-0.04	0.13	0.77
	β_{51} = 0	0.00	0.01	0.02	0.00	0.02	0.03	0.00	0.01	0.02	0.00	0.11	0.02
	λ_{11} = 1	0.25	0.45	1.00	0.23	0.44	1.00	0.16	0.31	1.00	0.12	0.27	1.00
	λ_{21} = 0.75	-0.06	0.33	0.83	-0.11	0.38	0.76	0.05	0.17	0.98	0.04	0.14	0.99
	λ_{31} = 0.25	-0.12	0.20	0.29	-0.14	0.18	0.23	-0.15	0.16	0.26	-0.15	0.14	0.28
	λ_{41} = 0	0.02	0.05	0.07	0.02	0.05	0.06	0.01	0.03	0.04	0.00	0.02	0.03
	λ_{51} = 0	0.00	0.01	0.02	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.01
Disease two ($p = 0.09$)	β_{12} = -2.5	-0.16	0.22	1.00	-0.25	0.26	1.00	-0.12	0.19	1.00	-0.11	0.18	1.00
	β_{22} = 1	0.07	0.11	1.00	0.12	0.13	0.99	0.05	0.10	1.00	0.05	0.09	1.00
	β_{32} = -0.75	-0.06	0.08	1.00	-0.06	0.11	0.99	-0.05	0.07	1.00	-0.05	0.07	1.00
	β_{42} = 0.3	0.00	0.05	0.99	-0.02	0.09	0.93	0.00	0.04	0.99	0.00	0.04	0.99
	β_{52} = 0	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01
	β_{62} = 0	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01
	λ_{12} = 0.8	0.20	0.19	1.00	0.29	0.22	1.00	0.16	0.17	1.00	0.15	0.15	1.00
	λ_{22} = 0.3	-0.03	0.17	0.70	-0.11	0.22	0.41	-0.01	0.15	0.79	0.01	0.13	0.86
	λ_{32} = 0.15	-0.12	0.06	0.12	-0.12	0.07	0.08	-0.12	0.07	0.13	-0.12	0.07	0.14
	λ_{42} = 0	0.00	0.01	0.02	0.00	0.02	0.03	0.00	0.01	0.01	0.00	0.01	0.01
	λ_{52} = 0	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.01
	λ_{62} = 0	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.01
	$\rho = 0.99$	-0.87	0.04	–	-0.96	0.01	–	-0.74	0.05	–	-0.63	0.06	–

Table 6.5: Simulation results with unknown assay accuracies. This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The average prevalence rate p for each disease is reiterated.

		DT			AT		
	Parameter	Bias	SSD	PI	Bias	SSD	PI
Disease one ($p = 0.03$)	β_{11} = -4	-0.52	0.75	1.00	-0.15	0.41	1.00
	β_{21} = -1.5	-0.21	0.35	1.00	-0.02	0.23	0.99
	β_{31} = 0.5	0.08	0.17	0.95	0.06	0.13	0.97
	β_{41} = 0.25	-0.04	0.13	0.70	-0.04	0.12	0.74
	β_{51} = 0	0.00	0.01	0.02	0.00	0.01	0.01
	λ_{11} = 1	0.21	0.39	1.00	0.12	0.28	1.00
	λ_{21} = 0.75	0.02	0.26	0.92	0.04	0.14	0.98
	λ_{31} = 0.25	-0.14	0.17	0.28	-0.16	0.14	0.27
	λ_{41} = 0	0.01	0.04	0.05	0.00	0.02	0.03
	λ_{51} = 0	0.00	0.01	0.02	0.00	0.01	0.01
	$S_{e(1):1}$ = 0.95	-0.03	0.04	–	0.00	0.03	–
	$S_{e(2):1}$ = 0.98	-0.03	0.02	–	-0.06	0.03	–
	$S_{p(1):1}$ = 0.98	-0.06	0.04	–	-0.03	0.03	–
	$S_{p(2):1}$ = 0.99	-0.04	0.02	–	-0.07	0.03	–
Disease two ($p = 0.09$)	β_{12} = -2.5	-0.31	0.50	1.00	-0.17	0.22	1.00
	β_{22} = 1	0.16	0.25	1.00	0.07	0.11	0.99
	β_{32} = -0.75	-0.12	0.18	1.00	-0.06	0.08	0.98
	β_{42} = 0.3	0.02	0.08	0.99	0.01	0.04	0.95
	β_{52} = 0	0.00	0.01	0.01	0.00	0.01	0.01
	β_{62} = 0	0.00	0.01	0.01	0.00	0.01	0.01
	λ_{12} = 0.8	0.29	0.27	1.00	0.19	0.17	1.00
	λ_{22} = 0.3	-0.03	0.16	0.72	-0.02	0.14	0.68
	λ_{32} = 0.15	-0.12	0.07	0.11	-0.12	0.07	0.12
	λ_{42} = 0	0.00	0.02	0.02	0.00	0.01	0.01
	λ_{52} = 0	0.00	0.01	0.01	0.00	0.01	0.01
	λ_{62} = 0	0.00	0.01	0.01	0.00	0.01	0.01
	$S_{e(1):2}$ = 0.95	-0.05	0.07	–	-0.01	0.04	–
	$S_{e(2):2}$ = 0.98	-0.04	0.06	–	-0.07	0.04	–
	$S_{p(1):2}$ = 0.98	-0.08	0.07	–	-0.04	0.03	–
	$S_{p(2):2}$ = 0.99	-0.05	0.06	–	-0.07	0.04	–
	$\rho = 0.99$	-0.81	0.07	–	-0.69	0.07	–

Table 6.6: Analysis of the Iowa chlamydia data. This summary includes the posterior mean estimate (Estimate), posterior standard deviation estimate (ESD), and the posterior probability of inclusion (PI).

	Parameter	Description	Estimate	ESD	PI
Gonorrhea	β_{11}	Intercept	-2.58	0.08	1.00
	β_{12}	Age	0.00	0.01	0.01
	β_{13}	Race	-0.04	0.04	0.31
	β_{14}	New partner	0.00	0.01	0.01
	β_{15}	Multiple partners	0.00	0.01	0.02
	β_{16}	Contact with STD	0.19	0.02	1.00
	β_{17}	Symptoms	0.00	0.01	0.01
	λ_{11}	Intercept	0.37	0.07	1.00
	λ_{12}	Age	0.00	0.01	0.03
	λ_{13}	Race	0.05	0.08	0.34
	λ_{14}	New partner	0.00	0.02	0.02
	λ_{15}	Multiple partners	0.00	0.01	0.02
	λ_{16}	Contact with STD	0.00	0.01	0.01
	λ_{17}	Symptoms	0.00	0.01	0.01
	$S_{e(1):1}$	Swab individual	0.95	0.01	–
	$S_{e(2):1}$	Urine individual	0.95	0.02	–
	$S_{e(3):1}$	Swab pool	0.95	0.01	–
	$S_{p(1):1}$	Swab individual	0.98	0.01	–
	$S_{p(2):1}$	Urine individual	0.99	0.01	–
	$S_{p(3):1}$	Swab pool	0.98	0.01	–
Chlamydia	β_{21}	Intercept	-1.44	0.03	1.00
	β_{22}	Age	-0.23	0.02	1.00
	β_{23}	Race	-0.04	0.03	0.60
	β_{24}	New partner	0.02	0.03	0.26
	β_{25}	Multiple partners	0.03	0.03	0.50
	β_{26}	Contact with STD	0.16	0.01	1.00
	β_{27}	Symptoms	0.01	0.02	0.11
	λ_{21}	Intercept	0.16	0.03	1.00
	λ_{22}	Age	0.00	0.01	0.01
	λ_{23}	Race	0.00	0.01	0.00
	λ_{24}	New partner	0.07	0.05	0.77
	λ_{25}	Multiple partners	0.01	0.01	0.01
	λ_{26}	Contact with STD	0.00	0.01	0.01
	λ_{27}	Symptoms	0.00	0.01	0.00
	$S_{e(1):2}$	Swab individual	0.99	0.01	–
	$S_{e(2):2}$	Urine individual	0.91	0.02	–
	$S_{e(3):2}$	Swab pool	0.99	0.01	–
	$S_{p(1):2}$	Swab individual	0.99	0.01	–
	$S_{p(2):2}$	Urine individual	0.99	0.01	–
	$S_{p(3):2}$	Swab pool	0.99	0.01	–

Table 6.7: Simulation configuration 1: First set of randomly selected SNPs, for all considered values of σ. Results were obtained through the GGDP and GDP models, as well as through applying the approach outlined in Armagan, Dunson, and Lee (2013). Presented results consist of the empirical bias (Bias), empirical mean-squared error (MSE), and sample standard deviation (SD) of the parameter estimates that were estimated as non-zero, as well as the percentage of the time that a coefficient was identified as being significant (Percent). Also included are the percentage of false discoveries (FDP) for each considered configuration and the average number of iterations the EM algorithm went through before convergence (Iter) with the average time (measured in seconds) being provided in parenthesis. The minor allele frequency is also reported (MAF).

			GGDP				GDP				Armagan et al. (2013)			
σ	Parameter	MAF	Bias	MSE	SD	Percent	Bias	MSE	SD	Percent	Bias	MSE	SD	Percent
0.5	$\beta_0 = 3.00$	–	0.017	0.041	0.203	100%	0.008	0.073	0.270	100%	-0.052	0.273	0.520	100%
	$\beta_1 = 3.50$	–	-0.003	0.001	0.034	100%	-0.003	0.001	0.034	100%	-0.001	0.001	0.034	100%
	$\beta_2 = 1.00$	–	-0.004	0.001	0.034	100%	-0.004	0.001	0.034	100%	0.000	0.001	0.034	100%
	$\beta_3 = 0.25$	21%	-0.021	0.002	0.043	100%	-0.013	0.003	0.048	100%	-0.023	0.005	0.068	100%
	$\beta_4 = 0.50$	13%	-0.007	0.002	0.044	100%	-0.003	0.003	0.056	100%	-0.020	0.008	0.087	100%
	$\beta_5 = 0.75$	11%	-0.010	0.002	0.046	100%	-0.008	0.003	0.056	100%	-0.048	0.012	0.100	100%
	$\beta_6 = 1.00$	20%	-0.004	0.001	0.037	100%	-0.002	0.002	0.042	100%	-0.024	0.007	0.078	100%
	$\beta_7 = 1.50$	13%	-0.007	0.002	0.044	100%	-0.006	0.003	0.057	100%	-0.019	0.010	0.096	100%
	$\beta_8 = 2.00$	10%	-0.003	0.001	0.033	100%	-0.002	0.002	0.040	100%	-0.018	0.004	0.062	100%
			Iter: 80(39s)			FDP: 0.29%	Iter: 152(40s)			FDP: 1.34%	Iter: 4818(419s)			FDP: 26%
1	$\beta_0 = 3.00$	–	0.094	0.189	0.424	100%	0.078	0.302	0.544	100%	0.062	1.138	1.066	100%
	$\beta_1 = 3.50$	–	-0.008	0.005	0.068	100%	-0.010	0.005	0.068	100%	0.001	0.005	0.067	100%
	$\beta_2 = 1.00$	–	-0.015	0.005	0.068	100%	-0.017	0.005	0.068	100%	-0.001	0.005	0.067	100%
	$\beta_3 = 0.25$	21%	-0.012	0.004	0.063	46%	-0.004	0.007	0.083	67%	-0.031	0.015	0.119	92%
	$\beta_4 = 0.50$	13%	-0.033	0.011	0.098	99%	-0.011	0.015	0.121	99%	-0.069	0.037	0.178	98%
	$\beta_5 = 0.75$	11%	-0.032	0.011	0.098	100%	-0.021	0.014	0.119	100%	-0.126	0.056	0.200	99%
	$\beta_6 = 1.00$	20%	-0.020	0.006	0.072	100%	-0.012	0.008	0.088	100%	-0.061	0.027	0.151	100%
	$\beta_7 = 1.50$	13%	-0.030	0.009	0.092	100%	-0.029	0.014	0.117	100%	-0.064	0.040	0.188	100%
	$\beta_8 = 2.00$	10%	-0.008	0.005	0.068	100%	-0.009	0.007	0.083	100%	-0.043	0.017	0.122	100%
			Iter: 89(39s)			FDP: 0.29%	Iter: 159(40s)			FDP: 1.35%	Iter: 5363(470s)			FDP: 26%
2	$\beta_0 = 3.00$	–	0.372	0.866	0.854	100%	0.302	1.396	1.143	100%	0.363	4.591	2.113	100%
	$\beta_1 = 3.50$	–	-0.041	0.021	0.140	100%	-0.047	0.022	0.141	100%	0.001	0.019	0.137	100%
	$\beta_2 = 1.00$	–	-0.072	0.027	0.149	100%	-0.081	0.029	0.150	100%	-0.008	0.019	0.138	100%
	$\beta_3 = 0.25$	21%	0.184	0.045	0.107	5%	0.151	0.040	0.131	19%	-0.013	0.047	0.217	71%
	$\beta_4 = 0.50$	13%	0.069	0.025	0.142	39%	0.095	0.044	0.188	58%	-0.054	0.088	0.292	78%
	$\beta_5 = 0.75$	11%	-0.033	0.034	0.182	73%	-0.003	0.051	0.226	80%	-0.194	0.141	0.322	79%
	$\beta_6 = 1.00$	20%	-0.086	0.029	0.149	97%	-0.043	0.032	0.174	97%	-0.168	0.124	0.310	98%
	$\beta_7 = 1.50$	13%	-0.097	0.046	0.190	100%	-0.086	0.059	0.226	100%	-0.146	0.171	0.386	100%
	$\beta_8 = 2.00$	10%	-0.027	0.019	0.137	100%	-0.018	0.029	0.168	100%	-0.100	0.073	0.252	100%
			Iter: 94(44s)			FDP: 0.30%	Iter: 170(45s)			FDP: 1.40%	Iter: 6257(539s)			FDP: 26%

Table 6.8: Data Application: Results include the estimated field and genetic effects on yield obtained by the GGDP and GDP models. NS indicates that the SNP was not selected by a particular model and the minor allele frequency of the selected SNPs is reported (MAF). The prediction error (CV_{error}) is also provided, and was computed via 5-fold cross validation.

Term	Chr	Ref	MAF	GGDP β	GDP β
$Intercept$				3.302	3.201
F_1				3.586	3.499
F_2				0.849	0.828
S_{64}	1	C	3%	-0.186	-0.257
S_{262}	1	T	14%	-0.388	-0.389
S_{664}	4	G	4%	NS	-0.318
S_{768}	5	A	12%	-0.285	-0.267
S_{838}	6	A	12%	-0.265	-0.254
S_{941}	7	T	22%	-0.180	NS
S_{1014}	8	T	5%	0.515	0.465
S_{1118}	10	T	16%	NS	0.234
S_{1215}	12	G	3%	0.199	0.191
CV_{error}				431.73	443.60

Table 6.9: Simulation results when $n = 200$ and $q_2 \in \{100, 200, 500\}$. The summary includes the empirical bias (Bias) and standard deviation (SD) of the MAP estimates, as well as the percent of times the significant variable remained in the model (Perc). The SNP coefficients are categorized according to their allelic frequencies (AF). The empirical false discovery proportion (FDP) for the truly unrelated variables is also included. The average times required to analyze each data set were 9.1, 12.6, and 46.9 seconds when $q_2 = 100, 200, 500$, respectively. This time includes the grid search over the various (α, η) settings.

		$q_2 = 100$			$q_2 = 200$			$q_2 = 500$		
AF	Parameter	Bias	SD	Perc	Bias	SD	Perc	Bias	SD	Perc
					Non-SNP coefficients					
	$\beta_0 = -1$	0.00	0.28	100%	-0.05	0.31	100%	-0.16	0.36	100%
	$\beta_{1,1} = 1$	-0.09	0.32	99%	0.03	0.36	99%	0.13	0.45	99%
	$\beta_{1,2} = 1$	-0.06	0.32	99%	0.02	0.34	99%	0.11	0.41	99%
					SNP coefficients					
Low	$\beta_{2,1} = 0.25$	-0.22	0.15	6%	-0.24	0.09	3%	-0.25	0.02	1%
	$\beta_{2,2} = 0.25$	-0.21	0.15	7%	-0.23	0.11	3%	-0.24	0.14	3%
	$\beta_{2,3} = 0.5$	-0.21	0.33	51%	-0.22	0.36	45%	-0.29	0.35	31%
	$\beta_{2,4} = 0.5$	-0.28	0.32	37%	-0.31	0.32	30%	-0.38	0.28	18%
	$\beta_{2,5} = 1$	-0.08	0.32	99%	-0.01	0.38	98%	0.04	0.43	97%
	$\beta_{2,6} = 1$	-0.10	0.37	96%	-0.07	0.43	94%	-0.24	0.55	76%
High	$\beta_{2,7} = 0.25$	-0.16	0.23	16%	-0.19	0.18	11%	-0.20	0.20	8%
	$\beta_{2,8} = 0.25$	-0.13	0.25	22%	-0.17	0.22	13%	-0.19	0.20	10%
	$\beta_{2,9} = 0.5$	-0.27	0.32	38%	-0.36	0.28	23%	-0.46	0.17	8%
	$\beta_{2,10} = 0.5$	-0.17	0.35	54%	-0.15	0.38	54%	-0.19	0.40	43%
	$\beta_{2,11} = 1$	-0.21	0.39	92%	-0.44	0.53	60%	-0.53	0.54	51%
	$\beta_{2,12} = 1$	-0.18	0.41	90%	-0.25	0.48	80%	-0.39	0.57	61%
		FDP: 3.6%			FDP: 3.2%			FDP: 1.8%		

Table 6.10: Simulation results when $n = 500$ and $q_2 \in \{100, 200, 500\}$. The summary includes the empirical bias (Bias) and standard deviation (SD) of the MAP estimates, as well as the percent of times the significant variable remained in the model (Perc). The SNP coefficients are categorized according to their allelic frequencies (AF). The empirical false discovery proportion (FDP) for the truly unrelated variables is also included. The average times required to analyze each data set were 30.2, 36.1, and 85.8 seconds when $q_2 = 100, 200, 500$, respectively. This time includes the grid search over the various (α, η) settings.

		$q_2 = 100$			$q_2 = 200$			$q_2 = 500$		
AF	Parameter	Bias	SD	Perc	Bias	SD	Perc	Bias	SD	Perc
					Non-SNP coefficients					
	$\beta_0 = -1$	0.05	0.15	100%	0.04	0.16	100%	0.01	0.15	100%
	$\beta_{1,1} = 1$	-0.07	0.15	100%	-0.07	0.16	100%	-0.04	0.17	100%
	$\beta_{1,2} = 1$	-0.07	0.15	100%	-0.06	0.15	100%	-0.03	0.17	100%
					SNP coefficients					
Low	$\beta_{2,1} = 0.25$	-0.19	0.14	19%	-0.22	0.10	9%	-0.24	0.05	2%
	$\beta_{2,2} = 0.25$	-0.20	0.13	15%	-0.21	0.12	11%	-0.23	0.08	4%
	$\beta_{2,3} = 0.5$	-0.09	0.20	91%	-0.12	0.23	83%	-0.18	0.24	70%
	$\beta_{2,4} = 0.5$	-0.14	0.22	80%	-0.17	0.24	73%	-0.26	0.25	53%
	$\beta_{2,5} = 1$	-0.10	0.17	100%	-0.09	0.19	100%	-0.09	0.19	100%
	$\beta_{2,6} = 1$	-0.08	0.16	100%	-0.07	0.18	100%	-0.10	0.19	99%
High	$\beta_{2,7} = 0.25$	-0.17	0.15	28%	-0.20	0.13	17%	-0.22	0.10	11%
	$\beta_{2,8} = 0.25$	-0.10	0.18	45%	-0.13	0.18	32%	-0.17	0.17	22%
	$\beta_{2,9} = 0.5$	-0.14	0.21	82%	-0.24	0.23	61%	-0.37	0.21	32%
	$\beta_{2,10} = 0.5$	-0.09	0.20	90%	-0.09	0.20	89%	-0.09	0.23	84%
	$\beta_{2,11} = 1$	-0.14	0.17	100%	-0.26	0.34	86%	-0.24	0.28	93%
	$\beta_{2,12} = 1$	-0.12	0.19	100%	-0.16	0.18	99%	-0.17	0.24	98%
			FDP: 2.8%			FDP: 2.4%			FDP: 1.3%	

Table 6.11: Summary of the analysis of the Colorectal cancer data. Presented results include the chromosome number (Chr) and coordinate (Coordinate) of the identified SNPs, the gene they lie on (Gene), reference allele (Ref), minor allele frequency (MAF), and estimated effect (Estimate).

Description	Chr	Coordinate	Gene	Ref	MAF	Estimate
Intercept						0.90
Gender						0.00
Age						-3.75
BMI						0.00
Smoking						1.32
S_3	3	57086348	ARHGEF3	G	0.07	2.40
S_{19}	16	81947156	PLCG2	C	0.08	0.85
S_{27}	10	129963848	intergenic	C	0.34	-1.32
S_{51}	5	98125016	RGMB	G	0.05	1.95
S_{58}	18	59822981	PIGN	TC	0.19	-1.39
S_{118}	5	164113078	intergenic	T	0.12	1.65
S_{128}	6	77328692	intergenic	A	0.04	1.22
S_{154}	17	45800299	intergenic	T	0.36	1.32
S_{172}	16	13018917	SHISA9	C	0.11	1.67
S_{200}	3	12816282	intergenic	A	0.03	2.13

Table 6.12: Simulation results with known assay accuracies ($S_{ej} = 0.95$ and $S_{pj} = 0.98$) under SSVS. This includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	IT			MPT			DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.08	0.26	1.00	-0.08	0.30	1.00	-0.03	0.22	1.00	-0.02	0.21	1.00
β_1 = -1.5	-0.01	0.20	1.00	-0.02	0.23	1.00	-0.02	0.18	1.00	0.00	0.18	1.00
β_2 = 0.5	0.02	0.09	0.99	0.04	0.14	0.97	0.02	0.08	0.99	0.01	0.08	0.99
β_3 = 0.25	-0.02	0.07	0.76	-0.05	0.10	0.58	-0.01	0.06	0.80	-0.01	0.06	0.81
β_4 = 0	0.00	0.03	0.03	0.00	0.03	0.03	0.00	0.03	0.03	0.00	0.03	0.02
β_5 = 0	0.00	0.03	0.03	0.00	0.03	0.03	0.00	0.03	0.02	0.00	0.03	0.02
λ_1 = 1	0.06	0.20	0.99	-0.02	0.32	0.93	0.01	0.18	0.99	0.02	0.18	0.99
λ_2 = 0.75	0.05	0.11	0.99	0.04	0.13	0.99	0.03	0.11	1.00	0.01	0.11	1.00
λ_3 = 0.25	-0.03	0.09	0.61	-0.04	0.13	0.49	-0.03	0.09	0.62	-0.04	0.09	0.62
λ_4 = 0	0.04	0.01	0.04	0.05	0.02	0.06	0.04	0.01	0.03	0.04	0.01	0.03
λ_5 = 0	0.04	0.01	0.03	0.04	0.01	0.04	0.04	0.01	0.03	0.04	0.01	0.03
λ_6 = 0	0.04	0.01	0.03	0.04	0.01	0.04	0.04	0.01	0.03	0.04	0.01	0.03
d_{11} = 1	0.21	0.45	–	0.14	0.57	–	0.09	0.36	–	0.11	0.39	–
d_{22} = 1.125	0.14	0.36	–	0.11	0.44	–	0.08	0.33	–	0.02	0.32	–
d_{33} = 0.109	0.02	0.09	–	0.04	0.15	–	0.01	0.08	–	0.00	0.07	–
d_{21} = 0.75	0.07	0.31	–	0.01	0.39	–	0.01	0.28	–	0.00	0.27	–
d_{31} = 0.125	-0.03	0.08	–	-0.06	0.08	–	-0.03	0.07	–	-0.03	0.07	–
d_{32} = 0.225	-0.03	0.10	–	-0.06	0.14	–	-0.03	0.10	–	-0.04	0.09	–

Table 6.13: Simulation results with known assay accuracies ($S_{ej} = 0.95$ and $S_{pj} = 0.98$) under NMIG. This includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	IT			MPT			DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.06	0.26	1.00	-0.07	0.30	1.00	-0.03	0.22	1.00	-0.04	0.21	1.00
β_1 = -1.5	-0.02	0.21	1.00	-0.02	0.22	1.00	-0.01	0.19	1.00	-0.01	0.18	1.00
β_2 = 0.5	0.02	0.10	0.99	0.04	0.14	0.97	0.02	0.08	0.99	0.01	0.08	0.99
β_3 = 0.25	-0.02	0.07	0.73	-0.05	0.10	0.56	-0.01	0.06	0.76	-0.01	0.06	0.78
β_4 = 0	0.00	0.03	0.03	0.00	0.03	0.04	0.00	0.03	0.03	0.00	0.03	0.03
β_5 = 0	0.00	0.03	0.03	0.00	0.03	0.04	0.00	0.03	0.03	0.00	0.03	0.03
λ_1 = 1	0.02	0.19	0.99	-0.02	0.31	0.93	0.02	0.17	0.99	0.02	0.17	0.99
λ_2 = 0.75	0.03	0.11	0.99	0.04	0.13	0.99	0.03	0.10	0.99	0.03	0.10	0.99
λ_3 = 0.25	-0.03	0.10	0.59	-0.03	0.13	0.50	-0.03	0.09	0.62	-0.03	0.09	0.63
λ_4 = 0	0.04	0.01	0.04	0.05	0.02	0.06	0.04	0.01	0.04	0.04	0.01	0.03
λ_5 = 0	0.04	0.01	0.03	0.04	0.02	0.04	0.04	0.01	0.03	0.04	0.01	0.03
λ_6 = 0	0.04	0.01	0.03	0.04	0.01	0.04	0.03	0.01	0.03	0.03	0.01	0.03
d_{11} = 1	0.12	0.39	–	0.13	0.56	–	0.10	0.37	–	0.10	0.36	–
d_{22} = 1.125	0.08	0.35	–	0.11	0.42	–	0.07	0.32	–	0.05	0.32	–
d_{33} = 0.109	0.01	0.09	–	0.04	0.16	–	0.01	0.07	–	0.01	0.07	–
d_{21} = 0.75	0.01	0.28	–	0.00	0.38	–	0.01	0.27	–	0.00	0.26	–
d_{31} = 0.125	-0.04	0.08	–	-0.06	0.07	–	-0.03	0.07	–	-0.03	0.07	–
d_{32} = 0.225	-0.04	0.11	–	-0.05	0.14	–	-0.04	0.09	–	-0.03	0.09	–

Table 6.14: Average estimated area under the receiver operator characteristic curve for the proposed model and the competing model which ignores both variable selection and random effects. The known assay accuracies ($S_{ej} = 0.95$ and $S_{pj} = 0.98$) setting (Known) is provided along with unknown assay accuracies (Unk).

		Proposed model				Competing model			
Assay		IT	MPT	DT	AT	IT	MPT	DT	AT
Known	SSVS	0.92	0.90	0.93	0.93	0.83	0.83	0.84	0.84
	NMIG	0.92	0.90	0.93	0.93	0.84	0.83	0.84	0.84
	Dirac	0.92	0.89	0.93	0.93	0.83	0.83	0.84	0.84
Unk	SSVS	–	–	0.93	0.93	–	–	0.84	0.84
	NMIG	–	–	0.93	0.93	–	–	0.84	0.84
	Dirac	–	–	0.93	0.93	–	–	0.84	0.84

Table 6.15: Simulation results with unknown assay accuracies under SSVS. This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.07	0.22	1.00	-0.06	0.21	1.00
β_1 = -1.5	-0.07	0.19	1.00	-0.04	0.18	1.00
β_2 = 0.5	0.02	0.08	0.99	0.02	0.08	0.99
β_3 = 0.25	-0.01	0.06	0.82	-0.01	0.06	0.82
β_4 = 0	0.00	0.03	0.02	0.00	0.03	0.02
β_5 = 0	0.00	0.03	0.02	0.00	0.03	0.02
λ_1 = 1	0.05	0.18	0.99	0.02	0.17	0.99
λ_2 = 0.75	0.05	0.10	1.00	0.04	0.11	1.00
λ_3 = 0.25	-0.03	0.09	0.62	-0.03	0.09	0.64
λ_4 = 0	0.04	0.01	0.03	0.04	0.01	0.03
λ_5 = 0	0.04	0.01	0.03	0.04	0.01	0.03
λ_6 = 0	0.04	0.01	0.03	0.04	0.01	0.03
d_{11} = 1	0.16	0.39	–	0.11	0.37	–
d_{22} = 1.125	0.14	0.33	–	0.08	0.32	–
d_{33} = 0.109	0.01	0.08	–	0.01	0.07	–
d_{21} = 0.75	0.06	0.28	–	0.01	0.26	–
d_{31} = 0.125	-0.03	0.07	–	-0.03	0.06	–
d_{32} = 0.225	-0.03	0.09	–	-0.03	0.09	–
$S_{e(1)}$ = 0.95	-0.02	0.03	–	0.00	0.01	–
$S_{e(2)}$ = 0.98	-0.01	0.01	–	0.00	0.01	–
$S_{p(1)}$ = 0.98	0.00	0.01	–	0.00	0.00	–
$S_{p(2)}$ = 0.99	0.00	0.00	–	0.00	0.01	–

Table 6.16: Simulation results with unknown assay accuracies under NMIG. This summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.06	0.23	1.00	-0.04	0.21	1.00
β_1 = -1.5	-0.05	0.19	1.00	-0.02	0.18	1.00
β_2 = 0.5	0.03	0.09	0.99	0.01	0.08	0.99
β_3 = 0.25	-0.01	0.06	0.78	-0.01	0.06	0.78
β_4 = 0	0.00	0.03	0.03	0.00	0.03	0.03
β_5 = 0	0.00	0.03	0.03	0.00	0.03	0.03
λ_1 = 1	0.03	0.18	0.99	0.03	0.17	0.99
λ_2 = 0.75	0.05	0.11	0.99	0.03	0.10	0.99
λ_3 = 0.25	-0.02	0.09	0.64	-0.03	0.08	0.63
λ_4 = 0	0.04	0.01	0.04	0.04	0.01	0.03
λ_5 = 0	0.04	0.01	0.03	0.04	0.01	0.03
λ_6 = 0	0.04	0.01	0.03	0.03	0.01	0.03
d_{11} = 1	0.14	0.39	–	0.12	0.36	–
d_{22} = 1.125	0.12	0.34	–	0.06	0.33	–
d_{33} = 0.109	0.02	0.08	–	0.01	0.07	–
d_{21} = 0.75	0.03	0.28	–	0.01	0.27	–
d_{31} = 0.125	-0.03	0.08	–	-0.03	0.07	–
d_{32} = 0.225	-0.02	0.10	–	-0.03	0.09	–
$S_{e(1)}$ = 0.95	-0.02	0.03	–	0.00	0.01	–
$S_{e(2)}$ = 0.98	-0.01	0.01	–	0.00	0.01	–
$S_{p(1)}$ = 0.98	0.00	0.01	–	0.00	0.00	–
$S_{p(2)}$ = 0.99	0.00	0.00	–	0.00	0.01	–

Table 6.17: Analysis of the Iowa chlamydia data when using the Dirac spike and informative priors on the testing accuracies. The summary includes the posterior mean estimate (Estimate), posterior standard deviation estimate (ESD), and posterior probability of inclusion (PI). The informative priors were developed based on the product literature and validation trials available on the Aptima Combo 2 assay and are given by: $S_{e(1)}, S_{e(3)} \sim$ Beta$(196, 13)$, $S_{e(2)} \sim$ Beta$(198, 12)$, $S_{p(1)}, S_{p(3)} \sim$ Beta$(1155, 29)$, and $S_{p(2)} \sim$ Beta$(1171, 14)$. The unstandardized effect estimate (β^{*}) is reported.

Parameter	Description	Estimate	ESD	PI
$\beta_0^\star$	Intercept	-0.576	0.087	1.00
$\beta_1^\star$	Age	-0.036	0.003	1.00
$\beta_2^\star$	Race	-0.159	0.057	0.94
$\beta_3^\star$	New partner	0.146	0.044	0.96
$\beta_4^\star$	Multiple partners	0.144	0.087	0.79
$\beta_5^\star$	Contact with STD	0.724	0.060	1.00
$\beta_6^\star$	Symptoms	0.026	0.051	0.23
λ_1	Intercept	0.166	0.034	1.00
λ_2	Age	0.000	0.000	0.00
λ_3	Race	0.000	0.001	0.00
λ_4	New partner	0.000	0.001	0.01
λ_5	Multiple partners	0.000	0.001	0.01
λ_6	Contact with STD	0.000	0.000	0.00
λ_7	Symptoms	0.000	0.000	0.00
$S_{e(1)}$	Swab individual	0.982	0.005	–
$S_{e(2)}$	Urine individual	0.937	0.016	–
$S_{e(3)}$	Swab pool	0.944	0.015	–
$S_{p(1)}$	Swab individual	0.974	0.004	–
$S_{p(2)}$	Urine individual	0.990	0.002	–
$S_{p(3)}$	Swab pool	0.989	0.002	–

Table 6.18: The six models (M1-M6) with the highest posterior probabilities (see Kuo and Mallick; 1998) under both informative and uninformative prior specifications for the testing accuracies. Here x denotes that a variable is included in the model and Frequency denotes how often the model is visited.

		Uninformative						Informative					
	Description	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Fixed effects	Intercept	x	x	x	x	x	x	x	x	x	x	x	x
	Age	x	x	x	x	x	x	x	x	x	x	x	x
	Race	x	x	x	x	–	x	x	x	x	x	–	x
	New partner	x	x	x	x	x	–	x	x	x	x	x	–
	Multiple partners	x	–	x	–	x	x	x	x	–	–	x	x
	Contact with STD	x	x	x	x	x	x	x	x	x	x	x	x
	Symptoms	–	–	x	x	–	–	–	x	–	x	–	–
Random effects	Intercept	x	x	x	x	x	x	x	x	x	x	x	x
	Age	–	–	–	–	–	–	–	–	–	–	–	–
	Race	–	–	–	–	–	–	–	–	–	–	–	–
	New partner	–	–	–	–	–	x	–	–	–	–	–	–
	Multiple partners	–	–	–	–	–	–	–	–	–	–	–	–
	Contact with STD	–	–	–	–	–	–	–	–	–	–	–	–
	Symptoms	–	–	–	–	–	–	–	–	–	–	–	–
	Frequency	0.51	0.16	0.12	0.07	0.03	0.02	0.56	0.13	0.13	0.06	0.03	0.02

Table 6.19: Robustness study: Summary includes the average bias of the posterior mean estimates (Bias), sample standard deviation of the posterior mean estimates (SSD), and the posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4. The parameter d_{ij} denotes the ijth element of $\mathbf{D}$.

	DT			AT		
Parameter	Bias	SSD	PI	Bias	SSD	PI
β_0 = -3	-0.06	0.21	1.00	-0.02	0.21	1.00
β_1 = -1.5	-0.05	0.18	1.00	0.00	0.18	1.00
β_2 = 0.5	0.03	0.08	0.99	0.01	0.08	0.99
β_3 = 0.25	0.01	0.04	0.99	0.00	0.04	0.99
β_4 = 0	0.00	0.01	0.03	0.00	0.01	0.03
β_5 = 0	0.00	0.01	0.03	0.00	0.01	0.03
λ_1 = 1	0.06	0.18	0.99	0.03	0.17	0.99
λ_2 = 0.75	0.06	0.10	1.00	0.03	0.10	1.00
λ_3 = 0.25	-0.03	0.12	0.77	-0.04	0.11	0.78
λ_4 = 0	0.00	0.01	0.01	0.00	0.00	0.01
λ_5 = 0	0.00	0.00	0.01	0.00	0.00	0.01
λ_6 = 0	0.00	0.00	0.01	0.00	0.00	0.01
d_{11} = 1	0.19	0.38	–	0.11	0.35	–
d_{22} = 1.125	0.17	0.32	–	0.06	0.31	–
d_{33} = 0.109	0.02	0.08	–	0.01	0.07	–
d_{21} = 0.75	0.08	0.27	–	0.01	0.25	–
d_{31} = 0.125	-0.04	0.08	–	-0.05	0.08	–
d_{32} = 0.225	-0.03	0.12	–	-0.04	0.11	–