Received: 18 January 2019 | Revised: 10 June 2019 | Accepted: 11 July 2019

DOI: 10.1002/bimj.201900050

Biometrical Journal

RESEARCH PAPER

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

Chase Joyner[1] | **Christopher McMahan**[1,3] | **James Baurley**[2,3] | **Bens Pardamean**[3]

[1]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA

[2]BioRealm LLC, Walnut, CA, USA

[3]Bioinformatics and Data Science Research Center, Bina Nusantara University, Kebon Jeruk, Indonesia

Correspondence
Christopher McMahan, School of Mathematical and Statistical Sciences, Clemson University, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA.
Email: mcmaha2@clemson.edu

Funding information
National Science Foundation, Grant/Award Number: OIA-1826715; National Institutes of Health, Grant/Award Numbers: R01 AI121351, R44 AA027675

Abstract

Recent advances in sequencing and genotyping technologies are contributing to a data revolution in genome-wide association studies that is characterized by the challenging large p small n problem in statistics. That is, given these advances, many such studies now consider evaluating an extremely large number of genetic markers (p) genotyped on a small number of subjects (n). Given the dimension of the data, a joint analysis of the markers is often fraught with many challenges, while a marginal analysis is not sufficient. To overcome these obstacles, herein, we propose a Bayesian two-phase methodology that can be used to jointly relate genetic markers to binary traits while controlling for confounding. The first phase of our approach makes use of a marginal scan to identify a reduced set of candidate markers that are then evaluated jointly via a hierarchical model in the second phase. Final marker selection is accomplished through identifying a sparse estimator via a novel and computationally efficient maximum a posteriori estimation technique. We evaluate the performance of the proposed approach through extensive numerical studies, and consider a genome-wide application involving colorectal cancer.

KEYWORDS

Bayes factors, EM algorithm, GWAS, MAP estimator, shrinkage prior

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 has led to the identification of common genetic variants that are predictive of a subject's predisposition toward colorectal cancer (Peters, Bien, & Zubair, 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 heritability" 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 & 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 behavior. As a solution, the current research paradigm is shifting to SNP assessment via joint models. This new direction also poses significant challenges; that is, 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., $p \gg n$) 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 (1996) proposed the least absolute shrinkage and selection operator (LASSO) that 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 (1996), many other proposals have been developed under other penalization schemes; for example, see Fan and Li (2001), Zou and Hastie (2005), Zou (2006), and Candes and Tao (2007). Extensions of penalized regression methods have been made to generalized linear models (GLMs), for example, Wu, Chen, Hastie, Sobel, and Lange (2009) and Friedman, Hastie, and Tibshirani (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 that make use of shrinkage priors; for example, the Bayesian LASSO (Park & Casella, 2008). In many instances, analytic and computational tractabilities are aided by the fact that shrinkage priors can be represented as scale mixtures of normals; for example, see Park and Casella (2008) and Armagan, Dunson, and Lee (2013). 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 (2015) proposed a hybrid Bayesian approach for quantitative traits that combined the marginal scan and joint modeling paradigms.

Motivated by the work of Yazdani and Dunson (2015) 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 (2015), 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. (2013) 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 GWAS 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.

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 $i = 1, \ldots, n$, with the event $Y_i = 1$, denoting that the individual is a case and $Y_i = 0$ otherwise. Similarly, we let E_{iq}, for $q = 1, \ldots, q_1$, 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 $q = 1, \ldots, q^*_2$, denote the qth SNP genotype of the ith individual. To evaluate both the confounding variables and genetic markers, we propose the following two-phase methodology.

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

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

where p_0 and p_q are binary data models (e.g., logistic or probit regression models) for the observed data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{\theta}_0$ 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 $q = 1, \ldots, q^*_2$. If B_{q0} is large, then there exists strong evidence in favor of p_q when compared to p_0; for example, $B_{q0} > 20$ 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. To avoid prior influence, it is standard to specify noninformative or vague priors that 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; that is, $\boldsymbol{\theta}_0 \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{q_1+1})$ 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 multidimensional integrals depicted in the numerator and denominator of (1) are analytically intractable and therefore have to be approximated. Many techniques for approximating such integrals have been proposed; for example, see Raftery (1996). Herein, we proceed to approximate the necessary integrals through the following Laplacian approximation:

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

where $\widetilde{\boldsymbol{\theta}}_q$ is the minimizer of $h(\boldsymbol{\theta}_q) = -\log\{p_q(\mathbf{Y} \mid \boldsymbol{\theta}_q)\pi_q(\boldsymbol{\theta}_q)\}$, $\mathbf{C}$ 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 3, we discuss a pragmatic approach that can be used to choose the value of q_2.

2.2 | Phase 2

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

$$g^{-1}\{P(Y_i = 1 \mid \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\} = \beta_0 + \mathbf{E}_i'\boldsymbol{\beta}_1 + \mathbf{S}_i'\boldsymbol{\beta}_2, \tag{3}$$

where $g(\cdot)$ is the link function. For the purposes of this work, we allow $g(\cdot)$ to take on two forms (i.e., logistic 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:

$$\beta_0 \mid T_0 \sim N(0, T_0),$$

$$\beta_{lq} \mid \alpha, \eta \sim \text{GDP}(\psi = \eta/\alpha, \alpha), \text{ for } q = 1, \ldots, q_l \text{ and } l = 1, 2,$$

where $\text{GDP}(\psi, \alpha)$ refers to the generalized double Pareto distribution outlined in Armagan et al. (2013). Under these prior choices, setting T_0 to be large provides a vague prior on β_0, while the hyperparameters $\alpha > 0$ and $\eta > 0$ govern the amount of shrinkage that 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. (2013) 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 hyperparameters and make use of model selection criteria (e.g., akaike information criterion (AIC), Bayesian information criterion (BIC), cross-validation, etc.) to choose the "optimal" configuration.

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 that 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 $i = 1, \ldots, n$. 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, Scott, and Windle (2013). Under either specification, the joint density of the observed and latent data is given by

$$\pi(\mathbf{Y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})\right\} \prod_{i=1}^{n} \xi(\omega_i), \tag{4}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$, and $\mathbf{X}_i = (1, \mathbf{E}_i', \mathbf{S}_i')'$. Under the probit link, $\mathbf{h} = (\omega_1, \ldots, \omega_n)'$, $\boldsymbol{\Omega} = \mathbf{I}$, 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, $\mathbf{h} = (\kappa_1/\omega_1, \ldots, \kappa_n/\omega_n)'$, $\kappa_i = Y_i - 1/2$, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$, 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. (2013), 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 $\mathbf{T}_l = (T_{l1}, \ldots, T_{lq_l})'$. Here, the rate parameterization 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

$$\begin{aligned}
Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}) = &-\frac{1}{2}E\{(\mathbf{h} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{h} - \mathbf{X}\boldsymbol{\beta}) \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}\}\\
&-\frac{1}{2}\beta_0^2 T_0^{-1} - \frac{1}{2}\sum_{l=1}^{2}\sum_{q=1}^{q_l}\beta_{lq}^2 E(T_{lq}^{-1} \mid \mathcal{D}, \boldsymbol{\beta}^{(d)}) + Q_r(\boldsymbol{\beta}^{(d)}),
\end{aligned} \tag{5}$$

where $Q_r(\boldsymbol{\beta}^{(d)})$ is a function that 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), the maximizer obtained in the M-step of the algorithm is given by

$$\boldsymbol{\beta}^{(d+1)} = (\mathbf{X}'\boldsymbol{\Omega}^{(d)}\mathbf{X} + \mathbf{D}^{(d)})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{(d)}\mathbf{h}^{(d)} = \text{argmax}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}), \tag{6}$$

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 forms of $\boldsymbol{\Omega}^{(d)}$ and $\mathbf{h}^{(d)}$ in (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 $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)'$ 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, El-Darzi, Vasilakis, and Millard (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) as well as to compute closed-form updates of the regression coefficients in the M-step given in (6).

From a computational perspective, the proposed approach has a few key attributes that 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 $p \gg n$, 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 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 an $n \times n$ matrix. Specifically, we may compute the necessary inversion in (6) as

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

where the inversion of $\mathbf{D}^{(d)}$ and $\boldsymbol{\Omega}^{(d)}$ 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, Kim, & Boyd, 2007) when fitting models for multiple specifications of the hyperparameters α and η.

3 | NUMERICAL STUDIES

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 2.2. The assessed characteristics include the method's ability to (a) identify significant covariates under various signal strengths, (b) accurately estimate the effect size of significant covariates, (c) classify covariates not related to the response as such, and (d) capably handle the complex data structures that are ubiquitous in GWAS. To accomplish this, data sets were simulated to mimic our motivating application; that is, 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 $T_0 = 1{,}000$ and we considered different values of the penalty parameters; that is, $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ and $\eta \in \{0.1, 0.2, 0.3\}$. These choices were made based on prior experience that 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 hyperparameter configurations. Thus, to choose the "best" from among them, we make use of

TABLE 1 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 s 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%		

the Bayesian information criterion (Neath & 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 1 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; that is, 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 that are truly zero but are identified to be nonzero by the MAP estimator. Table 2 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); that is, the model that 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; that is, 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 1 and 2 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 that are on the same order as the available sample size.

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

TABLE 2 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 s 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%		

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 consist of 173 observations that 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. Deoxyribonucleic acid (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, Walnut, California, USA). 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 1 provides a histogram of the Bayes factors associated with the 495,532 SNPs and Figure 2 provides a plot of the same across chromosomes. From this initial phase, and the results obtained in Section 3, we decided to focus attention on the top 200 SNPs; that is, the SNPs with largest associated Bayes factors. The sets of candidate SNPs are denoted as triangles in Figure 2. In Phase 2, we fit the following first-order model to the data:

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

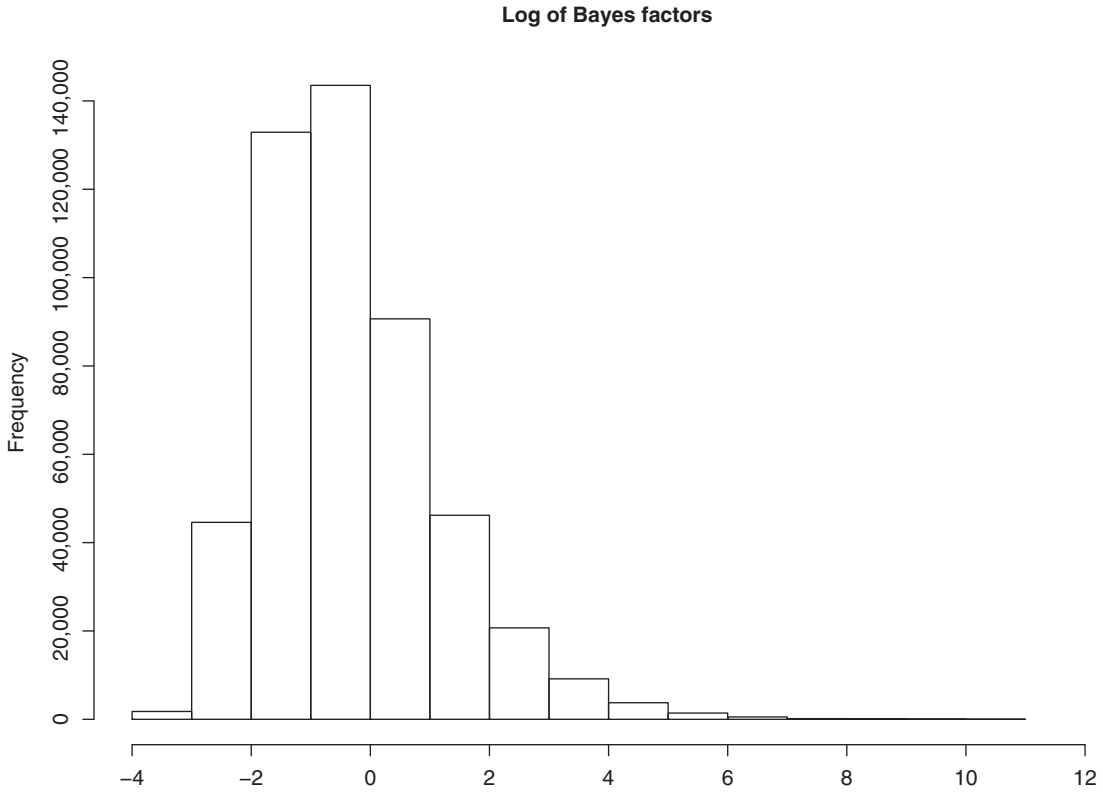

FIGURE 1 Histogram depicting the natural logarithm of the Bayes factors that were computed for each of 495,532 SNPs available in the CRC data

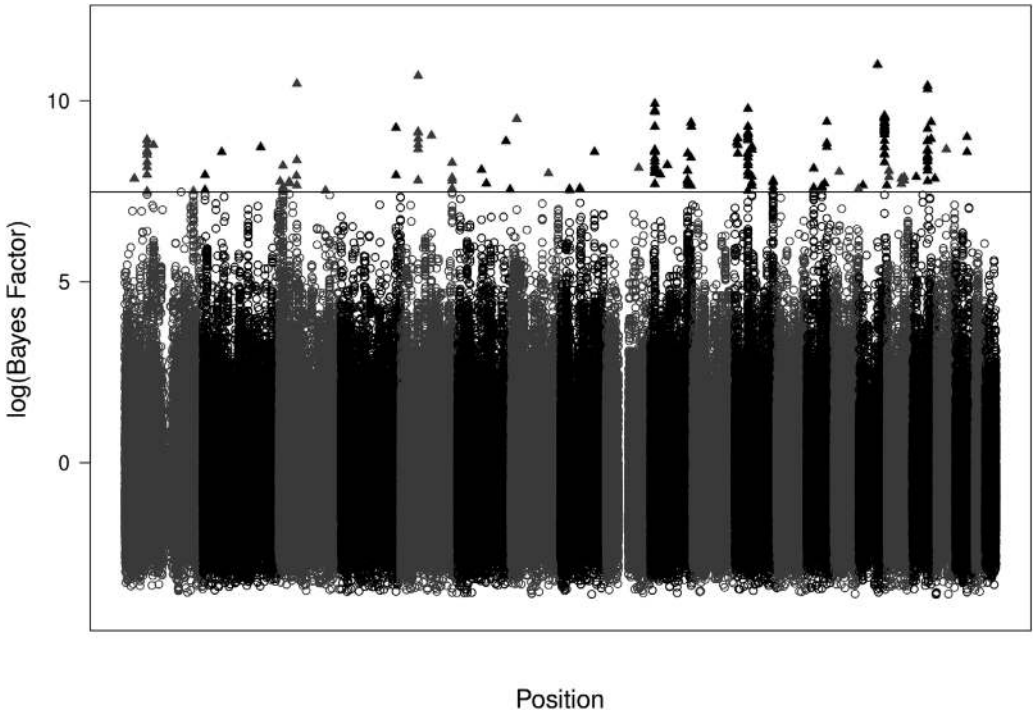

FIGURE 2 Plot of the natural logarithm of the Bayes factors that were computed for each of 495,532 SNPs verses their position in the genome. Each shade change represents the transition to a new chromosome and the black triangles above the horizontal line depict the 200 SNPs with the largest Bayes factors

TABLE 3 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

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, $\mathbf{E}_i = (E_{i1}, \ldots, E_{i4})'$, 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 hyperparameter-dependent MAP estimator for each considered configuration of (α, η), where $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ and $\eta \in \{0.1, 0.2, 0.3\}$ with $T_0 = 1,000$. Final model selection, as in Section 3, was guided by the Bayesian information criterion.

Table 3 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 that 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).

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 through an extensive numerical study, and was used to analyze the motivating cancer data. Through this application, 10 SNPs were identified to be associated with colorectal cancer via the proposed approach. To further disseminate this work, scripts written in R that 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 are 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; that is, once a set of candidate SNPs has been identified, additions in Phase 2 are not considered. For this reason, it could be of interest to merge the goals of Phases 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.

ACKNOWLEDGMENTS

The authors would like to thank the reviewer, associate editor, and editor for their helpful comments and suggestions. This research was partially supported by Grants R01 AI121351 and R44 AA027675 from the National Institutes of Health and Grant OIA-1826715 from the National Science Foundation. We acknowledge computing support from Amazon Web Services and NVIDIA.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID


Christopher McMahan https://orcid.org/0000-0001-5056-9615

REFERENCES

Abbi, R., El-Darzi, E., Vasilakis, C., & Millard, P. (2008). Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay. In *Proceedings of 2008 4th International IEEE Conference Intelligent Systems*, Vol. 1, pp. 3–9. IEEE9.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, *23*, 119–143.

Baurley, J. W., & Conti, D. V. (2013). A scalable, knowledge-based analysis framework for genetic association studies. *BMC Bioinformatics*, *14*, 312.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, … Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*, 291–295.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*, 2313–2351.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*, e1003348.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–20.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, *61*, 69–90.

Koh, K., Kim, S.-J., & Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, *8*, 1519–1555.

Liu, T.-H., Zheng, F., Cai, M.-Y., Guo, L., Lin, H.-X., Chen, J. W., … Xie, D. (2016). The putative tumor activator ARHGEF3 promotes nasopharyngeal carcinoma cell pathogenesis by inhibiting cellular apoptosis. *Oncotarget*, *7*, 25836–25848.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*, 199–203.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Peters, U., Bien, S., & Zubair, N. (2015). Genetic architecture of colorectal cancer. *Gut*, *64*, 1623–1636.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, *108*, 1339–1349.

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

Shi, Y., Chen, G.-B., Huang, X.-X., Xiao, C.-X., Wang, H.-H., Li, Y. S., … Guleng, B. (2015). Dragon (repulsive guidance molecule b, RGMb) is a novel gene that promotes colorectal cancer growth. *Oncotarget*, *6*, 20540–20554.

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

Wang, X., Philip, V. M., Ananda, G., White, C. C., Malhotra, A., Michalski, P. J., … Carter, G. W. (2018). A Bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset alzheimer's disease. *Genetics*, *209*, 51–64.

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

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*, 714–721.

Yazdani, A., & Dunson, D. B. (2015). A hybrid Bayesian approach for genome-wide association studies on related individuals. *Bioinformatics*, *31*, 3890–3896.

Zeggini, E., & Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, *10*, 191–201.

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

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

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

How to cite this article: Joyner C, McMahan C, Baurley J, Pardamean B. A two-phase Bayesian methodology for the analysis of binary phenotypes in genome-wide association studies. *Biometrical Journal*. 2020;62:191–201. https://doi.org/10.1002/bimj.201900050