G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2023;21(3):e28
https://doi.org/10.5808/gi.23047

Received: June 13, 2023
Revised: June 26, 2023
Accepted: June 27, 2023

*Corresponding author:
E-mail: htzhu@email.unc.edu

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Bayesian bi-level variable selection for genome-wide survival study

Eunjee Lee[1], Joseph G. Ibrahim[2], Hongtu Zhu[2]*
for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Information and Statistics, Chungnam National University, Daejeon 34134, Korea
[2]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

Mild cognitive impairment (MCI) is a clinical syndrome characterized by the onset and evolution of cognitive impairments, often considered a transitional stage to Alzheimer's disease (AD). The genetic traits of MCI patients who experience a rapid progression to AD can enhance early diagnosis capabilities and facilitate drug discovery for AD. While a genome-wide association study (GWAS) is a standard tool for identifying single nucleotide polymorphisms (SNPs) related to a disease, it fails to detect SNPs with small effect sizes due to stringent control for multiple testing. Additionally, the method does not consider the group structures of SNPs, such as genes or linkage disequilibrium blocks, which can provide valuable insights into the genetic architecture. To address the limitations, we propose a Bayesian bi-level variable selection method that detects SNPs associated with time of conversion from MCI to AD. Our approach integrates group inclusion indicators into an accelerated failure time model to identify important SNP groups. Additionally, we employ data augmentation techniques to impute censored time values using a predictive posterior. We adapt Dirichlet-Laplace shrinkage priors to incorporate the group structure for SNP-level variable selection. In the simulation study, our method outperformed other competing methods regarding variable selection. The analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data revealed several genes directly or indirectly related to AD, whereas a classical GWAS did not identify any significant SNPs.

Keywords: Bayesian variable selection, genome-wide association studies, group structure, linkage disequilibrium, survival analysis

Introduction

Mild cognitive impairment (MCI) is a clinical syndrome characterized by the onset and evolution of cognitive impairments. As 10%–15% of MCI patients develop Alzheimer's disease (AD) annually, MCI is commonly regarded as a transitional stage to AD. Identifying genetic characteristics among MCI patients who experience an accelerated progression to AD is important in enabling early diagnosis and facilitating drug discovery for AD. Genome-wide association studies (GWAS) are a standard tool for identifying single nucleotide polymorphisms (SNPs) associated with specific clinical conditions or outcomes. Researchers can delineate the genetic factors related to the rapid progression from MCI to AD as a phenotype in a GWAS by using the time of conversion from MCI to AD.

As approximately 500,000 to one million candidate SNPs exist, a GWAS deals with high-dimensional data, where the number of variables (SNPs in a GWAS) p is much greater than sample size n. A classical GWAS conducts several association tests, so called multiple testing, that examine an individual effect of each SNP on a clinical outcome. The

classical GWAS has two major limitations. First, GWAS has a multiple testing issue, which requires adequate control of false positives. Typically, the significance level of each test is adjusted by a Bonferroni correction. The significance level widely accepted to determine "genome-wide significant" association is 5×10^{-8} [1,2], which is a strict threshold and makes genome-wide significance difficult to be achieved. Second, the GWAS does not account for the intricate group structure among SNPs such as genes or linkage disequilibrium (LD) blocks. LD reflects how much an allele from a particular genetic variant is associated or inherited with an allele from another nearby genetic variant within the same population [3]. Incorporating the group information within the GWAS would increase statistical power by aggregating small effects of SNPs within a group.

To resolve the multiple testing issue, many statistical methods have been developed in terms of penalization [4-7], Bayesian variable selection [8,9], and sure independence screening strategy [10-12]. Researchers proposed variable selection methods by incorporating the group information to select genetic variants in both gene and SNP levels simultaneously [13-16]. While these methods are suitable for a continuous or binary outcome, only a limited number of studies for a time-to-event outcome are available. Bi et al. [17] developed a saddlepoint approximation implementation to correct p-values based on the Cox regression model. A Bayesian survival model with variable selection was proposed with application to GWAS [18]. Lin et al. [19] proposed kernel-machine SNP-set analysis to assess the group effect of each SNP-set on the survival time.

We propose a Bayesian bi-level variable selection (BBVS) method to detect SNPs associated with a time-to-event outcome by considering all the SNPs simultaneously and incorporating the group information of the SNP data, based on an accelerated failure time (AFT) model. Our method has two hierarchical levels of variable selection: the first level is group-wise and the second level is element-wise variable selection. In the first level, we identify important groups of variables by employing group inclusion indicators in the AFT model and update the censored event time from its predictive posterior distribution by data augmentation [18,20,21]. As this step generates posterior samples of censored time to event, their posterior mean will be used as an imputed value for the censored event time in the second level. In the second level, we only include variables in the selected groups in the first level as covariates in the regression model. To conduct element-wise variable selection, we adapt Dirichlet-Laplace shrinkage priors [22] to incorporate the group structure.

The rest of this paper is organized as follows. In the "Methods" section, we discuss our BBVS in the AFT model. In the Results and Discussion section, we present simulation study results to validate and compare the performance of BBVS with other group/bi-level selection methods. We discuss the real data analysis results for Alzheimer's Disease Neuroimaging Initiative (ADNI) data.

Methods

Accelerated failure time model

An AFT model is a parametric model to analyze a time-to-event outcome. While Cox regression postulates that covariates are multiplicatively related to the hazard, an AFT model assumes a direct relationship between time to event and covariates, which enables straightforward interpretation of regression coefficients. For the i-th subject, Y_i is survival time and $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{ip})'$ is a covariate vector. The first element of 1 allows estimation of the y-intercept. Subsequently, AFT model is given by $Y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})v_i$, $i = 1, \cdots, n$. It becomes the linear model in a log scale $\log Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$, $i = 1, \cdots, n$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ is a vector of $p + 1$ unknown regression coefficients including the y-intercept of β_0, and $\epsilon_i = \log v_i$ is an error term. Generally, the error term is assumed to follow the parametric distribution, such as normal distribution. The parametric AFT model is discussed extensively in [23-26] of the frequentist framework. Bayesian approaches were developed for the parametric AFT model [27-29]. In this paper, we consider a parametric Bayesian approach to model the error term ϵ_i with a normal distribution.

BBVS in accelerated failure time model

We propose a BBVS method on the AFT model. This method has two hierarchical levels of variable selection, the group-wise and the element-wise variable selection. It is motivated by natural grouping structures of SNPs, which can be captured by genes or LD blocks. By making use of the group structure in the model frame, we can efficiently select a small number of SNPs associated with a time-to-event outcome.

With predefined G blocks we can write our model as follows.

$$\log Y_i = \boldsymbol{x}_{i,0}'\boldsymbol{\beta}_0 + \sum_{g=1}^{G} \gamma_g \boldsymbol{x}_{i,g}'\boldsymbol{\beta}_g + \epsilon_i, i = 1, \cdots, n, \quad (1)$$

where $\boldsymbol{x}_{i,0} = (1, x_{i,1}^0, \cdots, x_{i,p_0-1}^0)$, $\boldsymbol{\beta}_0 = (\beta_{0,0}, \beta_{0,1}, \cdots, \beta_{0,p_0-1})$. For each g-th group of variables, $\boldsymbol{x}_{i,g} = (x_{i,1}^g, x_{i,2}^g, \cdots, x_{ik_g}^g)$, $\boldsymbol{\beta}_g = (\beta_{g,1}, \beta_{g,2}, \cdots, \beta_{g,k_g})$. Denote $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_G)$, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_G)$, where γ_g is an indicator variable having 0 or 1. When $\gamma_g = 1$, the g-th set of variables will be included in the model. If $\gamma_g = 0$, we remove the g-th group in the model construction. The covariates $x_{i,1}^0, \cdots, x_{i,p_0-1}^0$ are included in

https://doi.org/10.5808/gi.23047

the model to address their effects on the time to event. They can be clinical and demographic characteristics of subjects. The error term ϵ_i's are assumed to be independently distributed as $N(0,\sigma^2)$; hence, that the failure time Y_i follows a log-Normal distribution. When y_i is possibly right censored, we only observe $t_i = min(Y_i,c_i)$ and $v_i = I\{y_i<c_i\}$, where c_i is the censoring time. Here $w_i = log(y_i)$ can be considered as the augmented data such that

$$\begin{aligned} w_i &= \log(t_i) \;\; if \; \nu_i = 1, \\ w_i &> \log(t_i) \;\; if \; \nu_i = 0. \end{aligned} \tag{2}$$

Our bi-level variable selection method addresses two issues in the model (1): the selection of the relevant groups of SNPs and the imputation of the censored time to event y_i. In the first step, we identify important groups of variables by updating only the group inclusion vector γ and the censored time y_i from their posterior distributions. In the second step, the model (1) can be reduced by

$$log\, Y_i = \boldsymbol{x}_{i,0}^{'}\boldsymbol{\beta}_0 + \sum_{g=1}^{Q}\gamma_g \boldsymbol{x}_{i,g}^{*}{}^{'}\boldsymbol{\theta}_g + \epsilon_i, i = 1,\cdots,n, \tag{3}$$

where $\boldsymbol{x}_{i,g}^{*}$, $g = 1, 2, \cdots, Q$ are the Q selected groups in the first step, and $\boldsymbol{\theta}_g$, $g = 1, 2, \cdots, Q$ are the corresponding regression coefficient vectors. The censored time to event y_i is imputed by the mean of the posterior samples of w_i collected in the first step. It converts the AFT model to a usual log-linear regression problem.

We employ a shrinkage prior on the regression parameters $\boldsymbol{\theta}_g$ to enable the element-wise variable selection within $\boldsymbol{x}_{i,g}^{*}$, $g = 1, 2, \cdots, Q$. We consider a Dirichlet-Laplace (DL) prior proposed by Bhattacharya et al. [22] on the regression parameters and extend it to incorporate grouping information. As the regression parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}$ and the standard deviation σ of the error term are not of interests, the computational burden in the first step can be reduced by integrating out the irrelevant parameters, $\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma$ from the full posterior distribution. This kind of strategy has been employed in Sha et al. [20], although their variable selection has been conducted only in an element-wise fashion.

The first step: group-wise variable selection

In the first step, we consider the following conjugate priors.

$$\begin{aligned} \boldsymbol{\beta}_0|\sigma^2 &\sim N(0,\sigma^2 h_0 I_{p_0}) \\ \boldsymbol{\beta}_g|\sigma^2 &\sim N(0,c_0\sigma^2\textstyle\sum_g), g = 1,\cdots,G \\ \sigma^2 &\sim IG(v_0/2, v_0\sigma^2{}_0/2) \\ \gamma_j &\sim Bernoulli(p_j) \\ p_j &\sim Beta(a,b) \end{aligned}$$

In the prior, $X_0 = [\boldsymbol{x}_{1,0}, \cdots, \boldsymbol{x}_{n,0}]'$, $X_g = [\boldsymbol{x}_{1,g}, \cdots, \boldsymbol{x}_{n,g}]'$, $X = [X_0, X_1, \cdots, X_G]$, $\Sigma_g = (X_g^{'}X_g)^{-1}$ when $k_g \le n$ and $\Sigma_g = (X_g^{'}X_g+\lambda I_{kg})^{-1}$ when $k_g > n$ for the g-th group with size k_g. The prior on $\boldsymbol{\beta}_g$ is the Information Matrix (IM) or Information Matrix Ridge (IMR) prior proposed by Gupta and Ibrahim [30]. It is a generalization of Zellner's g-prior [31], while the IM prior is equal to the Zellner's g-prior in the Gaussian linear regression setting. The full posterior distribution of $(\boldsymbol{\beta}_g, \boldsymbol{\beta}, \gamma, \sigma^2)$ is given by,

$$\begin{aligned} L(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2|\boldsymbol{w},X) &\propto L(\boldsymbol{w}|X,\boldsymbol{\beta}_0,\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_0|\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\pi(\boldsymbol{\gamma})\pi(\sigma^2) \\ &\propto (\sigma^2)^{-n/2}exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(w_i - \boldsymbol{x}_{i,0}^{'}\boldsymbol{\beta}_0 - \sum_{g=1}^{G}\gamma_g \boldsymbol{x}_{i,g}^{'}\boldsymbol{\beta}_g\right)^2\right\} \\ &\quad \times (\sigma^2)^{-p_0/2}exp\left\{-\frac{1}{2h_0\sigma^2}\boldsymbol{\beta}_0^{'}\boldsymbol{\beta}_0\right\} \\ &\quad \times \prod_{g=1}^{G}(\sigma^2)^{-k_g/2}exp\left\{-\frac{1}{2c_0\sigma^2}\boldsymbol{\beta}_g'\Sigma_g^{-1}\boldsymbol{\beta}_g\right\} \times (\sigma^2)^{-\nu_0/2-1}exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \\ &\quad \times \prod_{g=1}^{G}p_g^{\gamma_g}(1-p_g)^{1-\gamma_g} \times \prod_{g=1}^{G}\frac{1}{B(a,b)}p_g^{a-1}(1-p_g)^{b-1} \end{aligned}$$

By integrating out $\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma^2$, we can obtain the posterior distribution of $\boldsymbol{\gamma}$:

$$\begin{aligned} L(\boldsymbol{\gamma}|w,X) &\propto \left\{\nu_0\sigma_0^2 + \boldsymbol{w}'\left(I + h_0X_0X_0' + c_0\sum_{g=1}^{G}\gamma_g X_g\Sigma_g X_g'\right)^{-1}\boldsymbol{w}\right\}^{-\frac{n+\nu_0}{2}} \\ &\quad \times \prod_{g=1}^{G}p_g^{\gamma_g}(1-p_g)^{1-\gamma_g} \end{aligned}$$

For a given $\boldsymbol{\gamma}_{(g)} = (\gamma_1, \cdots, \gamma_{(g-1)}, \gamma_{(g+1)}, \cdots, \gamma_G)$, the posterior distribution of γ_g is the Bernoulli distribution with success probability $\frac{A}{A+B}$, where

$$\begin{aligned} A &= f_t(\boldsymbol{w}|\nu_0, \sigma_0(A_{\gamma_{(g)}} + c_0\gamma_g X_g\Sigma_g X_g')) \times p_g, \\ B &= f_t(\boldsymbol{w}|\nu_0, \sigma_0 A_{\gamma_{(g)}}) \times (1-p_g), \end{aligned}$$

and $A_{\gamma_{(g)}} = I + h_0X_0X_0^{'} + c_0\sum_{k\neq g}^{G}\gamma_k X_k\Sigma_k X_k^{'}$. The function $f_t(\cdot\,|\nu,\sigma^2)$ denotes the probability density function of t-distribution with the degrees of freedom ν and the scale parameter σ^2. Then, update p_g from its posterior distribution $Beta(a+\gamma_g, b+1-\gamma_g)$. The marginal likelihood of the augmented data w can be derived as

$$L(\boldsymbol{w}|X,\boldsymbol{\gamma}) \propto \left\{1 + \frac{1}{\nu_0\sigma_0^2}\boldsymbol{w}'\left(I + h_0X_0X_0^{'} + c_0\textstyle\sum_{g=1}^{G}\gamma_g X_g\Sigma_g X_g'\right)^{-1}\boldsymbol{w}\right\}^{-\frac{\nu_0+n}{2}},$$

which is proportional to the truncated n-dimensional multivariate t-distribution with truncation given by (2) as follows.

$$\boldsymbol{w}|X,\boldsymbol{\gamma}\sim t_n\left[\nu_0,\mathbf{0},\sigma_0^2\left(I+h_0X_0X_0^{'}+c_0\sum_{g=1}^{G}\gamma_gX_g\Sigma_gX_g'\right)\right]$$

By using the full conditional distribution of w_i for a censored case v_i=0, the censored time w_i can be imputed by its posterior mean. Denote $H_\gamma = I + h_0X_0X_0' + c_0\sum_{g=1}^{G}\gamma_gX_g\Sigma_gX_g'$, where $h_{i,j}$ is a scalar element in i-th row, j-th column of H_γ and $H_{(i,j)}$ is the matrix H_γ without its i-th row and j-th column, and $\boldsymbol{h}_i^{(i)}$ is the i-th row of H_γ without its i-th element. Similarly, let $\boldsymbol{w}^{(i)}$ be the vector $\boldsymbol{w}$ without its i-th element. When w_i is censored, its full conditional distribution can be written as a truncated t location-scale distribution such that

$$w_i|\boldsymbol{w}^{(i)},X,\boldsymbol{\gamma}\sim t_{n+\nu_0-1}(\mu_{w_i},s_{w_i}),w_i>\log(t_i) \quad (4)$$

where μ_{w_i}, s_{w_i}, and $n+\nu_0-1$ are respectively the location, scale, and degrees of freedom parameters. The location and scale parameters are given by

$$\mu_{w_i} = \boldsymbol{h}_i^{(i)}H_{(i,j)}^{-1}\boldsymbol{w}_i^{(i)},$$

$$s_{w_i} = \sqrt{\left(h_{(i,i)}-\boldsymbol{h}_i^{(i)}H_{(i,i)}^{-1}\boldsymbol{h}_i^{(i)'}\right)\left(\nu_0\sigma_0^2+\boldsymbol{w}_i^{(i)}H_{(i,i)}^{-1}\boldsymbol{w}_i^{(i)}\right)/(n+\nu_0-1)}.$$

The censored w_i will be updated from (4) at each iteration and it will be imputed as their posterior mean in the element-wise selection step.

After running Gibbs sampling with M iterations, posterior inclusion probability can be calculated from the posterior sample of γ as their posterior mean, $\widehat{p_g}=\frac{1}{M}\sum_{m=1}^{M}\gamma_g^{(m)}$. The posterior inclusion probability $1-\widehat{p_g}$ can be considered as Bayesian q-values, or estimates of the local false discovery rate (FDR) [32,33], because they measure the probability of a false positive if the g-th group is "decided" to be included in the model. To select important groups, for some threshold p^*, we consider that any group with $\widehat{p_g}\geq p^*$ is relevant and will include them in our model. We determined the threshold p^* to control the average Bayesian FDR by using the method proposed by Morris et al. [34].

The second step: element-wise variable selection

In this step, we include all the variables of the Q selected groups in the first step and assume shrinkage priors on the regression parameters $\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_Q$ to achieve further sparsity in the element-wise level in the reduced model (3). As a shrinkage prior, the DL prior is assumed and extended to incorporate grouping information. The DL prior has been proposed in [22] as a novel form of shrinkage prior. Under the normal means setting

$$y_i=\theta_i+\epsilon_i,\epsilon_i\sim N(0,1),1\leq i\leq p,$$

the true signal θ_i has a DL prior, which has a hierarchical structure such that

$$\theta_j|\psi_j,\phi_j,\tau\sim N(\psi_j\phi_j{}^2\tau^2),\ \psi_j\sim Exp(1/2), \quad (5)$$
$$\boldsymbol{\phi}\sim Dir(a,\cdots,a),\tau\sim Gamma(pa,1/2),$$

where $\boldsymbol{\phi}=(\phi_1,\phi_2,\cdots,\phi_p)$. To efficiently control the global shrinkage, they introduced global (τ) and local ($\boldsymbol{\phi}$) scales, where the local scales have a joint structure such that they lie in the (p-1) dimensional simplex. Under the moderate-sized coefficients with sparse signal setting, their simulation study has shown that the DL prior outperforms least absolute shrinkage and selection operator (Lasso), Bayesian Lasso, empirical Bayes median, and point mass prior, while its performance is similar to that of Horseshoe prior.

In our model framework, we have prespecified grouping information. In order to get more flexibility depending on the grouping structure, we allow the hyperparameters ψ_j, ϕ_j, and τ in (5) to be group index(g)-dependent. In the selected group g , there are q_g variables and the total number of selected variables in the model (3) is $q=\sum_{g=1}^{Q}q_g$. Here, we impute w by the posterior mean $\widetilde{\boldsymbol{w}}$ obtained from the group-wise selection step. For g = 1, 2, ⋯, Q, the priors are set to be

$$\boldsymbol{\theta}_g|\sigma^2,\boldsymbol{\psi}_g,\boldsymbol{\phi}_g,\tau_g\sim N(0,\sigma^2\Sigma_g^*) \quad (6)$$

$$\sigma^2\sim IG(\nu_0/2,\nu_0\sigma_0^2/2)$$

$$\psi_{gj}\sim Exp(1/2),j=1,\cdots,q_g$$

$$(\phi_{g1},\cdots,\phi_{gq_g})\sim \underset{a}{Dir}(\underset{a}{\alpha_g},\cdots,\underset{a}{\alpha_g})$$

$$\tau_g\sim \underset{a}{gamma}(q_g\underset{a}{\alpha_g},1/2)$$

$$\underset{a}{\alpha_g}\sim \text{Discrete uniform from } \frac{1}{q_g} \text{ to } \frac{1}{2} \text{ with length 50,}$$

where $\Sigma_g^*=\text{diag}\left(\psi_{g1}\phi_{g1}^2\tau_g^2,\cdots,\psi_{gq_g}\phi_{gq_g}^2\tau_g^2\right)$, $\boldsymbol{\psi}_g=(\psi_{g1},\cdots,\psi_{gq_g}),\boldsymbol{\phi}_g=(\phi_{g1},\cdots,\phi_{gq_g})$. Here $IG(a,b)$ denotes the inverse agamma distribution with shape parameter α and the rate parameter b.

Denote $\boldsymbol{x}_{i*}^{'}=\left(\boldsymbol{x}^{*}{}_{i,1}^{'},\cdots,\boldsymbol{x}^{*}{}_{i,Q}^{'}\right),\boldsymbol{\theta}'=\left(\boldsymbol{\theta}_1',\cdots,\boldsymbol{\theta}_Q'\right)$, $\boldsymbol{\phi}'=\left(\boldsymbol{\phi}'{}_1,\cdots,\boldsymbol{\phi}'{}_Q\right)$, $\boldsymbol{\psi}'=\left(\boldsymbol{\psi}'{}_1,\cdots,\boldsymbol{\psi}'{}_Q\right)$, and $\boldsymbol{\tau}=(\tau_1,\cdots,\tau_Q)$. The design matrix is given by $X_g^*=\left[\boldsymbol{x}_{1,g}^*,,\cdots,\boldsymbol{x}_{n,g}^*\right]'$ for each g, and $X^*=\left[X_1^*,\cdots,X_Q^*\right]$ for all the groups. Σ^* is a block diagonal matrix with element matrices $\Sigma_1^*,\cdots,\Sigma_Q^*$. By combining (3) and (6), the posterior distribution can be obtained as

https://doi.org/10.5808/gi.23047

$$L(\boldsymbol{\beta}_0, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\tau} | \tilde{\boldsymbol{w}}, X^*) \quad (7)$$

$$\propto L(\tilde{\boldsymbol{w}} | X_*, \boldsymbol{\beta}_0, \boldsymbol{\theta}, \sigma^2) \pi(\boldsymbol{\beta}_0 | \sigma^2) \pi(\boldsymbol{\theta} | \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\tau}) \pi(\sigma^2) \pi(\boldsymbol{\phi}) \pi(\boldsymbol{\psi}) \pi(\boldsymbol{\tau})$$

$$\propto (\sigma^2)^{-n/2} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\widetilde{w_i} - \boldsymbol{x}_{i,0}^{'} \boldsymbol{\beta}_0 - \boldsymbol{x}_{i*}^{'} \boldsymbol{\theta} \right)^2 \right\}$$

$$\times exp\left\{ -\frac{1}{2h_0 \sigma^2} \boldsymbol{\beta}_0^{'} \boldsymbol{\beta}_0 \right\}$$

$$\times \det(\sigma^2 \Sigma^*)^{-1/2} exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{\theta}^{'} \Sigma^{*-1} \boldsymbol{\theta} \right\} \times (\sigma^2)^{-\nu_0/2-1} exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right)$$

$$\times exp\left(-\frac{\Sigma_{g=1}^{Q} \Sigma_{j=1}^{q_g} \psi_{gj}}{2} \right) \times \prod_{g=1}^{Q} \left(\frac{1}{B(\boldsymbol{\phi}_g)} \prod_{j=1}^{q_g} \phi_{gj}^{a_g - 1} \right)$$

$$\times \prod_{g=1}^{Q} \left\{ \tau_g^{q_g a_{g-1}} exp\left(-\frac{\tau_g}{2} \right) \right\},$$

where $B(\phi_g)$ denotes a multivariate Beta function. We propose a Gibbs sampler for posterior computation, which enables parameter estimation and variable selection simultaneously. The Gibbs sampler is computationally efficient and mixes rapidly. We first specified the hyperparameters h_0, σ_0, ν_0, a_1, ⋯, a_g at appropriate values. Starting from the initiation step, the Gibbs sampler for the model (3) and (7) proceeds as follows:

1. Update β_0 according to its full conditional distribution

$$p(\boldsymbol{\beta}_0 | -) \sim N_{p_0} \left(\left(X_0' X_0 + \frac{1}{h_0} \boldsymbol{I} \right)^{-1} X_0' (\tilde{\boldsymbol{w}} - X^* \boldsymbol{\theta}), \right.$$
$$\left. \sigma^2 \left(X_0' X_0 + \frac{1}{h_0} I \right)^{-1} \right).$$

2. Update $\boldsymbol{\theta}_g$ from its full conditional distribution $N_{q_g}(\widetilde{\boldsymbol{\mu}_g}, \tilde{\Sigma}_g)$, where

$$\widetilde{\boldsymbol{\mu}_g} = \left(X_g^{*'} X_g^* + \Sigma_g^{*-1} \right)^{-1} X_g^{*'} (\tilde{\boldsymbol{w}} - X_0 \boldsymbol{\beta}_0 - X_{(g)}^* \boldsymbol{\theta}_{(g)}),$$

$$\tilde{\Sigma}_g = \sigma^2 \left(X_g^{*'} X_g^* + \Sigma_g^{*-1} \right)^{-1}.$$

The design matrix $X_{(g)}^*$ is $\left[X_1^*, \cdots, X_{g-1}^*, X_{g+1}^*, \cdots, X_Q^* \right]$, and the regression coefficient vector $\boldsymbol{\theta}_{(g)}$ is $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{g-1}, \boldsymbol{\theta}_{g+1}, \cdots, \boldsymbol{\theta}_Q)$.

3. Let $N = n + q + p_0 + \nu_0$ and $\boldsymbol{\eta} = X_0 \boldsymbol{\beta}_0 + X^* \boldsymbol{\theta}$. Update σ^2 from

$$p(\sigma^2 | -) \sim IG\left(\frac{N}{2}, \frac{1}{2} \left\{ \nu_0 \sigma_0^2 + \| \tilde{\boldsymbol{w}} - \boldsymbol{\eta} \|^2 + \frac{\boldsymbol{\beta}_0' \boldsymbol{\beta}_0}{h_0} + \boldsymbol{\theta}' (\Sigma^*)^{-1} \boldsymbol{\theta} \right\} \right)$$

4. Independently sample ψ_{gj} from its full conditional distribution

$$p(\psi_{gj} | -) \sim IG\left(\frac{\phi_{gj} \tau_g \sigma}{|\theta_{gj}|}, 1 \right).$$

5. Update τ_g from its full conditional distribution, the generalized inverse Gaussian distribution (giG), such as

$$p(\tau_g | -) \sim giG\left(q_g \times a_g - q_g, 1, 2 \sum_{j=1}^{q_g} \frac{|\theta_{gj}|}{\phi_{gj} \sigma} \right).$$

6. Update ϕ_{gj}, where $\phi_{gj} = T_{gj} / T_g$ such that

$$p(T_{gj} | -) \sim giG\left(a_g - 1, 1, 2 \frac{|\theta_{gj}|}{\sigma} \right).$$

7. Update a_g from $\mathrm{MN}(1, \widetilde{p_1}/\tilde{p}, \cdots, \widetilde{p_{50}}/\tilde{p})$, where $\tilde{p} = \sum_{l=1}^{50} \widetilde{p_l}$ and

$$\tilde{p} = exp\left((u_l - 1) \sum_{j=1}^{q_g} log(\phi_{gj}) + (q_g u_l - 1) log(\tau_g) - log\, 5\, 0 \right).$$

As the DL prior does not give exactly zero coefficient value, an additional step is needed to select relevant variables. We followed a simple approach to choose important variables using k-means clustering [22]. Two clusters of $|\theta_j|$'s can exist, where (a) one cluster has nearly zero coefficient values while (b) another cluster has relatively bigger absolute coefficients away from zero. The clusters (a) and (b) can be considered as noise and signal, respectively. We cluster $|\theta_j|$'s at each Markov chain Monte Carlo iteration using k-means with $k = 2$ clusters. At each i-th iteration, the number of important variables h_i is set to be the smaller cluster size out of the two clusters. Subsequently, the number of important variables is finally estimated by taking the mode from the whole Markov chain Monte Carlo (MCMC) iterations, i.e., $H = \mathrm{mode}\{h_i\}$. The H largest elements of the absolute values of posterior medians $|\theta|$ are identified as the important variables.

ADNI-1 data

To reveal SNPs associated with the time of conversion to AD from MCI, we analyzed ADNI data obtained from the ADNI database (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. The initial 5-year ADNI study resulted in the ADNI-1 data.

We performed quality control (QC) steps on the raw genotype data to ensure that only high-quality data were included in the final analysis. QC procedures include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification,

(4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency, and (6) population stratification. The second line preprocessing steps include removal of SNPs with (1) more than 5% missing values, (2) minor allele frequency smaller than 5%, and (3) Hardy-Weinberg equilibrium p-value $<10^{-6}$. The remaining missing genotype variables were imputed as the modal value. After the QC procedures, 347 subjects and 494,564 SNPs remained in the current study. The above procedures were carried out in PLINK version 1.9. We also calculated the LD blocks to form the SNP-sets and remove SNP-sets with a single SNP. Eventually, 421,823 SNPs were left in our analysis grouped into 16,084 SNP-sets.

We study the subjects diagnosed with MCI at the baseline visit. If an MCI patient does not progress to AD within 48 months from the baseline, we define the time of conversion of the patient as "censored." For non-censored cases, the conversion time is determined by the difference between the baseline and the time of visit when the patient was diagnosed with AD.

Simulation data

We generated simulation data to examine the performance of the BBVS in the AFT model. To convey the correlation structure of SNP data in practice, our SNP data is simulated from the Hapmap projects 2009 phase III data [35]. For each subject, we randomly combined two haplotypes from the Centre d'étude du polymorphisme humain population to form its genotypes and used PLINK [36] to form SNP-sets by determining LD blocks. Among the blocks that were >30, we randomly selected 2,000 SNP-sets in each block, which results in about 86,000 SNPs. After removing those SNP data with duplicated columns, we have about 45,000 SNPs in total.

We considered two cases: non-censored data and censored data. In the non-censored case, the time to event outcome was generated from the model (1), where $\gamma_j = 1, j = 1, \cdots, 10$ and $\gamma_j' = 0, j' = 11, \cdots, 2{,}000$. Within the 10 relevant blocks, we randomly selected 10 SNPs and assumed an additive model. The additive model assumes that a uniform, linear increase in risk for each copy of the minor allele exists. The corresponding non-zero regression coefficients were generated from $N(-1, 0.5)$, which mimics the situation wherein a single copy of the minor allele decreases the time to event in relation to major allele. In the censored case, the censored event times were independently generated from a uniform distribution from 0 to c^*. The value of c^* was set to achieve a desired censoring rate. We replicated the simulation 50 times under the same setting. We assumed the inclusion indicator $\gamma_g \sim Beta(10,190)$, which gives average 5% of inclusion probability to reflect prior information that the important signal is sparse in the GWAS.

Results and Discussion

ADNI-1 data analysis

We applied BBVS on the ADNI-1 data to reveal SNPs associated with the time of conversion to AD from MCI. Other than the whole SNPs data, we also included demographic and clinical characteristics measured at the baseline, such as gender, age, handedness, marital status, education length, retirement, and Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) score. The first 5 principal components of the SNP were included to adjust for population stratification in the model [37]. The variable selection was only performed on the SNP data.

We determined the threshold a to control the average Bayesian FDR [34] and consider any group whose posterior inclusion probability is greater than that of a. In the ADNI data, the threshold is calculated by 0.941 (Fig. 1). In total, 19 SNP-sets were detected as important groups and 106 SNPs were identified by the elementwise-level selection. Fig. 2. shows the estimated coefficient values for 795 SNPs included in the 19 SNP-sets. We colored 106 SNPs selected in the element level in red.

Supplementary Figs. 1–3 show trace plots of the regression coefficients $\theta_{1,1}$, $\theta_{1,2}$, and $\theta_{1,3}$ of the first selected SNP-set for 5,000 iterations of the MCMC algorithm. They show fast convergence of the algorithm, indicating its good mixing properties.

We summarized the variable selection results of BBVS to present which genes are involved in Table 1. Among them, four genes have been reported in other studies to be related to AD directly or indirectly. Dipeptidyl-peptidase 10 (DPP10) is known to modulate the electrophysiological properties, cell-surface expression, and subcellular localization of voltage-gated potassium channels [38]. Chen et al. [39] demonstrated that aggregation of DPP10 was related to neurodegenerative disorders including AD, diffuse Lewy body disease, and fronto-temporal dementia. In addition, DPP10 had robust reactivity within neurofibrillary tangles and plaque-associated dystrophic neurites in AD brains, which suggested that it is involved in the pathology of AD [40]. All the findings indicate that DPP10 is associated with a risk to develop AD in a direct or indirect manner. THSD7B has been reported to be associated with age-related cognitive decline based on repeated measures of 17 cognitive tests [41]. In addition, several linkage mappings have identified VPS26A to be associated with AD [42]. Sidekick cell adhesion molecule 1 was reported as a susceptibility gene for hypertension in Japanese individuals [43], where hypertension moderately increased risk of AD [44].

https://doi.org/10.5808/gi.23047

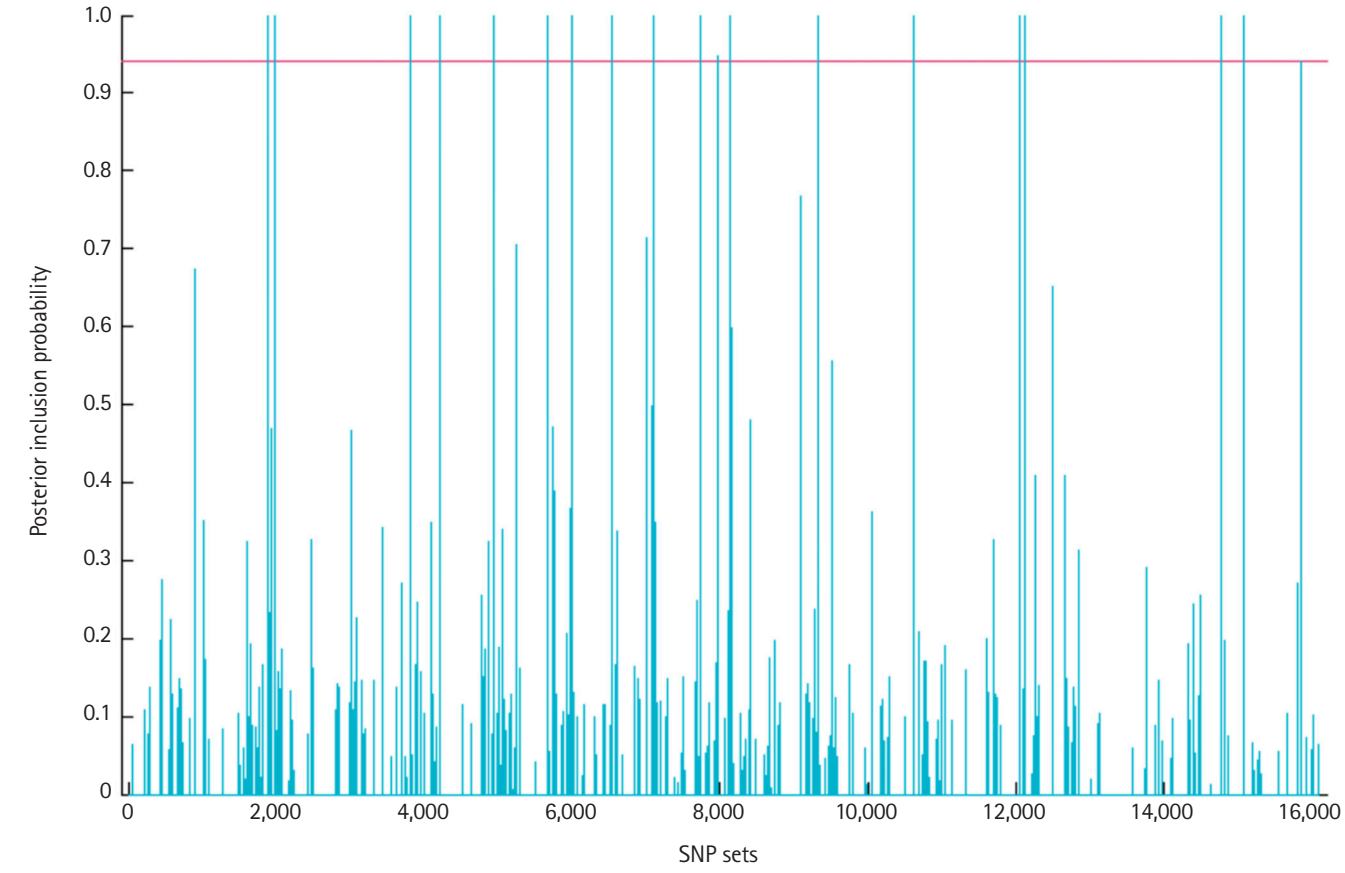

Fig. 1. Posterior inclusion probabilities of 16,106 SNP-sets. Our proposed method identified 19 important SNP-sets after Bayesian FDR correction. The solid line shows the FDR criteria, 0.941 in this data. SNP, single nucleotide polymorphism; FDR, false discovery rate.

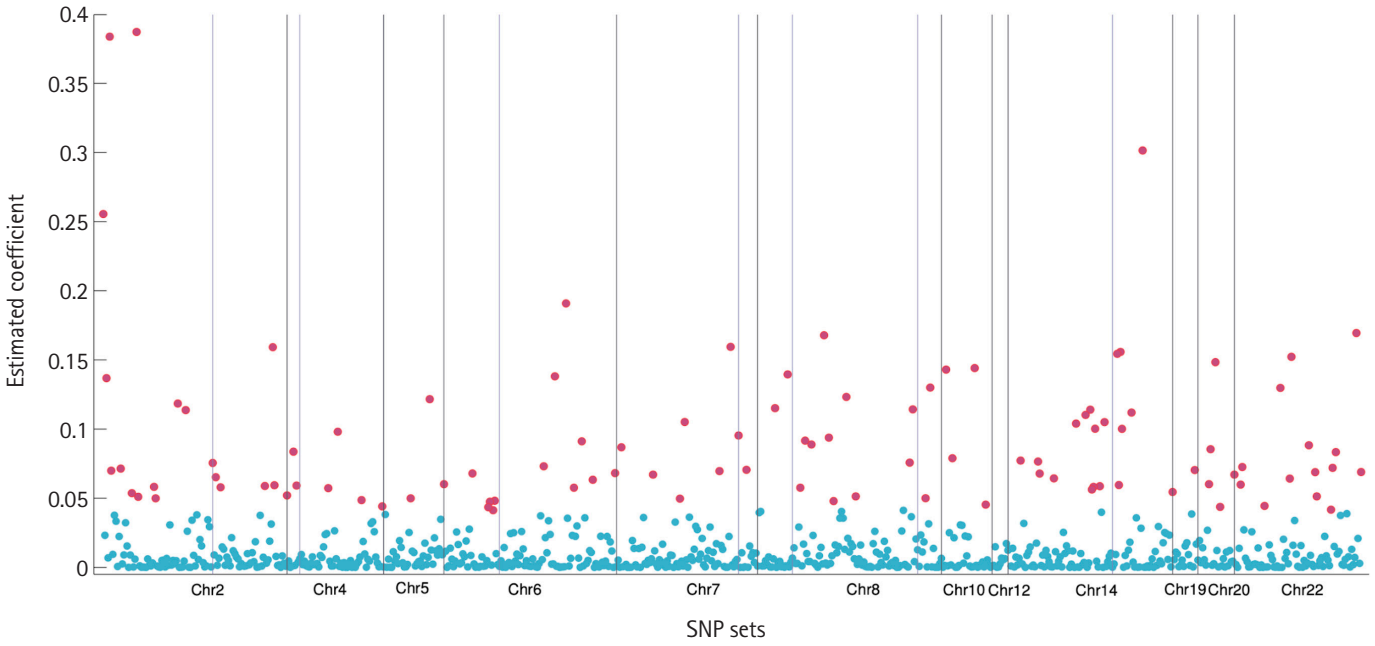

Fig. 2. Estimated coefficient values for 795 SNPs included in the 19 SNP-sets. We colored 106 SNPs selected in the element level in red. SNP, single nucleotide polymorphism.

Table 1. LD blocks and genes detected by BBVS

Chr	Begin (bp)	End (bp)	No. of SNPs	No. of selected	Genes
2	50596	50665	70	13	***DPP10***
2	53530	53576	47	6	***THSD7B***
4	104778	104785	8	2	*ATP8A1*
4	117728	117780	53	4	***FREM3***, *LOC101927636, GYPA*
5	135218	135255	38	3	-
6	154825	154859	35	5	-
6	166701	166774	74	7	-
7	181879	181955	77	7	*SDK1*
7	197664	197675	12	1	-
8	216741	216762	22	2	*PREX2*
8	224172	224250	79	11	*TNFRSF11B*, ***COLEC10***
8	228413	228427	15	2	***TRAPPC9***
10	261007	261038	32	4	*SRGN*, ***VPS26A***, *SUPV3L1, HKDC1*
12	294754	294763	10	0	*CLEC2A, KLRF2*
14	332351	332416	66	12	***HEATR5A***, *DTD2*, ***NUBPL***
14	334919	334956	38	7	-
19	394149	394164	16	1	*CPAMD8, HAUS8, MYO9B*
20	399491	399513	23	5	-
22	22468984	22671741	80	14	*VPREB1*, ***BMS1P20***

LD, linkage disequilibrium; BBVS, Bayesian bi-level variable selection; SNP, single nucleotide polymorphism.

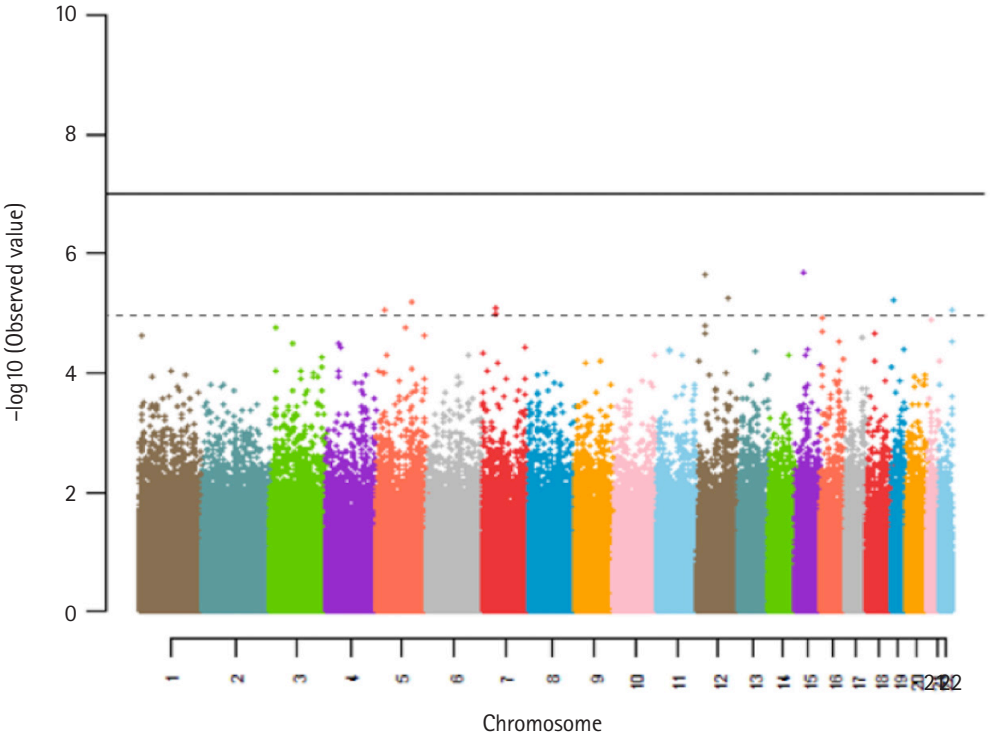

Fig. 3. A Manhattan plot with −log10(p-value) for the classical genome-wide association study. The solid and dotted lines show the 5×10^{-8} significance level and the 1×10^{-5} significance level, respectively.

https://doi.org/10.5808/gi.23047

For comparison purposes, we conducted two different types of GWASs: (1) a simple GWAS, multiple testing on each SNP and (2) kernel-machine SNP-set GWAS with the linear kernel [19]. Fig. 3. shows a Manhattan plot with –log 10(p-value) for the simple GWAS. The solid and dotted lines represent the genome-wide significance level and the suggestive significance level, respectively. Our study identified 9 SNPs at the 1×10^{-5} suggestive significance level, where none of them had been reported in previous GWASs. Supplementary Table 1 shows the p-values of 106 SNP selected by the element-wise variable selection of the proposed method. None of them were significant at the suggestive significance level. For the kernel-machine method, we considered three types of kernel functions such as the linear kernel, the identical by state (IBS) kernel, and the quadratic kernel. Table 2 shows the SNP-sets selected by the kernel-machine method at the 5×10^{-8} significance level. The selected SNP-sets vary with the type of kernel. The linear kernel, the IBS kernel, and the quadratic kernel selected 5, 8, and 6 SNP-sets, respectively. Two genes, calmodulin-binding transcription activator 1 (*CAMTA1*) and *RBFOX1*, were related to Alzheimer's disease. The linear kernel selected an SNP set located within *CAMTA1*. Huentelman et al. [45] identified SNPs within the *CAMTA1* gene that were significantly related to memory performance and memory-related regions on the human brain, which could be considered potential biomarkers of AD. *RBFOX1* was identified under the quadratic kernel function. Hooli et al. [46] reported that the gene co-segregates with disease status within early-onset familial AD and early or mixed-onset AD families. There were no overlapped SNPs among the three methods.

Simulation study

As the AFT model for non-censored data is the log-normal regression model, we can compare the performance of variable selection with other variable selection methods implemented based on the typical regression models. For competing methods, we considered the group Lasso (grLasso) [47], the group MCP(grMCP) [48], the group bridge (gBridge) [49], the group exponential lasso (gel) [50], the composite MCP (cMCP) [51] penaltues. The cMCP, gel, and gBridge penalties carry out bi-level selection, meaning that they carry out variable selection at the group level and at the level of individual covariates. The grLasso, grMCP, and grSCAD penalties carry out variable selection only at the group level, meaning that within a group, coefficients will either all be zero or all non-zero. We used Bayesian Information Criteria to select the tuning parameter value for each method.

We consider the following performance measurements: true positive rate (*TPR* or sensitivity), true negative rate (*TNR* or specificity), positive predictive value (*PPV*), and negative predictive value (*NPV*). They are defined as follows.

Table 2. LD blocks and the corresponding SNPs detected by the kernel-machine method

Chr	SNP	Gene	p-value		
			Linear	IBS	Quadratic
1	rs12128469, rs12402763, rs7543711, rs12563394, rs2301461, rs2301462	*CAMTA1*	5.00e-09	1.00e-04	6.00e-04
2	rs6545731, rs10169309		5.00e-09	2.00e-04	5.00e-09
2	rs2576778, rs880427	*FHL2*	4.00e-04	5.00e-09	2.10e-03
3	rs9288812, rs10511245, rs2053627		5.00e-09	3.00e-04	6.00e-04
3	rs6796883, rs293779	*CPNE9*	5.00e-04	5.00e-09	5.00e-03
3	rs307560, rs307558	*SYN2*	2.00e-04	5.00e-09	1.10e-03
3	rs6768031, rs1033222, rs1991443, rs1991442, rs1427840, rs11922896, rs6439279, rs748155, rs17275526, rs755568, rs1863916	*NEK11*	2.00e-04	5.00e-09	1.10e-03
5	rs6888634, rs2577531		5.00e-09	5.00e-09	2.00e-04
9	rs6475646, rs10733377		2.00e-04	4.00e-04	5.00e-09
13	rs11164144, rs944899		5.00e-09	1.00e-04	1.00e-04
13	rs9508716, rs1275190, rs1275191, rs9508717, rs1314940	*LINC00426*	5.00e-04	1.00e-03	5.00e-09
13	rs3764109, rs9530253	*KLF12*	2.00e-04	5.00e-09	4.00e-04
14	rs11850328, rs174994, rs8014403		8.00e-04	4.00e-04	5.00e-09
16	rs12149339, rs12933074, rs12934725, rs11861289	*RBFOX1*	2.00e-04	5.00e-09	5.00e-09
17	rs1990185, rs17772608, rs11077582, rs12951391, rs978425, rs16977009, rs12941303, rs2158917, rs12941651, rs7213040, rs16977023, rs12709255, rs17767678, rs7214582	AC003051.1	8.00e-04	5.00e-09	1.94e-02
21	rs2831525, rs6516819	LOC101927973	3.00e-04	7.00e-04	5.00e-09

LD, linkage disequilibrium; SNP, single nucleotide polymorphism; IBS, identical by state.

Table 3. Group-wise variable selection performance of BBVS and other competing methods

		TPR	TNR	PPV	NPV
Group-level	BBVS	1.000 (0.000)	1.000 (0.000)	0.996 (0.003)	1.000 (0.000)
	gBridge	0.986 (0.005)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	gel	0.980 (0.007)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	grMCP	0.998 (0.002)	1.000 (0.000)	0.972 (0.009)	1.000 (0.000)
	grSCAD	1.000 (0.000)	1.000 (0.000)	0.472 (0.007)	1.000 (0.000)
	grLASSO	1.000 (0.000)	0.990 (0.000)	0.348 (0.007)	1.000 (0.000)
	cMCP	1.000 (0.000)	0.963 (0.002)	0.144 (0.011)	1.000 (0.000)
Element-level	BBVS	0.686 (0.012)	0.999 (0.000)	0.616 (0.009)	1.000 (0.000)
	gBridge	0.643 (0.011)	0.999 (0.000)	0.503 (0.008)	1.000 (0.000)
	gel	0.651 (0.012)	0.999 (0.000)	0.441 (0.009)	1.000 (0.000)
	grMCP	0.306 (0.009)	0.998 (0.000)	0.165 (0.009)	0.999 (0.000)

BBVS, Bayesian bi-level variable selection; TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value.

$$TPR=\frac{TP}{10},\ TNR=\frac{TN}{1990},\ PPV=\frac{TP}{TP+FP},\ NPV=\frac{TN}{TN+FN}$$

where the *TP* and *TN* are the number of correctly identified significant variables and the number of correctly rejected non-significant variables, respectively. The *FP* and *FN* denote the number of identified non-significant variables and the number of rejected significant variables, respectively. Under the true model, $TP = 10$, $TN = 1990$, and $FP = FN = 0$, which implies that all the four rates are equal to one.

Table 3 shows the group-level and element-level variable selection results for the non-censored case. The average values of the performance measurements are presented with Monte Carlo standard errors in the parenthesis. Our method achieves the highest values of all the criteria,*TPR*, *TNR*, *NPV*, and *PPV* compared with other group penalty methods by removing the irrelevant groups consistently and selecting important groups very well. As the group penalties with only group-level selection especially grSCAD, grLasso tend to select groups more generously, they select important groups perfectly while the numbers of true positive cases are much bigger than other methods. The bi-level selection penalties, gBridge, and gel show comparative performance to our proposed method. In terms of the element-wise variable selection, our method yields the highest values of all the criteria,*TPR*, *TNR*, *NPV*, and *PPV* compared with other group penalty methods enabling bi-level selection. As the important signals are sparse, all the bi-level methods perform very well in terms of removing irrelevant signals.

The BBVS also shows satisfactory performance in terms of selecting important variables in the censored case. The average values of *TPR*, *TNR*, *PPV*, and *NPV* for the group-level selection from the 50 repetition of the simulation are 0.970, 1.000, 1.000, and 1. The corresponding Monte Carlo standard errors are 0.009, 0.000, 0.000, and 0.000. The average values of *TPR*, *TNR*, *PPV*, and *NPV* for the element-level selection are 0.634, 0.999, 0.589, and 1.000. The corresponding Monte Carlo standard errors are 0.012, 0.000, 0.011, and 0.000. Compared with non-censored cases, the performance of BBVS is satisfactory in the censored case as well.

Conclusion

The BBVS was developed to enable bi-level variable selection as incorporating grouping information within covariates in the high-dimensional setting. In the context of GWAS, our method addressed the challenging issues by making use of natural grouping information of SNPs in the group-level variable selection step. In addition, DL priors were adapted to reflect the grouping information in the element-wise variable selection.

The simulation studies showed that our proposed method outperformed other bi-level and group-level variable selection methods in the GWAS setting for a non-censored case. We applied BBVS on the ADNI-1 data to identify relevant SNP-sets associated with the time to develop AD within MCI patients. We identified 106 informative SNPs located within 10 genes, where four genes were directly and indirectly related to AD, while the simple form of GWAS only detected 3 SNPs that had not been reported in the literature. We need to analyze other AD data sets to see if the implicated genes are reproducible when we used different subjects in the future study. We also need to conduct a simulation study to compare the variable selection performance of BBVS with other survival models that enable variable selection for the high-dimensional data.

https://doi.org/10.5808/gi.23047

ORCID

Eunjee Lee: https://orcid.org/0000-0001-5268-1707
Joseph G. Ibrahim: https://orcid.org/0000-0003-2428-6552
Hongtu Zhu: https://orcid.org/0000-0002-6781-2690

Authors' Contribution

Conceptualization: EL, JGI, HZ. Data curation: EL. Formal analysis: EL. Funding acquisition: EL, JGI, HZ. Methodology: EL, JGI, HZ. Writing – original draft: EL. Writing – review & editing: EL, JGI, HZ.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This material was based on work partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022M3J6A1084843, No. NRF-2021R1C1C1013936). This work was also partially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2020-0-01441, No. RS-2022-00155857, Artificial Intelligence Convergence Research Center (Chungnam National University)). Part of this study has been published as a PhD thesis by the first author under the supervision of the co-authors (Lee E. Advanced Bayesian models for high-dimensional biomedical data. Ph.D. Dissertation. Chapel Hill: The University of North Carolina, 2016).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-1517.
2. International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299-1320.
3. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol 2012;8:e1002822.
4. Liu J, Wang K, Ma S, Huang J. Regularized regression method for genome-wide association studies. BMC Proc 2011;5 Suppl 9:S67.
5. St-Pierre J, Oualkacha K, Bhatnagar SR. Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data. Bioinformatics 2023;39:btad063.
6. Waldmann P, Meszaros G, Gredler B, Fuerst C, Solkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet 2013;4:270.
7. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 2009;25:714-721.
8. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat 2011;5:1780-1815.
9. Williams J, Ferreira MA, Ji T. BICOSS: Bayesian iterative conditional stochastic search for GWAS. BMC Bioinformatics 2022;23:475.
10. He Q, Lin DY. A variable selection method for genome-wide association studies. Bioinformatics 2011;27:1-8.
11. Li J, Zhong W, Li R, Wu R. A fast algorithm for detecting gene-gene interactions in genome-wide association studies. Ann Appl Stat 2014;8:2292-2318.
12. Wen C, Pan W, Huang M, Wang X. Sure independence screening adjusted for confounding covariates with ultrahigh dimensional data. Stat Sin 2018;28:293-317.
13. Kaplan A, Lock EF, Fiecas M; Alzheimer's Disease Neuroimaging Initiative. Bayesian GWAS with structured and non-local priors. Bioinformatics 2020;36:17-25.
14. Liu J, Huang J, Ma S, Wang K. Incorporating group correlations in genome-wide association studies using smoothed group Lasso. Biostatistics 2013;14:205-219.
15. Lock EF, Dunson DB. Bayesian genome- and epigenome-wide association studies with gene level dependence. Biometrics 2017;73:1018-1028.
16. Zhang X, Xue F, Liu H, Zhu D, Peng B, Wiemels JL, et al. Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies. BMC Genet 2014;15:130.
17. Bi W, Fritsche LG, Mukherjee B, Kim S, Lee S. A fast and accurate method for genome-wide time-to-event data analysis and its application to UK biobank. Am J Hum Genet 2020;107:222-233.
18. Lee KH. Bayesian variable selection in parametric and semiparametric high dimensional survival analysis. Ph.D. Dissertation. Columbia: University of Missouri, 2011.
19. Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, et al. Kernel

machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol 2011;35:620-631.
20. Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 2006;22:2262-2268.
21. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Am Stat Assoc 1987;82:528-540.
22. Bhattacharya A, Pati D, Pillai NS, Dunson DB. Dirichlet-Laplace priors for optimal shrinkage. J Am Stat Assoc 2015;110:1479-1490.
23. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons, 2011.
24. Lawless JF. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons, 2011.
25. Meeker WQ, Escobar LA, Pascual FG. Statistical Methods for Reliability Data. 2nd ed. New York: John Wiley & Sons, 2022.
26. Nelson WB. Accelerated Testing: Statistical Models, Test Plans, and Data Analysis. New York: John Wiley & Sons, 2009.
27. Bedrick EJ, Christensen R, Johnson WO. Bayesian accelerated failure time analysis with application to veterinary epidemiology. Stat Med 2000;19:221-237.
28. Christensen R, Johnson W. Modelling accelerated failure time with a Dirichlet process. Biometrika 1988;75:693-704.
29. Kuo L, Mallick B. Bayesian semiparametric inference for the accelerated failure-time model. Can J Stat 1997;25:457-472.
30. Gupta M, Ibrahim JG. An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. Stat Sin 2009;19:1641-1663.
31. Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (Goel PK, Zellner A, eds.). Amsterdam: Elsevier Science Publishers, 1986. pp. 233-243.
32. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 2004;5:155-176.
33. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 2003;31:2013-2035.
34. Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. Biometrics 2008;64:479-489.
35. International HapMap 3 Consortium; Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52-58.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.
37. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 2010;11:459-463.
38. Bezerra GA, Dobrovetsky E, Seitova A, Fedosyuk S, Dhe-Paganon S, Gruber K. Structure of human dipeptidyl peptidase 10 (DPPY): a modulator of neuronal Kv4 channels. Sci Rep 2015;5:8769.
39. Chen T, Shen XF, Chegini F, Gai WP, Abbott CA. Molecular characterisation of a novel dipeptidyl peptidase like protein: its pathological link to Alzheimers disease. Clin Chem Lab Med 2008;46:A13.
40. Chen T, Gai WP, Abbott CA. Dipeptidyl peptidase 10 (DPP10(789)): a voltage gated potassium channel associated protein is abnormally expressed in Alzheimer's and other neurodegenerative diseases. Biomed Res Int 2014;2014:209398.
41. De Jager PL, Shulman JM, Chibnik LB, Keenan BT, Raj T, Wilson RS, et al. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging 2012;33:1017.
42. Guerreiro RJ, Gustafson DR, Hardy J. The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. Neurobiol Aging 2012;33:437-456.
43. Oguri M, Kato K, Yokoi K, Yoshida T, Watanabe S, Metoki N, et al. Assessment of a polymorphism of *SDK1* with hypertension in Japanese individuals. Am J Hypertens 2010;23:70-77.
44. Skoog I, Gustafson D. Update on hypertension and Alzheimer's disease. Neurol Res 2006;28:605-611.
45. Huentelman MJ, Papassotiropoulos A, Craig DW, Hoerndli FJ, Pearson JV, Huynh KD, et al. Calmodulin-binding transcription activator 1 (*CAMTA1*) alleles predispose human episodic memory performance. Hum Mol Genet 2007;16:1469-1477.
46. Hooli BV, Kovacs-Vajna ZM, Mullin K, Blumenthal MA, Matthe-isen M, Zhang C, et al. Rare autosomal copy number variations in early-onset familial Alzheimer's disease. Mol Psychiatry 2014;19:676-681.
47. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Series B Stat Methodol 2006;68:49-67.
48. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat 2010;38:894-942.
49. Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for

https://doi.org/10.5808/gi.23047

variable selection. Biometrika 2009;96:339-355.
50. Breheny P. The group exponential lasso for bi-level variable selection. Biometrics 2015;71:731-740.
51. Breheny P, Huang J. Penalized methods for bi-level variable selection. Stat Interface 2009;2:369-380.