BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 516–523
doi:10.1093/bioinformatics/btq688

Genetics and population analysis

Advance Access publication December 14, 2010

The Bayesian lasso for genome-wide association studies

Jiahan Li[1,2], Kiranmoy Das[1,2], Guifang Fu[1,2], Runze Li[1,2] and Rongling Wu[1,2,*]

[1]Department of Statistics, Pennsylvania State University, State College, PA 16802 and [2]Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Despite their success in identifying genes that affect complex disease or traits, current genome-wide association studies (GWASs) based on a single SNP analysis are too simple to elucidate a comprehensive picture of the genetic architecture of phenotypes. A simultaneous analysis of a large number of SNPs, although statistically challenging, especially with a small number of samples, is crucial for genetic modeling.

Method: We propose a two-stage procedure for multi-SNP modeling and analysis in GWASs, by first producing a 'preconditioned' response variable using a supervised principle component analysis and then formulating Bayesian lasso to select a subset of significant SNPs. The Bayesian lasso is implemented with a hierarchical model, in which scale mixtures of normal are used as prior distributions for the genetic effects and exponential priors are considered for their variances, and then solved by using the Markov chain Monte Carlo (MCMC) algorithm. Our approach obviates the choice of the lasso parameter by imposing a diffuse hyperprior on it and estimating it along with other parameters and is particularly powerful for selecting the most relevant SNPs for GWASs, where the number of predictors exceeds the number of observations.

Results: The new approach was examined through a simulation study. By using the approach to analyze a real dataset from the Framingham Heart Study, we detected several significant genes that are associated with body mass index (BMI). Our findings support the previous results about BMI-related SNPs and, meanwhile, gain new insights into the genetic control of this trait.

Availability: The computer code for the approach developed is available at Penn State Center for Statistical Genetics web site, http://statgen.psu.edu.

Contact: rwu@hes.hmc.psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 16, 2010; revised on December 3, 2010; accepted on December 9, 2010

1 INTRODUCTION

Recent genotyping technologies allow the fast and accurate collection of genotype data throughout the entire genome for many subjects. By genome-wide association studies (GWASs), the genetic variants associated with a complex disease or trait, their chromosomal distribution and individual effects, can be identified. GWASs are based on either case–control cohorts to test the associations between SNPs and diseases or population cohorts to estimate genetic effects of SNPs on traits. In both cases, there are hundreds of thousands of SNPs genotyped on samples involving thousands of subjects. This typical problem, having the number of predictors far exceeding the number of observations, makes it impossible to analyze the data using traditional multivariate regression. In current GWASs, simple univariate linear regression that analyzes one SNP at a time is usually used and, by adjusting for multiple comparisons, the significance levels of the detected genes are then calculated (McCarthy *et al.*, 2008).

These single SNP-based GWASs have been instrumental for reproducibly detecting significants genes for various complex diseases or traits (Donnelly, 2008). However, such strategies have three major disadvantages, limiting the future applications of GWAS. First, because most complex traits are polygenic, a single SNP analysis can only detect a very small portion of genetic variation and, also, may not be powerful for identifying weaker associations (Hoggart *et al.*, 2008). Second, different genes may interact with each other to form a complex network of genetic interactions, which cannot be characterized from a single SNP analysis. Third, many GWASs analyze genetic associations separately for different environments, such as males and females, and then make an across-environment comparison in genetic effects. This analysis is neither powerful nor precise for the identification of gene–environment interactions. Because of these limitations, many authors have developed various approaches for simultaneously analyzing multiple SNPs for GWASs (Logsdon *et al.*, 2010; Wu *et al.*, 2009; Yang *et al.*, 2010), although most approaches focus on case–control cohorts.

There is a daunting need on the development of a variable selection model to identify SNPs with significant effects on quantitative traits in population cohorts and estimate all selected predictors simultaneously. Traditionally, a subset of predictors in a regression model is obtained by forward selection, backward elimination and stepwise selection, but these approaches are computationally expensive and unstable even when the number of predictors is not large. Recently, alternative approaches have been developed, including ridge regression, bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001). For the number of variables much larger than that of subjects, as commonly seen in GWASs, Fan and Lv (2008) proposed a two-stage procedure for variable selection by first suppressing the high dimensionality of response into its low-dimensional representation and then finding a subset of predictors that can predict the suppressed response. A similar two-stage approach was also developed by Paul *et al.* (2008).

*To whom correspondence should be addressed.

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

In this article, we for the first time integrate Paul *et al.*'s preconditioning procedure into LASSO to develop a two-stage strategy for identifying important SNPs in GWASs. In step one, we find a linear combination of predictors that are strongly correlated with the response by a supervised principle component analysis and get a consistent 'preconditioned' estimate of response variable. In step two, we implement the Bayesian lasso (Park and Casella, 2008) for variable selection based on the 'preconditioned' response that mitigates the observational noise. The Markov chain Monte Carlo (MCMC) algorithm is used to estimate all the parameters. The Bayesian hierarchical model is implemented to control an issue of over-fitting that arises when too many associations are included. Our model shows a great flexibility to fit many SNPs and many covariates at the same time. The statistical properties of the model were tested through simulation studies. We used a real GWAS dataset from the Framingham Heart Study (FHS) to validate the usefulness and utilization of the new model.

2 BAYESIAN GWAS MODEL

2.1 Preconditioning

When the number of predictors far exceeds the number of observations, preconditioning via a supervised principal component analysis is recommended to reduce the effect of observational noise on model selection (Paul *et al.*, 2008). In a GWAS of n subjects, we express a response variable y (assumed to be normally distributed) as a function of p SNPs genotyped throughout the entire genome using a linear model

$$y=W\mathbf{b}+\epsilon, \tag{1}$$

where $W=(w_1,\dots,w_n)^T$ is a $(n\times p)$ design matrix, $\mathbf{b}=(b_1,\dots,b_p)^T$ is the vector of regression coefficients and $\epsilon\sim N_n(0,\sigma^2 I_n)$ is the residual error.

The design matrix is reduced to one that consists of only those predictors whose estimated regression coefficients exceed a threshold θ in the absolute value. Thus, the reduced design matrix W_{reduced} consists of the j'-th column of W, where $j'\in\{j:|\hat{b}_j|>\theta\}$. The principal components of W_{reduced}, called supervised principal components, are computed. The first m supervised principal components can serve as independent variables in a linear regression model, from which a consistent predictor $\tilde{y}$ of the true response is obtained. In practice, we select θ by 5-fold cross-validation. Since only the first few components are useful for prediction, in the following examples we consider the first three principal components. Next, a standard variable selection procedure will be conducted for the preconditioned response variable $\tilde{y}$.

2.2 Lasso penalized regression

Given phenotypical measurements and genotype information, we could obtain the preconditioned response $\tilde{y}$ based on the generic form of linear regression (1). However in GWASs, a number of covariates, which are either discrete or continuous, may be measured for each subject. In order to estimate genetic effects precisely by adjusting for these covariates, a GWAS model that takes into account the effects of important covariates would be more appropriate. Therefore, we describe the preconditioned value $\tilde{y}_i$ of a quantitative trait for subject i as

$$\tilde{y}_i=\mu+\mathbf{X}_i^T\boldsymbol{\alpha}+\mathbf{Z}_i^T\boldsymbol{\beta}+\boldsymbol{\xi}_i^T\mathbf{a}+\boldsymbol{\zeta}_i^T\mathbf{d}+\epsilon_i, \quad i=1,\dots,n, \tag{2}$$

where μ is the overall mean, $\mathbf{X}_i$ is the d_1-dimensional vector of discrete covariates for subject i, $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_{d_1})^T$ is the vector of regression coefficients for discrete covariates, $\mathbf{Z}_i$ is the d_2-dimensional vector of continuous covariates for subject i, $\boldsymbol{\beta}=(\beta_1,\dots,\beta_{d_2})^T$ is the vector of regression coefficients for continuous covariates, $\mathbf{a}=(a_1,\dots,a_p)^T$ and $\mathbf{d}=(d_1,\dots,d_p)^T$ are the p-dimensional vectors of the additive and dominant effects of SNPs, respectively, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are the indicator vectors of the additive and dominant effects of SNPs for subject i, and ϵ_i is the residual error assumed to follow a $N(0,\sigma^2)$ distribution. The j-th elements of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are defined as

$$\xi_{ij}=\begin{cases} 1, & \text{if the genotype of SNP}\, j \text{ is } AA \\ 0, & \text{if the genotype of SNP}\, j \text{ is } Aa \\ -1, & \text{if the genotype of SNP}\, j \text{ is } aa, \end{cases}$$

$$\zeta_{ij}=\begin{cases} 1, & \text{if the genotype of SNP}\, j \text{ is } Aa \\ 0, & \text{if the genotype of SNP}\, j\text{is } AA \text{ or } aa. \end{cases}$$

Despite $p\gg n$ in the GWAS, most of the regression coefficients in (2) are expected to have no or only weak effects on the phenotype. To identify a few SNPs that may have notable effects and enhance prediction performance, we put L_1 lasso penalties on the sizes of additive effects and the dominant effects and encourage sparse solutions using

$$\sum_{j=1}^p|a_j|\le t, \quad \sum_{j=1}^p|d_j|\le t^*, \quad \text{for}\, t\ge 0, t^*\ge 0, \tag{3}$$

where t and t^* are a certain value chosen to penalize the additive and dominant effects, respectively. Thus, parameters in Equation (2) are estimated by the penalized least squares

$$\frac{1}{2}||\tilde{\mathbf{y}}-\boldsymbol{\mu}-X\boldsymbol{\alpha}-Z\boldsymbol{\beta}-\xi\mathbf{a}-\zeta\mathbf{d}||^2+\lambda\sum_{j=1}^p|a_j|+\lambda^*\sum_{j=1}^p|d_j|, \tag{4}$$

where $\tilde{\mathbf{y}}=(\tilde{y}_1,\dots,\tilde{y}_n)^T$, $\boldsymbol{\mu}=(\mu,\dots,\mu)^T$, $X=(X_1,\dots,X_n)^T$, $Z=(Z_1,\dots,Z_n)^T$, $\xi=(\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_n)^T$, $\zeta=(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_n)^T$, and λ and λ^* are tuning parameters or lasso parameters that control the degrees of shrinkage in the estimate of the genetic effects.

2.3 Bayesian hierarchical model and prior distributions

Noting the form of the L_1-penalty term in (4), Tibshirani (1996) suggested that lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e. double-exponential) priors. Therefore, when lasso penalties are imposed on the additive and dominant effects of SNPs, the conditional prior for a_j is a Laplace distribution with the scale parameter σ/λ:

$$\pi(\mathbf{a}|\sigma^2)=\prod_{j=1}^p\frac{\lambda}{2\sqrt{\sigma^2}}e^{-\lambda|a_j|/\sqrt{\sigma^2}}, \tag{5}$$

Similarly, the conditional Laplace prior for d_j is

$$\pi(\mathbf{d}|\sigma^2)=\prod_{j=1}^p\frac{\lambda^*}{2\sqrt{\sigma^2}}e^{-\lambda^*|d_j|/\sqrt{\sigma^2}}. \tag{6}$$

Since the Laplace distribution can be represented as a scale mixture of a normal distribution with an exponential distribution

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

(Andrews and Mallows, 1974), we have the following hierarchical representation of the penalized regression model:

$$
\begin{aligned}
&\tilde{\mathbf{y}}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{a},\mathbf{d},\sigma^2 \sim N_n(\boldsymbol{\mu}+X\boldsymbol{\alpha}+Z\boldsymbol{\beta}+\xi\mathbf{a}+\zeta\mathbf{d},\sigma^2 I_n),\\
&\boldsymbol{\alpha} \sim N_{d_1}(0,\Sigma_\alpha),\\
&\boldsymbol{\beta} \sim N_{d_2}(0,\Sigma_\beta),\\
&\mathbf{a}|\sigma^2,\tau_1^2,\ldots,\tau_p^2 \sim N_p(0,\sigma^2\text{diag}(\tau_1^2,\ldots,\tau_p^2)),\\
&\tau_1^2,\ldots,\tau_p^2|\lambda \sim \prod_{j=1}^{p}\exp(\frac{\lambda^2}{2}),\\
&\mathbf{d}|\sigma^2,\tau_1^{*2},\ldots,\tau_p^{*2} \sim N_p(0,\sigma^2\text{diag}(\tau_1^{*2},\ldots,\tau_p^{*2})),\\
&\tau_1^{*2},\ldots,\tau_p^{*2}|\lambda^* \sim \prod_{j=1}^{p}\exp(\frac{\lambda^{*2}}{2}),\\
&\sigma^2 \sim \pi(\sigma^2),\\
&\sigma^2,\tau_1^2,\ldots,\tau_p^2,\tau_1^{*2},\ldots,\tau_p^{*2} > 0.
\end{aligned}
$$

After integrating out $\tau_1^2,\ldots,\tau_p^2$ and $\tau_1^{*2},\ldots,\tau_p^{*2}$, the conditional priors on **a** and **d** have the desired forms (5) and (6), respectively. We assign conjugate normal priors to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ because they are low dimensional and not the parameters of interest. Finally, since the data are usually sufficient to estimate μ and σ, we can use an independent, flat prior $\pi(\mu)=1$ for μ and a non-informative scale-invariant prior $\pi(\sigma^2)=1/\sigma^2$ for σ^2.

The tuning parameters of the ordinary lasso can be prespecified by cross-validation, generalized cross-validation or the idea based on Stein's unbiased risk estimate. However, in the Bayesian lasso, λ and λ^* can be estimated along with other parameters by assigning appropriate hyperpriors to them. This procedure avoids the choice of lasso parameters and allows us to determine the amount of shrinkage from the data. In particular, we consider the conjugate gamma priors on $\lambda^2/2$ and $\lambda^{*2}/2$,

$$\pi\left(\frac{\lambda^2}{2}\right) \sim \text{Gamma}(a,b),$$

$$\pi\left(\frac{\lambda^{*2}}{2}\right) \sim \text{Gamma}(a^*,b^*).$$

where a, b, a^* and b^* are small values so that the priors are essentially non-informative. With this specification, lasso parameters can be treated similar to the other parameters and estimated by the Gibbs sampler.

3 POSTERIOR COMPUTATION AND INTERPRETATION

3.1 MCMC algorithm

We estimate the parameters by sampling from their conditional posterior distributions through the MCMC algorithm. The joint posterior distribution can be expressed as:

$$
\begin{aligned}
&\pi(\mu,\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{a},\tau_1^2,\ldots,\tau_p^2,\lambda,\mathbf{d},\tau_1^{*2},\ldots,\tau_p^{*2},\lambda^*,\sigma^2|\tilde{\mathbf{y}})\\
\propto\quad &\textstyle\prod_{i=1}^{n}\pi(\tilde{y}_i|\cdot)\pi(\mu)\pi(\sigma^2)\pi_(\boldsymbol{\alpha})\pi_(\boldsymbol{\beta})\\
&\textstyle\prod_{j=1}^{p}\pi(a_j|\tau_j^2)\pi(\tau_j^2|\lambda)\pi(\lambda)\pi(d_j|\tau_j^{*2})\pi(\tau_j^{*2}|\lambda^*)\pi(\lambda^*)
\end{aligned}
$$

Two-level hierarchical modeling allows us to easily derive the conditional posterior distributions of parameters and hyperparameters, from which the Gibbs sampler draws posterior samples. Conditional on the parameters $(\mathbf{a},\mathbf{d},\tau_1^2,\ldots,\tau_p^2,\tau_1^{*2},\ldots,\tau_p^{*2})$, the model is the standard linear regression and, thus, the conditional posterior distributions of $(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)$ are

$$\boldsymbol{\alpha}|\cdot \sim N_{d_1}\left(\Sigma'\left(\frac{\sum_{i=1}^{n}X_i(\tilde{y}_i-\mu-\mathbf{Z}_i^T\boldsymbol{\beta}-\boldsymbol{\xi}_i^T\mathbf{a}-\boldsymbol{\zeta}_i^T\mathbf{d})}{\sigma^2}\right),\Sigma'\right),$$

$$\text{with } \Sigma'=\left(\frac{\sum_{i=1}^{n}X_iX_i^T}{\sigma^2}+\Sigma_\alpha^{-1}\right)^{-1},$$

$$\boldsymbol{\beta}|\cdot \sim N_{d_2}\left(\Sigma''\left(\frac{\sum_{i=1}^{n}Z_i(\tilde{y}_i-\mu-\mathbf{X}_i^T\boldsymbol{\alpha}-\boldsymbol{\xi}_i^T\mathbf{a}-\boldsymbol{\zeta}_i^T\mathbf{d})}{\sigma^2}\right),\Sigma''\right),$$

$$\text{with } \Sigma''=\left(\frac{\sum_{i=1}^{n}Z_iZ_i^T}{\sigma^2}+\Sigma_\beta^{-1}\right)^{-1},$$

$$\sigma^2|\cdot \sim Inv-\chi^2\left(n,\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i-\mu-\mathbf{X}_i^T\boldsymbol{\alpha}-\mathbf{Z}_i^T\boldsymbol{\beta}-\boldsymbol{\xi}_i^T\mathbf{a}-\boldsymbol{\zeta}_i^T\mathbf{d})^2\right).$$

Conditional on the parameters $(\tau_1^2,\ldots,\tau_p^2,\tau_1^{*2},\ldots,\tau_p^{*2},\boldsymbol{\alpha},\boldsymbol{\beta})$, the model becomes the weighted linear regression, and thus the conditional posterior distributions of $(\mathbf{a},\mathbf{d})$ are

$$\mathbf{a}|\cdot \sim N\left(A_a^{-1}\xi(\tilde{y}_i-\mu-\mathbf{X}_i^T\boldsymbol{\alpha}-\mathbf{Z}_i^T\boldsymbol{\beta}-\boldsymbol{\zeta}_i^T\mathbf{d}),\sigma^2A_a^{-1}\right),$$

$$\text{with } A_a^{-1}=\left(\xi\xi^T+\text{diag}(\tau_1^2,\ldots,\tau_p^2)^{-1}\right)^{-1},$$

$$\mathbf{d}|\cdot \sim N\left(A_d^{-1}\zeta(\tilde{y}_i-\mu-\mathbf{X}_i^T\boldsymbol{\alpha}-\mathbf{Z}_i^T\boldsymbol{\beta}-\boldsymbol{\xi}_i^T\mathbf{a}),\sigma^2A_d^{-1}\right),$$

$$\text{with } A_d^{-1}=\left(\zeta\zeta^T+\text{diag}(\tau_1^{*2},\ldots,\tau_p^{*2})^{-1}\right)^{-1}.$$

Moreover, the full conditional for $\tau_1^2,\ldots,\tau_p^2,\tau_1^{*2},\ldots,\tau_p^{*2}$ are conditionally independent, with

$$\frac{1}{\tau_j^2}|\cdot \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}},\lambda^2\right),\quad j=1,\ldots,p,$$

and

$$\frac{1}{\tau_j^{*2}}|\ldots \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^{*2}\sigma^2}{\beta_j^2}},\lambda^{*2}\right),\quad j=1,\ldots,p.$$

Finally, with the conjugate priors Gamma(a,b) and Gamma(a^*,b^*), the conditional posterior distributions of the

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

hyperparameters are gammas

$$\lambda^2|\cdot \sim \text{Gamma}\left(p+a, \sum_{j=1}^{p} \frac{\tau_j^2}{2}+b\right),$$

and

$$\lambda^{*2}|\cdot \sim \text{Gamma}\left(p+a^*, \sum_{j=1}^{p} \frac{\tau_j^{*2}}{2}+b^*\right).$$

An efficient Gibbs sampler based on these full conditionals proceeds to draw posterior samples from each full conditional posterior distribution, given the current values of all other parameters and the observed data. This process continues until all chains converge. We use the potential scale reduction factor $\hat{R}$ to access the convergence (Gelman and Rubin, 1992). Once $\hat{R} < 1.1$ for all scalar estimands of interest, we continue to draw 15 000 iterations to obtain samples from the joint posterior distribution.

3.2 Posterior interpretation

The proposed MCMC algorithm for our Bayesian lasso model can provide posterior median estimates of the additive effects and dominant effects of individual SNPs, while adjusting for the effects of all other SNPs and covariates. Furthermore, using the posterior samples of **a**, **d**, and the observed genotypes, we can calculate the proportion of the phenotypic variance explained by a particular SNP, i.e. heritability, by

$$h_j^2 = \frac{2\hat{p}_1\hat{p}_0(\hat{a}_j+(\hat{p}_1-\hat{p}_0)\hat{d}_j)^2+4\hat{p}_1^2\hat{p}_0^2\hat{d}_j^2}{\text{var}(\tilde{y})}, \quad j=1,\ldots,p,$$

where $\hat{p}_1$ is the estimated allele frequency for A, and $\hat{p}_0$ is the estimated allele frequency for a, $\hat{a}_j$ is the median estimate of the additive effect for SNP j and $\hat{d}_j$ is the median estimate of the dominant effect for SNP j. Since heritability estimates are unitless, they could guide variable selection and identify SNPs that have relatively large effects on the phenotype.

4 RESULTS

4.1 Worked example

We used the newly developed model to analyze a real GWAS dataset from the FHS, a cardiovascular study based in Framingham, Massachusetts, supported by the National Heart, Lung and Blood Institute, in collaboration with Boston University (Dawber *et al.*, 1951). Recently, 550 000 SNPs have been genotyped for the entire Framingham cohort (Jaquish, 2007), from which 418 males and 559 females were chosen for our data analysis. These subjects were measured for body mass index (BMI) at different ages from 29 to 61 years. As is standard practice, SNPs with minor allele frequency <10% were excluded from data analysis. The numbers and percentages of non-rare allele SNPs vary among different chromosomes and ranges from 4417 to 28 771 and from 64% to 72%, respectively.

In principle, our approach can handle an extremely large number of SNPs at the same time. To save our computing time, however, we use those SNPs that cannot be neglected according to a simple single SNP analysis. We chose the phenotypic data of BMI in a middle measure age of each subject for a single SNP analysis, separately for males and females. Supplementary Figure S1 gives $-\log_{10}$ P-values for each SNP in the two sexes from which 1837 SNPs with a $-\log_{10}$ P-value of >3.5 in at least one sex was selected for Bayesian lassso analysis. Before this analysis, we imputed missing genotypes for a small proportion of SNPs (5.16%) according to the distribution of genotypes in the population. A preconditional analysis with $m=3$ and $\theta=0.426$ was used to mitigate observational noise, leading to the preconditioned phenotypes. Like original measures, the preconditioned BMI also displays a normal distribution (Fig. 1), which meets the normality assumption required by the new approach.

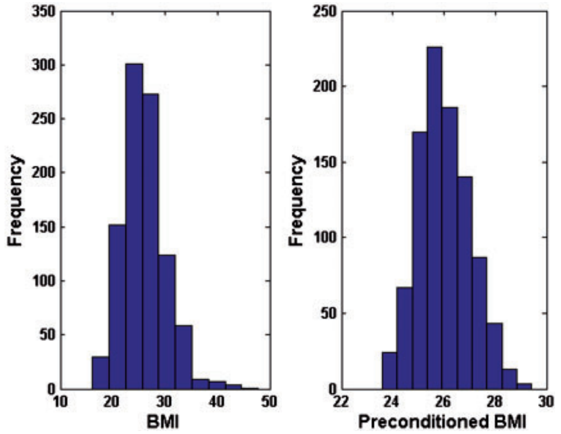

Fig. 1. The histograms of original and preconditioned BMI.

By treating the sex as a discrete covariate and age as a continuous covariate, we imposed lasso penalties on the additive effects $a_1,\ldots,a_p$ and dominant effects $d_1,\ldots,d_p$ to identify those SNPs with notable effects on BMI. We employ the proposed MCMC algorithms to estimate all parameters and implement variable selection, where $\Sigma_\alpha=1$, $\Sigma_\beta=1$ and all parameters in the conjugate gamma hyperpriors are 0.1. In unreported tests, we find that the posteriors are not sensitive to these prior specifications, as long as a and b are small values so that the priors are relatively flat (Park and Casella, 2008; Yi and Xu, 2008). Figure 2 plots the estimated additive and dominant effects of each SNP after adjusting for the effects of other SNPs and covariates. The heritability explained by each SNP is shown in Figure 3. The Bayesian hierarchical model automatically shrinks small coefficients to zero, and hence the posterior estimates of **a**, **d** and h_j^2 can guide variable selection. We claim that a genetic effect is significant if its 95% posterior credible interval does not contain zero. Alternatively, Hoti and Sillanpaa (2006) suggested to preset a threshold value, c, such that one SNP is included into the final model if the heritability explained by this SNP is greater than c. We usually report the SNPs with high heritabilities and, thus, this threshold can be chosen on more subjective grounds.

Table 1 tabulates the names and positions of SNPs with the heritability (h_j^2) greater than 0.5, as well as the estimated additive effects and heritabilities. We do not report the estimated dominant effects since they are relatively low in this example. The Bayesian lasso tends to shrink small effects of genes into zero. Assuming that $a=d=0.4$ for a SNP with allele frequencies of 0.5 in a population, the additive and dominant variances explained by this SNP is $\frac{1}{2}a^2=0.08$ and $\frac{1}{4}d^2=0.04$, respectively. Thus, there is a possibility

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

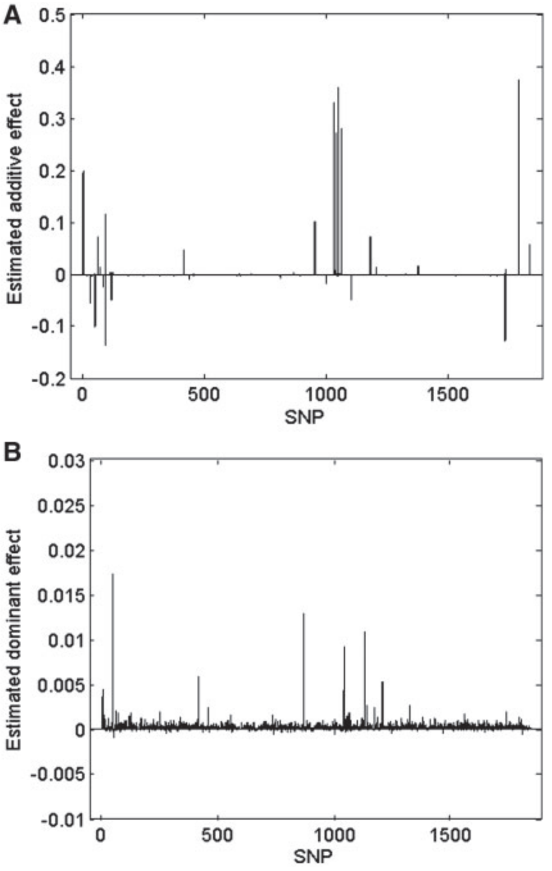

Fig. 2. Estimated additive (**A**) and dominant effects (**B**) of 1837 SNPs from the Framingham heart study.

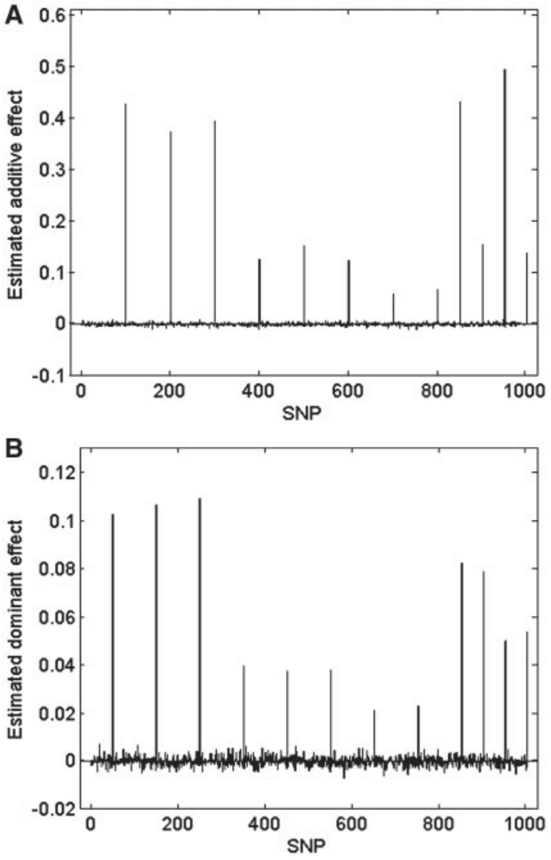

Fig. 3. Estimated additive (**A**) and dominant effects (**B**) based on 50 simulations.

that the dominant effects are shrunk to a greater extent than the additive effects if they are of similar size. This may partly explain why the dominant effects estimated for significant SNPs are much smaller than the additive effects.

The amount of shrinkage in the estimates of additive and dominant effects are quantified by two hyperparameters λ and λ^* determined from the data. The posterior medians for λ and λ^* are 54.474 and 54.523, respectively, with the 95% posterior intervals being [53.325, 55.626] and [53.359, 55.678], respectively. These suggest that the tuning parameters for the additive and dominant effects can be estimated precisely.

Since five significant SNPs are selected from chromosome 1, and four from chromosome 10, we will further examine the correlations among the significant SNPs from the same chromosome, as suggested by a referee. The correlation matrix of five significant SNPs from chromosome 1 is given by

$$\begin{pmatrix} 1.00 & 0.85^* & 0.86^* & -0.01 & 0.01 \\ 0.85^* & 1.00 & 0.78^* & -0.01 & 0.02 \\ 0.86^* & 0.78^* & 1.00 & -0.04 & 0.02 \\ -0.01 & -0.01 & -0.04 & 1.00 & -0.84^* \\ 0.01 & 0.02 & 0.02 & -0.84^* & 1.00 \end{pmatrix},$$

where star denotes significant correlations at the significance level 1%. Clearly, these SNPs can be classified into two groups, and within each group, SNPs are highly correlated. The correlation matrix of four significant SNPs from chromosome 10

$$\begin{pmatrix} 1.00 & 0.53^* & 0.48^* & 0.30^* \\ 0.53^* & 1.00 & 0.52^* & 0.53^* \\ 0.48^* & 0.52^* & 1.00 & 0.45^* \\ 0.30^* & 0.53^* & 0.45^* & 1.00 \end{pmatrix}$$

also suggested that these SNPs are closely linked to each other.

4.2 Computer simulation

The new approach is investigated through simulation studies. We generate data according to the model (2) with $\mu = 0$, $\sigma^2 = 10$ and $n = 500$. For ease of simulation, ξ_{ij} is derived from u_{ij}, where each u_{ij} has a standard normal distribution marginally, and $\rho = cov(u_{ij}, u_{ik}) = 0.1$. Then, to mimic a SNP with equal allele frequencies, we set

$$\xi_{ij} = \begin{cases} 1, & u_{ij} > c \\ 0, & -c \le u_{ij} \le c \\ -1, & u_{ij} < -c, \end{cases}$$

where $-c$ is the first quartile of a standard normal distribution. Finally, ζ_{ij} is derived from ξ_{ij}. We assume that there are 1000 SNPs from which 20 are significant for a phenotypic trait. The positions and additive and dominant effects of individuals are given in Table 2. It is assumed that the trait is measured at a subject-specific age, following the data structure of the FHS.

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

Table 1. Information about significant SNPs

Chromosome	Name	Position	Additive	Heritability (%)
1	ss66185476	8445140	0.15	0.74
1	ss66374301	8451728	0.19	1.36
1	ss66295856	8578082	0.20	1.35
1	ss66516012	198313489	−0.13	0.89
1	ss66364251	198321700	0.12	0.66
10	ss66311679	32719838	0.33	4.65
10	ss66293192	32903593	0.27	2.08
10	ss66303064	32995111	0.36	5.93
10	ss66128868	33407810	0.28	2.75
20	ss66171460	22580931	−0.13	0.78
22	ss66055592	23420006	0.33	5.13
22	ss66164329	23420370	0.37	6.70

Table 2. Genetic effects of 20 assumed SNPs for data simulation

Position	Additive	Position	Dominant
100	1.2	50	1.2
200	1.2	150	1.2
300	1.2	250	1.2
400	0.8	350	0.8
500	0.8	450	0.8
600	0.8	550	0.8
700	0.4	650	0.4
800	0.4	750	0.4
850	1.2	850	1.2
900	0.8	900	1.2
950	1.2	950	0.8
1000	0.8	1000	0.8

Figure 3 gives the estimated additive and dominant genetic effects of different SNPs over 50 simulations, and Figure 4 plots the heritability explained by each SNP. It is clear that lasso penalties shrink small genetic effects to zeros, resulting in sparse solutions of the regression coefficients. In general, the 20 assumed SNPs can be well identified and their additive and dominant effects well estimated. Also, two hyperparameters λ and λ^* whose influence the degree of shrinkage can be well estimated. In Supplementary Figure 2, the histograms of these two hyperparameters are shown.

Then, we carry out another simulation study to compare the performance of preconditioned Bayesian lasso, Bayesian lasso without preconditioning and the traditional single SNP analysis. Without loss of generality, only the additive model is considered. Specifically, we generate data on $n=200$ and $p=$ 500 or 1000 according to the model (2), with $\mu=0$, $\sigma^2=10$, $\rho=0.1$, $a_j=1$ for $1\leq j\leq 20$ and $a_j=0$ for $j>20$.

We apply three methods to the 100 simulated datasets: single SNP analysis (SSA), standard Bayesian lasso (B-lasso) and the Bayesian lasso applied to the preconditioned response from supervised principal components (PB-lasso). In single SNP analysis, we reject the null hypothesis that the genetic effect of an individual SNP equals to zero at the significance level of 5% with the FDR adjustment. For the Bayesian lasso and preconditioned Bayesian lasso, we reject the null hypothesis based on 95% Bayesian credible intervals. To ameliorate the bias of the parameter estimates introduced by lasso penalties, we always refit the linear regression model without the penalty term using only those SNPs selected by the model selection procedure.

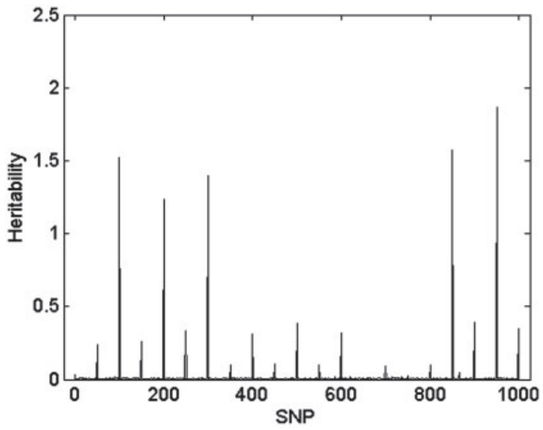

Fig. 4. Estimated heritability explained by each SNP based on 50 simulations.

For each estimated genetic effect obtained from each method, we calculate the average bias and empirical standard error over 100 simulations. Since the first 20 genetic effects are non-zeros with the same true value, in Table 3 we report the average values over the first 20 SNPs and over the rest of the SNPs separately. The standard error of each average is in parentheses. In the column labeled ‘Aver. Nonzeros’, we present the average number of non-zero coefficients correctly identified to be non-zero, or the average number of zero coefficients incorrectly estimated to be non-zero in 100 replications. In the column ‘Proportion of Correct-fit’, we present the proportion of replications that the exact true model was identified.

As can be seen from Table 3, the single SNP analysis tend to overestimate the genetic effect, since when we test a SNP for the association with the phenotype, we assume the genetic variation is solely due to this particular SNP, and ignore the effects from all other SNPs. Therefore, in terms of parameter estimates, model selection methods that simultaneously estimate the genetic effects associated with all SNPs outperform the traditional single SNP analysis. In terms of variable selection, although preconditioned Bayesian lasso has a slightly higher false positive rate due to the preconditioning step, it greatly improves the probability of correctly identifying regression coefficients with non-zero effects. Moreover, as the number of SNPs gets larger, single SNP analysis detected fewer important SNPs, since this method subjects to severe multiple comparison adjustment. However, preconditioned Bayesian lasso is still able to identify non-zero coefficients and zero coefficients correctly in almost every simulation. Supplementary Table 1 displays the simulation results when $\rho=0.5$, which are consistent with our findings.

Since our method is from the Bayesian perspective and is based on the Gibbs sampler, the computational time is relatively high. For example, for each replicate in this simulation study, averagely it takes 439 s when $p=1000$ and 109 seconds when $p=500$ on a 2.0 GHz PC.

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

Table 3. Simulation results for three methods based on 100 simulations

Method	Bias	Empirical Standard Error	Aver. non-zeros	Proportion of correct-fit
$n=200, p=500$, $\beta_1-\beta_{20}$				
SSA	4.17	1.99	16.62	0.08
	(0.21)	(0.19)	(1.51)	
B-lasso	0.07	0.34	18.28	0.18
	(0.04)	(0.02)	(1.36)	
PB-lasso	0.00	0.35	19.68	0.63
	(0.03)	(0.03)	(0.65)	
$n=200, p=500$, $\beta_{21}-\beta_{500}$				
SSA	0.44	0.43	0.78	0.46
	(0.05)	(0.07)	(1.09)	
B-lasso	0.00	0.04	0.95	0.42
	(0.01)	(0.01)	(0.94)	
PB-lasso	0.00	0.03	1.25	0.30
	(0.01)	(0.04)	(0.73)	
$n=200, p=1000$, $\beta_1-\beta_{20}$				
SSA	4.13	1.96	15.71	0.04
	(0.18)	(0.17)	(2.73)	
B-lasso	0.36	0.38	17.11	0.08
	(0.06)	(0.07)	(2.69)	
PB-lasso	0.00	0.36	19.24	0.51
	(0.04)	(0.03)	(1.81)	
$n=200, p=1000$, $\beta_{21}-\beta_{1000}$				
SSA	0.43	0.43	0.42	0.69
	(0.04)	(0.06)	(0.84)	
B-lasso	0.00	0.02	0.33	0.76
	(0.01)	(0.01)	(0.18)	
PB-lasso	0.00	0.02	1.17	0.56
	(0.00)	(0.01)	(1.38)	

5 DISCUSSION

When the number of predictors p is much larger than the number of observations n, highly regularized approaches, such as penalized regression models, are needed to identify non-zero coefficients, enhance model predictability and avoid over-fitting (Hastie *et al.*, 2009). The L_1 penalized regression or lasso is such one of the most popular techniques. In this article, we presented a Bayesian hierarchical model with lasso penalties to simultaneously fit and estimate all possible genetic effects associated with all SNPs in a GWAS, adjusting for both discrete and continuous covariates. Lasso penalties are imposed on the additive and dominant effects, and implemented by assigning double-exponential priors to their regression coefficients. It shrinks small effects toward zero and produces sparse solutions. In this framework, SNPs with significant genetic effects can be identified more accurately.

We fit the model in a fully Bayesian approach, employing the MCMC algorithm to generate posterior samples from the joint posterior distribution, which can be used to make various posterior inferences. Although computationally intensive, it is easy to implement and provides not only point estimates but also interval estimates of all parameters. The Bayesian lasso treats tuning parameters as unknown hyperparameters and generates their posterior samples when estimating other parameters. This technique avoids the choice of tuning parameters, and automatically accounts for the uncertainty in its selection that affects the estimation of the final model. In contrast, standard lasso algorithms usually select tuning parameters by K-fold cross-validation, which involves partitioning the whole dataset and refitting the model many times. This process may result in unstable tuning parameter estimates.

In order to improve the performance of lasso when p is greater than n, preconditioning is considered before variable selection. Preconditioning encourages the principal components of a reduced design matrix to be highly correlated with the response, and thus in most cases only the first or first few components tend to be useful for prediction. It denoises the response variable so that variable selection becomes more efficient. Our simulation demonstrated that when p greatly exceeds n, preconditioned Bayesian lasso could successfully identify almost all the SNPs with true genetic effects. By analyzing real data, our approach is shown to produce biologically relevant results. For example, the approach detected a significant SNP ss66171460 at position 22580931 of chromosome 20 associated with BMI. It is interesting to note that this SNP is within 500 kb of the FOXA2 (Forkhead Box A2) gene, an important genetic variant that regulates obesity (Wolfrum *et al.*, 2003).

One simulation example of Paul *et al.* (2008) implies that, in the context of GWASs, SNPs that are marginally independent of the phenotype could be screened out by preconditioning, but can be identified by standard variable selection techniques such as lasso or Bayesian lasso. In theory, if SNPs are correlated with the phenotype through marginal correlations, we believe the preconditioning step is worthwhile to identify more important SNPs. However, in reality, since different SNPs may display interactions, this approach may not work perfectly. In any case, this two-step variable selection procedure should always be advantageous over a single SNP analysis, because we are always testing the marginal correlation between the predictor and response when one SNP is analyzed at a time.

Motivated by Tibshirani (1996), Park and Casella (2008) developed the Bayesian lasso and demonstrated the diabetes data (Efron *et al.*, 2004) with $p=10$ and $n=442$. We applied the Bayesian lasso to the high-dimensional regression problem, and improved it by preconditioning. We have concentrated on the preconditioned Bayesian lasso method for continuous trait in GWASs. The proposed preconditioning procedure and MCMC algorithm can be readily extended to survival data analysis and lasso penalized logistic regression in case–control disease gene mapping. Also, we may look for gene–gene interaction effects after identifying main effects, as suggested by Wu *et al.* (2009). The model with a capacity to identify epistatic interactions will enables geneticists to decipher a detailed picture of the genetic architecture of a complex trait.

Funding: NSF/NIH Mathematical Biology grant (No. 0540745); NIDA; NIH grants (R21 DA024260 and R21 DA024266). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

Conflict of Interest: none declared.

REFERENCES

Andrews,D.F. *et al.* (1974) Scale mixture of normal distributions. *J. R. Stat. Soc. Ser. B*, **36**, 99–102.

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022

Dawber,T. *et al*. (1951) Epidemiological approaches to heart disease: the Framingham study. *Ame. J. Public Health*, **41**, 279–286.

Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **465**, 728–731.

Efron,B. *et al.* (2004) Least angle regression (with discussion). *Annu. Stat.*, **32**, 407–499.

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B*, **70**, 849–911.

Frank,I.E. and Friedman,J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

Gelman,A. and Rubin,D.B. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–511.

Hastie,T. *et al.* (2009) High-dimensional problems: $p > N$. *The Elements of Statistical Learning*, 2nd edn. Springer, New York.

Hoggart,C.*et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Hoti,F. and Sillanpaa,M.J. (2006) Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. *Heredity*, **97**, 4–18.

Jaquish,C. (2007) The Framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.*, **8**, 63.

Logsdon,B.A. *et al.* (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **27**, 11–58.

McCarthy,M. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Park,T. and Casella,G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Paul,D. *et al.* (2008) Preconditioning for feature selection and regression in high-dimensional problems. *Annu. Stat.*, **36**, 1595–1618.

Tibshirani,R. (1996) Regression shrinkage and selction via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Wolfrum,C. *et al.* (2003) Role of Foxa-2 in adipocyte metabolism and differentiation. *J. Clin. Invest.*, **112**, 345–356.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Rev. Genet.*, **42**, 565–569.

Yi,N. and Xu,S. (2008) Bayesian lasso for quantitative trait loci mapping. *Genetics*, **179**, 1045–1055.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.

Downloaded from https://academic.oup.com/bioinformatics/article/27/4/516/197666 by guest on 21 August 2022