*Bioinformatics*, 31(24), 2015, 3890–3896
doi: 10.1093/bioinformatics/btv496
Advance Access Publication Date: 30 August 2015
Original Paper

Genome analysis

# A hybrid bayesian approach for genome-wide association studies on related individuals

**A. Yazdani[1,*] and D. B. Dunson[2]**

[1]Human Genetic Center, University of Texas at Houston Health Science Center, Houston, USA and [2]Department of Statistical Science, Duke University, Durham, North Carolina USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 14, 2015; revised on August 15, 2015; accepted on August 18, 2015

## Abstract

**Motivation:** Both single marker and simultaneous analysis face challenges in GWAS due to the large number of markers genotyped for a small number of subjects. This large *p* small *n* problem is particularly challenging when the trait under investigation has low heritability.
**Method:** In this article, we propose a two-stage approach that is a hybrid method of single and simultaneous analysis designed to improve genomic prediction of complex traits. In the first stage, we use a Bayesian independent screening method to select the most promising SNPs. In the second stage, we rely on a hierarchical model to analyze the joint impact of the selected markers. The model is designed to take into account familial dependence in the different subjects, while using local-global shrinkage priors on the marker effects.
**Results**: We evaluate the performance in simulation studies, and consider an application to animal breeding data. The illustrative data analysis reveals an encouraging result in terms of prediction performance and computational cost.
**Contact:** Akram.Yazdani@uth.tmc.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Genome-wide association studies (GWAS) have been widely conducted in humans with the goal of identifying genetic factors predictive of disease. GWAS chips collect data on single nucleotide polymorphisms (SNPs) across the genome, focusing on common variants in which the minor allele frequency is at least 5%. The dominant focus in GWAS has been on independent screening methods, which consider the association between disease and each SNP separately while adjusting for false discoveries. This strategy tends to produce a small number of SNPs having modest effect sizes, failing to explain a substantial proportion of the known heritability in disease. Given the size of the multiple testing problems in which the number of SNPs ($p$) is dramatically larger than the number of individuals under study ($n$), it is not surprising that independent screening has failed to identify much signal in the data. As the identified SNPs typically explain a small proportion of the variability in the phenotype, GWAS has been unsuccessful at producing accurate predictive models in humans (see, e.g. Buckler *et al.*, 2009; Hoggart, 2008; Diabetes Genetics Initiative of Broad Institute of Harvard *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007; Weedon *et al.*, 2008).

Dramatic reduction in cost has driven the increased use of GWAS beyond humans to animal breeding studies. In animal breeding a major focus is on prediction of genetic merit (breeding value) when the pedigree structure is taken into account. Based on predictions early in life, animals having a high probability of developing desired traits will be differentially selected. When the focus is on prediction of traits instead of identification of the top individual variants, it has been suggested that all markers should be included instead of attempting variable selection (Meuwissen *et al.*, 2001).

Because in GWAS the number of SNPs exceeds the number of samples, simultaneous analysis requires using penalized or shrinkage approaches. The most popular approach is the Lasso (Tibshirani, 1996), which includes an $L_1$ penalty on the coefficients to induce simultaneous variable selection and shrinkage. However, a rich variety of alternative penalties have been proposed (Fan and Li, 2001; Zhang, 2010; Zou and Hastie, 2005) to provide desirable shrinkage behavior. Most of the point penalization estimates of marker effects correspond with the mode of a posterior distribution obtained under shrinkage priors. These priors typically can be expressed as local-global scale mixtures of normal distributions (see, e.g. Armagan *et al.*, 2013; Carvalho *et al.*, 2009; Griffin and Brown, 2011; Park and Casella, 2008). Carefully designed priors in this class shrink small signals towards zero while limiting shrinkage of larger signals. Compared with variable selection, such methods have computational advantages (avoid an intractable search over all possible subsets) and better accommodate lots of small but non-zero coefficients. In Bayesian analyses of animal breeding, it has been common to rely on mixture priors that include a mass at zero for zero coefficients (see, e.g. Gianola *et al.*, 2009). Such priors are tabulated in Supplementary Appendix A, Table 1.

Although the aforementioned simultaneous approaches have been applied to $p \gg n$ problems, there are some clear limitations in scaling computation to very large $p$, as well as issues in obtaining reliable results when $n$ is too small relative to $p$. In sparse signal processing problems involving large $p$ small $n$ linear regression, it has been discovered that there is often a phase transition that depends on the relative values of $n$ and $p$ and the true sparsity level. When $n$ is too small, so that one is on the wrong side of the phase transition, results are very unreliable. Such transitions have been characterized only in idealized cases, making assumptions on the design matrix that are not appropriate in genomic studies due to strong correlations that arise given the linkage disequilibrium (LD) structure. We hope to push back the phase transition by using a carefully structured Bayesian approach.

To combat the computational intractability of variable selection in massive dimensions, multistage approaches are recommended (see, e.g. Beattie *et al.*, 2002; Fan and Lv, 2008; Huanga *et al.*, 2010; Li *et al.*, 2011; Paul *et al.*, 2008; Shariati *et al.*, 2012). For family based studies, Kessler *et al.* (2014) proposed two-stage screening based on a Bayesian non-parametric regression model. Here, we instead propose a two-stage approach for a polygenic mixed model to predict phenotypic variation in the trait of interest from genomic information for data collected on genetically related individuals. In the first stage, we apply a Bayesian independent screening method, which takes into account familial dependence while examining SNPs one at a time. This Bayesian screening method has better performance than competitors in our experience, and is expected to select a superset of the important predictors (Ball, 2011; Stephens and Balding, 2009). In the second stage, we implement a Bayesian hierarchical model to control the level of sparsity and amount of shrinkage relative to the size of signals. We use the generalized double Pareto (GDP) prior (Armagan *et al.*, 2013) within a polygenic mixed model, extending previous implementations of the GDP to problems in which the samples are related. We evaluate the predictive performance and computational efficiency of the proposed approach called two-stage-GDP by conducting a simulation and real data analysis.

## 2 Model

The overwhelming majority of the literature on analysis of GWAS data focuses on unrelated individuals, while our interest is in animal breeding data from pedigrees. It is important to take the familial dependence structure into account in the analysis to avoid finding spurious associations (see for e.g. Kang *et al.*, 2010). A common model-based approach to incorporate familial dependence is random effects modeling. In particular, we let

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \tau^{-1}R). \tag{1}$$

Here, $\mathbf{y}$ is an $n$ vector of a quantitative trait measured on $n$ subjects (animals), $\boldsymbol{\beta}$ is a $p$ vector of additive genetic effects and $\mathbf{X}$ is an $n \times p$ matrix of genotypes measured at genetic markers, such that each SNP is coded as

$$\begin{cases} 2 & \text{if the genotype of the SNP is BB} \\ 1 & \text{if the genotype of the SNP is Bb} \\ 0 & \text{if the genotype of the SNP is bb,} \end{cases}$$

where B represents the allele with lower frequency. In model (1), $R$ is an $n \times n$ diagonal matrix accommodating heterogenous residual variance, which is estimated in a preliminary data preprocessing step, and $\mathbf{u}$ is an $n$ random vector such that

$$\mathbf{u} \sim N(\mathbf{0}, \tau_u^{-1}K),$$

where $K$ is an $n \times n$ relatedness matrix. Entries of $K$ are in (0, 1) and correspond to the degree of relationship between pairs of subjects, ranging from 0 for unrelated individuals to 1 for self-kinship. This matrix is calculated based on the pedigree measuring genetic similarity between subjects as $\sum_h (.5)^{L(h)}$, with the sum being over all paths $h$ connecting two subjects with unique common ancestors and $L(h)$ is the length of path $h$. If the pedigree is incomplete or not available, $K$ can be estimated based on genotype data (see for e.g. Oliehoek *et al.*, 2006).

### 2.1 First-stage

In this stage, we select a set of promising SNPs by testing the effect of each marker individually. By far the most common approach used in the literature relies on independent testing for association between each SNP and the phenotype using frequentist methods. This produces $P$-values for each SNP, and the set of significant SNPs can be obtained by choosing a small threshold on these $P$-values. This threshold is often chosen to maintain a desired false discovery rate (FDR). We instead rely on Bayes factors (BFs) measuring the weight of evidence in the data against the local null hypothesis $H_{0j}$ of no association between a given SNP and the phenotype. In this initial screening stage, we rely on Bayes factors in favor of the model with only the $j$th SNP included against the global null model that has no SNPs included. These BFs can be defined under model (1) as

$$\mathrm{BF}_j = \frac{\displaystyle\iint \mathrm{f}(\mathbf{y}, \mathbf{u} | H_{1j},\ \boldsymbol{\theta}_{1j}) \pi(\boldsymbol{\theta}_{1j} | H_{1j})\ d\mathbf{u}\ d\boldsymbol{\theta}_{1j}}{\displaystyle\iint \mathrm{f}(\mathbf{y}, \mathbf{u} | H_{0j},\ \boldsymbol{\theta}_{0j}) \pi(\boldsymbol{\theta}_{0j} | H_{0j})\ d\mathbf{u}\ d\boldsymbol{\theta}_{0j}},$$

where $\boldsymbol{\theta}_{kj} = (k\beta_j, \tau, \tau_u)$ is the set of parameters under $H_{kj}$ for $k \in \{0, 1\}$, $\pi(\boldsymbol{\theta}_{kj} | H_{kj})$ is its prior density, and $\mathrm{f}(\mathbf{y}, \mathbf{u} | H_{kj},\ \boldsymbol{\theta}_{kj})$ is the probability density of $(\mathbf{y}, \mathbf{u})$ given $\boldsymbol{\theta}_{kj}$.

The separate analysis of each SNPs makes it appropriate to specify the usual conjugate prior on the set of parameters $\boldsymbol{\theta}_{kj}$. In particular, we consider

$$\begin{gathered} \beta_j | \tau \sim N(0, \tau^{-1}), j = 1, \ldots, p, \\ \tau \sim \mathrm{Gamma}(a_1, b_1), \tau_u \sim \mathrm{Gamma}(a_2, b_2). \end{gathered} \tag{2}$$

To calculate the BF$_j$s, we first integrate out the parameters $\beta_j$s and the random effects $\mathbf{u}$ and obtain

$$\int_\tau\int_{\tau_u}(2\pi)^{-n/2}\frac{b_1^{a_1}b_2^{a_2}}{\Gamma(a_1)\Gamma(a_2)}\tau^{a_1-1}\tau_u^{a_2-1}\left|\frac{A_k}{\tau}+\frac{K}{\tau_u}\right|^{-1/2}$$
$$\times\exp(-\tau b_1-\tau_u b_2)\exp\left[-\frac{1}{2}\mathbf{y}^T\left(\frac{A_k}{\tau}+\frac{K}{\tau_u}\right)^{-1}\mathbf{y}\right]d\tau_u d\tau,$$

where $A_k = R + k(\mathbf{x}_j\mathbf{x}_j^T)$ corresponding to $H_{kj}$ for $k = 0$, 1 in the denominator and numerator, respectively. These integrals are analytically intractable. Hence, we approximate the integrals by applying Laplace transformation. To avoid having the mode on the boundary due to the constraints $\tau > 0$ and $\tau_u > 0$, the precision parameters are reparameterized with log transformation.

The calculated BF$_j$s provide a list of ranked SNPs for selecting the most promising markers. In the screening stage, our goal is to reduce the number of SNPs under consideration to a manageable number, while erring on the side of including too many SNPs. Usual thresholds on BFs, such as the common value of 10 recommended in (Jeffreys, 1961), are too small due to the massive multiple comparisons problem. On the other hand, thresholds that are fully adjusted for multiple comparisons, such as those recommended by Ball (2011) and Stephens and Balding (2009), will tend to select too few SNPs. Motivated by our goal of building a model for prediction, we instead rely on cross validation for threshold choice.

## 2.2 Second-stage

In this stage, we simultaneously include all SNPs selected in the first stage, while incorporating a shrinkage prior on their coefficients. In building a predictive model, it is necessary to characterize the simultaneous impact of the different SNPs on the phenotype. We find that including the first stage instead of incorporating all the SNPs in the simultaneous analysis improves performance in two respects. The first is a considerable computational speedup as the number of SNPs grows. The second is an improvement in accuracy of the resulting predictive model. This second improvement is due to the fact that most shrinkage priors are insufficiently flexible to handle truly massive-dimensional predictors that are not highly sparse in their effects. Shrinkage priors are designed to have separate parameters controlling concentration around zero and tail heaviness, but such control is insufficient in GWAS.

In the second stage, model (1) is rewritten as

$$\mathbf{y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \mathbf{u} + \tilde{\boldsymbol{\epsilon}}, \quad \tilde{\boldsymbol{\epsilon}} \sim N(0, \tau^{-1}\mathbf{R}), \tag{3}$$

where vector $\tilde{\boldsymbol{\beta}}$ represents the effects of $p_s$ selected SNPs and $\tilde{\mathbf{X}}$ is the corresponding design matrix.

### 2.2.1 Prior specification

We incorporate a local-global shrinkage prior into polygenic mixed model (3) as

$$\tilde{\boldsymbol{\beta}} \sim N(0, \tau^{-1}\Sigma_{\tilde{\boldsymbol{\beta}}}), \tag{4}$$

where $\Sigma_{\tilde{\boldsymbol{\beta}}} = \text{diag}\{\eta_j\}$. The variance of each coefficient, $\tilde{\beta}_j$, includes two parameters, $\tau$ and $\eta_j$, representing *global* and *local* shrinkage effects, respectively. In GWAS settings, $\tau^{-1}$ is expected to be close to zero, corresponding to a high degree of overall shrinkage to stabilize estimation and incorporate the expectation that most coefficients are close to zero. To avoid over-shrinkage of the most influential SNPs, the prior on $\eta_j$ is designed to have heavy tails. To obtain a flexible prior with the desired behavior, which also leads to computational advantages, we let

$$\eta_j \sim \exp(\xi_j^2/2).$$

Instead of presetting values for $\xi_j$s, it is appealing to assign a prior distribution to these parameters as

$$\xi_j \sim \text{Gamma}(c, d),$$

which induces a heavy-tailed marginal on $\eta_j$.

The above hierarchical shrinkage prior on the SNP coefficients corresponds to a representation of the GDP introduced by Armagan *et al.* (2013). To complete a specification of the prior, the precision parameters $\tau$ and $\tau_u$ are given gamma priors as in (2). The GDP has substantial advantages over simpler shrinkage priors, such as ridge and Bayesian Lasso, in terms of limiting over-shrinkage of the larger coefficients.

### 2.2.2 Posterior interpretation

Under the proposed model, the total phenotypic variance for the trait can be decomposed as

$$\text{V}_y = \sum_{j=1}^{p_s}\sum_{j'=1}^{p_s}\tilde{\beta}_j\tilde{\beta}_{j'}\,\text{cov}(\tilde{x}_j,\tilde{x}_{j'}) + s\tau^{-1} + s_u\tau_u^{-1}, \tag{5}$$

where $s$ and $s_u$ are the mean of diagonal elements in $R$ and $K$, respectively; i.e. $s = \frac{1}{n}\sum_{i=1}^n r_{ii}$ and $s_u = \frac{1}{n}\sum_{i=1}^n k_{ii}$ where $r_{ij}$ and $k_{ij}$ are the $ij$th elements of matrices $R$ and $K$. In the case that $K$ is calculated from a pedigree, $s_u$ is one. The proportion of the phenotypic variance explained by total genetic variance is called heritability, denoted by $h^2$. As is clear from (5), $h^2$ accounts for the covariance between markers as well. If we ignore the contribution from the covariance (Cai *et al.*, 2011), the proportion of the phenotypic variance explained by each marker is approximately

$$h_j^2 = \frac{\tilde{\beta}_j\text{var}(\tilde{x}_j)}{\text{V}_y}. \tag{6}$$

By relying on MCMC samples from the posterior for the parameters in our model, we can obtain posterior summaries for both total heritability $h^2$ and individual $h_j^2$ contributions.

### 2.2.3 MCMC algorithm

The joint posterior distribution is

$$\pi(\tilde{\boldsymbol{\beta}},\mathbf{u},\eta_1,\ldots,\eta_{p_s},\xi_1,\ldots,\xi_{p_s},\tau,\tau_u|\mathbf{y}) \propto \ \pi(\mathbf{y}|\ .\ )$$
$$\times\prod_{j=1}^{p_s}\pi(\tilde{\beta}_j|\eta_j,\tau)\pi(\eta_j|\xi_j)\pi(\xi_j|c,d)$$
$$\times\pi(\mathbf{u}|\tau_u)\pi(\tau|a_1,b_1)\pi(\tau_u|a_2,b_2),$$

where $\pi(\mathbf{y}|.)$ is the probability density of $\mathbf{y}$ given all parameters in the model. From this joint posterior, we obtained full conditional posterior densities for Gibbs sampling as

$$\tilde{\boldsymbol{\beta}}|\mathbf{y},\mathbf{u},\tau \sim N(\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}^p,\Sigma_{\tilde{\boldsymbol{\beta}}}^p)$$
$$\mathbf{u}|\mathbf{y},\tilde{\boldsymbol{\beta}},\tau,\tau_u \sim N(\boldsymbol{\mu}_{\mathbf{u}}^p,\Sigma_{\mathbf{u}}^p),$$
$$\tau|\mathbf{y},\tilde{\boldsymbol{\beta}},\mathbf{u},\boldsymbol{\mu},\Sigma_{\tilde{\boldsymbol{\beta}}} \sim \text{Gamma}(a_1^p,b_1^p),$$
$$\tau_u|\mathbf{u} \sim \text{Gamma}(a_2+n/2, \quad \mathbf{u}^TK^{-1}\mathbf{u}+b_2),$$
$$\xi_j|\tilde{\beta}_j,\tau \sim \text{Gamma}(c+1, \quad \tau^{1/2}|\tilde{\beta}_j|+d),$$
$$\eta_j^{-1}|\tilde{\beta}_j,\xi_j,\tau \sim \text{IN}-\text{Gaussian}\left(\left(\frac{\xi_j^2}{\tau\tilde{\beta}_j^2}\right)^{1/2},\xi_j^2\right)$$

where

$$\boldsymbol{\mu}^p_{\tilde{\boldsymbol{\beta}}} = (\tilde{\mathbf{X}}^T R^{-1}\tilde{\mathbf{X}} + \Sigma^{-1}_{\tilde{\boldsymbol{\beta}}})^{-1}\tilde{\mathbf{X}}^T R^{-1}(\mathbf{y} - \mathbf{u}),$$

$$\boldsymbol{\mu}^p_{\mathbf{u}} = (K^{-1}\tau_u + \tau R^{-1})^{-1}R^{-1}(\mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}),$$

$$\Sigma^p_{\tilde{\boldsymbol{\beta}}} = \tau^{-1}(\tilde{\mathbf{X}}^T R^{-1}\tilde{\mathbf{X}} + \Sigma^{-1}_{\tilde{\boldsymbol{\beta}}})^{-1},$$

$$\Sigma^p_{\mathbf{u}} = (K^{-1}\tau_u + \tau R^{-1})^{-1},$$

$$a^p_1 = \frac{n + p_s}{2} + a_1,$$

$$b^p_1 = \frac{1}{2}((\mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \mathbf{u})^T R^{-1}(\mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \mathbf{u}) + \tilde{\boldsymbol{\beta}}^T\Sigma^{-1}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{\beta}}) + b_1.$$

Gibbs sampling alternates sampling from these simple conditional distributions, with the draws converging to samples from the joint posterior. From these samples, posterior summaries for any marginal of interest can be calculated; e.g. the posterior mean provides the optimal point estimate under squared error loss.

## 3 Simulation study

To evaluate the prediction performance of two-stage-GDP, we conducted a simulation study. We first generated a pedigree based on a non-random mating scheme using the R package `synbreed`. This pedigree structure was then utilized in `simuPOP` (Peng and Amos, 2008), which generated genotype data for each individual taking into account their relatedness. To mimic our motivating real data, we selected the subset of individuals having the same gender for analysis. We then applied Rpackage `kinship2` to calculate the relatedness matrix based on the pedigree.

We considered four different scenarios for the effect sizes:

*Model-1:* 300 SNPs with non-zero regression coefficients generated from $U(-1,\ 1)$ for 100 individuals with 300 SNPs in total.

*Model-2:* 500 SNPs with non-zero regression coefficients generated from $U(-1,\ 1)$ for 100 individuals with 500 SNPs in total.

*Model-3:* 100 SNPs with non-zero regression coefficients generated from $U(1,\ 1.5)$and $U(-1.5, -1)$ for 200 individuals with 2000 SNPs in total.

*Model-4:* 400 SNPs with non-zero regression coefficients generated from $U(1,\ 1.5)$and $U(-1.5, -1)$ for 200 individuals with 4000 SNPs in total.

The first two models are designed for dense problems in which most of the predictors have small effects on the complex trait. The last two models are designed for sparse problems with different level of sparsity. We used cross validation (CV) to assess performance. CV is intrinsically a frequentist idea, but has become routinely used in assessing the frequentist operating characteristics of Bayesian methods (Ghosh *et al.*, 2007). For each scenario, we generated 20 datasets as training sets and 20 as validation sets. Evaluation is based on mean square prediction errors, MSPEs and correlations between observed values and predictive values. The predictions of future values are obtained from

$$\mathrm{E}\,(\mathbf{y}_f|\tilde{\mathbf{X}}_f, \mathbf{y}_o) = \tilde{\mathbf{X}}_f\,\mathrm{E}\,(\tilde{\boldsymbol{\beta}}|\mathbf{y}_o) + K_{f,o}K^{-1}_{o,o}\,\mathrm{E}\,(\mathbf{u}_o|\mathbf{y}_o),$$

where

$$\begin{bmatrix}\mathbf{u}_o\\ \mathbf{u}_f\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau_u^{-1}\begin{bmatrix}K_{o,o} & K_{o,f}\\ K_{f,o} & K_{f,f}\end{bmatrix}\right).$$

Index *f* and *o* denote the future data and observed data, respectively. To evaluate the predictive performance of the model and its computational efficiency, we made a comparison between two-stage-GDP and the widely applied method, SSVS, as well as some well-known shrinkage priors on $\tilde{\beta}_j$ including Bayesian Lasso, Student-*t* and one-stage GDP.

SSVS places a mixture of two normals on $\tilde{\beta}_j$ as $(1 - \gamma_j)N(0, \nu_j^2) + \gamma_j N(0, c_j^2\nu_j^2)$, where $\gamma_j \in \{0, 1\}$. Bayesian Lasso (BL) assigns a double-exponential prior to $\eta_j$ in (4) as $\eta_j \sim \exp(\xi^2/2)$, with $\xi^2/2 \sim \mathrm{Gamma}(c', d')$ following Park and Casella (2008). The student-*t* can be expressed as a scale mixture of normal distributions $\tilde{\boldsymbol{\beta}} \sim N(0, \Sigma_{\tilde{\beta}})$, with $\Sigma_{\tilde{\beta}} = \mathrm{diag}\{\eta_j\}$ and $\eta_j \sim \text{Inv-gamma}(c'', d'')$.

To set the hyperparameters for SSVS, we let $(\nu_j^2, c_j^2\nu_j^2) = (.001, 10)$, with $p(\gamma_j = 1) = 0.5$ in *Model-1* and *Model-2* and to 0.1 in *Model-3* and *Model-4*. In BL, the posterior is relatively insensitive to the prior on $\xi^2$ as long as $c'$ and $d'$ are small (Park and Casella, 2008), and we set these hyperparameters to 0.1. For the student-t prior, we fixed $c'' = 3$ as a small value greater than 2 as suggested by Frühwirth-Schnatter and Wagner (2011). We considered different values for $d''$ as 0.001; 0.01; 0.1 and found 0.01 as the best choice. For GDP hyperparameters, we considered $d = \sqrt{c + 1}$, since this choice ensures the continuity property and creates a trade-off between sparsity and tail-robustness. We increased *c* from 1 in *Model-1* and *Model-2* to 3 in *Model-3* and *Model-4* in order to allow more shrinkage for sparse models. We sent hyperparameters of $\tau$ and $\tau_u$ equal to $10^{-3}$.

Table 1 represents the average MSPEs of 20 simulated validation subsets for each model. The index numbers are the average of standard errors of MSPEs across 200 bootstrap samples of 20 standard errors. The correlation between predicted and observed values in validation sets are also represented in the table. The correlation gives insight into the accuracy of performance comparing different shrinkage approaches.

Table 1 shows that the t-prior performs better than the other approaches in dense problems, *Model-1* and *Model-2*. For the sparse cases, *Model-3* and *Model-4*, two-stage-GDP outperforms the other competitors. Although in *Model-3* two-stage-GDP is slightly better than SSVS, by increasing the sparsity level in *Model-4*, two-stage-GDP shows noticeably better predictive performance. In addition, its computational cost makes feasible the use of two-stage-GDP in GWAS.

The effect of sparsity level on prediction performance is presented in Figure 1. The figure shows mean of correlation between observed values and predicted values in 20 validation sets versus number of predictors in the model, while the number of predictors with non-zero effects in the model is 100 and $n = 100$. The performance of one-stage approaches are highly related to the *p*/*n* ratio. The correlations among observed values and predicted values are less than 0.7 for $p \geq 2000$. The computational time of these approaches versus the number of predictors in the model is presented in Figure 2 . Our programming code for *t*-priors, BL, GDP and two-stage-GDP is based on C++ in order to speed up the MCMC algorithm. The SSVS implemented in `JAGS` is also based on C++ programming.

Figures 1 and 2 represent one line corresponding to *t*-prior, BL and GDP since they almost have the same performance or running time with comparison to SSVS and two-stage-GDP. This line that is depicted using the average point of *t*-prior, BL and GDP quantities, provides better visualization for the comparison.

Available packages using frequentist approaches for linear mixed model with $p > n$ are restricted to specific cases, such as group structure. As those packages do not allow general linear mixed effect structure, such as that induced by familial dependence, we could not provide comparison with frequentist approaches.

**Table 1.** Average of MSPEs and average of correlation between predicted and observed values over 20 validation sets for SSVS, *t*-priors, BL, GDP and Two-stage-GDP

| *Model* | | SSVS | *t*-prior | BL | GDP | Two-stage-GDP |
|---|---|---|---|---|---|---|
| *1* | MSPE | $0.644_{(.029)}$ | $0.616_{(.035)}$ | $0.692_{(.038)}$ | $0.691_{(.036)}$ | $0.655_{(.0380)}$ |
| | Cor | 0.805 | 0.863 | 0.711 | 0.713 | 0.789 |
| *2* | MSPE | $0.604_{(.0277)}$ | $0.567_{(.050)}$ | $0.711_{(.041)}$ | $0.718_{(.035)}$ | $0.623_{(.037)}$ |
| | Cor | 0.786 | 0.819 | 0.640 | 0.643 | 0.757 |
| *3* | MSPE | $0.572_{(.018)}$ | $0.923_{(.024)}$ | $0.791_{(.029)}$ | $0.742_{(.032)}$ | $0.526_{(.030)}$ |
| | Cor | 0.884 | 0.631 | 0.742 | 0.780 | 0.9020 |
| *4* | MSPE | $0.995_{(.019)}$ | $1.045_{(.032)}$ | $0.992_{(.035)}$ | $0.935_{(.041)}$ | $0.610_{(.034)}$ |
| | Cor | 0.544 | 0.501 | 0.511 | 0.567 | 0.827 |

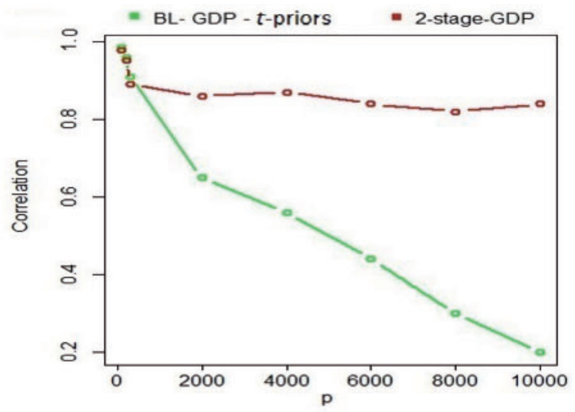

**Fig. 1.** Mean of correlation between observed values and predicted values versus number of predictors in 20 validation sets while the number of predictors with nonzero effects in the model is 100 and $n = 100$. Green line: *t*-prior/BL/GDP and Red line: two-stage-GDP

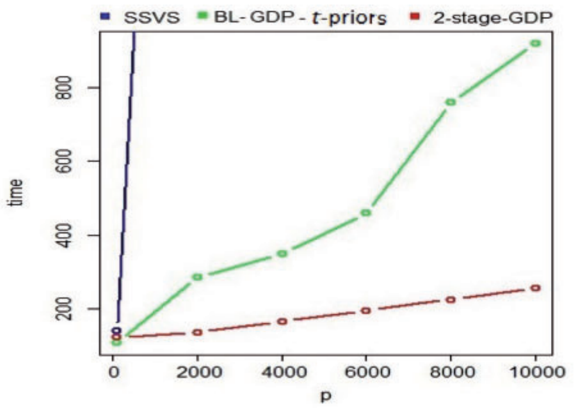

**Fig. 2.** Computational time versus the number of predictors or SNP in the model where $n = 100$. Blue line: SSVS, Green line: *t*prior/BL/GDP and Red line: two-stage-GDP

## 4 Application to real data

We applied two-stage-GDP on a GWAS dataset from an animal breeding study. The aim of study is to improve quantity and quality of milk production through investigating milk protein yield. The data contain 707, 962 SNPs genotyped for 607 Holstein Bulls. After quality control, excluding SNPs with minor allele frequency below 5%, samples with low genotyping efficiency, and violating Hardy-Weinberg Equilibrium test, the dataset contains 555, 651 SNPs. We further reduced the dimensionality of the set of SNPs by considering the fact that SNPs in the genome have groups of neighbors such that they are all nearly perfectly correlated with each other. Hence, the genotype of one SNP can perfectly predict those of correlated neighboring SNPs; i.e. one SNP can thereby serve as proxy for many others in the analysis. As the segments of SNPs in high linkage disequilibrium in cattle is longer than human because of inbreeding, we reduced the number of SNPs down to 135, 545 via a hierarchical clustering approach. In this agglomerative clustering, we defined the distance between two SNPs as $1 - r^2$, where $r^2$, the square of correlation between two SNPs, is 0.85.

After the preliminary analysis, we first evaluated the impact of each SNP by BF and selected 878 top-ranking SNPs using the algorithm in Supplementary Appendix B. In the second-stage, we first ran a 5-fold cross validation in order to set the hyperparameters. For hyperparameters of marker effects, we considered $d = \sqrt{c+1}$, since this choice ensures the continuity property and creates a trade-off between sparsity and tail-robustness (Armagan *et al.*, 2013). We ran 5-fold cross validation for different values of $c$: 1, 2.5, 3, 3.5. Generally, the rate of shrinkage increases along this path. For the other hyperparameters of the model, we set $(a_1, b_1) = (a_2, b_2)$ as $(0.001, 0.001)$, $(0.01,\ 0.01)$, $(0.1,\ 0.1)$, $(0.3,\ 0.3)$. Table 2 presents average of MSPEs for different sets of hyperparameters for 5-fold cross validation and the standard deviation from 50 bootstrap samples in subscript. It shows $c = 3$ and $(a_1,\ b_1) = (a_2,\ b_2) = (0.01\ ,\ 0.01)$ provides smallest MSPE.

In addition, we considered two other prior specifications for the second-stage analysis, *t*-priors and BL, in order to compare their predictive performance. The hyperparameters of *t*-priors and BL were chosen the same as in the simulation study.

Table 3 represents average of MSPEs of 10-fold cross validation. The index numbers are the average standard errors of MSPEs obtained by averaging 100 bootstrap samples of 10 MSPEs. We also calculated average of correlation between predicted and observed values in 10-fold cross validation. Although two-stage GDP shows slightly better predictive performance in terms of mean of MSPEs in 10-fold cross validation shown in Figure 3, the variation of 10-fold cross validation result based on two-stage-GDP is smaller than two-stage-BL and two-stage-*t*-priors.

The calculated total phenotypic variance from (5) for two-stage-GDP is 1.700. The total genetic variance contributed by the additive effects of the markers calculated from the first term of the right hand side in (5) is 0.483. Although the aim of this study is prediction, if one is interested in selecting the set of SNPs with the most contribution, heritability is a good criteria. To select a set SNPs based on heritability, Hoti and Sillanpää (2006) suggested presenting a threshold value, $t$, such that one SNP is included in the final model if the heritability explained by this SNP is greater than $t$. Under such an approach, the heritability for milk protein yield is expected to be small based on previous studies. Hence, instead of setting a threshold on heritability of each SNP, we considered a threshold on total heritability of a set of top SNPs ranked based on calculated $h_j^2$s. To this end, we first calculated heritability for each SNP by substituting the mean of posterior samples of $\beta_j$s in (5). This provided a list of ranked SNPs. We then selected a set of top SNPs with total heritability above 0.2. The estimated effect sizes and marginal heritabilities of 32 selected SNPs as well as their chromosomes' numbers are tabulated in Table 4.

Among these selected SNPs, 19 SNPs out of 32 have been found in known genes or nearby them. Figure 4 shows pieces of

**Table 2.** Average of MSPEs for different values of hyperparameters for 5-fold cross validation and the standard deviation from 50 bootstrap samples in subscript

| | C | | | |
|---|---|---|---|---|
| $(a_1, b_1) =$ <br> $(a_2, b_2)$ | 1 | 2.5 | 3 | 3.5 |
| (0.001,0.001) | $0.698_{(0.018)}$ | $0.539_{(0.012)}$ | $0.526_{(0.013)}$ | $0.529_{(0.010)}$ |
| (0.01, 0.01) | $0.686_{(0.019)}$ | $0.528_{(0.011)}$ | $0.519_{(0.011)}$ | $0.523_{(0.015)}$ |
| (0.1,0.1) | $0.736_{(.028)}$ | $0.564_{(0.018)}$ | $0.551_{(.022)}$ | $0.557_{(0.023)}$ |
| (0.3,0.3) | $0.845_{(0.028)}$ | $0.576_{(0.018)}$ | $0.572_{(0.058)}$ | $0.569_{(0.023)}$ |

**Table 3.** Average of 10-fold cross validation MSPEs and correlation of observed values and predicted values using two-stage methods, GDP, *t*-priors and BL. The subscripts are standard deviations based on 100 bootstrap samples

| *Model* | Two-stage-GDP | Two-stage-*t*-priors | Two-stage-BL |
|---|---|---|---|
| MSPE | $0.518_{(0.0276)}$ | $0.551_{(0.0362)}$ | $0.568_{(0.0344)}$ |
| Cor | 0.751 | 0.712 | 0.703 |

**Table 4.** Position of selected SNPs with their effect sizes and heritabilities

| Chr | $\beta_j$ | $h_j^2(\%)$ | Chr | $\beta_j$ | $h_j^2(\%)$ |
|---|---|---|---|---|---|
| 1 | −0.0562 | 0.131 | 13 | −0.0578 | 0.150 |
| 1 | 0.0617 | 0.192 | 13 | 0.0585 | 0.173 |
| 1 | −0.0983 | 2.930 | 13 | 0.0569 | 0.138 |
| 3 | −0.0580 | 0.120 | 14 | −0.0595 | 0.232 |
| 3 | −0.0572 | 0.156 | 17 | 0.0905 | 2.552 |
| 3 | −0.0572 | 0.138 | 18 | 0.0569 | 0.140 |
| 4 | −0.0595 | 0.204 | 18 | 0.0790 | 1.584 |
| 4 | −0.0591 | 0.206 | 22 | 0.0586 | 0.179 |
| 4 | 0.0796 | 1.529 | 25 | 0.0715 | 0.976 |
| 6 | 0.0707 | 0.906 | 25 | −0.0753 | 1.091 |
| 7 | −0.0726 | 0.950 | 25 | −0.0553 | 0.110 |
| 8 | 0.0571 | 0.136 | 28 | −0.0503 | 0.100 |
| 8 | 0.0606 | 0.798 | 28 | 0.0617 | 0.741 |
| 9 | 0.0636 | 0.821 | 28 | 0.0691 | 0.880 |
| 10 | 0.0566 | 0.140 | 28 | 0.0589 | 0.197 |
| 12 | −0.0581 | 0.182 | 29 | 0.0580 | 0.200 |

chromosome 1 and 10 as examples. The red regions in these pictures indicate the locations of the genes associated with milk protein yield that have been found through previous studies. The two vertical red lines represent the locations of two selected markers in our study. As it shows, the selected SNP in Chromosome 1 is in gene `PPP2R3A` and the one in Chromosome 10 is between genes `SAV1` and `NIN`.

We further examined the correlations between the SNPs that are selected from the same chromosome. The correlation matrices of these markers presented in Supplementary Appendix C reveal detected SNPs in most of the Chromosomes are weakly correlated. Although, the correlation matrix for chromosome 3 shows two markers closely link to each other. Hence, selecting those two markers might be due to linkage disequilibrium.

## 5 Discussion

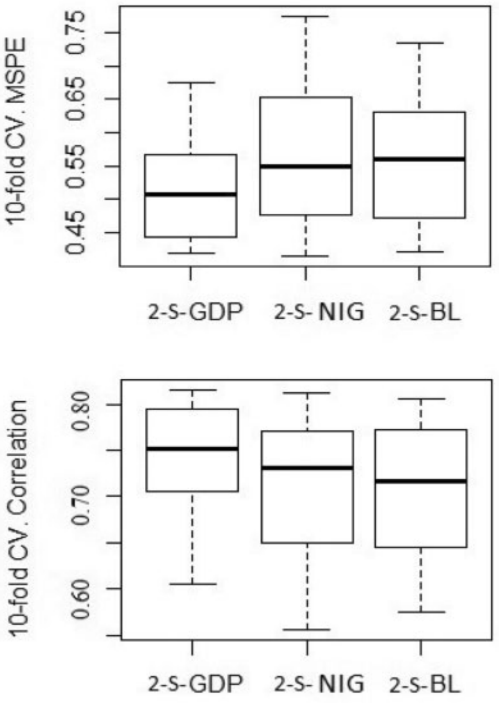

**Fig. 3.** Top: box plot of MSPE obtained through 10-fold cross validation, bottom: box plot of correlation between predicted and observed values in 10-fold cross validation

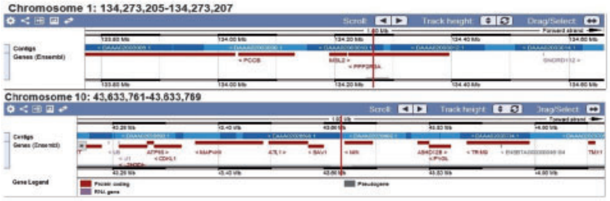

**Fig. 4.** Location of two selected markers from chromosome 1 and 10

In $p \gg n$ problems, current variable selection and mixture priors face problems in scaling to very large $p$ when $n$ in small relative to $p$. These problems include computational bottlenecks and insufficient flexibility. To solve these problems, multistage approaches have been proposed, but not yet in the case in which the subjects under study are related. We address this gap via new two-stage Bayesian approaches that account for familial dependence. The method is a hybrid of single marker and simultaneous analyses. In the first-stage, we select a superset of the most promising SNPs by evaluating the impact of each SNP individually while accounting for related samples. In the second stage, we simultaneously select and estimate the parameters of the model by placing a GDP prior on the SNP coefficients, again accounting for relatedness.

Our simulation analyses revealed that two-stage-GDP improves predictive performance in comparison to one-stage analysis when $n$ is too small relative to $p$. The computational cost of this proposed approach also makes it feasible for large scale problems like GWAS.

In the real data analysis, we made a comparison among different prior specifications in the second stage of analysis. This comparison represented that two-stage-GDP performs better than two-stage-*t*-priors and two-stage-BL. We then estimated breeding value for protein yield based on two-stage-GDP. Although the prediction accuracy was sufficient for the small sample size that we had, a small proportion of phenotypic variation is explained by SNPs. In problems that the sample size is not severely limited, splitting the data in two subsets and applying each stage on a subset of data (see for e.g. Wasserman and Roeder, 2009) may better recover this missing heritability, which may also arise from interactions.

## Acknowledgements

This makes use of data generated by *INNOVAGEN* Project. More information about this project is available from *http://centridiricerca.unicatt.it/dicdr–centridiricerca/biodna_2057. html?rdeLocaleAttr=en.* First author thanks Mohammad Shariati for his comments on this manuscript.

*Conflict of Interest*: none declared.

## References

Armagan,A. *et al.* (2013) Generalized double pareto shrinkage. *Statistica Sinica*, **23**, 119–143.

Ball,R.D. (2011) Experimental designs for robust detection of effects in genome-wide casecontrol studies. *Genetics*, **189**, 1497–1514.

Beattie,S.D. *et al.* (2002) A two-stage Bayesian model selection strategy for supersaturated designs. *Technometrics*, **44**, 55–63.

Buckler,E.S. *et al.* (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714–718.

Cai,X. *et al.* (2011) Fast empirical Bayesian lasso for multiple quantitative trait locus mapping. *BMC Bioinformatics*, **12**, 211.

Carvalho,C.M. *et al.* (2009) Handling sparsity via the horseshoe. *J. Mach. Learn. Res.*, **5**, 73–80.

Diabetes Genetics Initiative of Broad Institute of Harvard, MIT LUND University, Novartis Institutes for BioMedical Research. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **316**, 1331–136.

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

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849–8911.

Frühwirth-Schnatter,S. and Wagner,H. (2011) Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In: Bernardo,J.M. *et al.* Bayesian Statistics, **9**, 165–200. Oxford University Press, London.

Ghosh,J.K. *et al.* (2007) *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, LLC, New York.

Gianola,D. *et al.* (2009) Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**, 347–363.

Griffin,J.E. and Brown,P.J. (2011) Bayesian Hyper-lasso with non-convex penalization. *Aust. NZ J. Stat.*, **53**, 423–442.

Hoggart,C.J. (2008) Simultaneous analysis of all snps in genome-wide and resequencing association studies. *PLoS Genet.*, **4**, e1000130.

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

Huanga,H. *et al.* (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *PNAS*, **107**, 6823–6828.

Jeffreys,H. (1961) *Theory of Probability*. Oxford University Press, New York.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Kessler,D.C. *et al.* (2014) Learning phenotype densities conditional on many interacting predictors. *Bioinformatics*, **30**, 1562–1568.

Li,J. *et al.* (2011) The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.

Meuwissen,T.H.E. *et al.* (2001) Prediction of total genetic value using genomewide dense marker maps. *Genetics*, **157**, 1819–1829.

Oliehoek,P.A. *et al.* (2006) Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, **173**, 483–496.

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. *Ann. Stat.*, **36**, 1595–1618.

Peng,B. and Amos,C.I. (2008) Forward-time simulations of nonrandom mating populations using simupop. *Bioinformatics*, **24**, 1408–1409.

Shariati,M.M. *et al.* (2012) A two step bayesian approach for genomic prediction of breeding values. *BMC Proceedings*, **6**, S12. BioMed Central Ltd.

Stephens,M. and Balding,D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681–690.

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

Wasserman,L. and Roeder,K. (2009) High dimensional variable selection. *Ann. Stat.*, **37**, 2178.

Weedon,M.N. *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, **40**, 575–583.

Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature*, **447**, 661–678.

Zhang,C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat*, **38**, 894–942.

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