A Modification of the Lasso Method by Using the Bahadur Representation for the Genome-Wide Association Study

Lev V. Utkin
Peter the Great St.Petersburg Polytechnic University, St.Petersburg, Russia
E-mail: lev.utkin@gmail.com

Yulia A. Zhuk ITMO University, St.Petersburg, Russia
E-mail: zhuk_yua@mail.ru

Keywords: data analysis, feature selection, Lasso, Bahadur representation, genome-wide association study

Received: September 24, 2016

A modification of the Lasso method as a powerful machine learning tool applied to a genome-wide association study is proposed in the paper. From the machine learning point of view, a feature selection problem is solved in the paper, where features are single nucleotide polymorphisms or DNA-markers whose association with a quantitative trait is established. The main idea underlying the modification is to take into account correlations between DNA-markers and peculiarities of phenotype values by using the Bahadur representation of joint probabilities of binary random variables. Interactions of DNA-markers called the epistasis are also considered in the framework of the proposed modification. Various numerical experiments with real datasets illustrate the proposed modification.

Povzetek: Predstavljena je modifikacija metode strojnega učenja, imenovana Lasso.

1 Introduction

One of the important area for a successful application of the artificial intelligence, in particular, machine learning algorithms, is the computational biology which can be regarded as a basis for many engineering problems in biotechnology. An interesting task clearly illustrating the application of artificial intelligence to biotechnology problems is a genome-wide association study (GWAS). GWAS examines the association between phenotypes or quantitative traits and genetic variants or genotypes across the entire genome. In the machine learning framework, it can be regarded as one of the methods for a feature selection problem where features are the so-called single nucleotide polymorphisms (SNPs) or DNA-markers. As pointed out in [12, 15], there are some difficulties of solving this feature selection problem. First of all, the number of SNPs p is typically 10–100 times the number of individuals n in the training sample. This is the so called $p > n$ (or large p small n) problem, which leads to difficulty of an oversaturated model. Another difficulty is that SNPs may affect phenotype in a complicated and unknown manner. For example, some DNA-markers may interact in their effects on phenotype. This interaction is called the epistatic effect.

A huge amount of the statistical models and methods solving the SNP selection problem have been developed the last decades. A part of methods can be referred to as *filter* methods [1, 28] which use statistical properties of SNPs to filter out poorly informative ones. A review of filter methods in GWAS is proposed by Zhang et al. [53]. The t-test, Fisher criterion (F-statistics), χ^2-statistics, ANOVA tests are the well-known statistical methods for detecting differential SNPs between two samples in training data.

Another part of methods called *wrapper* methods generally provides more accurate solutions than the filter methods, but it is computationally demanding [24]. One of the well-known wrapper methods proposed by Guyon et al. [17] and called the Recursive Feature Elimination has been applied to the gene selection problem for cancer classification.

Filter methods and their modifications as well as wrapper methods may be efficient tools for solving the problems of GWAS. At the same time, a lot of methods of the feature selection use regression models. One of the pioneering and the most well-known papers devoted to the use of regression models in SNP selection has been written by Lander and Botstein [27]. Methods for constructing the corresponding regression models can be referred as *embedded* methods [25]. They performs feature selection in the process of model building and cover a lot of well-known approaches, including the Ridge regression, Least Absolute Shrinkage and Lasso techniques [41] which are the most popular and efficient tools in SNP selection problems. The main advantage of using the Lasso method is that it performs variable selection and classification or regression simultaneously. A lot of approaches using the Lasso method and its modifications have been developed for solving the SNP selection problem in the framework of the GWAS [13, 31, 35, 36, 40, 43]. Hayes [18] provided a comprehensive overview of statistical methods for GWAS in animals, plants, and humans. Various approaches to SNP selection with the Lasso algorithm and other methods can be also found in papers [16, 22, 33, 46].

The main aim of GWAS is to identify SNPs that are directly associated with a trait, i.e., the standard GWAS analyzes each SNP separately in order to identify a set of significant SNPs showing genetic variations associated with the trait. However, an important challenge in the analysis of genome-wide data sets is taking into account the so-called epistatic effect when different SNPs interact in their association with phenotype.

Campos et al. [12] explain some shortcomings of the standard GWAS. They write that the currently identified SNPs might not fully describe genetic diversity. For instance, these SNPs may not capture some forms of genetic variability that are due to copy number variation. Moreover, genetic mechanisms might involve complex interactions among genes and between genes and environmental conditions, or epigenetic mechanisms which are not fully captured by additive models. Many statistical approaches make sense under the assumption that only a few genes affect genetic predisposition. However, GWAS may be unsatisfactory for many important traits which may be affected by a large number of small-effect, possibly interacting, genes. Limitations and pitfalls of prediction analysis in the framework of the GWAS have been discussed in detail by Wray et al. [47] where it is shown how naive implementations can lead to severe bias and misinterpretation of results.

In fact, the epistatic effect can be viewed as gene-gene interaction when the action of one locus depends on the genotype of another locus. At the same time, there are different interpretations of the epistatic effect. A fundamental critical review of different definitions and interpretations of epistasis is provided by Cordell [11] where it is pointed out that there are many conflicting definitions of epistasis, which lead to certain problems in interpretations, namely, the statistical interaction may not correspond to the biological models of epistasis. As indicated by Wan et al. [44], there are mainly three different definitions of gene-gene interactions: functional, compositional and statistical epistasis. We consider only the statistical epistasis which can be regarded as the statistical deviation from the joined effects of two SNPs on the phenotype. At that, the individual SNPs may exhibit no marginal effects.

A lot of methods dealing with epistasis effect have been developed last decades [3, 30, 52, 49, 50, 51, 54]. Comprehensive and interesting reviews of methods detecting interacting the epistatic effect were provided by several authors [7, 45]).

Analyzing various modifications of the Lasso method applied to the GWAS problems, we can point out that many efficient modifications are based on applying special forms of the penalty function, which take into account some additional information about SNP markers and the corresponding phenotype values. Some interesting algorithms [33, 42] devoted to various penalty functions will be studied in the next section. The use of a specific additional information allows us to improve the GWAS and is considered in the paper.

In the present study, we modify the Lasso method by taking into account some peculiarities of the double haploid (DH) lines of barley which are very important in the plant biotechnology. According to the DH method, only two types of genotypes occur for a pair of alleles. From a statistical point of view, we solve a linear regression problem with binary explanatory variables. Our method is based on the well-known adaptive Lasso [56] and takes into account additional information about the correlation between SNPs, frequencies of alleles and expected phenotype values. We propose to use the Bahadur representation [2] by partially applying the ideas provided by Lee and Jun [29] where the authors propose to apply the Bahadur representation to classification problems. The Bahadur representation allows us to compute joint probabilities of SNPs by taking into account the correlation between binary random variables. That is another reason why we analyze only DH lines which produce the binary genotypes. In order to modify the adaptive Lasso, we propose to assign penalty weights in accordance with expected values of the phenotype with respect to a probability mass function somehow defined on the genotype values. In other words, computing the expected values of the phenotype in a special way is a main idea of the proposed method. We show that the proposed modification is directly extended on the case taking into account the epistatic effect.

2 The Lasso method

We analyze n double haploid (DH) lines of barley or a population from n individuals. From a statistical point of view, marker genotypes can be treated as qualitative explanatory variables, i.e., $X_j = (x_{1j}, ..., x_{nj})^{\mathrm{T}}$ is a predictor representing the j-th SNP, $j = 1, ..., p$. Here x_{ij} is a binary variable, i.e., $x_{ij} \in \{0, 1\}$. A quantitative trait of interest or a set of the phenotype values $y_i \in \mathbb{R}$, $i = 1, ..., n$, can be regarded as the response vector $Y = (y_1, ..., y_n)^{\mathrm{T}}$. We also denote $\mathbf{X} = [X_1, ..., X_p]$ is a genotype matrix for n lines or individuals or a predictor matrix in terms of statistics; $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, ..., x_{ip})$ is a vector of alleles corresponding to the i-th line, $i = 1, ..., n$.

First, we focus on the standard linear regression model

$$y = \sum_{i=1}^{p} \beta_i X_i + \beta_0 + \epsilon = \mathbf{X}\beta + \beta_0 + \epsilon. \tag{1}$$

Here ϵ is a noise variable with the zero-valued expectation; β_i is the SNP effect, $\beta = (\beta_1, ..., \beta_p)$.

Without loss of generality, we assume the predictors and the response are centered, and the predictors are standardized, that is

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1,\ X_i \in \mathbb{R}^p.$$

This implies that the intercept is not included in the regression function.

The Lasso is a regularization technique for simultaneous estimation and variable selection [41]. The Lasso estimates are defined from the following quadratic programming problem:

$$\beta = \arg\min_{\beta\in\mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2,$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

for some $s > 0$. The Lagrange formulation is

$$\beta = \arg\min_{\beta\in\mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a nonnegative regularization parameter. The second term is the L_1 penalty which is crucial for the success of the Lasso. The Lasso estimator is usually calculated at a grid of tuning parameters of λ, and a cross-validation procedure is subsequently used to select an appropriate value of λ.

The Lasso penalizes the regression coefficients by their L_1 norm. However, in order to improve the performance of the Lasso, the regression coefficients can be penalized individually. As a result, we write the weighted Lasso estimates as follows:

$$\beta = \arg\min_{\beta\in\mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $w_j > 0$, $j = 1,...,p$, are weights determined a priori in accordance with some rules. A larger weight w_j corresponds to a higher penalty and discourages the j-th predictor from the model. Conversely, a smaller weight w_j exerts less penalty and encourages selection of the corresponding predictor [55].

The penalized Lasso can be reformulated as the standard Lasso problem [6]. If we introduce new covariates and regression coefficients as

$$\widetilde{x}_{ij} = x_{ij}/w_j,\ i = 1,...,n,\ \widetilde{\beta}_j = \beta_j w_j,$$

then the weighted Lasso problem can be rewritten as follows:

$$\beta = \arg\min_{\beta\in\mathbb{R}^p} \left\|Y - \widetilde{\mathbf{X}}\widetilde{\beta}\right\|^2 + \lambda \sum_{j=1}^{p} \left|\widetilde{\beta}_j\right|,$$

where $\widetilde{\beta}$ and $\widetilde{\mathbf{X}}$ are the vector and the matrix with elements $\widetilde{\beta}_j$ and $\widetilde{x}_{ij}$, respectively.

Zou [56] proposed one of the methods for determining the weights w_j such that $w_j = 1/|\beta_{init,j}|$, where $\beta_{init,j}$ is a prior estimator of β_j, for example, the least square estimator. The corresponding Lasso problem is referred as the adaptive Lasso, and it has many nice properties improving the performance of the Lasso. Moreover, it can be a basis for constructing the boosting Lasso [6].

The Lasso has many interesting properties which make the method to be very popular. At the same time, Zou and Hastie [57] point out that in spite of success of the Lasso it has some limitations, in particular, if there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected. In other words, the Lasso tends to put all the weight on the selected variable. On the one hand, this is a shortcoming. Many methods have been proposed to overcome this obstacle, for example, the so-called the elastic net [57] where the estimates are defined by

$$\beta = \arg\min_{\beta\in\mathbb{R}^p} \left\|Y - \widetilde{\mathbf{X}}\widetilde{\beta}\right\|^2 + \lambda_1 \sum_{j=1}^{p} \left|\widetilde{\beta}_j\right| + \lambda_2 \left\|\widetilde{\beta}\right\|^2.$$

However, the elastic net requires to assign an additional parameter λ_2 whose value is a priori unknown. On the other hand, in contrast to the ridge regression which tends to select all of the correlated variables and make the corresponding coefficients to be equal, the Lasso selects a group of correlated variables and "isolates" it.

A special choice of the penalty term on the basis of some prior information about SNPs or about an exploited genome selection model itself may lead to a series of useful or important properties of the regression or classification model corresponding to the Lasso. Liu et al. [33] tried to apply the observed fact that there exists a natural grouping structure in SNPs and, more importantly, such groups are correlated. The authors proposed a new penalization method for group variable selection which can properly accommodate the correlation between adjacent groups. Their method referred to as smoothed group Lasso is based on a combination of the group Lasso penalty and a quadratic penalty on the difference of regression coefficients of adjacent groups. Liu et al. [33] assume that SNPs are divided into J groups, each with size d_j, $j = 1,...,J$, according to their physical locations and correlation patterns. As a results, the vector $\beta = (\beta_1,...,\beta_J)$ is defined for groups of SNPs, but not for separate SNPs, β_j is the vector of parameters corresponding to SNPs from the j-th group. The authors consider the quadratic loss function of the form:

$$l(\beta) = \left\|Y - \sum_{j=1}^{J} \mathbf{X}_j\beta_j\right\|^2.$$

Here $\mathbf{X}_j$ is an $n \times d_j$ matrix corresponding to the j-th group [33]. There are two main difficulties of using the above considered method. First, it is rather hard from the computation point of view. Second, we have to know a priori the grouping structure SNPs.

An interesting approach for dealing with correlated covariates was proposed by Tutz and Ulbricht [42]. Their method utilizes the correlation between predictors explicitly in the penalty term. Coefficients which correspond to pairs of covariates are weighted according to their marginal correlation. The correlation based penalty is given by

$$Q_\lambda(\beta) = \lambda \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}} \right\}.$$

Here ρ_{ij} denotes the empirical correlation between the i-th and the j-th predictors. If we have the positive correlation, i.e., $\rho_{ij} \to 1$, then the first term in the sum becomes dominant. When $\rho_{ij} \to -1$, then the second term becomes dominant. Both these cases lead to the approximate equality $\beta_i \approx \beta_j$. In case of uncorrelated predictors and $\rho_{ij} \to 0$, the corresponding model is reduced to the ridge regression.

Another model proposed by Park and Hastie [38] constructs sets of indicators representing all the available factors and all possible two-way interactions in order to fit gene-interaction models with the data consisting of genotype measurements and a binary response. The obtained grouped variables are used in the path-following algorithm for the group-Lasso method.

In order to take into account different probabilities of feature values, in particular, to take into account the allele frequency, Zhou et al. [55] proposed a weighted Lasso penalty in the Lasso method such that the weights are assigned in accordance with the following sources of prior knowledge. First, Zhou et al. [55] considered genotyping errors such that the unreliable variants should be penalized more. Second, they pointed out that the allele frequencies can be used in accordance with an idea of Madsen and Browning [34] where it was proposed to take the weight $w = 2\sqrt{\pi(1-\pi)}$ for a variant with population frequency π by arguing that this scheme assigns smaller penalties to rarer variants as suggested by classical population genetics theory.

3 The proposed method

3.1 Motivation for a new penalized method

The considered in previous sections modifications of the Lasso are efficient tools for solving the GWAS and SNP selection problems. Their performance has been experimentally shown by many authors [13, 22, 23, 31, 33]. However, every real application problem has some peculiarities whose accounting might improve the regression method. Let us mention these peculiarities.

1. First of all, our aim is not to find the "best" regression model for the given information, but to select SNPs which impact on the smallest (largest) values of the phenotype, for example, on the heading date early flowering of barley in the studied applications. This does not mean that the whole fitted regression model is not important for us. We have to combine two above aims. This can be done by introducing the weighted Lasso penalties of a special form. This form has to take into account in the first place the smallest values of phenotype. Smaller values of the phenotype produce larger weights, whereas larger values of the phenotype should be also considered. It would seem that we can assign the weights to phenotype values with respect to their closeness to the minimal phenotype value. However, the phenotype values are random and depend on environment conditions. Moreover, the smallest phenotype value does not mean that its value is caused by the corresponding genotype. This implies that we cannot assign weights to the available phenotype values. *The main idea underlying the method is to assign weights to expected values of the phenotype with respect to a probability mass function somehow defined on the genotype values.*

2. The genotype values corresponding to every SNP in the studied application make up a binary vector. The dependence of SNPs leads to dependence of the corresponding binary vectors which can be estimated.

3. The allele frequencies and correlations indirectly impact on the smallest values of the phenotype.

3.2 A method for computing weights for the Lasso

By extending the ideas proposed by aforementioned authors [34, 42, 55], we define the weighted Lasso penalty in a new way. The main idea is the following. We define the average contribution of every SNP to the mean phenotype value. These contributions or their function are nothing else but the weights w_k in the adaptive Lasso. They have to take into account the probabilities of alleles, the correlations between SNPs and the phenotype values. The next question is how to determine the average contribution of every SNP. It can be carried out as follows:

1. For every genotype vector $\mathbf{x}_j$ (the j-th line), we compute joint probabilities $\pi(x_{kj}, x_{ij})$ of all pairs (k, i) of SNPs by taking into account correlations between pairs of random variables (SNPs).

2. For every pair (k, i), we compute the mean phenotype value R_{ki} as the expectation of phenotypes with respect to the joint probabilities $\pi(x_{kj}, x_{ij})$ over all lines or individuals.

3. The average contribution of every, say k-th, SNP into the phenotype is computed by averaging the mean phenotype values R_{ki} over all $i = 1, ..., p$.

4. The weights or their function for the adaptive Lasso are defined by the average contributions.

Below we consider every step in detail.

3.3 Bahadur representation

The main idea for using the joint probability $\pi(x_{jk}, x_{ji})$ is to take into account the correlation between SNPs with indices k and i. For every pair of SNPs X_k and X_i, we have to determine the joint probability $\pi(x_{kj}, x_{ij})$, $i = 1, ..., p$, $i \neq k$, of the j-th individual. It can be computed by using the so-called Bahadur representation proposed by Bahadur [2]. The Bahadur representation takes into account the correlation between binary variables, and it can be written in the case of two binary variables with numbers k and i as

$$\begin{aligned}\pi(x_k, x_i) &= \pi_k^{x_k}(1-\pi_k)^{1-x_k} \cdot \pi_i^{x_i}(1-\pi_i)^{1-x_i} \\ &\times (1 + \rho_{ki} u_k u_i)\,. \qquad (2)\end{aligned}$$

Here π_k is the probability of an allele for the k-th SNP or its allele frequency, i.e., $\pi_k = \Pr\{x_k = 1\}$; ρ_{ki} is the correlation coefficient between the k-th and the i-th SNPs which is defined as $\rho_{ki} = \mathbb{E}\left[U_k U_i\right]$, where the random standardized variable U_k takes the values u_k such that there hold

$$U_k = \frac{X_k - \pi_k}{\sqrt{\pi_k(1-\pi_k)}},\ u_k = \frac{x_k - \pi_k}{\sqrt{\pi_k(1-\pi_k)}}.$$

Note that the first term in the right-hand side of the expression for $\pi(x_k, x_i)$ represents the joint probability mass function under condition that variables X_k and X_i are statistically independent. The second term includes the interaction from the first order up to the second. Note also that U_k should be evaluated by estimating π_k.

The corresponding estimates of parameters π_k, u_k, ρ_{ki} denoted as $\widehat{\pi}_k$, $\widehat{u}_k$, $\widehat{\rho}_{ki}$ are computed by means of the following expressions:

$$\widehat{\pi}_k = \sum_{l=1}^{n} x_{kl}/n,\ \ \widehat{\rho}_{ki} = \sum_{l=1}^{n} \widehat{u}_{kl}\widehat{u}_{il}/n,$$

where n is the number of individuals and

$$\widehat{u}_{kl} = \frac{(x_{kl} - \widehat{\pi}_k)}{\sqrt{\widehat{\pi}_k(1-\widehat{\pi}_k)}}$$

is the l-th observed value of variable U_k.

It should be noted that the Bahadur representation can be written also for joint probabilities of three, four, etc. variables. [32] mention a property of the Bahadur representation such that the joint probability distribution of any subset $x_1, x_2, ..., x_t$ can be written as follows:

$$\begin{aligned}\pi(x_1, ..., x_t) &= \prod_{i=1}^{t} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \\ &\times \left(1 + \sum_{Q \subset \{1,...,t\},\ |Q| \geq 2} \rho_Q \prod_{k \in Q} u_k \right).\end{aligned}$$

Here ρ_Q represents $\rho_{i_1,...,i_k}$ if $Q = \{i_1, ..., i_k\}$ and $|Q|$ denotes the number of elements in Q. The main disadvantage of the Bahadur representation is the large number of parameters and hard computations required for getting the probabilities. Therefore, we restrict our study only by probabilities of two variables.

It should be noted that the Bahadur representation has been used in some classification models. One of the interesting models for discriminant analysis of binary data was proposed by Lee and Jun [29]. The main contribution of [29] is that they proposed to take into account the correlation between variables or, more exactly, estimates of the correlation by means of the Bahadur representation.

There are pros and cons of using this model when the number of variables is larger than the number of observations. For example, Bickel and Levina [4] suppose that classification rules ignoring the correlation structure often perform better in this case. However, Lee and Jun [29] show by means of various experimental studies that the correlation should be taken into account in all cases.

In spite of arguments of [29] in defense of the correlation analysis for high-dimensional data, there is a risk of incorrect estimates of interactions of the large order. Moreover, it is practically impossible to compute the corresponding joint probabilities when the number of SNPs is rather large. Therefore, we propose an approach which partially uses joint probabilities of variables and partially takes into account the correlation between the variables.

3.4 Average contributions of SNPs

In order to determine the average contribution of the k-th SNP into the mean value of the phenotype, we consider all possible pairs of SNPs such that one of the SNPs in every pair is the k-th SNP, i.e., we are interesting in considering $p-1$ pairs of SNPs with numbers $(k,1), ..., (k, k-1), (k, k+1), ..., (k,p)$. Every pair, say (k,i), determines a mean phenotype value R_{ki} corresponding to this pair of SNPs as follows:

$$R_{ki} = \frac{\sum_{j=1}^{n} \pi(x_{kj}, x_{ij}) y_j}{\sum_{j=1}^{n} \pi(x_{kj}, x_{ij})}. \qquad (3)$$

In other words, we can compute the expected phenotype value under condition that every phenotype value y_j is produced by the subset of the genotypes corresponding to the k-th and the i-th SNPs. The measure R_{ki} can be regarded as a contribution of the k-th and the i-th SNPs to the mean phenotype value.

Then the contribution of the k-th SNP denoted by $\widetilde{R}_k$ into the mean phenotype value can be determined through averaging the measures R_{ki}, i.e., it is computed as

$$\widetilde{R}_k = \frac{1}{p-1} \sum_{i=1, i \neq k}^{p} R_{ki}. \qquad (4)$$

It is obvious that the smaller values of the measure $\widetilde{R}_k$ give us significant or top ranked SNPs and exert less penalty w_k, i.e., we can introduce an increasing function g such that

$$w_k = g\left(1/\left|\beta_{init,k}\right|\right).$$

One of the possible functions which will be used in numerical experiments is

$$w_k = \left(\frac{\widetilde{R}_k - \min_{k=1,...,p} \widetilde{R}_k}{\max_{k=1,...,p} \widetilde{R}_k - \min_{k=1,...,p} \widetilde{R}_k} \right)^{-q}. \quad (5)$$

Here q is a positive real which defines how changes of the difference between $\widetilde{R}_k$ and $\min_{k=1,...,p} \widetilde{R}_k$ impact on changes of weights w_k. The number q can be regarded as a tuning parameter whose optimal value can be obtained by means of the cross-validation procedure.

In sum, the obtained weights take into account the correlation between SNPs, the allele frequencies, binary data and the fact that the smallest (largest) values of the phenotype are more important in comparison with other values because we are looking for the SNPs which impact on the values of some trait with predefined properties, for example, the heading date of barley should be as small as possible. At the same time, we do not need to directly use the obtained weights and to implement the adaptive Lasso algorithm. It has been mentioned in the previous section that the adaptive Lasso can be transformed to the standard Lasso by means of introducing new covariates $\widetilde{x}_{ij} = x_{ij}/w_j$.

Finally, we write the following SNP selection algorithm.

Algorithm 1 The SNP selection algorithm.

```
Require: $Y = (y_1, ..., y_n)^{\mathrm{T}}$ is the response vector (phenotype values), $\mathbf{X} = [X_1, ..., X_p]$ is the binary predictor matrix (genotype values).
Ensure: $\beta = (\beta_1, ..., \beta_p)$ is the vector of the regression coefficients (degrees of the SNP effect).
  repeat
    $k \leftarrow 1$
    Compute joint probabilities $\pi(x_{jk}, x_{ji})$, $i = 1, ..., p$, $i \neq k$, for all $j = 1, ..., n$, by means of the Bahadur representation (2)
    Compute the mean phenotype values $R_{ki}$, for all $i = 1, ..., p$, $i \neq k$, by means of (3)
    Compute the average mean phenotype value $\widetilde{R}_k$ by means of (4)
    Compute the weights $w_k$ by means of (5)
    Compute new variables $\widetilde{x}_{ik} = x_{ik}/w_k$, $i = 1, ..., n$.
  until $k > p$
  Compute $\beta^{\text{opt}}$ by using the standard Lasso with $\widetilde{\beta}$ and $\widetilde{\mathbf{X}}$ instead of $\widetilde{\beta}$ and $\mathbf{X}$.
  Compute $\beta_k = \widetilde{\beta}_k / \widetilde{R}_k$, $k = 1, ..., p$.
```

Let us indicate the main virtues of the proposed method. First of all, it does not require to develop special algorithms for solving the optimization problem for computing the vector of regression coefficients β. The obtained problem is solved as the standard Lasso algorithm after reformulating the penalized Lasso.

Second, the method is rather general because we could change the weights in (5) in accordance with our goal. For example, in one of the applications, we have aimed to minimize the mean heading date of barley as the mean phenotype value. However, we could aim to maximize, for example, the amount of grain protein. In this case, we change (5) by taking decreasing function g as follows:

$$w_k = \left(\frac{\max_{k=1,...,p} \widetilde{R}_k - \widetilde{R}_k}{\max_{k=1,...,p} \widetilde{R}_k - \min_{k=1,...,p} \widetilde{R}_k} \right)^{-q}.$$

Here the larger values of the measure $\widetilde{R}_k$ give us more significant SNPs and exert less penalty w_k.

Third, we consider not only correlations between SNPs, but also joint probabilities accounting for correlations. The joint probabilities are more informative in comparison with the correlation coefficients.

Fourth, we have simplified procedures for computing the joint probabilities. This substantially reduces the computation time.

3.5 The proposed method with epistatic effect

A lot of studies devoted to the epistatic effect (see, for example, [5]) consider extension of the so-called main effect model (1) on the interaction model which can be written as

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sum_{i<j,\ i,j=1,...,p} \beta_{ij} X_i X_j + \beta_0 + \epsilon. \quad (6)$$

The second term in (6) corresponds to pairwise interactions whose number is $p(1-p)/2$. Here β_{ij} is the parameter characterizing the epistatic interaction effect of a pair SNPs with indices i and j. Now the weighted Lasso estimates can be written as follows:

$$(\beta, \beta^*) = \arg \min_{\beta \in \mathbb{R}^p} \left\| Y - \mathbf{X}\beta - \mathbf{X}^* \beta^* \right\|^2 + \lambda \sum_{j=1}^{p} w_j \left| \beta_j \right| + \lambda \sum_{i<j,\ i,j=1,...,p} w_{ij} \left| \beta_{ij} \right|,$$

where $\beta^* = (\beta_{12}, ..., \beta_{p-1,p})$ is the additional vector characterizing the epistatic interaction effect of every pair of SNPs; $\mathbf{X}^* = (X_1 X_2, ..., X_{p-1} X_p)$ is the vector of covariates corresponding to pairwise interactions; $w_{ij} > 0$ are weights penalizing the additional parameters β_{ij}, $i < j$, $i, j = 1, ..., p$, in accordance with the rules of the adaptive Lasso [55].

It can be seen from the previous section that the weight or contribution of the pair of the k-th and the i-th SNPs into the phenotype values can be determined by the mean phenotype value R_{ki} obtained by means of (3). It is interesting to note that, in contrast to the k-th SNP contribution $\widetilde{R}_k$ obtained in a heuristic way (5), the value R_{ki} is the expectation of the phenotype with respect to the probability mass function $\pi(x_{kj}, x_{ij}) / \sum_{j=1}^{n} \pi(x_{kj}, x_{ij})$. So, the weight w_{ij} can be directly computed as

$$w_{ij} = \left(\frac{R_{ij} - \min_{ij} R_{ij}}{\max_{ij} R_{ij} - \min_{ij} R_{ij}} \right)^{-q}.$$

In order to take into account the interactions and to implement the method for epistasis detection, we apply a two-stage procedure (see the Screen and Clean method proposed by Wu et al. [48] for example). The first stage is for constructing the main effect model and searching for marginal significant SNPs by using the proposed penalized Lasso method with weights w_k determined from (5). Then only top ranked SNPs and pairs of SNPs composed from the significant ones are used in the interaction penalized Lasso model. The main idea here is to again use the Bahadur representation, namely, the mean phenotype values R_{ki} computed by means of (3). This is a very important place because we do not need to repeatedly compute the mean phenotype values. They have been computed during construction of the main effect model.

We do not provide here an algorithm for computing the optimal vectors β and β^* because it is just a simple extension of the algorithm given above.

4 Numerical experiments

The Lasso method in numerical experiments is regarded as a special case of a general problem solved by means of the R-package “glmnet” developed by Friedman et al. [14]. The tuning parameter λ is computed by using the function cv.glmnet() with 10-fold cross validation.

Below we use indices of SNPs instead of their full titles for short.

4.1 Data sets

Numerical experiments are carried out on three populations of double haploid (DH) lines of barley:

1. The first dataset consists of 93 DH lines of barley described in [8] and [9]. Phenotyping and genotyping data are available at Oregon Wolfe Barley Data (OWBD) and GrainGenes Tools. The lines are analyzed with respect to seven phenotypic traits: spike length (SL) in cm; grain number (GN); floret number (FS); hundred grain weight (HGW) in g of 100 grains; plant height (PH) in cm; spike number (SN); heading date (HD) in days. The linkage map consists of 1328 markers (SNPs).

2. The second dataset consists of 92 DH lines of barley obtained from the Dicktoo x Morex cross and described by several authors [20, 19, 37]. Phenotyping and genotyping data are available at http://wheat.pw.usda.gov/ggpages/DxM/ . We analyze the lines with respect to two phenotypic traits: heading date with and without vernalization with an 8-h light/16-h dark photoperiod regime. The linkage map consists of 117 markers.

3. The third population dataset includes 150 DH lines of barley obtained from the Steptoe x Morex cross [10, 21]. Phenotyping and genotyping data are available at http://wheat.pw.usda.gov/ggpages/SxM. The linkage map consists of 223 markers. The lines are analyzed with respect to the heading date (HD) trait, which is measured in 16 environments, and grain yield (GY) trait, which is measured in 6 environments.

4.2 Missing data

Missing marker data in all the datasets are estimated by means of the following heuristic procedure which can be regarded as some modification of the well-known method of K-nearest neighbors. Suppose the vector X_i corresponding to the i-th SNP has a missing value at the k-th position, i.e., X_{ik} is missing. By using the specific Hamming distance between the vector X_i and all vectors X_j, $j = 1, ...p$, $j \neq i$, we select K nearest neighbors $X_{i_1}, ..., X_{i_K}$ or K closest vectors. In order to take into account the missing values, they are excluded from computing the Hamming distance. That is why we use the specific Hamming distance in order to compare vectors with different numbers of missing elements, i.e., we compute the distance per one element of X_i. The imputed value is that represents the maximum of the K values at the k-th position of all the nearest neighbors $X_{i_1 k}, ..., X_{i_K k}$.

4.3 Error measure

From each of the (synthetic or real) data sets we randomly select two distinct subsets: a training data set of n examples to learn the model, and a test data set of n_{test} instances to evaluate the performance of the algorithms. The performance is assessed by means of the mean square residual (RMSR), which is defined by

$$\text{RMSR} = \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{f}(\mathbf{x}_i))^2}{n_{test}},$$

where $\hat{f}$ is the function estimated by the proposed method, and $\hat{f}(\mathbf{x}_i)$ is the predicted value of the phenotype value y_i for each $i \in \{1, \dots, n_{test}\}$. The error measure RMSR is computed from repeatedly random drawing training and test data sets and by averaging over the runs. The smaller the values of the average error measure are, the better the corresponding method. We use the one-fold-cross-validation, i.e., $n_{test} = 1$. This is because the number of lines is very small in comparison with the number of SNPs and we cannot reduce them.

4.4 The first dataset

First, we investigate DH lines of barley from OWBD. Values of the RMSR for the first dataset are shown in Table 1, where the first column corresponds to seven traits analyzed, columns 2-5 illustrate the RMSR by using only 40 top ranked SNPs. At that, we study cases when the accuracy is determined for all lines (All lines) and for the first 10 lines with the smallest values of phenotypes (First 10 lines). Abbreviations S.L. and P.L. denote the standard and new proposed Lasso methods, respectively. One can see that the proposed method provides better accuracy for the most traits. It does not mean that it can be successful in all cases. It is seen from Table 1 that the proposed method by traits PH and HD does not outperform the standard Lasso. Perhaps, another function determining the weights w_k from R_k could provide better results, but we did not find it. In addition, we can observe from Table 1 that use only of top ranked SNPs gives outperforming results in comparison with taking all SNPs for modelling GWAS. The same can be said about considering all lines and the first 10 lines.

Table 2 illustrates how the error measures depend on the reduced number of top ranked SNPs which are used for constructing the GWAS for the spike length trait. We take the fixed value of $q = 0.25$. It can be seen from Table 2 that the optimal number of top ranked SNPs is 40. It is interesting to observe also that the standard Lasso weakly depends on the SNP number.

Table 3 is similar to Table 1, but RMSRs in Table 3 are obtained by taking into account the epistatic effect. By comparing Tables 1 and 3, we can see that the consideration of epistasis allows us to construct a more accurate model. Moreover, the proposed method outperforms the standard Lasso even for traits PH and HD which distinguished from other traits and illustrated worse results with the proposed method (see Table 1). This is a very important fact showing that joint probabilities of pairs of SNPs as well as correlations between SNPs may improve the GWAS.

Table 4 shows the top ranked SNPs or their pairs with the largest 10 weights β obtained by means of the standard Lasso and the proposed method. Moreover, Table 4 shows the chromosomes where the corresponding SNPs are located. One can see that the largest weight has a pair of SNPs 997×1279. This implies that impact of the epistatic effect is very significant. It is interesting to note that the both methods select this pair of SNPs as the most significant one.

4.5 The second dataset

Let us study the dataset consisting of 92 DH lines of barley obtained from the Dicktoo x Morex cross. Tables 5 and 6 contain error measures for the Dicktoo x Morex dataset by considering two traits mentioned above. At that, Table 5 is obtained without taking into account the epistatic effect. In Table 6, the results are represented under condition of epistasis. Comparison of the tables shows that the use of condition of epistasis allows us to get outperforming results.

Table 7 shows the top ranked SNPs or their pairs with the largest 10 weights β obtained by means of the standard Lasso and the proposed method for the heading date without vernalization.

4.6 The third dataset

The third dataset consists of 150 DH lines of barley obtained from the Steptoe x Morex cross. Values of the RMSR for the third dataset are shown in Table 8. It can be seen from the table that the proposed method provides outperforming results. Table 9 shows also reduced values of the RMSR for the case of taking into account the epistatic effect. Comparing Tables 8 and 9, we can conclude that the model taking into account the epistatic effect significantly improves the regressor accuracy when the model is constructed by using only 40 top ranked SNPs. Moreover, the standard Lasso method also shows better results when the epistatic effect is considered.

It should be noted that the results given in Tables 8 and 9 is obtained for a certain value of q, namely, for $q = 0.8$. However, it is interesting to analyze how the value q impact on numerical results by using the third dataset. Figs. 1-4 depict the difference D between RMSRs of the proposed and standard Lasso methods for the HD trait. The larger the values of D are, the better the corresponding proposed method. The positive values of D say that the proposed method outperforms the standard Lasso for the corresponding values of q. It can be seen from Figs. 1-4 that there is an optimal value of q for every condition of the model use such that D achieves its maximum at this q. For example, it follows from Fig. 1 that the best results by using only top ranked SNPs can be obtained by $q = 0.8$. If we use all SNPs and analyze the first examples, then the optimal value of q is 0.5 (see Fig. 2). The same conclusions can be inferred from pictures illustrating the methods taking into account the epistatic effect (see Figs. 3-4).

Table 10 shows the top ranked SNPs and their pairs with the largest 10 weights β obtained by means of the standard Lasso and the proposed method for the grain yield trait.

It is interesting to note that the use of t-statistics for computing weights β of SNPs by the same parameters for GY trait gives the following 10 top ranked SNPs:

$$\begin{array}{cccccc} 82 & 81 & 83 & 84 & 85 & 79 \\ 86 & 130 & 129 & 80 & . & \end{array}$$

One can see that the most top ranked SNPs concentrated around the SNP with index 82. This is the obvious interaction of genes in a group of SNPs located at the same chromosome.

5 Conclusion

The results of numerical experiments and the logic underlying the proposed method have demonstrated that the proposed method outperforms the standard Lasso for many real datasets. Moreover, it takes into account the epistatic effect or the SNP-SNP interaction. It should be noted that the proposed method is very simple from a computation point of view. It does not require to develop a special software. The standard software (package “glmnet” in R) can be used for the method.

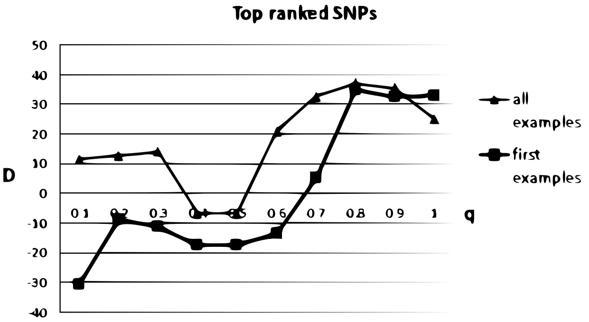

Figure 1: Difference between RMSRs of the standard and proposed Lasso methods for top SNPs.

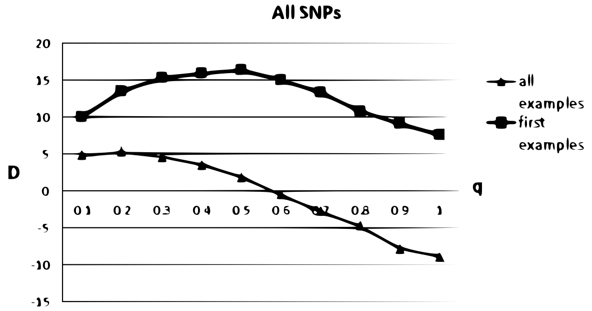

Figure 2: Difference between RMSRs of the standard and proposed Lasso methods for all SNPs.

It can be seen from the Bahadur representation that one of the crucial elements of the proposed method is a set of correlation coefficients between SNPs. It should be noted that they often use in GWAS as additional information. However, the correlation coefficients do not contain all probabilistic information about impacts of SNPs on values of a phenotype. The joint probabilities taking into account the correlation between SNPs can be viewed as a way for constructing association between SNPs and traits.

We have analyzed DH populations of barley. According to the DH method, only two types of genotypes occur for a pair of alleles, i.e., every x_{ij} takes only two values. At the same time, in diploid method, three genotypes occur, i.e., every x_{ij} takes three values. In this case the Bahadur representation cannot be applied, but the Sarmanov-Lancaster expansion [26, 39] can be used [1]. This is a direction for further research.

Of course, we have used a heuristic procedure by taking pairs of SNPs for computing R_{ki} by means of (3). We could consider joint probabilities of three and more SNPs. However, the increase of SNP numbers for computing the joint probabilities is impossible when the total number of SNPs is rather large. In this way, we can propose a multi-step procedure when the large set of top ranked SNPs is consequently determined by computing the joint probabilities of SNP pairs at the first step, then by computing the joint probabilities of SNP triples but from the reduced set obtained at the previous step. This procedure can be continued. At that, we could use the ridge regression in order to avoid a situation when a very small number of SNPs are obtained at some step. However, this is a direction for further research. The above modification may be very useful when the number of lines or individuals is small.

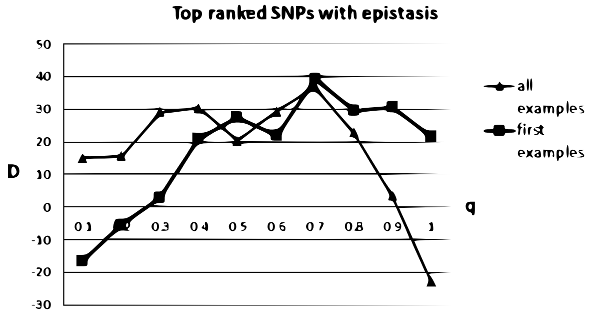

Figure 3: Difference between RMSRs of the standard and proposed Lasso methods for top ranking SNPs with epistasis.

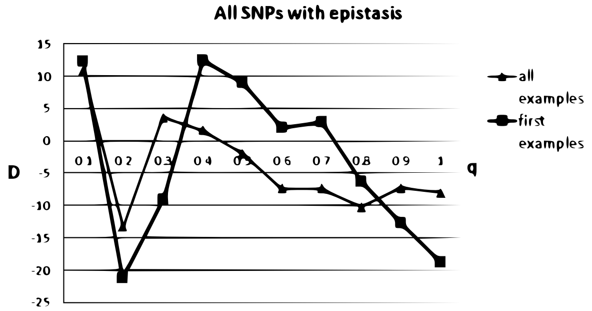

Figure 4: Difference between RMSRs of the standard and proposed Lasso methods for all SNPs with epistasis.

[1] A rather simple presentation of the Sarmanov-Lancaster expansion and its usage can be found in the paper I. Goodman and D.H. Johnson, Multivariate dependence and the Sarmanov-Lancaster expansion, 2005, http://www-ece.rice.edu/~igoodman/papers/goodman-johnson05.pdf

Table 1: RMSRs for the standard and proposed Lasso for OWBD.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
SL	1.760	0.583	1.596	0.515	3.280	4.429	2.594	3.845
GN	75.62	71.62	61.16	57.96	140.6	139.3	103.8	110.0
FS	77.86	79.59	50.39	40.78	162.0	158.9	75.68	74.60
HGW	0.134	0.121	0.094	0.085	0.209	0.186	0.147	0.137
PH	24.92	24.92	16.23	16.23	237.5	237.5	147.5	147.5
SN	17.96	16.53	8.847	8.285	27.71	27.92	12.73	12.66
HD	31.08	31.08	29.12	29.12	130.3	130.3	82.50	82.50

Table 2: RMSRs for the standard and proposed Lasso for OWBD by different numbers of top ranked SNPs.

	All lines		First 10 lines	
SNP numbers	S.L.	P.L.	S.L.	P.L.
20	1.412	1.116	1.396	1.177
40	1.392	0.550	1.392	0.443
60	1.393	0.680	1.393	0.611
80	1.393	1.548	1.393	1.569

Table 3: RMSRs for the standard and proposed Lasso for OWBD with epistasis.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
SL	0.815	0.724	0.697	0.652	2.866	3.638	2.622	2.630
GN	51.67	52.97	45.25	42.23	85.15	86.13	77.27	77.43
FS	44.99	70.42	33.19	27.53	72.00	102.1	51.17	50.86
HGW	0.077	0.056	0.069	0.052	0.156	0.107	0.118	0.098
PH	46.08	33.69	43.13	39.97	254.2	247.0	219.8	249.1
SN	12.21	10.42	6.842	5.274	24.26	23.84	13.34	10.71
HD	31.08	29.05	29.12	25.79	130.3	133.4	82.50	78.47

Table 4: Top ranked SNPs and their weights for the standard and proposed Lasso for OWBD HD with epistasis.

S.L.			P.L.		
SNP	chromosome	β	SNP	chromosome	β
997 × 1279	6 × 6	3.421	997 × 1279	6 × 6	3.401
138	1	3.314	903 × 325	5 × 2	3.193
896	5	3.176	1101	6	-2.764
1101 × 1152	6 × 6	2.750	138	1	2.661
734	4	-2.683	896	5	2.634
670 × 273	4 × 2	-2.542	1101 × 1152	6 × 6	2.583
324 × 903	2 × 5	2.128	725	4	-2.493
1096 × 976	6 × 6	-1.877	670 × 273	4 × 2	-2.012
997 × 526	6 × 3	1.826	734	4	-1.629
903 × 325	5 × 2	1.706	1101 × 976	6 × 6	-1.447

Table 5: RMSRs for the standard and proposed Lasso for Dicktoo-Morex without epistasis.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
unvernalized	43.74	44.68	18.13	17.87	63.57	60.77	26.57	27.43
vernalized	79.56	78.37	27.95	26.49	108.3	106.7	26.46	25.81

Table 6: RMSRs for the standard and proposed Lasso for Dicktoo-Morex with epistasis.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
unvernalized	34.33	38.85	24.08	17.63	58.63	58.28	30.46	27.81
vernalized	38.41	62.26	17.28	17.11	124.0	119.8	33.32	32.14

Table 7: Top ranked SNPs and their weights for the standard and proposed Lasso for Dicktoo-Morex HD with epistasis.

S.L.			P.L.		
SNP	chromosome	β	SNP	chromosome	β
112	7	-7.860	112	7	-7.238
22	2	6.387	110	7	-5.242
110	7	-5.296	20	2	4.426
20	2	4.074	22	2	4.377
113	7	-2.26	113	7	-3.55
51	3	-1.684	21	2	2.645
59 × 84	4 × 5	-1.394	50	3	-2.398
84	5	-1.261	49 × 113	3 × 7	1.627
19	2	1.002	84	5	-1.174
49	3	-0.980	33 × 50	2 × 3	-1.037

Table 8: RMSRs for the standard and proposed Lasso for Steptoe-Morex without epistasis.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
HD	79.05	41.95	78.68	43.73	46.62	51.37	57.49	46.66
GY	104.0	78.99	95.81	74.24	137.3	140.4	151.3	163.9

Table 9: RMSRs for the standard and proposed Lasso for Steptoe-Morex without epistasis.

	Top ranked SNPs				All SNPs			
	All lines		First 10 lines		All lines		First 10 lines	
Trait	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.	S.L.	P.L.
HD	77.15	40.31	78.52	39.14	44.12	51.52	50.15	47.26
GY	87.48	23.00	78.63	18.61	156.5	161.8	192.8	171.1

Table 10: Top ranked SNPs and their weights for the standard and proposed Lasso for the Steptoe-Morex GY with epistasis.

S.L.			P.L.		
SNP	chromosome	β	SNP	chromosome	β
82	3	9.562	82	3	13.158
53	2	-7.067	53	2	-6.687
81	3	5.278	222 × 114	7 × 4	5.355
29	1	-4.62	29	1	-5.089
42 × 53	2 × 2	4.126	68	2	4.599
20	1	-3.693	20	1	-3.543
111	4	3.510	108	4	3.338
68	2	2.377	154 × 19	5 × 1	-3.156
72	2	2.348	154 × 45	5 × 2	3.028
203	7	-1.675	108 × 105	4 × 3	2.881

References

[1] W. Altidor, T.M. Khoshgoftaar, J. Van Hulse, and A. Napolitano (2011) Ensemble feature ranking methods for data intensive computing applications. In B. Furht and A. Escalante, editors, *Handbook of Data Intensive Computing*, pages 349–376. Springer, New York.

[2] R.R. Bahadur (1961) A representation of the joint distribution of response to n dichotomous items. In H. Solomon, editor, *Studies in Item Analysis and Prediction*, pages 158–168. Stanford University Press, Palo Alto, CA.

[3] A.L. Beam, A. Motsinger-Reif, and J. Doyle (2014) Bayesian neural networks for detecting epistasis in genetic association studies. *BMC Bioinformatics*, 15(368):1–12.

[4] P.J. Bickel and E. Levina (2004) Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

[5] J. Bocianowski (2014) Estimation of epistasis in doubled haploid barley populations considering interactions between all possible marker pairs. *Euphytica*, 196(1):105–115.

[6] P. Buhlmann and S. van de Geer (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin Heidelberg.

[7] L. Chen, G. Yu, C.D. Langefeld, D.J. Miller, R.T. Guy, J. Raghuram, X. Yuan, D.M. Herrington, and Y. Wang (2011) Comparative analysis of methods for detecting interacting loci. *BMC Genomics*, 12:344:1–23.

[8] Y. Chutimanitsakun, R.W. Nipper, A. Cuesta-Marcos, L. Cistue, A. Corey, T. Filichkina, E.A. Johnson, and P.M. Hayes (2011) Construction and application for qtl analysis of a restriction site associated dna (rad) linkage map in barley. *BMC Genomics*, 12:4:1–13.

[9] L. Cistue, A. Cuesta-Marcos, S. Chao, B. Echavarri, Y. Chutimanitsakun, A. Corey, T. Filichkina, N. Garcia-Marino, I. Romagosa, and P.M. Hayes (2011) Comparative mapping of the oregon wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics*, 122(7):1399–1410.

[10] T.J. Close, P.R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J.T. Svensson, S. Wanamaker, S. Bozdag, M.L. Roose, M.J. Moscou, S. Chao, R.K. Varshney, P. Szucs, K. Sato, P.M. Hayes, D.E. Matthews, A. Kleinhofs, G.J. Muehlbauer, J. DeYoung, D.F. Marshall, K. Madishetty, R.D. Fenton, P. Condamine, A. Graner, and R. Waugh (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, 10:582:1–13.

[11] H.J. Cordell (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468.

[12] G. de los Campos, D. Gianola, and D.B. Allison (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880–886.

[13] Z. Feng, X. Yang, S. Subedi, and P.D. McNicholas (2012) The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(2):629–636.

[14] J.H. Friedman, T. Hastie, and R. Tibshirani (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[15] M.E. Goddard, N.R. Wray, K. Verbyla, and P.M. Visscher (2009) Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, 24(4):517–529.

[16] X. Gu, G. Yin, and J.J. Lee (2013) Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary Clinical Trials*, 36(2):642 – 650.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

[18] B. Hayes (2013) Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*, 1019:149–169.

[19] P. Hayes, F.Q. Chen, A. Corey, A. Pan, T.H.H. Chen, E. Baird, W. Powell, W. Thomas, R. Waugh, Z. Bedo, I. Karsai, T. Blake, and L. Oberthur (1997) The dicktoo x morex population. In PaulH. Li and TonyH.H. Chen, editors, *Plant Cold Hardiness*, pages 77–87. Springer US.

[20] P.M. Hayes, T. Blake, T.H.H. Chen, S. Tragoonrung, F. Chen and.A. Pan, and B. Liu (1993) Quantitative trait loci on barley (Hordeum vulgare L.) chromosome 7 associated with components of winterhardiness. *Genome*, 36(1):66–71.

[21] P.M. Hayes and O. Jyambo (1993) Summary of QTL effects in the steptoe x morex population. *Barley genetics newsletter*, 23:98–143.

[22] A. Huang, S. Xu, and X. Cai (2013) Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. *BMC Genetics*, 14(5):1–14.

[23] O. Kohannim, D.P. Hibar, J.L. Stein, N. Jahanshad, Xue Hua, P. Rajagopalan, A.W. Toga, C.R. Jack Jr., M.W. Weiner, G.I. de Zubicaray, K.L. McMahon, N.K. Hansell, N.G. Martin, M.J. Wright, P.M. Thompson, and The Alzheimer's Disease Neuroimaging Initiative (2012) Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, 6:1–13.

[24] R. Kohavi and G.H. John (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

[25] T.N. Lal, O. Chapelle, J. Weston, and A. Elisseeff (2006) Embedded methods. In *Feature extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 137–165. Springer, Berlin Heidelberg.

[26] H.O. Lancaster (1958) The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29(3):719–736.

[27] E.S. Lander and D. Botstein (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199.

[28] I.-H. Lee, G.H. Lushington, and M. Visvanathan (2011) A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics*, 1(11):1–8.

[29] S.-H. Lee and C.-H. Jun (2011) Discriminant analysis of binary data following multivariate Bernoulli distribution. *Expert Systems with Applications*, 38(6):7795–7802.

[30] J. Li, B. Horstman, and Y. Chen (2011) Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics*, 27(13):i222–i229.

[31] Z. Li and M.J. Sillanpaa (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, 125(3):419–435.

[32] B.G. Lindsay, G.Y. Yi, and J. Sun (2011) Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105.

[33] J. Liu, J. Huang, S. Ma, and K. Wang (2013) Incorporating group correlations in genome-wide association studies using smoothed group LASSO. *Biostatistics*, 14(2):205–219.

[34] B.E. Madsen and S.R. Browning (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

[35] C.M. Mutshinda and M.J. Sillanpaa (2010) Extended bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*, 186(3):1067–1075.

[36] J.O Ogutu, T. Schulz-Streeck, and H.-P. Piepho (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6(2):1–10.

[37] A. Pan, P.M. Hayes, F. Chen, T.H.H. Chen, T. Blake, S. Wright, I. Karsai, and Z. Bedo (1994) Genetic analysis of the components of winterhardiness in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, 89(7-8):900–910.

[38] M.Y. Park and T. Hastie (2006) Regularization path algorithms for detecting gene interactions. Technical Report 2006-13, Department of Statistics, Stanford University.

[39] O.V. Sarmanov (1958) The maximum correlation coefficient (symmetric case). *Dokl. Akad. Nauk SSSR*, 120:715–718.

[40] S. Subedi, Z. Feng, R. Deardon, and F.S. Schenkel (2013) SNP selection for predicting a quantitative trait. *Journal of Applied Statistics*, 40(3):600–613.

[41] R. Tibshirani (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[42] G. Tutz and J. Ulbricht (2009) Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253.

[43] M.G Usai, A. Carta, and S. Casu (2012) Alternative strategies for selecting subsets of predicting SNPs by LASSO-LARS procedure. *BMC Proceedings*, 6(2):1–9.

[44] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N.L.S. Tang, and W. Yu (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340.

[45] Y. Wang, G. Liu, M. Feng, and L. Wong (2011) An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics*, 27(21):2936–2943.

[46] H. Warren, J.-P. Casas, A. Hingorani, F. Dudbridge, and J. Whittaker (2014) Genetic prediction of quantitative lipid traits: Comparing shrinkage models to gene scores. *Genetic Epidemiology*, 38(1):72–83.

[47] N.R. Wray, Jian Yang, B.J. Hayes, A.L. Price, M.E. Goddard, and P.M. Visscher (2013) Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14:507–515.

[48] J. Wu, B. Devlin, S. Ringquist, M. Trucco, and K. Roeder (2010) Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiol.*, 34(3):275–285.

[49] J. Xia, S. Visweswaran, and R.E. Neapolitan (2012) Mining epistatic interactions from high-dimensional data sets. In D.E. Holmes and L.C. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms*, pages 187–209. Springer, Berlin Heidelberg.

[50] P. Yang, J. Ho, A. Zomaya, and B. Zhou (2010) A genetic ensemble approach for gene-gene interaction identification. *BMC Bioinformatics*, 11(1):524.

[51] C. Yao, D.M. Spurlock, L.E. Armentano, C.D. Page Jr., M.J. VandeHaar, D.M. Bickhart, and K.A. Weigel (2013) Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *Journal of Dairy Science*, 96(10):6716–6729.

[52] N. Yi, B.S. Yandell, G.A. Churchill, D.B. Allison, E.J. Eisen, and D. Pomp (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344.

[53] X. Zhang, S. Huang, Z. Zhang, and W. Wang (2012) Chapter 10: Mining genome-wide genetic markers. *PLoS Computational Biology*, 8(12):e1002828.

[54] Y. Zhang, B. Jiang, J. Zhu, and J.S. Liu (2011) Bayesian models for detecting epistatic interactions from genetic data. *Annals of Human Genetics*, 75(1):183–193.

[55] H. Zhou, D.H. Alexander, M.E. Sehl, J.S. Sinsheimer, and K. Lange (2011) Penalized regression for genome-wide association screening of sequence data. In *Pacific Symposium on Biocomputing*, pages 106–117. World Scientific Publishing.

[56] H. Zou (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

[57] H. Zou and T. Hastie (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.