Theor Appl Genet (2012) 125:419–435
DOI 10.1007/s00122-012-1892-9

REVIEW

Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection

Zitong Li · Mikko J. Sillanpää

Received: 14 November 2011 / Accepted: 27 April 2012 / Published online: 24 May 2012

Abstract Quantitative trait loci (QTL)/association mapping aims at finding genomic loci associated with the phenotypes, whereas genomic selection focuses on breeding value prediction based on genomic data. Variable selection is a key to both of these tasks as it allows to (1) detect clear mapping signals of QTL activity, and (2) predict the genome-enhanced breeding values accurately. In this paper, we provide an overview of a statistical method called least absolute shrinkage and selection operator (LASSO) and two of its generalizations named elastic net and adaptive LASSO in the contexts of QTL mapping and genomic breeding value prediction in plants (or animals). We also briefly summarize the Bayesian interpretation of LASSO, and the inspired hierarchical Bayesian models. We illustrate the implementation and examine the performance of methods using three public data sets: (1) North American barley data with 127 individuals and 145 markers, (2) a simulated QTLMAS XII data with 5,865 individuals and 6,000 markers for both QTL mapping and genomic selection, and (3) a wheat data with 599 individuals and 1,279 markers only for genomic selection.

Communicated by R. Varshney.

Z. Li · M. J. Sillanpää
Department of Mathematics and Statistics,
University of Helsinki, PO Box 68, 00014 Helsinki, Finland

M. J. Sillanpää (✉)
Department of Mathematical Sciences,
University of Oulu, PO Box 3000, 90014 Oulu, Finland
e-mail: mjs@rolf.helsinki.fi

M. J. Sillanpää
Department of Biology, University of Oulu,
PO Box 3000, 90014 Oulu, Finland

M. J. Sillanpää
Biocenter Oulu, Oulu, Finland

M. J. Sillanpää
Department of Agricultural Sciences,
University of Helsinki, Helsinki, Finland

Introduction

The use of DNA marker information for plant (or animal) breeding has become increasingly popular. The traditional Marker assisted selection (MAS) method (Dekkers and Hospital 2002), which uses only a small number of markers to predict breeding values, may have poor predictive ability due to the fact that only a limited proportion of genetic variance can be captured by the markers (Goddard and Hayes 2007). An alternative approach, known as genomic selection (Meuwissen et al. 2001; Heffner et al. 2009; Piepho 2009) can utilize genome-wide information that results in enhanced predictive ability. A fundamental requirement for implementing genomic selection is that a dense set of markers through the whole genome need to be genotyped, thus guaranteeing that most QTLs are in linkage disequilibrium (LD) with at least one marker. This has become increasingly possible because of recent advances in laboratory techniques (Bernardo and Yu 2007). Another closely related topic called QTL/association mapping aims at finding the genomic loci which are associated with the phenotypes and estimating their effect sizes. When a dense set of markers are used, nearby markers can approximately represent QTLs (Xu 2003).

In quantitative genetics, a multiple linear regression model is often used to describe the relationship between phenotypes and markers. All marker effects can be

simultaneously estimated from the model, and then based on the estimates, one can perform (1) *hypothesis testing*: to identify the QTL/association signals; (2) *prediction*: to calculate the genomic breeding values for a new data set. For oligogenic traits, only a small proportion of markers are associated with the trait variation, and most markers have zero effects in theory. Including all of them into the model may lead to poor accuracy of the estimated marker effects, and reduce the reliability of both hypothesis testing and prediction. In an even worse case, when the number of markers is larger than the number of individuals in the sample, the model is over-saturated, and ordinary least squares estimation is not applicable. This motivates us to implement variable selection (Broman and Speed 2002; Sillanpää and Corander 2002), to include only a trait-associated subset of markers into the model. One popular way for achieving this goal is to implement a penalized regression approach, named LASSO (least absolute shrinkage and selection operator) (Tibshirani 1996), which can provide sparsity inducing estimation of regression coefficients by adding ℓ_1 penalty functions to the traditional least squares. LASSO, and its extensions including Elastic net (Zou and Hastie 2005) and Adaptive LASSO (Zou 2006) have been used in various QTL mapping or genomic selection studies (Chen and Cui 2010; Cho et al. 2010; Wang et al. 2010; Usai et al. 2009; Harris and Johnson 2010). Furthermore, it is also well known that LASSO has a Bayesian interpretation (Tibshirani 1996), that is the ℓ_1 penalty function is equivalent to a double exponential or Laplace prior distribution. Figueiredo (2003) and Park and Casella (2008) suggested a Bayesian hierarchical model using a scale mixture parametrization to mimic LASSO, under which a Markov Chain Monte Carlo (MCMC) can be implemented to provide shrinkage estimation. Bayesian LASSO has been applied both to identify QTLs (Yi and Xu 2008; Li et al. 2011) and to predict genomic breeding values (De Los Campos et al. 2009; Legarra et al. 2011).

In this article, we provide an overview of LASSO, its extensions including Elastic net and Adaptive LASSO, as well as its related Bayesian model, and their applications in QTL mapping and genomic selection. Although detecting QTL signals and predicting genomic values are different problems in principle, our view is that both of them can be solved efficiently using the LASSO and its related methods. The structure of the article is the following. We first describe the multiple regression model which we consider for both QTL mapping and genomic selection problems. We then summarize the theory and computations involved in the LASSO, and its generalizations, and finally, we describe the example data analyses using the North American Barley data (Tinker et al. 1996), the QTL/MAS XII simulated data (Lund et al. 2009), and the wheat data from International Maize and Wheat Improvement Center (CIMMYT) (Crossa et al. 2010), respectively.

Model and problems

Because of the continuous nature of quantitative traits, it is intuitive to use a multiple linear regression model to describe the relationship between trait values and marker loci. A typical regression model is:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{1}$$

where y_i $(i = 1, \ldots, n)$ is the phenotypic value of the ith individual in the mapping population, β_0 is the intercept, x_{ij} is the genotypic value of the jth marker for individual i, β_j is the effect of marker j, and e_i is the random error assumed to follow a normal distribution $N(0, \sigma_e^2)$ with mean zero and variance σ_e^2 independently for $i = 1, \ldots, n$. The mutually independent residual terms e_i $(i = 1, \ldots, n)$ represent all unknown factors that may contribute to the trait variation. Note that the assumption of independent errors with the constant variance is often a simplification in plant breeding (see Burgueño et al. 2012; Piepho et al. 2012). The genotype value x_{ij} is defined as

$$x_{ij} = \begin{cases} 1 & \text{if marker genotype is AA,} \\ 0 & \text{if marker genotype is AB,} \\ -1 & \text{if marker genotype is BB.} \end{cases} \tag{2}$$

In QTL mapping with a dense set of markers, we are interested in estimating the marker effects $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$, and deciding which markers are in linkage disequilibrium (LD) with QTLs. Therefore, QTL mapping can be regarded as a variable selection problem. On the other hand, in genomic selection, we are not interested in the exact location of the QTLs, but we are mainly concerned with calculating the genome-enhanced breeding values (GEBV)

$$\text{GEBV} = \boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}}, \tag{3}$$

where $\boldsymbol{x}_{\text{new}}$ is a matrix containing genotype values from new individuals who do not have their phenotypes measured yet, and $\hat{\boldsymbol{\beta}}$ are certain estimates of the regression coefficients $\boldsymbol{\beta}$. When the true breeding values (TBV) for the new individuals are available (as is the case for simulated data), the correlation coefficient between TBV and GEBV $\text{cor}(\text{TBV}, \text{GEBV})$ can be used to measure the prediction accuracy. Therefore, genomic selection can be viewed as a prediction problem. Since it is necessary to estimate the parameters based on the training data sets before prediction, how to estimate the regression coefficients $\boldsymbol{\beta}$ accurately is a primary question. The ordinary

least squares (OLS) method, which is a common way to estimate regression coefficients in statistics has two deficiencies: (1) Although an OLS estimator gives an unbiased estimate of the regression coefficients, it often shows large variance, which will cause some inaccuracy in the predicted values. (2) OLS estimation is not available for the situations where the number of explanatory variables is larger than the number of observations, the so called $p > n$ problem. Instead, a relatively stable estimator will be beneficial to both QTL identification and genomic breeding value prediction problems, since it often helps to increase the prediction accuracy, and also the power to detect QTL signals in hypothesis testing. In addition, in genomic selection or QTL mapping with a dense set of markers, it is common to have the number of markers being genotyped much larger than the number of individuals. Therefore, it is necessary to seek a regression estimator with small variance, and which is able to handle situations where $p > n$. Finally, when the marker density is high, QTL may only be present in a few marker intervals and most of the markers may have zero or close to zero effects in theory. In this case, it is preferable to obtain a sparse model in order to find the markers associated with the traits easier. Next, we provide an overview of the LASSO-penalized regression method (Tibshirani 1996) and its generalizations called Elastic net (Zou and Hastie 2005) and Adaptive LASSO (Zou 2006), which are able to do the parameter estimation and variable selection simultaneously.

LASSO and its extensions

Theory

The LASSO regression can be specified as estimating the regression coefficients $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$, by minimizing the penalized sum of squares $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$. Here the sum of absolute values (ℓ_1 norm) of the regression coefficients $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penalty function, and $\lambda \geq 0$ is called shrinkage factor, which needs to be specified by the analyst. By adding the penalty function to the residual sum of squares and setting λ larger than zero, the LASSO is able to shrink the least square estimators towards zero, and reduce the variances. Furthermore, different from an older penalized regression approach called Ridge regression (Hoerl and Kennard 1970), which adopts ℓ_2 norm penalty function $\lambda \sum_{j=1}^{p} \beta_j^2$, the LASSO is able to shrink some of the regression coefficients exactly to zero because of the non-differentiable property of the ℓ_1 norm penalty. The sparsity of the model is determined by the value of the shrinkage factor. Therefore, the LASSO can also be regarded as a variable selection method for the regression model.

Compared with Ridge regression, LASSO has two disadvantages: (1) when multicollinearity is shown to occur among the explanatory variables, LASSO tends to select only a single variable from a group of highly correlated variables. (2) When $p > n$, at most n explanatory variables can be selected into the model. In a dense set of markers, it is common that those markers are highly correlated among them. Particularly, in a genomic selection problem, the limitation that only n variables can be selected by maximum may cause discarding of some important markers, that are contributing to the prediction accuracy, out of the model. To overcome these limitations, Elastic net (Zou and Hastie 2005) adopts a penalty function with the convex combination of ℓ_2 and ℓ_1 norms as

$$P_\alpha(\boldsymbol{\beta}) = \lambda[(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|], \quad (4)$$

where

$$\begin{cases} \alpha = 0 & : \text{Ridge regression}, \\ 0 < \alpha < 1 & : \text{Elastic net}, \\ \alpha = 1 & : \text{LASSO}. \end{cases} \quad (5)$$

It can be seen that Elastic net penalty is a compromise between ℓ_2 and ℓ_1 norm penalties. Since $P_\alpha(\boldsymbol{\beta})$ is non-differentiable at the zero point, Elastic net keeps the good property of LASSO that it can shrink regression coefficients exactly to zero. A difference from LASSO is that, when a group of markers with high pairwise correlations is present, Elastic net tends to select all of them into the model and assign equal regression coefficients to them. Furthermore, Elastic net is able to select more than n explanatory variables when p is larger than n. Therefore, Elastic net combines the advantages from both LASSO and Ridge regression.

From another point of view, a good variable selection method should satisfy the following oracle properties (Fan and Li 2001): (1) *Consistency of variable selection* choosing the correct non-zero coefficients into the model with probability tending to 1 with increasing sample size; and (2) *unbiasness of parameter estimation* assuming the true subset of explanatory variables with non-zero effects, the estimator of each non-zero coefficient follows asymptotic normality similarly as the OLS estimator. LASSO has been shown not to follow the oracle properties by Fan and Li (2001), and Zou (2006). First, it has been proven that LASSO only does consistent variable selection under the so-called irrepresentable condition (Zhao and Yu 2006), or an equivalent neighborhood stability condition (Meinshausen and Bühlmann 2006). Those are rather restrictive conditions that many data sets in practice may not satisfy.

Second, LASSO tends to underestimate the regression coefficients. For improvement, Fan and Li (2001) proposed a non-convex penalized regression method called smoothly clipped absolute deviation (SCAD) penalty regression. Due to its non-convex feature, the computation might be challenging. Zou (2006), on the other hand, proposed a two-step procedure named Adaptive LASSO for improvement of standard LASSO which still maintains the ease of computation. In the first step, a standard estimation methods such as the OLS is applied to the data, so that the initial estimates $\hat{\beta}_{\text{init},j}$ are obtained. In the second step, a weighted ℓ_1 norm penalized function is specified by

$$\lambda \sum_{j=1}^{p} w_j |\beta_j|, \tag{6}$$

where $w_j = 1/|\hat{\beta}_{\text{init},j}|$ $(j = 1, \ldots, p)$, and based on (6) the adaptive LASSO estimates can be computed. By assigning such data dependent weights to each local (marker specific) penalty term, Adaptive LASSO enlarges the penalties for variables with zero coefficients and relaxes the shrinkages for those with non-zero coefficients, and therefore it reduces some bias. Zou (2006) showed that Adaptive LASSO enjoys oracle properties under a weaker condition than the irrepresentable condition. Meanwhile, Zou (2006) demonstrated that the same algorithms that fit the LASSO problem can be used for Adaptive LASSO with the same computation costs. Furthermore, for a high dimensional data where $p > n$, the OLS estimates are not available for weight calculation. The estimates of Ridge regression (Zou 2006), Marginal regression (Huang et al. 2008) and standard LASSO (Bühlmann and van de Geer 2011) have been suggested as alternative choices. Since in practice, the performance of Adaptive LASSO might be sensitive to the choices of those initial estimates, a better strategy is to re-estimate the weights based on the current estimates of the regression coefficients in each iterative step. This can be implemented under both frequentist (Bühlmann and Meier 2008) and Bayesian (Sun et al. 2010; Mutshinda and Sillanpää 2010; Li and Sillanpää 2012) frameworks. More discussion on the Bayesian interpretation of LASSO is presented below. Other improved approaches, which address the problems of LASSO's biased estimation or variable selection consistency, include LASSO-OLS hybrid (Efron et al. 2004), Relaxed LASSO (Meinshausen 2007) and Thresholded LASSO (Zhou 2010).

Various algorithms for finding the LASSO (or Elastic net and Adaptive LASSO) solution have been developed, including a homotopy algorithm (Osborne et al. 2000), the least angle square algorithm (LARS) (Efron et al. 2004), and a coordinate descent algorithm (Friedman et al. 2007). The coordinate descent algorithm, which is perhaps the most efficient algorithm for LASSO computation, is used in our example analyses. A description of the algorithm can be found in the appendix.

Selection strategies for shrinkage factor

The LASSO and its extension methods provide a potential opportunity to select markers which are highly correlated with the phenotype and discard those with negligible effects. An important remaining issue is how to choose the shrinkage factor λ, which decides (1) the number of markers remaining in the model, and (2) the level of shrinkage for the marker effects. In the following, we discuss two possible strategies for selecting an 'optimal' shrinkage factor from the perspective of prediction and variable selection, respectively.

Cross validation

Cross validation (CV) (see Hastie et al. 2009) is perhaps the most common criterion for deciding the LASSO shrinkage factor. It aims to find a model that has the best predictive ability. In a CV procedure, we first need to randomly divide data into V non-overlapping roughly equal-sized parts with approximate m individuals in each part. In turn, we take each single part as the validation data denoted by $\boldsymbol{x}_v$ and $\boldsymbol{y}_v$ $(v \in \{1, \ldots, V\})$, and the remaining $V - 1$ parts as the training data denoted by $\boldsymbol{x}_{-v}$ and $\boldsymbol{y}_{-v}$. The model is used to fit the training data with a specific choice of λ and then is applied to the validation data to obtain the prediction of $\boldsymbol{y}_v$ as $\hat{\boldsymbol{y}}_v(\lambda)$. The averaged predictive ability of the model can be evaluated by

$$P_{\text{CV}}(\lambda) = \frac{1}{V} P(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v(\lambda)), \tag{7}$$

where the function $P(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v(\lambda))$ is a certain metric of the prediction accuracy. For a regression model, the mean squared prediction error is often used as the metric of the prediction accuracy, which is defined as

$$P(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v(\lambda)) = \frac{1}{m} (\boldsymbol{y}_v - \hat{\boldsymbol{y}}_v(\lambda))^T (\boldsymbol{y}_v - \hat{\boldsymbol{y}}_v(\lambda)), \tag{8}$$

In this case we are interested in obtaining an optimal shrinkage factor $\hat{\lambda}$ which minimizes the averaged prediction error in (7).

Bayesian information criterion

We are looking forward to a shrinkage factor leading to a model that is able to (1) give accurate predictions, or (2) identify the true model structure (i.e. detect QTLs). However, in practice, a shrinkage factor chosen by the cross validation often leads to a model with too many non-zero effects (Bühlmann and van de Geer 2011; Xu 2007), so that

the true QTL signals are not clear. Therefore, an alternative criterion is required, which may choose a larger λ in order to produce a sparser model. Zou et al. (2007) suggested a Bayesian information criterion (BIC) as a good model selection criterion for achieving this goal. The BIC criterion aims to find λ that minimizes the following score function

$$\mathrm{BIC} = \log \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2}{n} + \frac{\log(n)}{n} \mathrm{df}(\lambda). \tag{9}$$

The first term is simply the log sum of squares function measuring the model fit, and the second term includes the degree of freedom $\mathrm{df}(\lambda)$, which can be estimated by the the number of non-zero regression coefficients estimated from LASSO. Therefore, BIC can be regarded as a compromise between model fitting and model complexity. Zou et al. (2007) showed that BIC is a more suitable criterion for determining λ than other approaches such as cross validation, Mallow's C_p (Efron et al. 2004) and Akaike information criterion (AIC) (Akaike 1974) from the perspective of variable selection, since it tends to give a sparser model. Sun et al. (2010) discussed using BIC to select hyperparameters for Bayesian Adaptive LASSO in QTL mapping.

Chen and Chen (2008) claimed that BIC was still a too liberal criterion for the high dimensional data, and they developed extended Bayesian information criterion (EBIC), which can achieve more conservative variable selection. EBIC combined with LASSO has been applied to QTL mapping in Chen and Cui (2010).

LASSO and hypothesis testing

As we mentioned earlier, LASSO usually does not perform consistent variable selection. Nevertheless, LASSO was demonstrated to have a nice screening property meaning that by properly setting the shrinkage factor and by assuming that all the true non-zero regression coefficients are sufficiently large, there is high probability that the set of explanatory variables selected by LASSO contains the true set of effective variables (Meinshausen and Bühlmann 2006). At the same time, it is likely that the set of markers selected by LASSO with non-zero estimated effects contain some false selected signals. Perhaps, a hypothesis testing can be used for error control in order to judge the true QTL signals based on the LASSO solution (Wu et al. 2009), as what have frequently been done in many single locus analyses. Unfortunately, unlike the OLS method, LASSO cannot directly provide a test statistic or confidence interval for each estimate, and therefore a hypothesis testing cannot be directly applied. Recently, Wasserman and Roeder (2009) proposed a two-stage procedure for high-dimensional data. The data is first randomly split to two equal sized parts by individuals. Then in a *screening* step, LASSO is performed on the first part of the data (with the cross validation to select the shrinkage factor). In a *cleaning* step, the standard OLS is implemented on the second part of the data with non-zero variables selected from the screening step, and the p values from a t test can be obtained for each marker. A drawback of this single-split method is that the result depends very much on how the data is split. For improvement, Meinshausen et al. (2009) suggested a multi-split method, in which the previous procedure is repeated many times, and an empirical distribution of the p value for each marker can be obtained. Finally, an overall p value can be constructed based on that empirical distribution. Meinshausen et al. (2009) demonstrated that the multi-split method can be used for both family-wise error (FWER) control and false discovery rate (FDR) control. The results from their data analyses also indicate that this procedure has a good performance in terms of error control for both high and low dimensional data. Furthermore, Meinshausen et al. (2009) also claimed this procedure cannot only be applied to LASSO, but also to many other variable selection methods such as Adaptive LASSO and Elastic net.

Finally, one should be aware that the above-mentioned concepts such as "correct non-zero coefficients" or "true set of effective variables" are rather idealized quantities which are used only to demonstrate the statistical properties of LASSO and the corresponding multi-split method for hypothesis testing. In practice, a QTL is usually not exactly located on any marker, so that we can only seek a subset of markers which are linked to QTLs, and use their locations to approximate the QTL positions. In that case, the definition of "correct non-zero coefficients" is not quite obvious.

Bayesian interpretation of LASSO

Tibshirani (1996) showed that the ℓ_1 norm penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ is proportional to the logarithm of the product of p independent double exponential distributions (or Laplace distributions) with rate λ, so that LASSO estimates can be regarded as the posterior mode estimates under the double exponential prior distributions for the regression coefficients as

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp(-\lambda|\beta_j|). \tag{10}$$

Note the likelihood function can be specified as

$$p(\boldsymbol{y}|\boldsymbol{\beta},\sigma_e^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma_e^2}\right), \tag{11}$$

so that under a Bayesian posterior distribution, the residual variance σ_e^2 needs to be taken into account, and be treated as a parameter just like regression coefficients (in the standard LASSO, it is also possible to include σ_e^2 into the model, but this is not usually considered).

Inspired by the fact that a double exponential distribution can be written as a scale mixture of normals

$$\frac{\lambda}{2}\exp(-\lambda|\beta|)=\int_0^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\beta^2/2\sigma^2\right)\times\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2\sigma^2}{2}\right)\mathrm{d}\sigma^2, \tag{12}$$

a hierarchical posterior model with the following prior settings:

1. $p(\beta_0)\propto 1$,
2. $p(\sigma_e^2)\propto\frac{1}{\sigma_e^2}$,
3. $p(\beta_j)\propto N(\beta_j|0,\sigma_j^2)$,
4. $p(\sigma_j^2)\propto \mathrm{Exp}(\sigma_j^2|\frac{\lambda^2}{2})$,

have been proposed (see Park and Casella 2008 and Yi and Xu 2008).

A Gibbs sampling algorithm can be used to simulate dependent samples from the posterior distribution. Next, the posterior mean of these samples can be used as the point estimates of the marker effects, and in addition a credible interval for each marker effect can also be constructed and used to judge QTL signals (Kyung et al. 2010; Li et al. 2011). Different from standard LASSO in which the shrinkage factor λ needs to be selected explicitly, in Bayesian LASSO, a prior distribution can be assigned to λ^2, so that the value of λ can be estimated as well as other model parameters. A typical choice is to use a conjugate gamma prior Gamma(a,b) for λ^2, where $a>0$ and $b>0$ are predetermined hyperparameters. For example, $a>0$ and $b>0$ can be set to be small values (say 10^{-4}), so that the priors are non-informative (Li et al. 2011).

In addition, Bayesian hierarchical models of Elastic net and Adaptive LASSO can be built correspondingly in a similar manner (see, for example, Li and Lin 2010, and Mutshinda and Sillanpää 2010), which is beyond the scope of the discussion in this review.

Finally, based on the double exponential prior (10), a biological interpretation has been given to the shrinkage factor λ in LASSO. Legarra et al. (2011) noticed that a relationship between the variance of marker effects and λ can be represented as $\mathrm{Var}(\beta)=\frac{2}{\lambda^2}$. Therefore, λ plays a key role to determine the shape of the distribution of the SNP effects in the LASSO model. In addition, a further rough relationship can be established between λ and the genetic variance σ_g^2 in a population as $\mathrm{Var}(\beta)=\frac{2}{\lambda^2}=\frac{\sigma_g^2}{2\sum_{j=1}^{p}q_j(1-q_j)}$, where q_j is the allele frequency of the first allele at the jth marker.

LASSO for mixed models

So far, we have considered using the pure genetic marker information to construct a linear regression model, which assumes equal relatedness among the sample individuals. One problem in plant (or animal) breeding is that the group relatedness/pedigree structure often exists among the samples (i.e. a subgroup of individuals may be more closely related than others), so that the above-mentioned assumption is not quite reasonable. Omitting relatedness in the model may be causative for false positives or for reduced predictive ability (see Sillanpää 2011; Solberg et al. 2009). A linear mixed effect model has been suggested to take the relatedness/pedigree structure into account (Jannink et al. 2001; Yu et al. 2006), which can be represented in the following matrix form:

$$\boldsymbol{y}=\beta_0+\boldsymbol{X\beta}+\boldsymbol{Zu}+\boldsymbol{e}, \tag{13}$$

where the intercept β_0, the genetic fixed effects $\boldsymbol{\beta}=[\beta_1,\ldots,\beta_p]^{\mathrm{T}}$, the design matrix containing genotypes $\boldsymbol{X}=[x_{ij}]$ $(i=1,\ldots,n, j=1,\ldots,p)$ and the error terms $\boldsymbol{e}=[e_1,\ldots,e_n]^{\mathrm{T}}$ follow the previous definitions. In addition, $\boldsymbol{u}=[u_1,\ldots,u_n]^T$ is a vector of random effects that follows a multivariate normal distribution MVN$(0,\boldsymbol{A}\sigma_a^2)$, where $\boldsymbol{A}$ is an additive genetic relationship matrix which can be constructed based on the pedigree information and σ_a^2 is the variance of the random additive genetic effects $\boldsymbol{u}$. Finally, $\boldsymbol{Z}$ is the design matrix for the random effects. A LASSO estimator under a mixed effect model can be defined as

$$\hat{\boldsymbol{\beta}},\hat{\sigma}_a^2,\hat{\sigma}_e^2=\arg\min_{\boldsymbol{\beta},\sigma_a^2,\sigma_e^2}[\log|\boldsymbol{V}|+(\boldsymbol{y}-\beta_0-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}-\beta_0-\boldsymbol{X\beta})+\lambda\sum_{j=1}^{p}|\beta_j|], \tag{14}$$

where $\boldsymbol{V}=\sigma_a^2\boldsymbol{ZAZ}^{\mathrm{T}}+\sigma_e^2\boldsymbol{I}$. In addition, mixed Adaptive LASSO and mixed Elastic net can be defined similarly. More complicated non-convex programming algorithms can be used to compute the solution paths under these models. Wang et al. (2010) proposed a mixed Adaptive LASSO procedure for analyzing QTL effects. In a simulation study, they showed that a mixed Adaptive LASSO model performed better than LASSO and Adaptive LASSO with only fixed effects to control the false positives caused by the population structure among the data.

On the other hand, in Bayesian LASSO settings, we may consider the distribution MVN$(0, \boldsymbol{A}\sigma_a^2)$ as the prior distribution of random effects $\boldsymbol{u}$. Under such circumstances, their variance σ_a^2 can be assigned with an non-informative prior $p(\sigma_a^2) \propto \frac{1}{\sigma_a^2}$, so that $\boldsymbol{u}$ can be incorporated into the hierarchical model, and be sampled as the other parameters using MCMC. De Los Campos et al. (2009) applied such a Bayesian mixed LASSO method for predicting phenotypic values in wheat and mouse data sets. They compared three models using (1) sole pedigree information, (2) sole marker information, (3) both information sources, and they found that the Bayesian mixed LASSO with both pedigree and markers included showed the best predictive ability.

Software tools

The coordinate descent algorithm for computing the LASSO, Elastic net and Adaptive LASSO solution path can be implemented by the `Glmnet` software package (Friedman et al. 2010; Simon et al. 2011). The package has been incorporated into both R and Matlab with the core of the coordinate descent algorithm written in Fortran. It can be efficiently implemented even on a large scale data set, and it should be a suitable tool for both QTL mapping and genomic selection. Specifically, `Glmnet` is not only designed for Gaussian linear regression model, but also for binomial, multinomial (with unordered response), Poisson and Cox's proportional hazard regression models. Thus, potentially, the discrete trait data and survival data can be analyzed by `Glmnet` as well. Besides, various R packages for implementing other algorithms including LARS for solving the LASSO problem have been summarized in Hesterberg et al. (2008). For tools especially involved in discrete traits, see Ayers and Cordell (2010). Software tools are also available for Bayesian LASSO. Pérez et al. (2010) developed `R/BLR` package which implements both Bayesian LASSO (De Los Campos et al. 2009) and Bayesian Ridge regression for genomic selection. The `R/BLR` uses mixed models and allows inclusion of pedigrees in addition to the markers. In addition, another package `R/fGWAS` developed by Li et al. (2011), implements Bayesian LASSO for genome-wide association studies. Finally, the software tools for implementing mixed LASSO mentioned in the last section are generally limited. Some R codes for implementing mixed Adaptive LASSO on QTL mapping are available from the first author of Wang et al. (2010). More efforts are needed to develop mixed LASSO software tools for genomic selection.

Example analyses

In our case study, we consider the following three data sets: (1) a small scale real data set from North American Barley study (Tinker et al. 1996), (2) a large scale simulated data from XII QTL MAS Workshop (Lund et al. 2009; Crooks et al. 2009), and (3) a CIMMYT wheat data (Crossa et al. 2010). The first two data sets are used for both QTL mapping and genomic selection, and the third data set is only used for genomic selection.

Barley data

This is the well-known North American Barley data (Tinker et al. 1996). The mapping population consists of 145 doubled haploid lines ($n = 145$), each grown in a range of environments. A total of 127 markers were genotyped, covering 1270 cM of the barley genome, with the average distance between markers of 10.5 cM. For the marker data, one genotype (AA) is coded as 1 and the other (BB) as -1. Around 5.05 % of genotype data were missing. Missing marker data were handled in all methods as the preprocessing of that data set once before the analysis. In the preprocessing, each missing marker genotype is replaced by its conditional expectation estimated from flanking markers with known genotypes. Detailed information of this method can be found in Haley and Knott (1992), or in Siegmund and Yakir (2007). Seven traits including yield, heading, maturity, height, lodging, kernel weight, and test weight were measured for each plant. We selected kernel weight as the phenotype used in the analysis. Before the analysis, the average phenotypic value for each line was calculated over such environments which were not missing, and was used as a phenotype for the analysis similarly as in Xu (2003).

QTL mapping

The LASSO, Elastic net and Adaptive LASSO are implemented for QTL problem using `Matlab/Glmnet` (Friedman et al. 2010), and the Bayesian LASSO was implemented by our own Matlab code (Li and Sillanpää 2012). For LASSO, Elastic net and Adaptive LASSO, we first implemented those methods on the full data set to obtain the estimates of the marker effects, and then we used the multi-split methods of Meinshausen et al. (2009) to calculate the p values for each locus (only the FWER control is considered). We set the significance threshold as 0.05, so that a p value smaller than 0.05 indicates the corresponding locus as a QTL. We considered both fivefold cross validation and BIC as the criteria to choose the shrinkage parameters. The shrinkage factor selection for LASSO was done in a standard way as introduced above. For Elastic net, we followed a strategy from Zou and Hastie (2008) to find a combination of α and λ. First, a grid of values of α was defined as $\{0.1, 0.2, \ldots, 0.8, 0.9\}$. For each of them, we used either CV or BIC to find the optimal

value $\hat{\lambda}$ out of a set of 100 λs (automatically generated by the `Glmnet` program), and recorded the corresponding prediction error. Next, an optimal value of α denoted as $\hat{\alpha}$ was selected which gives the smallest prediction error. For Adaptive LASSO, we performed a standard LASSO with CV for tuning λ and used these LASSO estimates to construct the weights. Since the LASSO estimators are sparse, we modified the weights as

$$w_j = \frac{1}{|b_{\text{lasso},j}| + \frac{1}{n}}, \tag{15}$$

similarly as Zou and Zhang (2009). Finally, for Bayesian LASSO, we assigned a non-informative prior $\text{Gamma}(0.0001, 0.0001)$ to λ^2, and this prior setting was used through our example analyses. Specifying the parameters in the gamma prior as small values were suggested in Li et al. (2011). We totally generated 65,000 dependent samples from the Gibbs sampler, where the first 5,000 samples were considered as burn-in, the remaining were stored in every 30th, so that eventually we obtained 2,000 samples. The convergence was assessed by visual inspection of the trace plots of several parameters. The posterior mean of those samples was used as the point estimates for the marker effects. Furthermore, the 95 % credible intervals (CI) were also calculated, and if the CI for a marker did not contain zero, the corresponding marker was judged to be linked to a QTL. For simplicity, we used the abbreviation forms as L-CV, L-BIC ,EN-CV, EN-BIC, AL-CV, AL-BIC to represent LASSO, Elastic net and Adaptive LASSO with either CV or BIC to select the value of λ, and BL for Bayesian LASSO, respectively.

Figure 1 shows the estimated marker effects using all the methods we mentioned above, and Table 1 shows the markers which are judged to be linked to QTLs and their corresponding p values (or credible intervals). Clearly, as expected we can see the different patterns of the estimates obtained by different methods. Compared to LASSO, Elastic net tends to assign average effects to the markers which are highly correlated. On the other hand, Adaptive LASSO tends to provide a sparser estimate than LASSO, and does not shrink those marker effects with large absolute values as much as LASSO does. Furthermore, it is obvious that the BIC method can lead to sparser models than fivefold cross validation. Also in the hypothesis testing procedure, more QTL signals are detected if BIC is used for the shrinkage factor selection in the screening step for LASSO and Elastic net. Finally, the estimates from BL show a quite different pattern. BL does not shrink any marker effect exactly to zero, probably due to the fact (1) the hierarchical (scale mixture) posterior model is not exactly the same as the true LASSO model and (2) we consider the posterior mean instead of the posterior mode as the point estimate. However, the QTLs detected in Bayesian LASSO does not differ much from those obtained by the other methods. In fact, markers 2, 12, and 102, which were found to be significant in most approaches, were close to the major QTLs detected by the interval mapping approach in Tinker et al. (1996).

Genomic selection

Here we consider the cross validation error (see Table 2) as the measure of predictive ability. For L-CV, EN-CV and AL-CV, we naturally report the lowest CV-error corresponding to the optimal shrinkage factor $\hat{\lambda}$ we chose. On the other hand, in L-BIC, EN-BIC and AL-BIC, we performed the fivefold CV only based on the optimal shrinkage factor selected by BIC to evaluate their predictive abilities. For each method, we repeatedly ran the fivefold CV for 100 times and report the average performance. Usai et al. (2009) also averaged over many CV replications. This may help to reduce the biasness of a prediction error estimate caused by randomly splitting data in a CV procedure. AL-CV tends to give the smallest CV error, followed by AL-BIC, and then EN-CV. On average, the prediction tends to be optimized at $\hat{\alpha} = 0.5570$ for EN-CV, meaning that it takes equivalent strength from ℓ_1 and ℓ_2 norm penalties. Furthermore, Bayesian LASSO tends to provide slightly better prediction than L-CV, although it does not shrink any marker effect to exact zero. It is also interesting to see that L-BIC, EN-BIC ($\hat{\alpha} = 0.9$) and AL-BIC shows less predictive abilities than the corresponding methods with CV to determine the shrinkage factors, although they tend to produce much sparser models. Finally, we further implemented a Ridge Regression-BLUP (RR-BLUP) method on this data using the R package `rrBLUP` (Endelman 2011). The method RR-BLUP refers to a special case of Ridge regression with the shrinkage factor fixed to be $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$, the ratio of the residual variance and the genetic variance of each marker. In the `rrBLUP` package, those variance components were estimated by the restricted maximum likelihood (REML) approach (Patterson and Thompson 1971). In this example, the predictive performance of RR-BLUP is better than L-BIC and is equivalent with EN-BIC, but is worse than all the other approaches.

Simulated data

This data set was originally simulated for XII QTL MAS Workshop 2008, Uppsala (http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html). A population of individuals with seven recorded generations was simulated. The first recorded generation consists of 15 males and 150 females. The other six generations with 1,500 individuals

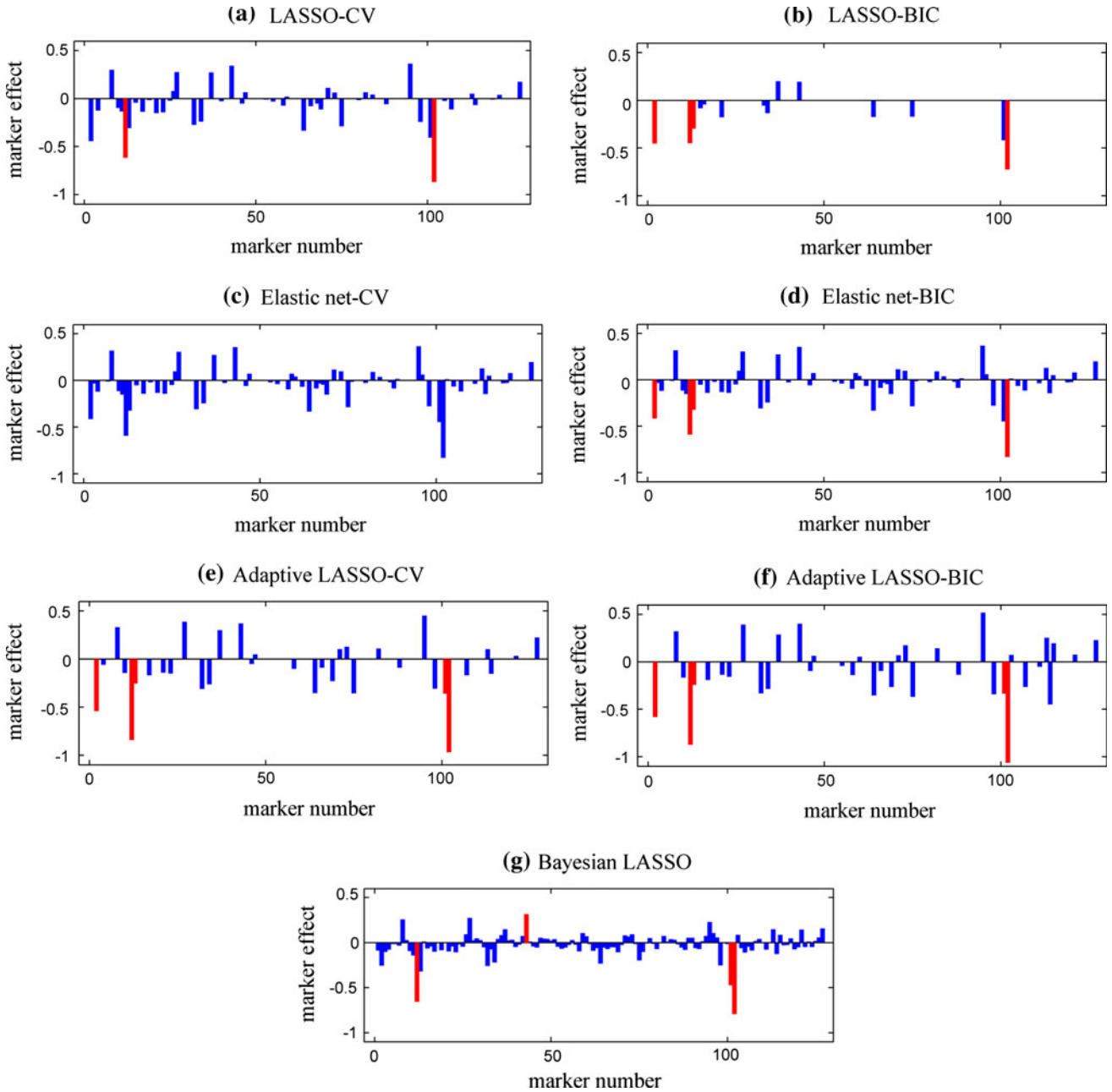


Fig. 1 In barley data, the estimated coefficients, regarded as the marker effects are plotted against marker locations along the genome for **a** L-CV, **b** L-BIC, **c** EN-CV, **d** EN-BIC, **e** AL-CV, **f** AL-BIC, and **g** BL. A *red solid line* indicates the location of a QTL detected by the multi-split method of Meinshausen et al. (2009) for LASSO, Elastic net and Adaptive LASSO, and by credible interval test for Bayesian LASSO

for each were simulated by crossing 15 males and 150 females randomly selected from the last generation. The first four generations consisting of 4,665 individuals were considered as the training data set. For each of the last three generations, 400 individuals were chosen randomly, and the total 1,200 individuals constructed the validation data set. For each individual, 6,000 biallelic markers evenly distributed in six chromosomes with length of 100 cM for each, were genotyped. The constant genetic distance between every two markers was 0.1 cM. In addition, 50 simulated QTLs with their genomic locations, additive effects, and genetic variances were known. 15 of them are considered as major QTLs (M-QTL), and the remaining are defined as secondary QTLs (S-QTL). Finally, the true genomic breeding values for individuals in the training data set were also listed on the website given above to be used for evaluating the prediction accuracy. More information of the data set is available in Lund et al. (2009).

Table 1 The indices of markers which are detected as QTLs and the corresponding p values (in brackets) by the multi-split method of Meinshausen et al. (2009) for LASSO, Elastic net and Adaptive LASSO, and those detected as QTLs and the corresponding 95 % CI (in brackets) by Bayesian LASSO for the barley data

Methods	QTL signals and p values (or CI)
L-CV	$2(1.10 \times 10^{-3})$, $102(2.00 \times 10^{-4})$
L-BIC	$2(1.18 \times 10^{-2})$, $12(4.00 \times 10^{-4})$, $13(2.00 \times 10^{-4})$, $102(2.20 \times 10^{-3})$
EN-CV	None
EN-BIC	$2(1.22 \times 10^{-2})$, $12(9.20 \times 10^{-3})$, $13(1.00 \times 10^{-4})$, $102(1.80 \times 10^{-3})$
AL-CV	$2(4.10 \times 10^{-3})$, $12(6.93 \times 10^{-7})$, $13(1.44 \times 10^{-4})$, $101(1.07 \times 10^{-5})$, $102(3.17 \times 10^{-9})$
AL-BIC	$2(2.06 \times 10^{-2})$, $12(2.96 \times 10^{-6})$,$13(4.26 \times 10^{-6})$, $101(1.49 \times 10^{-7})$, $102(1.73 \times 10^{-9})$
BL	2([−1.1370 − 0.1606]), 43([0.0540 0.5792]), 101([−0.8958 − 0.0445]), 102([−0.3347 − 1.2456])

Table 2 The fivefold cross validation errors (CVE) for each method and the number of non-zero markers (NNM) selected by them averaged over 100 runs for the barley data

Methods	CVE	NNM
L-CV	1.4382	51.25
L-BIC	1.8902	15
EN-CV	1.3069	60.56
EN-BIC	1.5254	15
AL-CV	0.9367	37.43
AL-BIC	1.1009	20
BL	1.4282	127
RR-BLUP	1.5379	127

QTL mapping

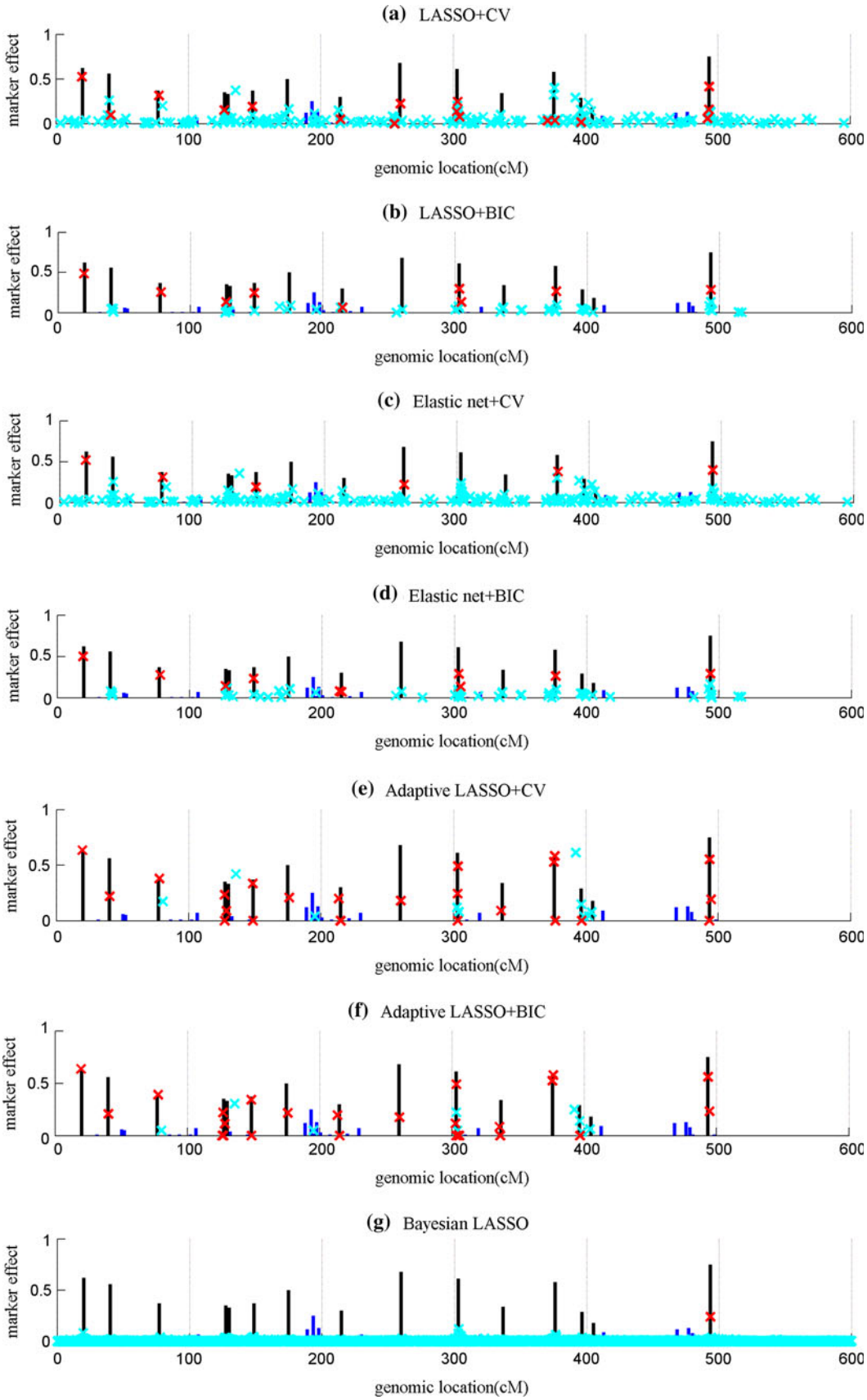
We applied the above mentioned four methods (but not RR-BLUP) on the simulated training data with 4,665 individuals. Here we did not take the relationships among individuals into consideration. For each method, we used the same strategies to determine the tuning parameters, and judge the QTL signals as we did in the barley data analysis. The estimated marker effects are shown in Fig. 2. L-CV and EN-CV tends to produce over two hundred non-zero estimates which cover almost all the genomic regions of true QTL effects (including both M-QTL and S-QTL), but in addition they also produce many spurious signals (i.e. non-zero estimates that are located far from the true simulated QTLs). In contrast, L-BIC and EN-BIC produce fewer non-zero effects. In addition, AL-CV and AL-BIC tend to produce even fewer non-zero effects, but they miss at S-QTL positions. In addition, LASSO and Elastic net tend to underestimate the effect sizes of M-QTLs, but Adaptive LASSO tend to give more precise estimates. Finally, Bayesian LASSO tends to significantly underestimate almost all the QTL effects, and cannot provide clear separations between QTLs and non-QTLs.

Based on Crooks et al. (2009), if a reported QTL is located within 5 cM of a simulated M-QTL, it is considered as a correct identification of that QTL. The mapped positions of QTLs found by different methods are shown in Table 3. The multi-split testing method with AL-BIC and AL-CV succeeded in identifying 13 M-QTLs, followed by L-CV detecting 11 M-QTLs, L-BIC and EN-BIC detecting 8 M-QTLs, and EN-CV detecting 6 M-QTLs. Note that for each of the above mentioned approaches, all the reported QTLs are within 5 cM of at least one M-QTL, and sometimes there are more than one reported QTL located within 5 cM of a M-QTL. Crooks et al. (2009) also reported the performance of several methods. According to them, the 'best' method called LDHap identified 11 M-QTLs, and the 'worst' methods LDLA1 and LDLA2 detected 7 M-QTLs. The results from the above methods seem to be competitive with the previous published results. In addition, the 95% credible interval based on Bayesian LASSO only reported one QTL, which is located close to M15. The poor performance of the credible interval approach is likely due to high collinearity among markers in the data.

Genomic selection

After obtaining the estimates of the marker effects from training data with 4,665 individuals, the TBV can be calculated based on the validation set with 1,200 individuals. We consider (1) the correlation coefficient between TBV and GEBV as cor(TBV,GEBV), (2) the regression coefficient of regressing TBV on GEBV as b(TBV,GEBV) (see Meuwissen et al. 2001; Usai et al. 2009), and (3) the

Fig. 2 In QTLMAS XII simulated data, the estimated coefficients ▶ regarded as the marker effects (*cross*) are compared to the simulated QTL effects (*solid line*) for **a** L-CV, **b** L-BIC, **c** EN-CV, **d** EN-BIC, **e** AL-CV, **f** AL-BIC, and **g** BL. *Black solid lines* represent the major QTLs, and *blue solid lines* represent the secondary QTLs. *Red crosses* represents the location of QTLs detected by the multi-split method of Meinshausen et al. (2009) for LASSO, Elastic net and Adaptive LASSO, and by credible interval test for Bayesian LASSO


(a) LASSO+CV
marker effect
genomic location(cM)
(b) LASSO+BIC
marker effect
genomic location(cM)
(c) Elastic net+CV
marker effect
genomic location(cM)
(d) Elastic net+BIC
marker effect
genomic location(cM)
(e) Adaptive LASSO+CV
marker effect
genomic location(cM)
(f) Adaptive LASSO+BIC
marker effect
genomic location(cM)
(g) Bayesian LASSO
marker effect
genomic location(cM)

averaged cross validation error (repeated over 10 times for BL, and 100 times for the others) only based on the training data to evaluate the predictive ability of different methods. The results were summarized in Table 4. L-CV produced the highest cor(TBV,GEBV), followed by EN-CV (on average $\hat{\alpha} = 0.5670$). On the other hand, the cross validation error of EN-CV is the smallest. Both L-CV and EN-CV have the b(TBV,GEBV) close to 1, meaning that they are able to produce nearly unbiased prediction of true breeding values. Furthermore, AL-CV and AL-BIC showed equivalent predictive ability, and their results were slightly worse than L-CV and EN-CV. L-BIC and EN-BIC ($\hat{\alpha} = 0.9$) produced lower prediction accuracies than L-CV and EN-CV. In addition, Bayesian LASSO does not give quite high prediction accuracy for this simulated data, probably due to the fact that it provides too biased estimates of both M-QTLs and S-QTLs. In addition, note that we assigned the non-informative prior Gamma$(0.0001, 0.0001)$ to λ^2 as we did in the barley data, which might not be an optimal choice for genomic selection. Alternatively, we may use the CV to choose a λ^2 or hyperparameters giving the highest predictive accuracy, but that would require huge computation time.

This QTLMAS XII simulated data set has been intensively used for evaluating prediction performances of other methods based on the marker regression model. Compared with the correlation coefficients cor(TBV,GEBV) reported in related literatures include 0.75 by RR-BLUP (Usai et al. 2009), 0.84 by BayesA (Usai et al. 2009), 0.84–0.87 by Bayesian MCMC methods as reported in Lund et al. (2009), 0.85–0.87 by EMBayesB and 0.85 by ICE (Shepherd et al. 2010), the results obtained from LASSO, Elastic net and Adaptive LASSO are also quite competitive. Finally, note that, Usai et al. (2009) also applied the standard LASSO method on the same QTLMAS XII simulated data. The cor(TBV,GEBV) obtained by them is slightly higher than ours, because Usai et al. (2009) proposed a cross validation procedure different from ours. First of all, in Eq. (7), they used $P(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v(\lambda)) = \text{cor}(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v(\lambda))$ as the metric of the prediction in CV instead of the prediction error as we used. Second, they also took the information of each individual's pedigree and family relationship into account when defining training and validation data. Here, we focused more on a comparison of different methods, and therefore we simply used a more standard approach in statistics to perform cross validation.

Table 4 The number of non-zero marker effects (NNM), the cross validation error (CVE), the prediction accuracy based on the correlation coefficient between TBV and GEBV (PA) and the regression coefficient of regressing TBV on GEBV (RC) for L-CV, L-BIC, EN-CV, EN-BIC, AL-CV, AL-BIC, and BL for the simulated data

Methods	NNM	CVE	PA	RC
L-CV	230.11	3.2503	0.8845	1.0711
L-BIC	50	3.4503	0.8221	1.3980
EN-CV	268.69	3.2323	0.8775	1.0557
EN-BIC	64	3.3937	0.8466	1.3385
AL-CV	25.78	3.3015	0.8663	1.1138
AL-BIC	26	3.3051	0.8667	1.1146
BL	6000	3.3234	0.7688	0.8591

Wheat data

To further test the predictive ability of the involved methods for plant breeding, we consider a wheat data set, which is available from the R package BLR (Pérez et al.

Table 3 The locations (cM) of the QTLs identified by (a) L-CV, (b) L-BIC, (c) EN-CV, (d) EN-BIC, (e) AL-CV, (f) AL-BIC, and (g) BL, which is located within 5cM of the M-QTLs for the simulated data

Simulated QTL			Estimated QTL						
QTL	Chr	Loc	L-CV	L-BIC	EN-CV	EN-BIC	AL-CV	AL-BIC	BL
M1	1	20.0	19.6	19.6	19.6	19.6	19.6	19.6	
M2	1	40.0	41.4				40.2	40.2	
M3	1	77.2	77.8	77.8	77.8	77.8	77.8	77.8	
M4	2	27.4	27.0	27.0		27.0	27.1	27.1	
M5	2	30.0							
M6	2	48.6	48.3	48.3	48.3	48.3	48.4	48.4	
M7	2	74.9					75.8	75.8	
M8	3	14.9	14.9	14.9		14.9	14.9	14.9	
M9	3	60.0	60.2		60.2		60.2	60.2	
M10	4	3.2	3.3	3.3		3.3	3.2	3.3	
M11	4	36.9					35.9	36.4	
M12	4	76.1	76.9	76.5	76.5	76.5	76.5	76.5	
M13	4	96.5	96.7				96.7	96.7	
M14	5	5.1							
M15	5	93.5	93.5	93.5	93.5	93.5	93.5	93.5	93.5

If more than one QTLs were found to be located within 5 cM of a M-QTL, then only the nearest QTL is reported. The multi-split method of Meinshausen et al. (2009) was used to judge QTLs for LASSO, Elastic net and Adaptive LASSO, and credible interval test was used for Bayesian LASSO

2010). See also Crossa et al. (2010). The wheat data set contains a set of 599 CIMMYT wheat lines, which were genotyped using 1447 Diversity Array Technology (DArT) markers. After removing markers with a minor allele frequency lower than 0.05, the remaining 1,279 markers were used for analyses. The trait was a 2-year average grain yield in four environments. Here we treated four environments independently, and analyzed the trait corresponding to each of them separately. In addition to the above mentioned methods, we also analyzed the data by RR-BLUP and mixed Bayesian LASSO (MBL) with both genotype and pedigree information. MBL is implemented by the R package BLR (Pérez et al. 2010). The predictive ability of each method was evaluated by cross validation (repeated over 100 times) similarly as in the barley data example. The results are summarized in Table 5. The orders of predictive performance for those methods are similar through the four environments. Different from barley and simulated data examples, here RR-BLUP has better predictive performance than L-CV. EN-CV also produces smaller cross validation error than L-CV. Interestingly, for the four environments, the predictions are optimized at $\alpha = 0.30$, $\alpha = 0.32$, $\alpha = 0.37$ and $\alpha = 0.32$ respectively on average, indicating the ℓ_2 (Ridge regression) penalty is preferable for this data. The performance of BL and MBL is quite competitive as well. MBL provides the best predictive performance among all the methods for the first environment, and provides the second best for the remaining. The fact that MBL had better predictive ability than BL shows the importance of including polygenic effects into the model. These results jointly support the polygenic architecture of the trait. Surprisingly, AL-CV shows identical predictive ability to MBL for the first environment, and its performance is the best for the other three. Note that the behavior of AL-CV is quite different from MBL and RR-BLUP, because it usually selects less than 60 markers into the model. Finally, mixed LASSO, mixed Elastic net and mixed Adaptive LASSO were not performed due to the lack of the relevant software tools. The performance of these methods are remaining for future investigation.

Discussion

In this article, we have focused on discussing four penalized regression methods including LASSO and its extensions Elastic net, Adaptive LASSO and Bayesian LASSO with their applications to QTL mapping and genomic selection. Although in principle these methods can be used for both identifying a correct subset of markers linked to QTLs and predicting genome-enhanced breeding values. These two tasks are quite different, and should be treated separately.

The results from our example analyses indicate that a model with good predictive ability may select a relatively large number of markers with non-zero effects, which contains some noisy signals in addition to the true QTLs. More general discussions and theoretical justifications about this point can be found in Bühlmann and van de Geer (2011). Specifically, a simulated study performed by Habier et al. (2007) showed that not only markers which are in LD with QTLs, but also the markers capturing the genetic relationships can contribute to the prediction accuracy in genomic selection. In practice, cross validation is a promising tool to determine the optimal penalty of a method for prediction when the individuals are not so correlated with each other. If a close genetic relationships exist among the individuals (e.g. individuals with offsprings) in the training data, the predictive accuracy estimated by CV might be upward biased (Habier et al. 2007; Dekkers 2010), and the optimal shrinkage factor determined by CV may not be the best for predicting the breeding values of individuals from new generations.

Table 5 The fivefold cross validation errors (CVE) for each method and the number of non-zero markers (NNM) selected by them averaged over 100 runs for the wheat data

Methods	Environment1		Environment2		Environment3		Environment4	
	CVE	NNM	CVE	NNM	CVE	NNM	CVE	NNM
L-CV	0.7938	137.37	0.8108	122.33	0.8906	87.21	0.8255	74.83
L-BIC	0.9082	8	0.9996	1.96	0.9997	2.53	0.9047	14.75
EN-CV	0.7550	212.2	0.7745	181.82	0.8570	148.24	0.7965	136.77
EN-BIC	0.9025	9	1.0006	2.76	1.0011	3.69	0.9046	15.77
AL-CV	0.7299	27.96	0.6179	57.55	0.7013	37.56	0.6987	36.91
AL-BIC	0.7475	20	0.6812	41.02	0.7990	22.53	0.7499	24.22
BL	0.7498	1279	0.7612	1279	0.8655	1279	0.7885	1279
MBL	0.7281	1279	0.7579	1279	0.8251	1279	0.7620	1279
RR-BLUP	0.7469	1279	0.7613	1279	0.8649	1279	0.7924	1279

In this case, CV can be more carefully designed so that the individuals from the latest generation are only used in the validation set in order to obtain a model for better predicting the genomic breeding values of new generations (Usai et al. 2009). Alternatively, the use of mixed LASSO by including both pedigree and marker information into the model may also be beneficial to analyze such kind of data.

On the other hand, in QTL/association mapping, we are interested in detecting the markers which are only in LD with the QTLs. LASSO and Elastic net do not hold the variable selection consistency in general, and they often tend to select some noisy signals. Although Adaptive LASSO has better theoretical properties for consistent variable selection, the difficulty about how to choose an optimal λ for correctly identifying the QTLs is still remaining. BIC might be more suitable for this purpose as a criterion to determine λ, since it can usually lead to a model with fewer estimated QTLs compared to CV. However, Chen and Chen (2008) argued that BIC is still too liberal for the variable selection in high dimensional data ($p > n$ problem). A more reliable way to judge QTLs is performing hypothesis testing, and obtain a p value for each marker effect to control the false positive errors. Since constructing a test statistic based on the estimates of LASSO, Elastic net and Adaptive LASSO is difficult, Wasserman and Roeder (2009) and Meinshausen et al. (2009) suggested two-step procedures for obtaining p values. Both of them showed that their approaches can yield asymptotic error controls under certain conditions. However, the multi-split method of Meinshausen et al. (2009) should be a more stable approach than the single-split approach of Wasserman and Roeder (2009). In our example analyses, we tested the multi-split method, and obtained a good performance on identifying the QTLs with large effects and controlling false positives by tolerating some false negatives. However, it requires a repeating procedure that increases the time costs for computation, and may not be optimal for large scale data. An alternative fast approach is a two-step procedure proposed by Chen and Cui (2010), where in the second stage, EBIC is used for selecting QTLs. However, this achieves asymptotic FDR control only when the consistent variable selection of LASSO is assumed, which is more restrictive than the screening property assumed by Meinshausen et al. (2009). Other multi-stage strategies for testing QTLs based on LASSO or Adaptive LASSO estimates are provided in Wu et al. (2009) and Sun et al. (2010), but they are rather heuristic approaches that lack theoretical justifications, and should be used carefully. Another recently proposed method of interest is stability selection (Meishausen and Bühlmann 2010; Alexander and Lange 2011). In summary, identifying QTL signals is a more challenging problem than predicting genomic breeding values under a LASSO procedure, and it seems that currently there is no standard methodology for this task. Therefore, the problem of post-LASSO QTL identification needs more investigation in future.

The performance of a method for both variable selection and prediction will also depend on the data. According to Tibshirani (1996), LASSO would perform better than Ridge regression for shrinkage estimation on a data with a small portion of variables having large effects and the others with negligible effects, and Ridge regression would be more suitable for the data with many variables having small effects. From the theoretical perspective, Bühlmann and van de Geer (2011) discussed that LASSO are not able to select many explanatory variables with very small effects. Moreover, LASSO can only select at most n variables when $p > n$. Thus, if it happens that there are many QTLs with small effects in a data set and their cumulative contribution to the trait is large, LASSO may work poorly on detecting causal genes and estimating genomic breeding values. From the application perspective, Daetwyler et al. (2010) compared the performance of Bayes B method, a Bayesian variable selection method (Meuwissen et al. 2001) to GBLUP (BLUP with realized relationship matrix), which has been shown to be equivalent to RR-BLUP (Habier et al. 2007; VanRaden 2008), for genomic selection on a series of simulated data sets. They concluded that when the total number of QTLs was small, variable selection methods such as Bayes B had better performance than GBLUP, but when the total number of QTLs was large, GBLUP tended to be dominant over Bayes B. In a related simulation study, Clark et al. (2011) also found that GBLUP produced slightly higher prediction accuracy than Bayes B when the data consisted of a large number of QTLs with small effects. As a variable selection method, LASSO may have a behavior similar to the Bayes B method, and may not be suitable for a data set with large number of QTLs. Some evidences of this behavior may also be found in the results of our example analyses. LASSO shows better predictive ability than RR-BLUP for the simulated data with only 50 QTLs among 6,000 markers, but shows worse predictive ability for the wheat data. In addition, Usai et al. (2009) analyzed a mouse data set (Valdar et al. 2006), where they reported that for the traits "weight growth slope" and "body length", RR-BLUP showed slightly better predictive ability than LASSO. Therefore, it is important to choose a suitable method based on the prior knowledge on genetic architecture. Particularly, the Elastic net method may be applicable to a data set with no pre-knowledge of genetic architecture available, since it provides a possibility to determine an optimal combination of ℓ_1 and ℓ_2 norm penalties. Such an example can be found in Harris and Johnson (2010). Finally, in all three example analyses, we also observed that the Adaptive LASSO method performs

constantly well. This is surprising because Adaptive LASSO always selects fewer markers with non-zero effects than LASSO, but can have good predictive performance even when RR-BLUP is superior to LASSO. Empirical evaluation of this method deserves more attention.

Finally, the MCMC-based Bayesian LASSO and its related approaches have also gained increasing interest and have been applied to several QTL mapping and genomic selection studies. Compared to standard LASSO, an advantage of Bayesian LASSO is that it can provide the interval estimates in addition to the point estimates, which can be used for identifying QTLs. In addition, relatively non-informative priors can be assigned to the shrinkage factor, so that tuning can be avoided. However, we should be aware that the posterior mean estimates obtained from the Bayesian LASSO can be quite different from the LASSO estimates. First, Bayesian LASSO cannot produce as sparse model as LASSO can (Sun et al. 2010). Second, from the results of our data examples, Bayesian LASSO tends to underestimate those marker effects even more severely than LASSO. In fact, Park and Casella (2008) showed empirically that the Bayesian LASSO estimates tended to be a compromise between the LASSO and ridge regression estimates. Thus, it may not be an optimal method for data having only a small number of QTLs with large effects. This point of view needs to be further studied.

Acknowledgments We thank Daniel Blande, Mahlako Makgahlela and Crispin Mutshinda for giving constructive suggestions on the manuscript. We are also grateful to two anonymous referees for their valuable comments. This work was supported by the Finnish Graduate School of Population Genetics, and by research grants from the Academy of Finland and the University of Helsinki's Research Funds.

Appendix: Coordinate descent algorithm

Initially, the marker data are assumed to be standardized and phenotype data to be centered so that $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$, and $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$. The Elastic net problem (LASSO: $\alpha = 1$, Ridge regression: $\alpha = 0$) can be specified as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2N}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\left[(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right]\right\}. \quad (16)$$

The principle of the coordinate descent is that when minimizing the Elastic net target function, the algorithm updates each component β_j successively in the direction giving the largest decrease of the objective function by fixing all other components. Assuming the current estimate of β_j is $\beta_j^{(0)}$, and we have already updated the estimate of $\beta_1, \beta_2, \ldots, \beta_{j-1}$ as $\beta_1^{(1)}, \beta_2^{(1)}, \ldots, \beta_{j-1}^{(1)}$, the estimate of $\beta_j^{(1)}$ can be updated as

$$\beta_j^{(1)}(\lambda) = \frac{S(\beta_j^{(0)} + \frac{1}{N}\sum_{i=1}^{n} x_{ij}r_i, \lambda\alpha)}{1 + \lambda(1-\alpha)}, \quad (17)$$

where $S(a, b)$ is the thresholding function defined as

$$S(a, b) = \text{sign}(a) \cdot \max(|a| - b, 0), \quad (18)$$

and $r_i = y_i - \sum_j^p x_{ij}\beta_j$ for $i = 1, \ldots, n$ is the residual, which should be updated as $r_i = r_i + x_{ij}(\beta_j^{(0)} - \beta_j^{(1)})$ when $\beta_j^{(1)}$ is ready. The algorithm updates each component of $\boldsymbol{\beta}$ in a cyclic manner as $1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$, until the solutions converge.

The coordinate descent algorithm can be used for Adaptive LASSO as well. In each iteration, we use the update function:

$$\beta_j^{(1)}(\lambda) = S\left(\beta_j^{(0)} + \frac{1}{N}\sum_{i=1}^{n} x_{ij}r_i, \frac{\lambda}{|\hat{\beta}_{\text{init},j}|}\right), \quad (19)$$

where $\hat{\beta}_{\text{init},j}$ are certain initial estimates, for example, from OLS or standard LASSO.

For more information, see Friedman et al. (2007) and (2010).

References

Akaike H (1974) New look at the statistical model identification. IEEE T Autom Contr 19:716–723

Alexander DH, Lange K (2011) Stability selection for genome-wide association. Genet Epidemiol 35:722–728

Ayers KL, Cordell HJ (2010) SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34:879–891

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc B 64:641–656

Bühlmann P, Meier L (2008) Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" (authors Zou H and Li R). Ann Stat 36:1534–1541

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, New York

Burgueño J, DeLos Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719

Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95:759–771

Chen J, Cui W (2010) A two-phase procedure for QTL mapping with regression models. Theor Appl Genet 121:363–372

Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, Han BG, Kim H, Ott J, Park T (2010) Joint identification of multiple genetic

variants via elastic-net variable selection in a genome-wide association analysis. Ann Hum Genet 74:416–428
Clark SA, Hickey JM, van der Werf JHJ (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18
Crooks L, Sahana G, De Koning DJ, Lund MS, Carlborg Ö (2009) Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping. BMC Proc 3:S2
Crossa J, DeLos Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724
Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031
Dekkers JCM (2010) Use of high-density marker genotype for genetic improvement of livestock by genomic selection. CAB Reviews 5
Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3:22–32
DeLos Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385
Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32:407–451
Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package `rrBLUP`. Plant Genome 4:250–255
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Figueiredo MAT (2003) Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell 25:1150–1159
Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1:302–332
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1
Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330
Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397
Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324
Harris BL, Johnson DL (2010) SNP selection using Elastic net, with application to genomic selection. In 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany. http://www.kongressband.de/wcgalp2010/assets/pdf/0282.pdf
Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning. Springer, New York
Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12
Hesterberg T, Choi NH, Meier L, Fraley C (2008) Least angle and ℓ_1 penalized regression: a review. Stat Surv 2:61–93
Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
Huang J, Ma S, Zhang CH (2008) Adaptive Lasso for sparse high-dimensional regression models. Stat Sin 18:1603–1618
Jannink JL, Bink MCAM, Jansen RC (2001) Using complex plant pedigrees to map valuable genes. Trends Plant Sci 6:337–342
Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and Bayesian Lassos. Bayesian Anal 2:369–412
Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S (2011) Improved Lasso for genomic selection. Genet Res 93:77–87
Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Anal 5:151–170
Li Z, Sillanpää MJ (2012) Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231–249
Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian LASSO for genome-wide association studies. Bioinformatics 27:516–523
Lund MS, Sahana G, De Koning DJ, Su G, Carlborg Ö (2009) Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proc 3:S1
Meinshausen N (2007) Relaxed LASSO. Comput Stat Data An 52:374–393
Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34:1436–1462
Meinshausen N, Bühlmann P (2010) Stability selection. J Roy Stat Soc B 72:417–473
Meinshausen N, Meier L, Bühlmann P (2009) P-values for high-dimensional regression. J Am Stat Assoc 104:1671–1681
Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Mutshinda CM, Sillanpää MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186:1067–1075
Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA J Numer Anal 20:389–404
Park T, Casella G (2008) The Bayesian LASSO. J Am Stat Assoc 103:681–686
Patterson HD, Thompson R (1971) Recovery of inter-block information with block sizes are unequal. Biometrika 58:545–554
Pérez P, DeLos Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. Plant Genome 3:106–116
Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176
Piepho HP, Ogutu JO, Schulz-Streeck T, Estaghvirou B, Gordillo A, Technow F (2012) Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci 52:1093–1104
Shepherd RK, Meuwissen THE, Woolliams JA (2010) Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinforma 11:529
Siegmund D, Yakir B (2007) The statistics of gene mapping. Springer, Berlin
Sillanpää MJ (2011) Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106:511–519
Sillanpää MJ, Corander J (2002) Model choice in gene mapping: what and why. Trends Genet 18:301–307
Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 39:5
Solberg TR, Sonesson AK, Woolliams JA, Ødegard J, Meuwissen THE (2009) Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genet Sel Evol 41:53
Sun W, Ibrahim JG, Zou F (2010) Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics 185:349–359
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58:267–288
Tinker NA, Mather DE, Rossnagel BG, Kasha KJ, Kleinhofs A et al (1996) Regions of the genome that affect agronomic performance in two-row barley. Crop Sci 36:1053–1062

Usai MG, Goddard ME, Hayes BJ (2009) LASSO with cross-validation for genomic selection. Genet Res 91:427–436

Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R, Flint J (2006) Genetic and environmental effects on complex traits in mice. Genetics 174:959–984

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang D, Eskridge KM, Crossa J (2010) Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. J Agric Biol Envir S 16:170–184

Wasserman L, Roeder K (2009) High dimensional variable selection. Ann Stat 37:2178–2201

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

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801

Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63:513–521

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

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhao P, Yu B (2006) On model selection consistency of LASSO. J Mach Learn Res 7:2541–2563

Zhou S (2010) Thresholded Lasso for high dimensional variable selection and statistical estimation. arXiv:1002.1583v2

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

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

Zou H, Hastie T (2008) Model building and feature selection with genomic data. In Liu H and Motoda H, editors, Computational Methods of Feature Selection, chapter 20, pp 393–411. Chapman & Hall, London

Zou H, Zhang H (2009) On the adaptive elastic-net with a diverging number of parameters. Ann Stat 37:1733–1751

Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the lasso. Ann Stat 35:2173–2192