Communications for Statistical Applications and Methods
2014, Vol. 21, No. 4, 349–361

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.4.349
Print ISSN 2287-7843 / Online ISSN 2383-4757

Comparison of Lasso Type Estimators for High-Dimensional Data

Jaehee Kim[1,a]

[a]Department of Statistics, Duksung Women's University, Korea

Abstract

This paper compares of lasso type estimators in various high-dimensional data situations with sparse parameters. Lasso, adaptive lasso, fused lasso and elastic net as lasso type estimators and ridge estimator are compared via simulation in linear models with correlated and uncorrelated covariates and binary regression models with correlated covariates and discrete covariates. Each method is shown to have advantages with different penalty conditions according to sparsity patterns of regression parameters. We applied the lasso type methods to Arabidopsis microarray gene expression data to find the strongly significant genes to distinguish two groups.

Keywords: Adaptive Lasso, elastic net, fused lasso, high-dimensional data, lasso, ridge.

1. Introduction

High-dimensional data refers to a situation where the number of unknown parameters p are estimated to be larger than the number of samples n in the data ($p >> n$). High-dimensional data arise in the areas like information technology, bioinformatics, astronomy and brain research. Classical statistical inference cannot be used for high-dimensional problems. High-dimensional statistical inference is impossible without additional assumptions or restrictions to a certain class of models. For example, least-squares fitting of a linear model having many unknown parameters than observations is ill-posed. A well-posed framework for fitting is based on assuming structural smoothness. Shifting the focus from smoothness to sparsity for high-dimensional data opens the way for many more applications that involve complex data.

The lasso, proposed by Tibshirani (1996), is an acronym for Least Absolute Shrinkage and Selection Operator. The lasso estimates a vector of regression coefficients by minimizing the residual sum of squares subject to a constraint on L_1-norm of the coefficient vector. The lasso estimator typically has one or more zero elements and shares characteristics of both shrinkage estimation and variable selection. The method is rather general and can be used in a broad variety of models since the lasso is a penalized likelihood approach. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and gives interpretable models with lower dimensional data. The form of this penalty encourages sparse solutions (with many coefficients equal to 0). There are various penalties reflecting the coefficient features for example sparsity of their differences. Bühlmann and van de Geer (2011) provided methodological concepts and mathematical theory for high-dimensional statistics.

In this paper, we compare lasso type estimators in some high-dimensional situation setups. We understand the features of penalties and compare the selection methods in linear models and generalized linear models via simulations. In Section 2 we describe lasso type estimators with their properties. Section 3 gives the simulation results and a real data application. Finally Section 4 concludes.

This research was supported by Duksung Women's University 2013 Research Fund.

[1] Department of Statistics, Duksung Women's University, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr

Published online 31 July 2014 / journal homepage: http://csam.or.kr

2. Penalized Regression Problems in Linear Models

Consider a linear regression with standardized predictors x_{ij} and centered response values y_i for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ with p-dimensional covariates such as

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{2.1}$$

where ϵ_i's are identically independently distributed with $E[\epsilon_i] = 0$ and independent of x_{ij}.

For simplicity and without loss of generality, we assume that the intercept is zero with centering such as $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i = 0$. Using the matrix and vector notation, we use

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

with response vector $\mathbf{y}_{n\times 1}$, design matrix $\mathbf{x}_{n\times p}$, parameter vector $\boldsymbol{\beta}_{p\times 1}$ and error vector $\boldsymbol{\epsilon}_{n\times 1}$. The only unusual aspect of the linear model (2.1) is that $p >> n$.

2.1. Ridge estimator

If the multicollinearity exists, $(\mathbf{x}'\mathbf{x})$ is hard to invert accurately for the OLS (ordinary least squares) estimate. For large enough k, $(\mathbf{x}'\mathbf{x}) + \lambda\mathbf{I}$ can be inverted very accurately. Instead of $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$, Hoerl and Kennard (1970) suggested the ridge estimator $\boldsymbol{\beta}_R = (\mathbf{x}'\mathbf{x} + \lambda\mathbf{I})^{-1}\mathbf{x}'\mathbf{y}$ to get better regression estimator than the OLS estimator. Large values of λ correspond to increased bias but lower variance, so a value of λ must be chosen to balance bias against variance. Equivalently, the ridge estimator solves L_2-penalized regression problem of finding β_j's to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{2.3}$$

Equivalently the ridge estimator is represented as

$$\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 \right). \tag{2.4}$$

2.2. Lasso

If $p > n$, the OLS estimator is not unique and will heavily overfit the data. Thus, a form of complexity regularization is necessary. It became very popular for high-dimensional estimation problems for its statistical accuracy for prediction and variable selection coupled with its computational feasibility.

The lasso, proposed by Tibshirani (1996), solves the L_1-penalized regression problem of finding β_j's to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.5}$$

That is, the parameters in (2.2) are estimated with the lasso as

$$\hat{\boldsymbol{\beta}}_{Lasso}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right), \tag{2.6}$$

where $\lambda \geq 0$ is a penalty parameter. The estimator has the property that it does variable selection in the sense that $\hat{\beta}_j(\lambda) = 0$ for some j's and $\hat{\beta}_j(\lambda)$ can be thought as a shrunken least squares estimator.

This is equivalent to minimizing the sum of squares with a constraint of the form $\sum_{j=1}^{p} |\beta_j| \leq s$. It is similar to ridge regression, which has constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$. Because of the form of the L_1-penalty, the lasso does variable selection and shrinkage, whereas ridge regression, in contrast, only shrinks. If a more general penalty of the form is considered as $(\sum_{j=1}^{p} |\beta_j|^q)^{1/q}$ then the lasso uses $q = 1$ and ridge regression has $q = 2$. Subset selection emerges as $q \to 0$, and the lasso uses the smallest value of q (*i.e.* closest to subset selection) that yields a convex problem. Convexity is very attractive for computational purposes. The optimization in (2.6) is convex, enabling efficient computation of the estimator. An equivalence holds since $\|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|_2^2$ is convex in $\boldsymbol{\beta}$ with convex constraint $\|\boldsymbol{\beta}\|_1 \leq R$.

Generalized linear models (GLM) are a unified framework containing many extensions of linear models. Logistic regression for binary responses or Poisson regression for count data are examples of GLM.

Since penalizing with the squared error loss in linear regression is similar to penalizing the negative log-likelihood due to the convexity of the negative log-likelihood. Penalizing the negative log-likelihood with the L_1 norm, still called the lasso.

Consider a classification problem involving a binary response variable $y \in \{0, 1\}$. We can use the lasso which yields an estimate of the conditional class probability $f(\mathbf{x}_i) = P[y = 1|\mathbf{x}_i] = E[y|\mathbf{x}_i]$:

$$\log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = f_\lambda(\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}(\lambda) = \sum_{j=1}^{p} \beta_j(\lambda) x_{ij} \tag{2.7}$$

with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$.

In practice, the tuning parameter λ is usually chosen via some cross-validation scheme that aims for prediction optimality. However L_1-penalty has two limitations: first, the number of selected features is bounded by the number of samples. Second, it tends to select only one feature from a group of correlated features and drops others.

Interesting approaches to correct the lasso's overestimation behavior, researchers provided other lasso type methods. Each method is explained in the following subsections.

2.3. Adaptive lasso

When the lasso is inconsistent for variable selection, Zou (2006) proposed a new version of the lasso, the adaptive lasso, in which adaptive weights are used for penalizing different coefficients in the L_1 penalty. To correct lasso's overestimation behavior, Zou (2006) replaced the L_1 penalty by a re-weighted version. If $\|\hat{\beta}_{init,j}\|_1$ is large, the adaptive lasso employs a small penalty for the jth coefficient β_j which implies less bias. The adaptive lasso is defined as a two-stage procedure for a linear model (2.1) as

$$\hat{\boldsymbol{\beta}}_{Adapt}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right), \tag{2.8}$$

where $\mathbf{w} = (w_1, \ldots, w_p)'$ is a known weight vector. Zou (2006) showed that if the weights are data-dependent and cleverly chosen, then the weighted lasso can have the oracle properties.

Typically, cross-validation is used to select the tuning parameter, denoted by $\hat{\lambda}_{init,CV}$. The initial estimator is $\hat{\boldsymbol{\beta}}_{init} = \hat{\boldsymbol{\beta}}(\hat{\lambda}_{init,CV})$ from (2.6). For the second stage, cross-validation is done again to select the tuning parameter, λ in the adaptive lasso (2.8). Proceeding this way is computationally cheaper

since we optimize twice over a single parameter instead of simultaneous optimization over two tuning parameters. The adaptive lasso yields a sparse solution and it can be used to reduce the number of false positives (selected variables which are not relevant) from the first stage.

Suppose that $\hat{\boldsymbol{\beta}}$ is a root n-consistent estimator, for some $\gamma > 0$, $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$ for example, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. In case collinearity is a concern, we can try $\hat{\boldsymbol{\beta}}_{Ridge}$ from the best ridge regression fit, because it is more stable than $\hat{\boldsymbol{\beta}}_{OLS}$.

2.4. Fused lasso

Tibshirani *et al.* (2005) proposed the fused lasso, a generalization of lasso penalty that is designed for problems with features that can be ordered in some meaningful way. The fused lasso penalizes the L_1-norm of both the coefficients and their successive differences. Thus it encourages sparsity of the coefficients and sparsity of differences such as local constancy of the coefficient profile. The fused lasso is defined for a linear model (2.1) as

$$\hat{\boldsymbol{\beta}}_{Fused}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right). \tag{2.9}$$

The first constraint encourages sparsity in the coefficients; the second encourages sparsity in their differences, *i.e.* flatness of the coefficient profiles β_j as a function of j. The penalty $\lambda_2 \sum |\beta_j - \beta_{j-1}| \leq s_2$ gives a piecewise constant solution, and this corresponds to a simple averaging of the features.

2.5. Elastic net

To overcome the limitations of lasso, Zou and Hastie (2005) proposed double penalization using a linear combination of L_1 and L_2 penalties which is called Elastic Net:

$$\hat{\boldsymbol{\beta}}_{Enet}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2 \right), \tag{2.10}$$

where $\lambda, \lambda_2 \geq 0$ are two regularization parameters.

The elastic net estimator can be interpreted as a stabilized version of the lasso. Since the lasso is not well-defined unless the bound on the L_1-norm of the coefficients is smaller than a certain value. The lasso may select one (but typically not both of the penalties) for strongly correlated covariates. Non-selected variable can then be approximated as a linear function of the selected ones. However, in terms of interpretation, strongly correlated variables among the selected variables may be considered in the model.

If there is a group of variables among which the pairwise correlations are very high, the lasso tends to select one variable from that group. Similar to the lasso, the elastic net simultaneously does automatic variable selection, continuous shrinkage and selection of groups of correlated variables like a stretchable fishing net which retains 'all the big fish'. The algorithm LARS-EN is used to solve the elastic net efficiently based on LARS of Efron *et al.* (2004).

2.6. Selection of tuning parameters

We now discuss how to choose an optimal tuning parameter in lasso type estimation. Five-fold cross-validation of Efron and Tibshirani (1993) can be used for the prediction error. The lasso is indexed in terms of the normalized parameter $s = t/\sum \hat{\beta}_j^0$, and the prediction error is estimated over a grid of values of s from 0 to 1. The value $\hat{s}$ yielding the lowest estimated prediction error is selected.

A second method to estimate t may be derived from a linear approximation to the lasso estimate. The generalized cross-validation style statistic is

$$\mathrm{GCV}(t) = \frac{1}{n}\frac{\mathrm{rss}(t)}{(1 - p(t)/n)^2}, \tag{2.11}$$

where $\mathrm{rss}(t)$ is the residual sum of squares for the constrained fit with constraint t and $p(t)$ is the number of effective parameters in the constrained fit by the lasso. The third method is based on Stein's unbiased estimate of risk (Stein, 1981).

There are well-established methods to choose such tuning parameters (Friedman *et al.*, 2001, Chapter 7). If only training data are available, ten-fold cross-validation (CV) is a popular method to estimate the prediction error and comparing different models. We use five-fold CV in the simulation and for the real data analysis. Note that there are two tuning parameters in the elastic net, so we need to cross-validate on a two-dimensional surface. Typically we first pick a (relatively small) grid of values for λ_2, say $(0, 0.01, 0.1, 1, 10, 100)$. Then, for each λ_2, the LARS-EN algorithm produces the entire solution path of the elastic net. The other tuning parameter (λ, s or k) is selected by ten-fold CV. The chosen λ_2 is the one giving the smallest CV error. For each λ_2, the computational cost of ten-fold CV is the same as 10 OLS fits. Thus two dimensional CV is computationally thrifty in the usual $n > p$ setting. In the $n \ll p$ case, the cost grows linearly with p and is still manageable. Practically, early stopping is used to ease the computational burden.

In practice, the tuning parameter λ is usually chosen via some cross-validation scheme aiming for prediction optimality. Such prediction optimality is often in conflict with variable selection where the goal is to recover the underlying set of active variables. We often need a large penalty parameter than for a good prediction. It is generally difficult to choose a proper amount of regularization to identify the true active set.

The lasso constraint $\sum |\beta_j| \le t$ is equivalent to the addition of a penalty term $\lambda \sum |\beta_j|$ to the residual sum of squares. $|\beta_j|$ is proportional to the log-density of the double-exponential distribution. Therefore the lasso estimate can be derived as the Bayes posterior mode under independent double-exponential priors for the β_j's. Park and Casella (2008) suggested the Bayesian Lasso estimation and provide interval estimates that guide variable selection. Recently Kyung *et al.* (2010) provided the Bayesian estimation for various lasso type tuning parameters.

3. Empirical Study

We illustrate the lasso, adaptive lasso, fused lasso and elastic net on some simulated data from linear models and generalized linear models with uncorrelated and correlated covariates.

We consider some high-dimensional data setups to compare the estimation and selection performance with various penalties in the linear regression models and generalized linear models that include the binary regression.

This simulation compares lasso type estimators in the regression models with sparse data or in a high-dimensional situation.

3.1. Simulation

We simulate data from the true linear model

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \sigma\epsilon, \quad \epsilon \sim N(0, 1).$$

Firstly, we simulated 100 data sets from a linear model (2.1) with $p = 100$ and $n = 10$ observations with $\sigma = 3$. The predictor $\mathbf{x}$ were generated identically and independently from $N(0, \sigma^2)$. The errors $\boldsymbol{\epsilon}$'s are iid from $N(0, \sigma^2)$. We consider 6 sparsity patterns such as many zeros except the first a few or the last a few non-zero parameter values, many zeros with intermittent non-zero parameter values, and many zeros with the first a few non-zero parameter values which are the same.

(1) Let $\boldsymbol{\beta} = (0.3, 1.5, 1, 0, \ldots, 0)$ have 96 zeros out of 100 coefficients.

(2) Let $\boldsymbol{\beta} = (1, 2, 3, 4, 5, 0, \ldots, 0)$ have 95 zeros.

(3) Let $\boldsymbol{\beta} = (2, 0, \ldots, 0, 2, 0, \ldots, 0, \ldots, 2, 0, \ldots, 0)$ have 90 zeros. Nine 0's are between 2's.

(4) Let $\boldsymbol{\beta} = (1, 2, 3, 4, 0, \ldots, 0, 4, 3, 2, 1)$ have 92 zeros.

(5) Let $\boldsymbol{\beta} = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, \ldots, 0)$ have 90 zeros and ten 2's.

(6) Let $\boldsymbol{\beta} = (2, 0, \ldots, 0, 2, 0, \ldots, 0, 2, 0, \ldots, 0)$ have 97 zeros and three 2's.

Secondly, we simulated 100 data sets consisting of $n = 10$ observations with $p = 20$ regression parameters for the explanatory variables. The predictor $\mathbf{x}$ were generated correlated. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is set to be $\text{corr}(x_{ki}, x_{kj}) = 0.5^{|i-j|}$, for $k = 1, \ldots, n,\ i, j = 1, \ldots, p$. Also we consider 6 sparsity patterns as follows:

(i) Set $\boldsymbol{\beta} = (0.3, 1.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$.

(ii) Set $\boldsymbol{\beta} = (1, 2, 3, 4, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$.

(iii) Set $\boldsymbol{\beta} = (2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0)$.

(iv) Set $\boldsymbol{\beta} = (1, 2, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 3, 2, 1)$.

(v) Set $\boldsymbol{\beta} = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0)$.

(vi) Set $\boldsymbol{\beta} = (2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0)$.

In the third, we consider the GLM with binary responses. we simulated 100 data sets consisting of $n = 20$ observations with $p = 100$ regression parameters for the explanatory variables from (1) to (6). For the binary regression models, we generate data $y \sim \text{Bernoulli}\{\text{p}(\mathbf{x}\boldsymbol{\beta})\}$ where $p(u) = \exp(u)/(1 + \exp(u))$. The first two components of $\mathbf{x}$ are identically and independently distributed as a Bernoulli distribution with success probability 0.5. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is set to be $\text{corr}(x_{ki}, x_{kj}) = 0.5^{|i-j|}$, for $k = 1, \ldots, n,\ i, j = 3, \ldots, p$.

The tuning parameters are computed by five-fold cross-validation based in this simulation. We calculate the performance measure for model parameter estimation including deviance, $\text{ME}(\boldsymbol{\beta})$, and prediction error (PE) and test error.

As an overall bias for the regression coefficients estimation,

$$\text{ME}(\boldsymbol{\beta}) = \frac{1}{p} \sum_{i=1}^{p} \left(\beta_i - \hat{\beta}_i\right)^2.$$

The prediction error is calculated as

$$\mathrm{PE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 .$$

Test errors are computed after the coefficients are estimated with the training set generated from the same model.

Since the true model includes some zero's, the median and mean of the estimated number of zero's are reported as "med zero" and "mean zero" respectively. Sensitivity and specificity refer to the proportion of true non-zero coefficients and that of true zero coefficients which are detected by each method.

Table 1 gives the simulation results according to the different sparse pattern setups (1)–(6) of regression parameters with uncorrelated covariates in the linear models with $n = 10 < p = 100$. Ridge estimators do not give zero estimates while other lasso type estimators give some zeros when the true model contains some zeros. Table 2 is the result with the same setup as Table 1 except $n = 20$. As the sample size increases, test error and prediction error of the ridge method tend to increase while the lasso type method do not. Each method has a tendency to have more specificity except the ridge method with the bigger sample size. ME($\boldsymbol{\beta}$) of each method is reduced according to the sample size increase.

Therefore the sparse data need some estimators with the suitable penalty due to the sparsity pattern. According to test error, different estimators are winners in different underlying models. According to prediction error, fused lasso has the least values overall. Table 3 gives the simulation results according to regression parameter setups (i)–(vi) with correlated covariates in the linear models with $n = 10 < p = 20$. The results in Table 3 have a similar pattern to Table 1.

For the binary regression, Table 4 gives the simulation results in models (i)–(vi) with correlated covariates with $n = 20 < p = 100$. Lasso has the least test error in model (i) and (ii). Adaptive lasso has the least test error in model (v) and (vi). As expected the ridge method has is far less specificity in each model than others. Elastic net has the highest specificity in model (ii).

The optimal penalty might differ due to the sparsity patterns. We should select the appropriate penalty for the pattern of parameters.

3.2. Application to Arabidopsis microarray data

We applied the various lasso type estimators to select genes with Arabidopsis microarray gene expression data. Arabidopsis plants are especially in response to the small compound DFPM (Kim *et al.*, 2011). DFPM (5-(3,4-dichlorophenyl) furan-2-piperidine-1-ylmethanethione) was identified as an inhibitor of ABA signal transduction, which is a major plant hormone controlling resistant mechanism against abiotic stress condition. We aimed to get a clue for the function of DFPM in the induction of certain signal transduction in Arabidopsis by comparing the gene expressions of DFPM treated samples and control samples. Each treatment has only three repetitions.

Plants were grown for 12 days on agar plates (1X MS salts, 0.1% MES, 1% sucrose, pH 5.8) and treated with 30 μM DFPM in 6 well plates for 6 hours before RNA extraction. For microarray analyses, Affymetrix ATH1 oligonucleotide arrays were used according to the procedures by UCSD microarray facility (Kim *et al.*, 2011).

At the first step, 562 significant genes out of 22,810 are screened using False Discovery Rate (FDR) control in Benjamin and Hochberg (1995) multiple testing procedure with $q = 0.05$. In the

Table 1: Comparison of lasso type methods for variable selection with uncorrelated covariates and $n = 10$, $p = 100$ in linear models in 100 repetitions

(1)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	4.42	4.07	2.91	1.3	3.06
λ_2				0.83	2.00
deviance	25.714	50.036	23.568	13.298	25.175
med zero	1.00	93.00	92.00	90.00	93.00
mean zero	0.93	93.22	92.42	83.73	92.56
ME(β)	0.031	0.016	0.014	0.02	0.014
sensitivity	1.000	0.590	0.613	0.767	0.607
specificity	0.010	0.948	0.941	0.856	0.942
test error	29.711	15.043	12.937	13.414	12.982
PE	2.570	5.007	2.359	1.332	2.518
(2)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.89	7.22	3.65	0.98	3.68
λ_2				1.34	2.00
deviance	246.76	182.705	49.177	26.516	50.493
med zero	0.00	89.00	88.00	76.00	88.00
mean zero	0.28	89.24	88.21	67.02	88.32
ME(β)	0.518	0.543	0.556	0.325	0.558
sensitivity	1.000	0.364	0.384	0.850	0.380
specificity	0.003	0.906	0.896	0.698	0.897
test error	494.842	487.231	493.12	217.573	494.499
PE	24.676	18.276	4.920	2.654	5.052
(3)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.57	7.31	4.01	0.94	3.68
λ_2				1.37	2.00
deviance	162.706	184.639	54.512	36.961	49.309
med zero	0.00	90.00	89.00	75.00	89.00
mean zero	0.20	89.68	88.59	66.93	88.53
ME(β)	0.371	0.476	0.491	0.545	0.494
sensitivity	1.000	0.207	0.218	0.402	0.221
specificity	0.002	0.908	0.897	0.677	0.897
test error	320.003	446.468	458.134	463.521	459.190
PE	16.270	18.469	5.455	3.698	4.934
(4)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.46	7.15	3.40	0.86	3.09
λ_2				1.37	2.00
deviance	224.998	181.350	46.528	22.701	41.796
med zero	0.00	89.00	88.00	64.50	88.00
mean zero	0.19	89.13	88.24	59.15	88.15
ME(β)	0.56	0.662	0.682	0.595	0.686
sensitivity	0.998	0.269	0.272	0.766	0.276
specificity	0.002	0.905	0.896	0.623	0.895
test error	530.794	637.198	654.964	509.365	657.559
PE	22.504	18.142	4.657	2.271	4.184
(5)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.87	7.57	3.96	0.38	3.48
λ_2				1.75	2.00
deviance	175.176	192.756	53.652	13.175	48.677
med zero	0.00	89.00	88.00	48.50	88.00
mean zero	0.22	89.44	88.40	45.14	87.93
ME(β)	0.372	0.449	0.469	0.179	0.475
sensitivity	0.999	0.222	0.243	0.951	0.250
specificity	0.002	0.907	0.898	0.496	0.894
test error	333.096	395.158	418.842	107.993	420.201
PE	17.517	19.280	5.368	1.318	4.871
(6)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	3.32	6.09	3.39	1.53	3.23
λ_2				0.91	2.00
deviance	42.190	108.991	36.723	28.640	33.640
med zero	0.00	92.00	91.00	88.00	91.00
mean zero	0.64	92.24	90.78	79.23	90.59
ME(β)	0.109	0.093	0.092	0.103	0.091
sensitivity	1.000	0.533	0.557	0.657	0.573
specificity	0.007	0.936	0.922	0.806	0.921
test error	81.656	60.719	60.393	72.425	58.162
PE	4.218	10.903	3.675	2.866	3.367

Table 2: Comparison of lasso type methods for variable selection with uncorrelated covariates and $n = 20$, $p = 100$ in linear models in 100 repetitions

(1)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	3.87	2.2	2.09	1.13	1.95
λ_2				0.64	2.00
deviance	95.184	26.546	23.3	20.112	20.399
med zero	1.0	90.0	90.0	89.0	90.0
mean zero	0.9	90.2	89.9	87.8	89.5
ME(β)	0.029	0.003	0.003	0.003	0.003
sensitivity	1.000	0.717	0.723	0.907	0.727
specificity	0.010	0.922	0.919	0.902	0.915
test error	27.635	3.727	3.404	2.946	3.202
PE	4.760	1.329	1.166	1.007	1.021
(2)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.29	4.57	2.78	0.62	2.48
λ_2				0.57	2.00
deviance	984.652	205.812	74.954	12.020	68.970
med zero	0.00	83.00	83.00	77.00	82.00
mean zero	0.23	82.93	82.50	75.73	81.70
ME(β)	0.474	0.144	0.133	0.026	0.127
sensitivity	1.000	0.793	0.767	0.967	0.780
specificity	0.002	0.862	0.856	0.795	0.848
test error	415.980	118.467	110.560	22.316	107.482
PE	49.234	10.291	3.749	0.603	3.450
(3)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.24	6.90	3.54	1.29	3.89
λ_2				1.31	2.00
deviance	596.767	451.805	133.180	100.553	143.585
med zero	0.00	82.00	80.00	66.00	80.00
mean zero	0.23	81.30	79.50	58.70	79.70
ME(β)	0.339	0.368	0.373	0.412	0.370
sensitivity	1.000	0.520	0.543	0.627	0.520
specificity	0.003	0.850	0.833	0.611	0.832
test error	298.326	334.459	337.095	343.937	336.327
PE	29.838	22.592	6.661	5.029	7.183
(4)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.68	5.18	2.19	0.67	2.97
λ_2				1.02	2.00
deviance	832.628	323.222	78.226	40.259	101.237
med zero	0.0	80.0	78.0	68.0	79.0
mean zero	0.2	80.4	78.5	63.0	79.6
ME(β)	0.500	0.279	0.271	0.101	0.270
sensitivity	1.000	0.658	0.675	0.975	0.679
specificity	0.002	0.844	0.825	0.683	0.837
test error	523.728	293.112	283.623	86.679	286.058
PE	41.622	16.167	3.913	2.014	5.065
(5)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.67	7.43	3.18	0.52	3.62
λ_2				1.63	2.00
deviance	413.010	465.129	113.23	41.984	139.051
med zero	0.0	82.0	79.0	59.5	79.0
mean zero	0.27	81.87	79.1	56.4	79.5
ME(β)	0.334	0.351	0.357	0.026	0.357
sensitivity	1.000	0.453	0.483	1.000	0.477
specificity	0.003	0.849	0.821	0.627	0.825
test error	291.930	302.124	302.808	15.235	305.377
PE	20.649	23.258	5.661	2.101	6.950
(6)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	3.34	3.30	1.71	1.58	2.54
λ_2				0.14	2.00
deviance	219.867	75.860	19.105	22.320	43.979
med zero	1.0	90.5	87.0	84.5	88.5
mean zero	0.8	90.1	87.3	85.9	88.6
ME(β)	0.102	0.026	0.02	0.019	0.023
sensitivity	1.000	0.900	0.967	0.933	0.900
specificity	0.008	0.926	0.899	0.884	0.910
test error	70.262	9.015	6.803	6.689	7.244
PE	10.995	3.796	0.956	1.118	2.201

Table 3: Comparison of lasso type methods for variable selection with correlated covariates and $n = 10$, $p = 100$ in linear models in 100 repetitions

(i)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.19	1.46	1.51	1.31	1.54
λ_2				0.29	2.00
deviance	36.999	8.369	8.861	9.475	8.988
med zero	0.0	15.0	14.0	14.0	15.0
mean zero	0.0	14.3	14.4	13.6	14.6
ME(β)	0.114	0.012	0.014	0.020	0.013
sensitivity	0.997	0.803	0.797	0.850	0.793
specificity	0.001	0.811	0.814	0.774	0.822
test error	26.543	3.437	3.675	3.556	3.518
PE	3.700	0.837	0.886	0.948	0.899
(ii)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	0.50	1.84	1.36	0.81	1.34
λ_2				0.46	2.00
deviance	287.844	37.294	18.503	13.286	19.803
med zero	0.0	11.0	11.0	9.0	11.0
mean zero	0.0	10.8	10.6	8.1	10.8
ME(β)	1.569	0.400	0.328	0.226	0.344
sensitivity	1	0.880	0.904	0.954	0.878
specificity	0.002	0.684	0.676	0.529	0.681
test error	375.841	79.694	56.328	32.665	69.407
PE	28.784	3.730	1.851	1.329	1.980
(iii)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	0.72	3.69	1.77	0.69	2.19
λ_2				1.82	2.00
deviance	223.240	157.605	48.062	81.581	59.559
med zero	0.0	10.0	10.0	0.0	10.0
mean zero	0.0	10.2	9.7	1.7	9.8
ME(β)	1.271	2.189	2.278	1.786	2.19
sensitivity	0.999	0.588	0.621	0.927	0.615
specificity	0.000	0.608	0.599	0.103	0.595
test error	223.627	418.678	403.967	247.934	398.499
PE	22.324	15.760	4.806	8.158	5.955
(iv)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	0.49	3.79	2.04	0.75	2.09
				0.86	2.00
deviance	282.881	148.493	52.774	29.714	54.591
med zero	0.0	10.0	10.0	5.5	10.0
mean zero	0.0	10.6	9.9	5.2	10.0
ME(β)	1.752	1.666	1.575	1.092	1.611
sensitivity	1.000	0.742	0.775	0.938	0.764
specificity	0.000	0.718	0.675	0.393	0.683
test error	417.238	363.126	330.233	195.178	340.085
PE	28.292	14.851	5.277	2.972	5.460
(v)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	0.20	3.43	2.07	0.54	2.04
λ_2				1.6	2.0
deviance	92.529	132.696	49.275	21.548	47.346
med zero	0.0	10.0	9.0	1.0	9.0
mean zero	0.0	9.7	9.6	2.1	9.4
ME(β)	1.040	2.314	2.398	0.500	2.337
sensitivity	0.999	0.670	0.678	0.998	0.685
specificity	0	0.729	0.720	0.268	0.710
test error	266.150	534.422	540.327	49.889	517.181
PE	9.253	13.272	4.928	2.155	4.736
(vi)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.36	2.44	1.75	1.10	1.85
another				0.44	2.00
deviance	76.626	49.212	26.169	21.524	26.718
med zero	0.0	13.0	12.0	11.0	13.0
mean zero	0.0	12.8	12.5	10.0	12.5
ME(β)	0.408	0.201	0.184	0.230	0.180
sensitivity	1	0.823	0.867	0.930	0.853
specificity	0.001	0.722	0.714	0.578	0.714
test error	65.557	27.946	24.011	32.828	22.750
PE	7.664	4.921	2.617	2.153	2.672

Table 4: Comparison of variable selection with various penalties with $n = 20$, $p = 100$ in binary regression

(1)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	3.53	1.98	1.31	0.62	2.14
λ_2				0.51	2.00
deviance	15.532	15.394	15.722	12.918	15.076
med zero	5.0	96.0	96.0	91.0	96.0
mean zero	5.5	95.2	95.5	84.8	94.9
ME(β)	0.032	0.031	0.03	0.026	0.031
sensitivity	0.990	0.290	0.293	0.510	0.297
specificity	0.056	0.959	0.963	0.860	0.957
test error	0.517	0.402	0.418	0.504	0.492
PE	0.133	0.291	0.285	0.114	0.136
(2)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.75	0.49	0.47	0.16	0.47
λ_2				0.24	2.00
deviance	9.609	7.843	8.332	5.684	8.536
med zero	2.0	91.0	91.0	84.5	92.0
mean zero	3.2	91.1	91.4	74.7	91.6
ME(β)	0.537	0.498	0.501	0.480	0.499
sensitivity	0.992	0.490	0.490	0.716	0.474
specificity	0.034	0.932	0.936	0.772	0.937
test error	0.511	0.294	0.308	0.495	0.510
PE	0.069	0.356	0.346	0.048	0.064
(3)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.28	1.73	1.51	0.41	2.05
λ_2				1.57	2.00
deviance	11.772	18.493	19.122	14.289	19.188
med zero	3.5	97.0	97.0	71.5	97.5
mean zero	4.0	95.3	95.8	56.3	95.7
ME(β)	0.393	0.394	0.395	0.390	0.395
sensitivity	0.980	0.124	0.117	0.490	0.122
specificity	0.043	0.962	0.967	0.569	0.966
test error	0.494	0.471	0.471	0.518	0.500
PE	0.087	0.277	0.274	0.122	0.168
(4)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.28	1.58	1.45	0.20	0.52
λ_2				0.29	2.00
deviance	11.705	14.168	13.084	7.970	11.455
med zero	4.0	97.0	93.5	77.5	92.5
mean zero	4.2	94.8	93.6	63.5	92.5
ME(β)	0.589	0.574	0.569	0.556	0.569
sensitivity	1.000	0.175	0.188	0.725	0.288
specificity	0.046	0.959	0.947	0.666	0.943
test error	0.470	0.375	0.375	0.460	0.440
PE	0.082	0.325	0.335	0.064	0.090
(5)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	1.43	1.00	0.70	0.08	0.98
λ_2				0.35	2.00
deviance	8.609	13.427	11.573	5.009	13.216
med zero	2.0	93.0	91.0	70.0	92.5
mean zero	2.8	93.1	91.8	58.3	92.7
ME(β)	0.388	0.385	0.382	0.310	0.383
sensitivity	0.999	0.303	0.343	0.900	0.313
specificity	0.032	0.957	0.947	0.637	0.954
test error	0.497	0.406	0.398	0.507	0.491
PE	0.060	0.302	0.311	0.041	0.109
(6)	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	2.47	1.68	0.99	0.55	1.29
λ_2				0.60	2.00
deviance	12.254	13.567	12.193	11.249	12.520
med zero	3.0	94.0	93.0	88.0	93.5
mean zero	4.5	93.6	92.7	80.1	93.2
ME(β)	0.117	0.109	0.111	0.112	0.110
sensitivity	0.993	0.460	0.520	0.590	0.483
specificity	0.047	0.948	0.942	0.813	0.945
test error	0.491	0.404	0.382	0.495	0.502
PE	0.093	0.311	0.316	0.088	0.102

Table 5: Comparison of parameter estimates by lasso type methods to Arabidopsis microarray data with $n = 6$, $p = 562$ in binary regression

	ridge	lasso	adaptive lasso	fused lasso	elastic net
λ	5.000	0.12	0.17	0.10	0.10
λ_2				0.94	2.00
number of zeros	6	544	546	534	543
number of non-zeros	556	18	16	28	19
deviance	0.535	0.627	0.884	1.378	0.500

second step, with these 562 genes, the lasso type estimations are done to select more strongly involved genes that classify the control group versus DFPM group with the binary response variable.

For λ values, 200 grids are generated from 0.1 to 5 and calculated for each grid. Table 5 shows the result including parameter estimates and the number of non-zero parameters chosen by each method. It also provides that lasso selects 18 genes and adaptive lasso selects 16 genes. The fused lasso and elastic net also select 28 and 19 genes respectively. Based on elastic net estimation, 19 genes are decided to discriminate the groups. These selected genes are important to distinguish the groups, control versus treatment.

4. Conclusion

Modern statistics deals with large and complex data sets, and consequently with models that contain a large number of parameters. With this huge number of parameters, the lasso type estimators reduce the number of non-zero parameters to make a compact model. We explore and compare various lasso type estimators. These estimators are useful for the high dimensional data inference. The encouraging results in this research suggest that absolute value constraints might prove to be useful in a wide variety of statistical estimation problems. Further study is expected to investigate the penalties for appropriate estimators according to sparsity pattern and data environment.

References

Benjamin, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of Royal Statistical Society B*, **57**, 289–300.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*, Springer, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics*, **32**, 407–451.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London.

Friedman, J., Hasti, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Springer, New York.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems, *Technometrics*, **12**, 69–82.

Kim, T. H., Hauser, F., Ha, T., Xue, S., Boehmer, M., Nishimura, N., Munemasa, S., Hubbard, K., Peine, N., Lee, B. H., Lee, S., Robert, N., Parker, J. E. and Schroeder, J. I. (2011). Chemical genetics reveals negative regulation of abscisic acid signaling by a plant immune response pathway, *Current Biology*, **21**, 990–997.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos, *Bayesian Analysis*, **5**, 369–412.

Park, T. and Casella, G. (2008). The Bayesian lasso, *Journal of the American Statistical Association*, **103**, 681–686.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *The Annals of Statistics*, **9**, 1135-1151.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of Royal Statiatical Society B*, **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of Royal Statistical Society B*, **67**, 91–108.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

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

Received June 6, 2014; Revised July 14, 2014; Accepted July 14, 2014