Journal of the Korean Data & Information Science Society
2013, **24**(5), 1077–1088

http://dx.doi.org/10.7465/jkdi.2013.24.5.1077
한국데이터정보과학회지

A small review and further studies on the LASSO

Sunghoon Kwon[1] · Sangmi Han[2] · Sangin Lee[3]

[1]Department of Applied Statistics, Konkuk University
[23]Department of Statistics, Seoul National University

Received 4 July 2013, revised 12 August 2013, accepted 23 August 2013

Abstract

High-dimensional data analysis arises from almost all scientific areas, evolving with development of computing skills, and has encouraged penalized estimations that play important roles in statistical learning. For the past years, various penalized estimations have been developed, and the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) has shown outstanding ability, earning the first place on the development of penalized estimation. In this paper, we first introduce a number of recent advances in high-dimensional data analysis using the LASSO. The topics include various statistical problems such as variable selection and grouped or structured variable selection under sparse high-dimensional linear regression models. Several unsupervised learning methods including inverse covariance matrix estimation are presented. In addition, we address further studies on new applications which may establish a guideline on how to use the LASSO for statistical challenges of high-dimensional data analysis.

Keywords: High dimension, LASSO, penalized estimation, review.

1. Introduction

Consider a risk function $R(\boldsymbol{\theta}) = EL(\mathbf{z}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a p-dimensional parameter vector of interest and $L(\mathbf{z}, \boldsymbol{\theta})$ is a loss function of a random vector $\mathbf{z}$ and parameter $\boldsymbol{\theta}$. Given n independent copies, $\mathbf{z}_i, i \le n$, of $\mathbf{z}$, penalized estimation minimizes the penalized empirical risk function,

$$Q^\lambda(\boldsymbol{\theta}) = \textstyle\sum_{i=1}^n L(\mathbf{z}_i, \boldsymbol{\theta})/n + \sum_{j=1}^p J^\lambda(|\theta_j|),$$

for a penalty function J^λ that is often indexed by a tuning parameter (vector) $\lambda > 0$. The penalized estimator $\hat{\boldsymbol{\theta}}^\lambda$ is the global minimizer of Q^λ, that is,

$$\hat{\boldsymbol{\theta}}^\lambda = \arg\min_{\boldsymbol{\theta}} Q^\lambda(\boldsymbol{\theta}).$$

[1] Corresponding author: Assistant professor, Department of Applied Statistics, Konkuk University, Seoul 143-701, Korea. E-mail: shkwon0522@konkuk.ac.kr
[2] Ph.D. candidate, Department of Statistics, Seoul National University, Seoul 151-747, Korea.
[3] Post-doc., Data Science for Knowledge Creation Research Center, Seoul National University, Seoul 151-747, Korea.

If the penalty function is convex then $\hat{\boldsymbol{\theta}}^{\lambda}$ is unique. Examples are the least absolute shrinkage and selection operator (LASSO) penalty $J^{\lambda}(t) = \lambda|t|$ (Tibshirani, 1996) and the ridge penalty $J^{\lambda}(t) = \lambda t^2$. However, if the penalty function is concave then there may be many local minimizers and the global minimizer itself is not unique. Examples are the smoothly clipped absolute deviation (SCAD) penalty $dJ^{\lambda}(|t|)/d|t| = \min\{\lambda, (a\lambda - |t|)_+/(a+1)\}, a > 2$ (Fan and Li, 2001), the bridge penalty $J^{\lambda}(|t|) = \lambda|t|^{\gamma}, 0 < \gamma \leq 1$ (Huang *et al.*, 2008a) and the minimax concave (MC) penalty $dJ^{\lambda}(|t|)/d|t| = (\lambda - |t|/a)_+, a > 1$ (Zhang, 2010). Here, $x_+ = xI(x \geq 0)$. The penalty functions above have their own characteristics in parameter estimation and model selection, and often we need to choose an appropriate penalty function carefully, that can carry out the goal of data analysis well.

In general, penalized estimation is known to give high prediction accuracy from the shrinkage effect on the non-zero elements of the estimator, and increases interpretability since the fitted model is sparse enough (Tibshirani, 1996). However, the most important property of the penalized estimation is that we can apply the penalized estimation even when the model is high-dimensional, where the dimension of the parameter of interest is much larger than the sample size, that is, $p > n$. For example, detecting important genes in gene expression data often requires high-dimensional models since they include thousands of genes as independent variables (Alon *et al.*, 1999; Dudoit *et al.*, 2002; Lee and Lee, 2012). Most classical estimations unfortunately have limitations for high-dimensional situations since the estimators are not identifiable in general and sometimes cannot be constructed numerically.

In this paper, we first give a small review of existing studies on the penalized estimation. For convenience, we will focus on the LASSO since, among penalties, the LASSO has shown outstanding ability, earning the first place on the development of penalized estimation. Although the LASSO suffers from a certain theoretical disadvantage (Zou, 2006; Zhao and Yu, 2006; Leng *et al.*, 2006; Meinshausen and Yu, 2009), it is relatively easy to implement the penalized estimator numerically, since the optimization problem is convex (Efron *et al.*, 2004; Friedman *et al.*, 2007; Rosset and Zhu, 2007; Park and Hastie, 2007). Second, we give new challenging topics on high-dimensional data analysis. Given current research direction of penalized estimation, various areas still remain to be applied and studied with the LASSO. For example, unsupervised grouping method is one of interesting problems but requires new idea of application.

The rest of the paper is organized as follows. Section 2 presents a survey of existing results on the LASSO and Section 3 follows introducing some new topics on the use of the LASSO. Some concluding remarks are in Section 4.

2. A review on the LASSO and its applications

2.1. Variable selection

Consider a sparse linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is a response vector, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a sparse true regression coefficient vector, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is an $n \times p$ design matrix and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ is a random error vector. Here, the sparsity of the model implies that there exists a non-empty subset $\mathcal{N}^* = \{j : \beta_j^* = 0\} \neq \emptyset$. Hence the model includes one or more noisy predictive variables which requires identifying the correct index set $\mathcal{A}^* = \{j : \beta^*_j \neq 0\}$. In this case, we can use the LASSO penalty $J^\lambda(|t|) = \lambda|t|$ for estimating $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \lambda\|\boldsymbol{\beta}\|_1 \right\},$$

for some $\lambda > 0$. Since the LASSO penalty is non-differentiable at the origin, the estimator $\hat{\boldsymbol{\beta}}^\lambda$ of $\boldsymbol{\beta}^*$ must have sparsity (Fan and Li, 2001; Donoho and Johnstone, 1994) yielding a nonempty subset $\hat{\mathcal{A}}^\lambda = \{j : \hat{\beta}^\lambda_j \neq 0\}$ indexed by λ. Hence we can use $\hat{\mathcal{A}}^\lambda$ as an estimator of $\mathcal{A}^* = \{j : \beta^*_j \neq 0\}$, which implies we can do parameter estimation and variable (model) selection simultaneously (Tibshirani, 1996). This, in fact, shows a nice statistical implementation since we may not use known model comparison measures such as Akaike or Beysian information criteria (Akaike, 1973; Schwarz, 1978), together with stepwise variable selection methods such as forward selection and backward elimination, which shows far more unstable results on model selection (Breiman, 1996).

Statistical properties of the LASSO have been studied by many authors. Zhao and Yu (2006) and Meinshausen and Yu (2009) proved that the LASSO requires a certain condition to select the correct set of nonzero true coefficients, provided that these coefficients are bounded away from zero at a certain rate. Hence, the LASSO does not have selection consistency in general which agrees with the results of Leng *et al.* (2006) and Zou (2006). Zhang and Huang (2008) proved that the LASSO selects a model whose size, the number of predictive variables in the model, is $O(q)$ at most, where $q = \|\boldsymbol{\beta}^*\|_0$ is the number of nonzero true coefficients, and includes all coefficients of greater order than the bias of the selected model, achieving $(\log p/n)^{1/2}$-consistency under a sparse Riesz and weak sparsity conditions.

In general, the LASSO selects more predictive variables than the number of true variables due to the shrinkage effect but achieves an optimality in minimax sense, producing high prediction accuracy. See, Raskutti *et al.* (2011), Bickel *et al.* (2009) and Zhang (2009) for some sharp minimax rates of the LASSO in high-dimensional models. These results hold similarly through other methods below, and hence we skip theoretical properties of the LASSO when we introduce methodologies. We refer to Zhang and Zhang (2012) for a well organized review of penalized estimation including the LASSO for variable selection in high-dimensional linear regression models.

2.2. Adaptive variable selection

One main deficiency of the LASSO is to conflict between correct variable selection and optimal prediction (Leng *et al.*, 2006) since the order of tuning parameter λ varies in each purpose. To overcome this problem, Zou (2006) proposed the adaptive LASSO:

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \lambda {\textstyle\sum_{j=1}^p} w_j|\beta_j| \right\},$$

for some $\lambda > 0$, where $\mathbf{w} = (w_1, \ldots, w_p)^T$ is a weight vector obtained through samples. By using a data-dependent weight vector $\mathbf{w}$, the adaptive LASSO can achieve the oracle property (Fan and Li, 2001; Fan and Peng, 2004; Zou and Zhang, 2009); asymptotic equivalence between a penalized estimator and the oracle least square estimator obtained by true predictive variables only.

The key issue in using the adaptive LASSO is how to obtain the weight vector $\mathbf{w}$. For example, Zou (2006) used the inverse of absolute value of the ordinary least square estimator to construct $\mathbf{w}$ when $p < n$. Huang *et al.* (2008b) proved that marginal linear regression can be used to obtain a weight vector under partial orthogonality conditions even when $p > n$, and Zhou *et al.* (2009) suggested to use the two-stage adaptive Lasso for consistent model selection in linear and Gaussian graphical models under the restricted eigenvalue conditions (Bickel *et al.*, 2009).

Although there are many non-convex penalties such as the SCAD and MC that have statistical advantages in selection consistency, they suffer from bad local minimizers and it is much hard to identify the theoretical optimal penalized estimators among them (Kim and Kwon, 2012; Zhang and Zhang, 2012; Zhang, 2010). Hence the adaptive LASSO is a useful practical alternative in variable selection since the problem is convex that is easy to solve and the solution is unique having the oracle property.

2.3. Grouped variable selection

Assume that there is a known group structure among predictive variables so that it might be possible to derive an advantage from the group information by selecting groups, and sometimes, selecting both groups and variables simultaneously. In this case, the sparse linear regression model in (2.1) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \textstyle\sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k^* + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\mathbf{X}_k$ is $n \times p_k$ design matrix of the kth group, $\boldsymbol{\beta}_k^* = (\beta_{k1}^*, \ldots, \beta_{kp_k}^*)^T$ is a corresponding true regression coefficient vector and K is the number of groups in the model. The model includes an example where a linear regression model is extended to include groups of dummy variables of categorical predictive variables. In this case, selecting groups of dummy variables is interesting and often gives higher prediction accuracy than selecting variables without using the group structure. Further, if possible, simultaneous selecting both groups of dummy variables and dummy variables in selected groups is also challenging.

For the problem, Yuan and Lin (2006) proposed the group LASSO:

$$\hat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \textstyle\sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k\|_2^2/2n + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_2 \right\}, \tag{2.3}$$

for some $\lambda = (\lambda_1, \ldots, \lambda_K)^T, \lambda_k > 0, k \leq K$, which is equivalent to the LASSO when each group consists of just one predictive variable. From the L_2-norm inside the penalty, the group LASSO selects groups and all the variables in selected groups are included in the fitted model, which implies the group LASSO does group selection and parameter estimation simultaneously.

In fact, the criterion in (2.3) can be a special case of the inner-outer composite criterion introduced by Huang *et al.* (2012):

$$\hat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \textstyle\sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k\|_2^2/2n + \sum_{k=1}^{K} J_{\lambda_k}^{O} \big(\sum_{j=1}^{p_k} J_{\gamma}^{I}(|\beta_{kj}|) \big) \right\},$$

for some $\lambda = (\lambda_1, \ldots, \lambda_K, \gamma)^T$, $\lambda_k > 0, k \leq K$, $\gamma > 0$, where $J_{\lambda_k}^{O}$ is the kth outer penalty for group selection and J_{γ}^{I} is the inner penalty for variable selection. The group LASSO uses the bridge penalty (Huang *et al.*, 2008a), $J_{\lambda_k}^{O}(|t|) = \lambda p_k^{1-\nu}|t|^{\nu}$ with $\nu = 1/2$, as the outer penalty, and the ridge penalty $J^I_\gamma(t) = \gamma|t|^2$ as the inner penalty, where the inner tuning parameter is fixed with $\gamma = 1$.

As we mentioned above, the group LASSO cannot select variables within the selected groups while selecting groups. To improve the group LASSO, Friedman *et al.* (2010) proposed the sparse group LASSO:

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \textstyle\sum_{k=1}^K \mathbf{X}_k\boldsymbol{\beta}_k\|_2^2 + \sum_{k=1}^K J^G_{\lambda_k}(\|\boldsymbol{\beta}_k\|_2) + \sum_{k=1}^K \sum_{j=1}^{p_k} J^V_\gamma(|\beta_{kj}|) \right\},$$

for some $\lambda = (\lambda_1, \ldots, \lambda_K, \gamma)^T$, $\lambda_k > 0, k \leq K$, $\gamma > 0$, where $J^G_{\lambda_k}$ is penalty for group selection and J^V_γ for variable selection. By adding the LASSO penalty, $J^V_\gamma(t) = \gamma|t|$, to the group LASSO penalty, the sparse group LASSO gives a clearer way of controlling variable and group selection, although there are two different tuning parameters that cause higher computational cost. We refer to Huang *et al.* (2012) for a nice review of penalized estimations for group and variable selection in high-dimensional linear regression models.

2.4. Structured variable selection

The group LASSO controls L_1-norms of coefficient vectors when there is a known group structure as (2.2). However it is often more challenging to find an unknown group structure itself in the model (2.1). For example, Tibshirani *et al.* (2005) proposed the fused LASSO:

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \textstyle\sum_{j=2}^p |\beta_j - \beta_{j-1}| \right\}, \tag{2.4}$$

for some $\lambda = (\lambda_1, \lambda_2)^T, \lambda_1 > 0, \lambda_2 > 0$. The special form of the penalty produces a successive group structure of estimators, achieving variable selection. Since the LASSO penalty imposed on both parameters and their successive differences, the fused LASSO constructs a group structure keeping the order of variables. Note that the idea of the fused LASSO can be generalized by considering a network structure $\mathcal{E}$, where $(s,t) \in \mathcal{E}$ denotes possible connectivity between variables $\mathbf{X}_s$ and $\mathbf{X}_t$. In this case, the penalized estimator can be obtained with the second penalty $\lambda_2 \sum_{(s,t)\in\mathcal{E}} |\beta_s - \beta_t|$, and we can construct a sparse (sub)network structure $\mathcal{E}' \subset \mathcal{E}$ since connectivity is considered over the network structure $\mathcal{E}$ only. Note that in this example, we consider variables as a group if they have the same coefficients in the fitted model.

The fused LASSO was generalized by Tibshirani and Taylor (2011):

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \textstyle\sum_{k=1}^K \lambda_k \|\mathbf{D}^{(k)}\boldsymbol{\beta}\|_1 \right\},$$

for some $\lambda = (\lambda_1, \ldots, \lambda_K)^T$, $\lambda_k > 0, k \leq K$, where $\mathbf{D}^{(k)}$ is a $m_k \times p$ specified structure matrix for the kth penalty that depends on applications or geometric behaviors of $\boldsymbol{\beta}$, and K is the number of structures we want to specify. For example, the fused LASSO corresponds to a choice of $K = 2$, where $\mathbf{D}^{(1)} = I$ with $m = p$ and $\mathbf{D}^{(2)}$ whose elements satisfy $D^{(2)}_{jj} = -1$, $D^{(2)}_{j(j+1)} = 1$ and $D^{(2)}_{jk} = 0, k \neq j, j+1$ with $m = p - 1$. Another example is the linear and polynomial trend filtering when $K = 1$ and $\mathbf{D}^{(1)}$ satisfies $D^{(1)}_{jj} = -1$, $D^{(1)}_{j(j+1)} = 2$, $D^{(1)}_{j(j+2)} = -1$ and $D^{(1)}_{jk} = 0, k \neq j, j+1, j+2$ with $m = p - 2$. In this case, the penalty becomes $\lambda\|\mathbf{D}^{(1)}\boldsymbol{\beta}\|_1 = \lambda \sum_{j=2}^{p-1} |2\beta_j - (\beta_{j-1} - \beta_{j+1})|$.

In linear regression models, sometimes, we need to fix a certain heredity structure. For example, an interaction term can be included in the model only if the corresponding main terms are also included in the model, which often referred as strong heredity. For this issue, Choi *et al.* (2010b) proposed to use the strong heredity interaction model:

$$(\hat{\boldsymbol{\beta}}^{\lambda}, \hat{\boldsymbol{\eta}}^{\lambda}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\eta}\|_2^2/2n + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\eta}\|_1 \big\},$$

for some $\lambda = (\lambda_1, \lambda_2)^T, \lambda_1 > 0, \lambda_2 > 0$, where $\mathbf{V} = (\mathbf{V}_{12}, \mathbf{V}_{13}, \ldots, \mathbf{V}_{(p-1)p})$ and $\boldsymbol{\eta} = (\eta_{12}, \ldots, \eta_{(p-1)p})^T$ are the design matrix and parameter vector for interactions, respectively. In the model, they reparameterize the parameter vector β as $\eta_{jk} = \delta_{jk}\beta_j\beta_k, j < k$, so that the estimator keeps the strong heredity constraint automatically. That is, whenever $\hat{\beta}_j^{\lambda} = 0$ or $\hat{\beta}_k^{\lambda} = 0$, $\hat{\eta}_{jk}^{\lambda} = 0$, and vice versa, $\hat{\eta}_{jk}^{\lambda} \neq 0$ if both $\hat{\beta}_j^{\lambda} \neq 0$ and $\hat{\beta}_k^{\lambda} \neq 0$.

2.5. Inverse covariance and partial correlation matrix estimation

Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p) = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$, where $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are n independent copies of $\mathbf{z} = (z_1, \ldots, z_p)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ for some $p \times p$ covariance matrix $\boldsymbol{\Sigma} \succ 0$. Here, $\boldsymbol{\Sigma} \succ 0$ denotes $\boldsymbol{\Sigma}$ is positive definite. When $p > n$, the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \mathbf{Z}^T\mathbf{Z}/n$ does not have an inverse matrix. Hence, if we are interested in estimating the inverse covariance matrix (concentration or precision matrix), $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, we need to find another estimator instead of inverting $\hat{\boldsymbol{\Sigma}} = \mathbf{Z}^T\mathbf{Z}/n$. For the problem, the penalized estimator $\hat{\boldsymbol{\Omega}}^{\lambda}$ has been studied by many authors (Yuan, 2008; Yuan, 2010; Banerjee *et al.*, 2008; Friedman *et al.*, 2008):

$$\hat{\boldsymbol{\Omega}}^{\lambda} = \arg\min_{\boldsymbol{\Omega}\succ 0} \big(\log|\boldsymbol{\Omega}| + \mathrm{trace}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}) + \lambda \textstyle\sum_{j\neq k} |\Omega_{jk}| \big),$$

for some $\lambda > 0$, where $|\boldsymbol{\Omega}|$ is determinant of $\boldsymbol{\Omega}$. The estimator $\hat{\boldsymbol{\Omega}}^{\lambda}$ is positive definite and execute model selection and estimation simultaneously as the LASSO. Banerjee *et al.* (2008) shows that the problem is equivalent to

$$\max_{\boldsymbol{\Omega}\succ 0} \min_{\|\mathbf{U}\|_\infty \leq \lambda} \big(\log|\boldsymbol{\Omega}| - \mathrm{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \mathbf{U}\boldsymbol{\Omega}) \big),$$

for some $\lambda > 0$, where $\|\mathbf{U}\|_\infty = \max_{jk} |U_{jk}|$. This equivalence leads to an efficient optimization algorithm called the graphical LASSO developed by Friedman *et al.* (2008).

Under normality assumption, nonzero elements of the inverse covariance matrix $\boldsymbol{\Omega}$ imply conditional dependency between corresponding variable pairs conditional on the other variables (Edward, 2000). Let $\boldsymbol{\Theta}$ be the partial correlation matrix whose (j,k) entry Θ_{jk} is the partial correlation between x_j and $x_k, j \neq k$, given the others, that is, $\Theta_{jk} = \mathrm{Corr}(z_j, z_k | z_l, l \neq j, k)$. It is well known that $\Theta_{jk} = -\Omega_{jk}/\sqrt{\Omega_{jj}\Omega_{kk}}$ and the linear regression model $z_k = \sum_{j\neq k} \zeta_{kj} z_j + \epsilon_k$ satisfies $\zeta_{kj} = -\Omega_{kj}/\Omega_{kk} = \Theta_{kj}\sqrt{\Omega_{jj}/\Omega_{kk}}$. Hence, we can identify nonzero elements of partial correlation matrix by estimating $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^T, \ldots, \boldsymbol{\zeta}_p^T)^T$, where $\boldsymbol{\zeta}_k = (\zeta_{k1}, \ldots, \zeta_{kp})^T$ and $\zeta_{kk} = -1$, using the penalized estimator (Meinshausen and Bülmann, 2006),

$$\hat{\boldsymbol{\zeta}}^{\lambda} = \arg\min_{\boldsymbol{\zeta}}, \textstyle\sum_{k=1}^{p} \big(\|\mathbf{Z}_k - \sum_{j\neq k} \zeta_{kj}\mathbf{Z}_j\|_2^2 + \lambda \sum_{j\neq k} |\zeta_{kj}| \big), \tag{2.5}$$

for some $\lambda > 0$, subject to $\zeta_{kk} = -1, k \leq p$.

The estimator $\hat{\boldsymbol{\zeta}}^{\lambda}$ enables us to construct a kind of network structure among variables, and then we can find the estimator $\hat{\boldsymbol{\Theta}}^{\lambda}$ of $\boldsymbol{\Theta}$ by using the estimated network structure. However, it is of interest also to directly estimate partial correlation itself as in Peng *et al.* (2009):

$$(\hat{\boldsymbol{\zeta}}^{\lambda}, \hat{\boldsymbol{\omega}}^{\lambda}) = \arg\min_{\boldsymbol{\zeta},\mathbf{w}} \sum_{k=1}^{p} \big(\|\mathbf{Z}_k - \sum_{j\neq k} \zeta_{kj} \sqrt{\omega_{jj}/\omega_{kk}} \mathbf{Z}_j\|_2^2 + \lambda \sum_{j\neq k} |\zeta_{kj}|\big),$$

for some $\lambda > 0$, subject to $\zeta_{kk} = -1, k \leq p$. We refer to Pourahmadi (2011) for a survey of the progress made in modeling covariance matrices from the perspectives of generalized linear models for high-dimensional data.

2.6. Sparse principal component and factor analysis

The penalized estimation can be applied to principal component analysis (PCA) and factor analysis (FA). In real applications, the sparse loadings greatly help us to interpret the results from the PCA and FA especially when the dimension is much larger than the sample size.

For example, Zou *et al.* (2006) introduced the sparse PCA, using a regression type reformulation of the optimization problem of PCA. Given k, they modified the first k principal components to have sparse loadings by applying the LASSO penalty in the estimation. Let $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k)$ and $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k)$ be $n \times p$ matrices. The sparse loadings for the first k principal components can be obtained by optimizing the following problem:

$$(\hat{\boldsymbol{\Phi}}^{\lambda}, \hat{\boldsymbol{\Psi}}^{\lambda}) = \arg\min_{\boldsymbol{\Phi},\boldsymbol{\Psi}} \ \sum_{i=1}^{n} \|\mathbf{z}_i - \boldsymbol{\Phi}\boldsymbol{\Psi}^T \mathbf{z}_i\|_2^2 + \lambda_1 \sum_{j=1}^{k} \|\boldsymbol{\psi}_j\|_2^2 + \lambda_2 \sum_{j=1}^{k} \|\boldsymbol{\psi}_j\|_1,$$

for some $\lambda = (\lambda_1, \lambda_2)^T, \lambda_1 > 0, \lambda_2 > 0$, subject to $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \mathbf{I}_{k\times k}$. Given $\boldsymbol{\Phi}$, the problem becomes k independent elastic net problems (Zou and Hastie, 2005) of finding $\hat{\boldsymbol{\psi}}_j^{\lambda}, j \leq k$, whose elements are allowed to be exactly zero, and given $\boldsymbol{\Psi}$, we can find exact solution $\hat{\boldsymbol{\Phi}}^{\lambda}$ by the singular value decomposition. See next section for another example of sparse PCA, which is more intuitive to understand.

Another example for FA was proposed by Choi *et al.* (2010a). Let $\mathbf{Z} = \mathbf{L}\mathbf{F} + \mathbf{E}$, where $\mathbf{L}$ is the $n \times q$ unobserved factor matrix, $\mathbf{F}$ is the $q \times p$ factor loading matrix and $\mathbf{E}$ is the $n \times p$ random error matrix with mean zero and diagonal covariance matrix $\boldsymbol{\Sigma}$. To obtain sparse factor loadings, they maximizes penalized maximum likelihood under the normality assumption:

$$(\hat{\boldsymbol{\Sigma}}^{\lambda}, \hat{\mathbf{F}}^{\lambda}) = \arg\max_{\boldsymbol{\Sigma},\mathbf{F}} \big(\log|(\boldsymbol{\Sigma}^2 + \mathbf{F}^T\mathbf{F})| + \mathrm{trace}(\mathbf{Z}^T\mathbf{Z}/n)(\boldsymbol{\Sigma}^2 + \mathbf{F}^T\mathbf{F})^{-1}) + \lambda\|\mathbf{F}\|_1\big),$$

for some $\lambda > 0$, where $\|\mathbf{F}\|_1 = \sum_{ij} |F_{ij}|$. They also developed the generalized expectation-maximization algorithm for obtaining the penalized estimator, which solves the LASSO penalized least squares iteratively. We refer to Zou *et al.* (2006) and Choi *et al.* (2010a) for more details of sparse PCA and FA, respectively.

3. New applications to recent topics

In this section, we introduce some possible applications of the LASSO to recent topics of machine learning research. Most examples are one of penalized estimations which may establish simple guidelines how to use the LASSO for statistical challenges of high-dimensional data analysis.

3.1. Unsupervised method of grouping variables

Grouping variables give a disjoint network structure. As we see above, the fused LASSO recovers a successive linkage of predictive variables keeping their orders in the linear regression model in (2.1). That is, the k predictive variables, $\mathbf{X}_j, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_{j+k-2}$ and $\mathbf{X}_{j+k-1}$ form a local group in the sense that $\hat\beta_j^\lambda = \cdots = \hat\beta_{j+k-1}^\lambda \neq \hat\beta_s^\lambda$ for all $s < j$ and $s \geq j+k$. The estimated network structure from the fused LASSO has an interpretation; if some elements of the fused LASSO are all the same then their effects on the target vector $\mathbf{y}$ are equivalent to each other, so that we can consider the corresponding predictive variables as local network members.

However, if the problem is not supervised, that is, if there is no target variable, then the problem is not trivial. One possible approach may be sparse partial correlation estimation in (2.5). Since some elements of the partial correlation matrix estimator are exactly zero, we can construct a network structure based on the non-zero elements of the estimator. However, there is no guarantee that the estimated network structure is disjoint, and hence this is not a method of grouping variables as the fused LASSO in (2.4). To the author's knowledge, this problem has not been studied yet, and is an open problem. We will discuss in the end of this subsection what statistical application can be addressed from this problem in high-dimensional data analysis.

Before proceeding, we introduce another penalized estimation for sparse PCA proposed by Jolliffe *et al.* (2003). Assume that we have $n \times p$ sample matrix $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$. The p sparse principal components $\mathbf{c}_j, j \leq p$ can be constructed as linear combinations of p variables $\mathbf{Z}_j, j \leq p$:

$$\mathbf{c}_j = \mathbf{Z}\hat{\boldsymbol\alpha}_j^\lambda = \textstyle\sum_{k=1}^p \mathbf{Z}_j \hat\alpha_{jk}^\lambda,$$

where $\hat{\boldsymbol\alpha}_j^\lambda$ successively maximizes the variance

$$\boldsymbol\alpha_j^T \mathbf{Z}^T \mathbf{Z} \boldsymbol\alpha_j = \|\mathbf{Z}\boldsymbol\alpha_j\|_2^2,$$

for some $\lambda > 0$, subject to $\boldsymbol\alpha_j^T\boldsymbol\alpha_j = 1$, $\boldsymbol\alpha_h^T\boldsymbol\alpha_j = 0, h < j$ and $\|\boldsymbol\alpha_j\|_1 \leq \lambda$. It is easy to see that the penalized estimator $\hat{\boldsymbol\alpha}_j^\lambda$ is sparse so that the principal component becomes to have sparse representation: $\hat{\mathbf{c}}_j^\lambda = \sum_{k=1}^p \hat\alpha_{jk}^\lambda \mathbf{Z}_j I(\hat\alpha_{jk}^\lambda \neq 0)$. As the authors pointed out, the sparse PCA effectively ignores any small coefficients as zero, so that the interpretation becomes easier.

Using the idea of sparse PCA, we introduce a new idea of penalized estimation for grouping variables. We consider the following successive penalized estimation of maximizing

$$\boldsymbol\alpha_j^T \mathbf{Z}^T \mathbf{Z} \boldsymbol\alpha_j = \|\mathbf{Z}\boldsymbol\alpha_j\|_2^2, j \leq p,$$

for some $\lambda_j > 0$, subject to $\boldsymbol\alpha_j^T\boldsymbol\alpha_j = 1$, $\|\boldsymbol\alpha_j\|_1 \leq \lambda_j$ and $\big(\cup_{h<j} \operatorname{supp}(\boldsymbol\alpha_h)\big) \cap \operatorname{supp}(\boldsymbol\alpha_j) = \emptyset$, where $\operatorname{supp}(\boldsymbol\alpha_j) = \{k : \alpha_{jk} \neq 0\}$. Given $j \leq p$, the optimization problem is quite easy to solve since it is a L_1-constrained problem with one equality L_2-constraint. After some iterations, we can find a $q < p$ such that $\operatorname{supp}(\boldsymbol\alpha_j) = \emptyset$ for all $j > q$. Further, all the principal components must be represented as linear combinations of disjoint set of variables, which implies the variables are grouped. Note that the proposed method fails to have the orthogonality of the original PCA since $\mathbf{c}_j^T\mathbf{c}_k \neq 0$ in general and so does the sparse PCA. Hence, to investigate the properties of the principal components from the proposed method is also of interest, which can be a future study.

Before going to next subsection, we give a nice statistical application of the proposed method. Since the estimated group structures are disjoint, we can estimate the covariance matrix $\mathbf{\Sigma}$ of samples by a block diagonal matrix $\hat{\mathbf{\Sigma}}^{\lambda}$. The diagonal matrices of $\hat{\mathbf{\Sigma}}^{\lambda}$ are simply groupwise sample covariance matrices: $\mathbf{Z}_{\mathcal{A}_k}^T \mathbf{Z}_{\mathcal{A}_k}/|\mathcal{A}_k|$ for $k \leq q$, where $\mathcal{A}_k$ is the index set of variables in the kth principal component $\mathbf{c}_k$, $|\mathcal{A}_k|$ is the cardinality of $\mathcal{A}_k$ and $\mathbf{Z}_{\mathcal{A}_k}$ is the submatrix of $\mathbf{Z}$ that consists of the variables in $\mathcal{A}_k$.

3.2. High dimensional clustering

Assume that we have n samples $\mathbf{z}_i \in \mathbb{R}^p, i \leq n$, and we want to construct K-clusters, $\mathcal{S}_k, k \leq K$, that satisfy the followings: $\mathcal{S}_k \subset \mathcal{S}, k \leq K$, where $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, $\mathcal{S}_k \cap \mathcal{S}_j = \emptyset, k \neq j \leq K$ and $\cup_{k=1}^K \mathcal{S}_k = \mathcal{S}$. For example, the K-means clustering method constructs K clusters $\mathcal{S}_k, k \leq K$, by minimizing

$$L(\mathcal{S}_1, \ldots, \mathcal{S}_K) = \textstyle\sum_{k=1}^K \sum_{\mathbf{z}_i \in \mathcal{S}_k} \|\mathbf{z}_i - \hat{\boldsymbol{\mu}}_k\|_2^2,$$

with respect to $\mathcal{S}_k, k \leq K$, where

$$\hat{\boldsymbol{\mu}}_k = \arg\min_{\boldsymbol{\mu}_k} \textstyle\sum_{\mathbf{z}_i \in \mathcal{S}_k} \|\mathbf{z}_i - \boldsymbol{\mu}_k\|_2^2 = \sum_{i=1}^n \mathbf{z}_i I(\mathbf{z}_i \in \mathcal{S}_k)/|\mathcal{S}_k|.$$

Note that the K-means clustering method depends on the distances between samples even when p is far lager than n. Considering the high-dimensional situation, we can expect that some of variables does not highly affect the cluster structure, that is, there are noisy variables in the samples. In this case, it is hard to have a desirable clustering results unless we identify useless variables while clustering.

For this issue, we propose penalized K-means clustering method that improves the original K-means clustering method by using L_1-constraint. Given $\lambda_k > 0, k \leq K$, consider the problem of minimizing

$$L(\mathcal{S}_1, \ldots, \mathcal{S}_K) = \textstyle\sum_{k=1}^K \sum_{\mathbf{z}_i \in \mathcal{S}_k} (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_k^{\lambda_k})^T \mathrm{diag}(\hat{\boldsymbol{\omega}}_k^{\lambda_k})(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_k^{\lambda_k}),$$

with respect to $\mathcal{S}_k, k \leq K$, where

$$(\hat{\boldsymbol{\mu}}_k^{\lambda_k}, \hat{\boldsymbol{\omega}}_k^{\lambda_k}) = \arg\min_{\boldsymbol{\mu}_k, \boldsymbol{\omega}_k} \textstyle\sum_{\mathbf{z}_i \in \mathcal{S}_k} (\mathbf{z}_i - \boldsymbol{\mu}_k)^T \mathrm{diag}(\boldsymbol{\omega}_k)(\mathbf{z}_i - \boldsymbol{\mu}_k). \tag{3.1}$$

subject to $\|\boldsymbol{\omega}_k\|_1 = \lambda_k$ and $\omega_{kj} \geq 0$ for all $k \leq K$ and $j \leq p$. Note that the optimization problem in (3.1) is a linear equality constrained linear optimization problem which is easy to solve. Further, by the linear equality constraint, the solution $\hat{\boldsymbol{\omega}}_k^{\lambda_k}$ must be sparse yielding sparse $\hat{\boldsymbol{\mu}}_k^{\lambda_k}$. This shows the key idea of the proposed method that we can ignore the marginal distances by using the sparsity of $\hat{\boldsymbol{\omega}}_k^{\lambda_k}$. As the K-means clustering algorithm, the proposed method may depend on the initial choice of $\hat{\boldsymbol{\mu}}_k^{\lambda_k}$, and require a proof of convergence which is a future study.

4. Concluding remarks

In this paper, we briefly range over a number of recent advances of the LASSO in high-dimensional data analysis such as variable selection in sparse linear regression models. These methods have been extended to various statistical models such as generalized linear models (Fan and Peng, 2004; Kwon and Kim, 2011; Kim *et al.*, 2011) and survival analysis (Fan and Li, 2002; Hwang *et al.*, 2011). However, we skip some details on other important topics such as tuning parameter selection and optimization algorithm.

As statistical applications become increasingly sophisticated and progressive, developing efficient algorithms plays a key role in the processes. Definitely, algorithms have been evolving rapidly for the past years. However, it is still insufficient to handle various high-dimensional statistical applications. Further, it is much hard to ensure global competitiveness without achieving computational efficiency, considering high-dimension models and huge number of samples. We refer to Tibshirani (1996), Efron *et al.* (2004), Rosset and Zhu (2007) Friedman *et al.* (2007), Yuan and Lin (2006), Friedman *et al.* (2008), Tibshirani *et al.* (2005) and Tibshirani and Taylor (2011) as examples.

Another critical issue is probably how to choose tuning parameters that are optimal in a certain sense. A conventional way of choosing is the training, validating and testing procedure (Hastie *et al.*, 2001), but sometimes the K-fold cross validation method is preferred, when a sample size is small. However, the model that minimizes the K-fold cross validation error must overfit as shown by Wang *et al.* (2007). Hence, if the objective of data analysis is to identify the true model, the cross validation is not a good choice and the increasing computational cost must be concerned also.

As alternatives, consistent tuning parameter selection methods have been studied. Typical examples are information criteria proposed by Wang *et al.* (2009), Wang *et al.* (2007), Wang and Zhu (2011), Hirose *et al.* (2012) and Fan and Tang (2012). All these methods are one of slight modifications of generalized information criterion proposed by Shao (1997) to apply the common idea to penalized estimations; they are different only in orders (depending only on model size p and sample size n) of weights of the degrees of freedom.

However, as studied by Kim *et al.* (2012), there are quite large number of consistent model selection criteria indexed by the weight in each criterion. The weight must depend on not only n and p but also many unknown things such as model sparsity, minimal signal strength and error variance. Hence, it is quite hard to choose an optimal criterion among them case by case which is a reason why various forms of weights as above exist. Hence, to authors' knowledge, an adaptive selection criterion such as Ye (1998) and Shen and Ye (2002) is much worth studying that can cover various situations in a uniform way, which is a nice further study.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings Second International Symposium on Information Theory*, 267-281.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA*, **96**, 6745-6750.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, **9**, 485-516.

Bickel, P. J., Ritov, Y. A. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, **37**, 1705-1732.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350-2383.

Choi, J., Zou, H. and Oehlert, G. (2010a). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface*, **3**, 429-436.

Choi, N. H., Li, W. and Zhu, J. (2010b). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, **105**, 354-364.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkages. *Biometrika*, **81**, 425-455.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

Edward, D. (2000). *Introduction to graphical modelling*, Second edition, Springer, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348.1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928-961.

Fan, Y. and Tang, C. Y. (2012). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society B*, **75**, 671-683

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**, 302-332.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736.*

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning*, Springer, New York.

Hirose, K., Tateishi, S. and Konishi, S. (2012). Tuning parameter selection in sparse regression modeling. *Computational Statistics and Data Analysis*, **59**, 28-40.

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, **27**, 481-499.

Huang, J., Horowitz, J. L. and Ma, S. (2008a). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, **36**, 587-613.

Huang, J., Ma, S. and Zhang, C.-H. (2008b). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.

Hwang, C., Kim, M. S. and Shim, J. (2011). Variable selection in ℓ_1 penalized censored regression. *Journal of the Korean Data & Information Science Society*, **22**, 951-959.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, **12**, 531-547.

Kim, Y., Jun, C.-H. and Lee, H. (2011). A new classification method using penalized partial least squares. *Journal of the Korean Data & Information Science Society*, **22**, 931-940.

Kim, Y. and Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315-325.

Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, **13**, 1037-1057.

Kwon, S. and Kim, Y. (2011). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, **22**, 629-653.

Lee, S. and Lee, K. (2012). Detecting survival related gene sets in microarray analysis. *Journal of the Korean Data & Information Science Society*, **23**, 1-11.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, **16**, 1273-1284.

Meinshausen, N. and Bülmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**, 1436-1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representation for high-dimensional data. *The Annals of Statistics*, **37**, 246.270.

Park, M. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B*, **69**, 659-667.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, **104**, 735-746.

Pourahmadi, M. (2011). Covariance estimation: The glm and regularization perspectives. *Statistical Science*, **26**, 369-387.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over L_q-balls. *IEEE Transactions on Information Theory*, **57**, 6979-6994.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, **35**, 1012-1030.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221-242.

Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210.221.

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

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, **39**, 1335-1371.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society B*, **71**, 671-683.

Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, **102**, 1141-1151.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120-131.

Yuan, M. (2008). Efficient computation of l1 regularized estimates in gaussian graphical models. *Journal of Computational and Graphical Statistics*, **17**, 809-826.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, **99**, 2261-2286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, **68**, 49-67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894-942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567-1594.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, **27**, 576-593.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *The Annals of Statistics*, **37**, 2109-2144.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Reserach*, **7**, 2541-2563.

Zhou, S., van de Geer, S. and Bülmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515.*

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 the Royal Statistical Society B*, **67**, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.

Zou, H. and Zhang, H. (2009). On the adaptive elastic net with a diverging number of parameters. *The Annals of Statistics*, **37**, 1733-1751.