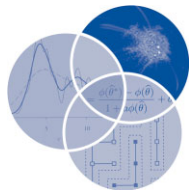
Advanced Review



Variable selection using L_q penalties

Xingye Qiao*

High-dimensional data analysis emerges in many different areas of scientific research and social practice. Variable selection is an important task in statistical analysis for high-dimensional data. The traditional all-subset-selection method can be ideal for variable selection, but it is computationally intractable in the high-dimensional setting. In both regression and classification contexts, sparsity penalties are commonly used for the purpose of defining the complexity of the resulting model so as to achieve variable selection. The current article reviews a few important milestones and some recent works in the area of variable selection, especially those methods which use L_q norm and their variants as the penalty term. In particular, we review the L_q penalty, nonconvex penalties, among others. In ultrahigh-dimensional data analysis, independence learning is often used for the purpose of dimension reduction. Theoretical results, methodological developments and computational innovations in regard to these methods are discussed. These variable selection techniques can be easily extended to problems beyond linear regression, such as classification, quantile regression, etc. Lastly, an interesting and promising research trend is the combination of multiple methods and we review several successful methods which fall into this category.

How to cite this article:
WIREs Comput Stat 2014, 6:177–184. doi: 10.1002/wics.1299

Keywords: high-dimensional learning; independence learning; penalized likelihood; linear regression; variable selection

INTRODUCTION

Nowadays, high- and ultrahigh-dimensional data arise from almost every aspect of the human society, ranging from scientific areas such as health care, biological and genetic sciences, medical sciences including medical imaging, geology, physical science, chemistry, and meteorology, to social practices such as traffic control, social network, business, homeland security, logistics, and supply chain management and finance. When analyzing these high- and ultrahigh-dimensional data, statisticians aim for both interpretation of the statistical model and its power of accurate estimation, prediction, and classification. Donoho[1] shows the pressing need for developing high-dimensional data analysis methodology and theory. Fan and Li[2] and Fan and Lv[3] give comprehensive overviews of statistical challenges with high-dimensionality and provide viable solutions to variable selection with high dimensions and ultrahigh dimensions.

One difficulty of high-dimensional data analysis is the presence of multicollinearity: variables are (nearly) linearly dependent from each other, which makes traditional statistical procedures numerically unstable and many times not identifiable. Multicollinearity is more likely to occur when the dimensionality is large, and occurs almost surely when the dimensionality is greater than the sample size. Consequences of multicollinearity include overfitting (models perform badly when being generalized to the out-of-sample data), mis-modeling (irrelevant

*Correspondence to: qiao@math.binghamton.edu

Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

variables are identified as relevant variables by mistake) and data piling in classification problem (data vectors pile upon each other when projected to the classification direction.) Regularization directly tackles the multicollinearity problem. In particular, sparsity penalties have been used to shrink the parameter space from which we search for the optimal solution. Herein, a *de facto* assumption we adopt is that data lie in low-dimensional structures within a high-dimensional space. With this assumption, sparsity becomes a very appealing property because it encourages concise representations of the data, which improve the model interpretation. For a statistical method which is capable of doing variable selection, two types of properties are desired. The first types of properties focus on the statistical accuracy in estimation, prediction, and classification. The second types of properties focus on the accuracy of model selection, variable selection, tuning parameter selection and so on. A number of researchers have made great progress on both ends. Moreover, in recent decades, we have also witnessed the improvement of the computational efficiency of these new statistical methods, compared to the much slower traditional model selection methods, thanks to the marriage of statistics and machine learning. For many procedures, we can even obtain the whole solution path corresponding to a wide range of parameter values. The current article takes a few snapshots in the journey of developing methods and theory for variable selection using L_p penalties and their variants. For data with ultrahigh dimensions, the independence learning method provides an initial simplification of the problem which we briefly discuss here. Although we start with the canonical linear regression problem as a working example, the variable selection techniques introduced here can be easily extended to other statistical problems. Some promising research directions will be reviewed in the end of the article.

BACKGROUND AND NOTATIONS

We first introduce the background of the underlying statistical problem as well as some general notations to be used throughout the article.

For a general vector in the p-dimensional Euclidean space $\boldsymbol{v} \in \mathbb{R}^p$, let v_j be the jth component of $\boldsymbol{v}$ and $|\boldsymbol{v}|_q = \left(\sum_{j=1}^p |v_j|^q\right)^{1/q}$ be the L_q norm of $\boldsymbol{v}$, where $0 \le q \le \infty$. Note that the L_0 norm of $\mathbf{v}$ is the number of nonzero components in $\boldsymbol{v}$ and the L_∞ norm of $\mathbf{v}$ is $\max(v_1, \ldots, v_p)$.

We consider the linear regression model as a working example, although other topics such as quantile regression and classification will be discussed later as well. We consider the canonical linear regression setting,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is the vector of the response variable, $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$ is the n by p design matrix, whose ith row, $\boldsymbol{x}_i \in \mathbb{R}^p$, represents the ith observation in the sample, $\boldsymbol{\epsilon}$ is an n-dimensional vector for the random error, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the coefficient vector. The ordinary least square solution to the linear regression problem minimizes the sum of squared errors, that is,

$$\boldsymbol{\beta}_{\mathrm{OLS}} = \underset{\boldsymbol{\beta}^*}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - \boldsymbol{x}_i^{\mathrm{T}} \beta^*\right)^2 = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{Y}.$$

We denote the support set of the true coefficient vector $\boldsymbol{\beta}$ as $\mathcal{S} = \left\{1 \le j \le p : \beta_j \ne 0\right\}$ and $\mathcal{S}^{\mathrm{c}}$ the complement of $\mathcal{S}$ in $\{1, \ldots, p\}$. We refer $\mathcal{S}$ to the signals or relevant variables and $\mathcal{S}^{\mathrm{c}}$ to the noises or irrelevant variables. Moreover, we let $\mathbf{X}_{\mathcal{S}}$ be the columns of $\mathbf{X}$ corresponding to the signals $\mathcal{S}$ and $\mathbf{X}_{\mathcal{S}^{\mathrm{c}}}$ be the columns of $\mathbf{X}$ for the noise variables. Similarly, $\boldsymbol{\beta}_{\mathcal{S}}$ is the vector of the nonzero coordinates of $\boldsymbol{\beta}$.

CLASSICAL MODEL SELECTION

Given $\boldsymbol{\beta}^\dagger$, an estimate to $\boldsymbol{\beta}$, with model size (number of nonzero coordinates) d, that is, $|\boldsymbol{\beta}^\dagger|_0 = d$, we define the sum of squared errors $\mathrm{SSE}_d \equiv \mathrm{SSE}\left(\boldsymbol{\beta}^\dagger\right) = \sum_{i=1}^n \left(Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^\dagger\right)^2$. We consider the assumption that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, where σ^2 is an unknown equal variance, and $\mathbf{I}$ is an identify matrix with appropriate dimensions. Many classical model selection methods have been proposed to minimize an ultimate criterion, such as the adjusted R^2 criterion, or some prediction error criteria. For example, many methods attempt to estimate the mean squared prediction bias across all observations of the sample. Among these methods, some well known ones include $\mathrm{PRESS} \equiv \sum_{i=1}^n \left(Y_i - \hat{Y}_{i(i)}\right)^2$ where $\hat{Y}_{i(i)}$ is the prediction of Y_i based on a reduced sample without the ith observation, Mallow's C_p criterion $\equiv \mathrm{SSE}_d/s_p^2 - (n - 2d)$ where s_p^2 is the mean squared error estimated under the full model, $\mathrm{AIC} \equiv n\log(\mathrm{SSE}_d/n) + 2d$ and $\mathrm{BIC} \equiv n\log(\mathrm{SSE}_d/n) + \log(n)d$. Readers are advised to refer to classic statistical textbook on linear regression, such as Faraway,[4] for further references.

A statistician who aims to find the best (or correct) model should hopefully be able to calculate one such measure for all possible linear regression models with different combinations of variables. For models with size d, there are 'p choose d' possible models.

This amounts to $2^p - 1$ possible models in total. For small p, the computational burden is reasonable. For moderate p, it may be doable, given that the sample size is not beyond manageable. A common practice is to avoid trying all possible models using forward selection, backward selection, stepwise selection, etc. These schemes aim to mimic the global optimal model through a single path with fewer evaluations involved. As approximations, they may not have the guarantee to achieve the global optimality. As is obvious, for large p, it is infeasible to evaluate for all or even a substantial subset of all candidate models.

REGULARIZATION AND L_q PENALTIES

For high-dimensional data, the ordinary least square approach to linear regression encounters some challenges. In addition to the potential computational issue in model selection, collinearity is a notorious problem. In high dimensions ($p \gg n$), $\mathbf{X}^T\mathbf{X}$ is not invertible with probability 1 and hence the least square solution is not available. In the classification setting, a similar issue can lead to the so-called data-piling phenomenon, see for example, Refs 5 and 6.

For the regression problem, we consider a penalized least square method now. A regularization term, used to describe complexity of the model, is added to the scaled sum of squared errors criterion to form the new objective function. This leads to

$$\boldsymbol{\beta}^{\dagger} = \underset{\boldsymbol{\beta}^*}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* \right)^2 + \lambda \cdot p\left(\boldsymbol{\beta}^* \right) \right\}, \tag{2}$$

where $p(\boldsymbol{\beta})$ is the additional regularization term, also known as the penalty term. See a general introduction to regularization in statistics in Ref 7. Often we consider p additive penalties, one for each of the p coefficients, that is, we let $\lambda \cdot p(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_{\lambda_j}(\beta_j)$, where $p_{\lambda_j}(\cdot) : \mathbb{R} \mapsto \mathbb{R}^+$.

For an instance, the ridge regression by Hoerl and Kennard[8], which employees the L_2 norm as the penalty term $p(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_2^2 = \sum_{j=1}^{p} \beta_j^2$, has been very successful in addressing the collinearity issue. In particular, the ridge regression makes use of the invertible $(\mathbf{X}^T\mathbf{X} + a\mathbf{I})^{-1}$, where $a > 0$ depends on the regularization parameter λ, in place of $(\mathbf{X}^T\mathbf{X})^{-1}$ in the ordinary least square solution $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

The L_0 norm of the coefficient vector $|\boldsymbol{\beta}|_0 = \sum_{j=1}^{p} 1_{[\beta_j \neq 0]}$ is a natural choice for the penalty term $p(\boldsymbol{\beta})$ since it directly counts the number of nonzeros in $\boldsymbol{\beta}$ and hence, implies a selection of variables. In particular, the use of the objective function $\frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} 1_{[\beta_j \neq 0]}$ favors a model with small sum of squared errors and small model size. However, the penalized least square method with the L_0 norm as the penalty term is computationally intractable because of the discontinuity of the function $1_{[\theta \neq 0]}$ at $\theta = 0$. This makes it impossible to employ any efficient optimization technique to solve this problem. One way to overcome this technical difficulty is to relax the noncontinuous and nonconvex L_0 norm to a continuous and convex function, such as the L_1 norm, $p(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1 = \sum_{i=1}^{p} |\beta_j|$. The L_1 penalized least square regression has been studied in the seminal LASSO paper by Tibshirani[9]. The L_1 regularization has also been used in Dantzig selector by Candes and Tao[10]. Figure 1 shows the penalty functions for L_0 norm, L_1 norm, $L_{0.5}$ norm, and L_2 norm. An obvious advantage of the L_1 penalty (over the L_0 penalty) is that it is continuous and convex. It turns out that LASSO can be very efficiently solved using various algorithms. The computation of LASSO and other penalized least square methods will be discussed in a later section.

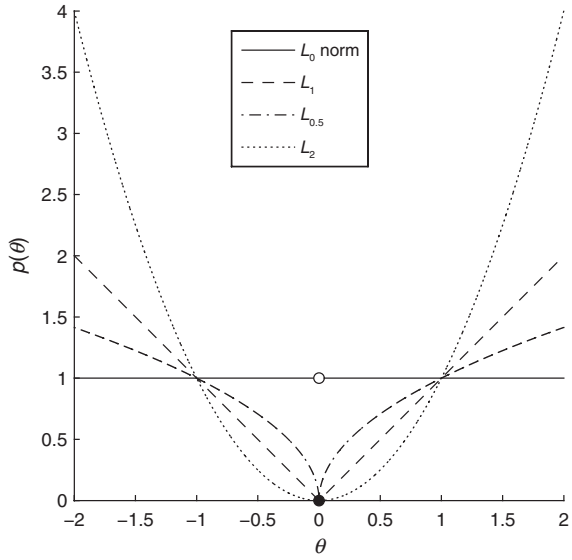


FIGURE 1 | The penalty functions for L_0 norm, L_1 norm, $L_{0.5}$ norm, and L_2 norm.

A broader framework for such relaxation of the L_0 norm to the L_1 norm is described by Frank and Friedman[11] as bridge regression, using the L_q norm $|\boldsymbol{\beta}|_q$ for $0 < q \leq 2$, in which the penalty term $p(\boldsymbol{\beta})$ in Eq. (2) is chosen to be $\sum_{j=1}^{p} |\beta_j|^q$. This framework bridges the L_0 regression and the L_2 ridge regression, hence the name *bridge regression*. Theoretical properties of the bridge regression estimator have been studied by Knight and Fu.[12]

There are rich developments on theoretical properties of L_1 regularization, such as those by Donoho and Johnstone,[13] Donoho et al.,[14] Chen et al.,[15] Zhao and Yu,[16] Candes et al.,[17] Zhang and

Huang,[18] Bickel et al.,[19] and Wainwright.[20] Many of these works concern the problem of model selection consistency. For an instance, the model selection *sign* consistency is a desirable property of a model selection method. It states that with high probability, the estimated coefficient vector resulting from the method of interest can fully recover the sign of the true coefficients $\boldsymbol{\beta}$ in Eq. (1). Often, some irrepresentable-type condition is needed in order to show the model selection sign consistency. For example, Zhao and Yu[16] consider the strong (or weak) irrepresentable condition, $|\mathbf{X}_{\mathcal{S}^c}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\left(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)^{-1}\mathrm{sign}(\boldsymbol{\beta}_{\mathcal{S}})|_{\infty} \leq 1-\eta$ (or <1 respectively), for a constant $\eta > 0$. Intuitively, these irrepresentable conditions imply that the noise variables must not be well explained by the signal variables so that they can be easily identified and eliminated. However, irrepresentable conditions are very difficult to verify in practice. As a matter of fact, they turn out to be very stringent in high dimensions. However, such conditions are nearly necessary for the LASSO method to be selection consistent.

NONCONVEX PENALTIES AND IMPROVEMENTS OF L_q PENALTIES

While penalties such as $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^q$ have been studied by various groups of researchers, Fan and Li[21] systematically analyze penalized least square problem (as a special case of the penalized likelihood method) with a nonconvex penalty function (or, in a more general case, nonconcave penalization for a maximum log-likelihood approach.) They raise the concern about the bias effect of a penalized estimator. Moreover, they advocate three desirable properties of a penalized estimator: *sparsity*, *unbiasedness*, and *continuity*. They articulate the mathematical conditions of the penalty function under which these three conditions hold. Motivated by these properties, they propose the SCAD penalty $\sum_{j=1}^{p} p^{\mathrm{SCAD}}\left(\left|\beta_j\right|\right)$, where $p^{\mathrm{SCAD}}(\cdot)$ is singular at the origin and nonconvex over $(0,\infty)$. In particular, the derivative of the penalty $p^{\mathrm{SCAD}}(\theta)$ for $\theta > 0$ is $\dot{p}^{\mathrm{SCAD}}(\theta) = \min\left\{1, (\gamma - \theta/\lambda)_+/(\gamma-1)\right\}$ for some $\gamma > 2$ where $(a)_+ = \max(a, 0)$ is defined as the positive part of $a \in \mathbb{R}$. To obtain more insights to this penalty function, note that a penalty function $p(\theta)$ must be singular at 0 to have the sparsity and continuity properties; hence in the case of L_q penalties, q must be less than or equal to 1. On the other hand, a sufficient condition of the unbiasedness property is that $\dot{p}(\theta) = 0$ for large $|\theta|$; hence the L_q norm does not satisfy this condition. On both ends, this SCAD penalty satisfies both conditions.

Fan and Li[21] define a new theoretical property called the *oracle property*. Simply put, an estimator with the oracle property, (1) has zero coordinates for those noise variables and (2) mimics the estimator that would have been obtained by an oracle who knows exactly which variables are signals. Fan and Li[21] prove the oracle property for the nonconvex SCAD penalty. Their result holds for finite-dimensional settings. Fan and Peng[22] show the oracle property results in moderate-dimensional settings such as $p = o(n^{1/5})$ or $p = o(n^{1/3})$. Zou[23] proves that LASSO does not have the oracle property and proposes to address this issue by adaptively weighting the penalties imposed for different variables. In particular, he proposes to use $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j|\beta_j|$ where the weight w_j for the jth variable is $|\widehat{\beta}_j|^{-\gamma}$ for some positive γ. Here $\widehat{\beta}_j$ is a root-n consistent estimator of β_j such as the ordinary least square estimator. This adaptive version of LASSO has been shown to have the oracle property.

Zhang[24] proposes the minimax concave penalty (MCP), $\sum_{j=1}^{p} p^{\mathrm{MCP}}\left(\left|\beta_j\right|\right)$. The MCP is another choice for nonconvex penalty which can achieve the three basic properties and the oracle property. The motivation of the MCP is that it provides the convexity of the penalized loss in sparse regions *to the greatest extent* while keeping the certain amount of concavity for the purpose of variable selection and unbiasedness. To compare it with the SCAD penalty (*cf.* Figure 1 in Ref 24), we note that the derivative of the MCP penalty $p^{\mathrm{MCP}}(\theta)$ is $\dot{p}^{\mathrm{MCP}}(\theta) = \left(1 - \theta/\gamma_1\right)_+$ for some $\gamma_1 > 0$. When the parameters in both penalties are chosen so that $\gamma_1 = \lambda\gamma$, then both derivatives have zero value when $\theta \geq \gamma_1 = \lambda\gamma$; hence both satisfy the sufficient condition for the estimator to be unbiased. When $\theta = 0$, both are 1; however, the derivative of MCP starts off decreasing right away when θ is greater than 0 while that of the SCAD penalty remain as a constant until $\theta = \lambda$.

COMPUTATION

As mentioned before, the LASSO method is very attractive in its computation. Efron et al.[25] have developed the least angle regression (LARS) algorithm to solve LASSO. The LARS algorithm provides the whole solution path as a function of the parameter λ. This solution path is, as a matter of fact, piecewise linear in λ. Later on, Rosset and Zhu[26] provide a general condition under which the solution to a penalized likelihood method is piecewise linear.

Incidentally, another very efficient algorithm, called *coordinate descend*, has been developed, studied and applied to various settings by, for example, Fu,[27] Wu and Lange,[28] Fan and Lv,[29] and

Mazumder et al.,[30] among many others. The coordinate descend approach sequentially optimizes each coordinate separately, holding all other coordinates as constants, and updates coordinates cyclically. It is a general technique which can be applied to problems beyond the LASSO method, although the LASSO has a very simple closed form solution for each iteration when the coordinate descent algorithm is applied.

On the nonconvex penalty side, Fan and Li[21] suggest the local quadratic approximation (LQA) scheme by iteratively, locally approximating the penalty function by a quadratic function. In particular, the penalty $p(\theta)$ in $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} p\left(\left|\beta_j\right|\right)$ is approximately by the quadratic function $p\left(\theta^{(0)}\right) + \frac{1}{2}\left\{\dot{p}\left(\theta^{(0)}\right)/\theta^{(0)}\right\}\left(\theta^2 - \theta^{(0)^2}\right)$ where $\theta^{(0)}$ is an initial value of $\theta = |\beta_j^{(0)}|$ close to the true value of $|\beta_j|$. However, the LQA algorithm has some drawbacks which is commonly shared with many backward stepwise variable selection methods. To address such issue, Zou and Li[31] propose the local linear approximation (LLA), $p(\theta) \approx p\left(\theta^{(0)}\right) + \dot{p}\left(\theta^{(0)}\right)\left(\theta - \theta^{(0)}\right)$. They further advocate to stop the iteration early (at step one) because the results after the first step can already provide sparse representation of the model and good estimation of the coefficients. Early stopping can help reduce the computational burden and avoid the potential local minimal problem in minimizing the penalized sum of squares. For the MCP method, Zhang[24] also develops the PLUS algorithm to achieve penalized linear unbiased selection.

More recently, lots of attention have been drawn to the relation between multiple local solutions and the global optimal solution. Many of the aforementioned methods are capable of reaching a local optimal solution which is not necessarily the global optimality. Fan et al.[32] provide a unified theory to show how to obtain the oracle solution via the LLA algorithm. Zhang and Zhang[33] present a general theoretical framework showing that under appropriate conditions, the global solution of nonconvex regularization leads to desirable recovery performance and conditions under which the global solution corresponds to a unique sparse local solution. Incidentally, Wang et al.[34] prove that a calibrated algorithm produces a consistent solution path which contains the oracle estimator with high probability. Wang et al.[35] propose an approximate path-following algorithm for a nonconvex regularization problem and provide rates of convergence of arbitrary local solution obtained by the algorithm.

INDEPENDENCE LEARNING

While variable selection using a penalty term for the sake of sparsity, such as the L_q norm, SCAD or MCP, is efficient in high-dimensional or moderate-dimensional settings, its performance may not be as good for ultrahigh-dimensional data. Here ultrahigh dimensions may be referred to as the case where $\log(p)$ is at the same order as n^{ξ} for some $\xi > 0$.

A common practice is to conduct fast dimensionality reductions. Fan and Lv[36] propose to use the marginal information (particularly, correlations between dependent variables and the response variable) to screen out irrelevant variables and reduce the data from ultrahigh dimensions to high dimensions or moderate dimensions. After the dimension reduction, the variable selection methods introduced in the previous few sections can be applied. The name *independence learning* is due to the fact that only marginal information is used, implying the independence among variables. Fan and Lv[36] have proved the *sure screening property*, which states that with high probability, only irrelevant variables will be screening out. Motivated by the theory in Fan and Lv,[36] Wang[37] investigates the forward regression method. Along the similar line of employing the marginal information, Huang et al.[38] use the marginal bridge estimators for selecting variables.

The regularization after retention (RAR) method by Weng et al.[39] is another instance of independence learning. But it uses the marginal information at a different direction from that of Fan and Lv.[36] In particular, the RAR method is a two-step procedure where a subset of the signals are retained in the first step using marginal correlations and a penalized least square problem is considered in the second step where penalties are imposed on only those not retained in the first step. The RAR method possesses the model selection sign consistency under a weaker condition than the irrepresentable condition of LASSO and has been shown to have very good finite sample performance compared to the LASSO and the sure screening method. Note that RAR is different from the screening method because the variables not retained in the first step will be reevaluated in the second step, whereas they are discarded by the screening method after the first step.

CLASSIFICATION, QUANTILE REGRESSION, AND OTHERS

These variable selection methods can be easily extended to applications beyond linear regression, such as survival data analysis using penalized partial likelihood, longitudinal data, semiparametric regression modeling, among many others. We introduce some works in classification and quantile regression here.

For simplicity, we focus on large margin classifiers, exemplified by the support vector machine by Vapnik,[40] Vapnik,[41] and Cortes and Vapnik.[42] See an excellent introduction in Cristianini and Shawe-Taylor.[43] Bradley and Mangasarian[44] and Zhu et al.[45] introduce two versions of L_1 penalized support vector machine. Liu et al.[46] propose L_q support vector machine where the best q in the L_q penalty is automatically chosen by data. Zou and Yuan[47] select groups of variables in the support vector machine when grouping information is given. Zhang et al.[48] consider variable selection for multicategory classification problem by using sup-norm regularization.

For ultrahigh-dimensional data, Fan and Fan[49] advocate the use of independence classification rule after a variable screening step (called annealing variables), much in the similar spirit of the sure independence screening for regression by Fan and Lv.[36] In particular, marginal two-sample t-tests based on each variable individually are considered to select potentially important variables. They also establish the condition under which all true signals are selected.

For quantile regression, Wu and Liu[50] demonstrate the oracle properties of the SCAD and adaptive-LASSO penalized quantile regressions. Instead of using the LLA algorithm, they solve the optimization related to the SCAD penalty using the difference convex algorithm (DCA). See Refs 51, 52, and 53 for more instances where DCA is used.

In the linear mixed effect model, Fan and Li[54] use a nonconcave penalized profile likelihood methods for the fixed effect and a group variable selection strategy to select and estimate random effects.

COMBINATION OF TWO PENALTIES

There is a trend to combine multiple penalties to achieve enhanced performance of variable selection and estimation. The logic behind is quite natural: it is rare that a single penalty possesses all the desired properties for all situations, thus one may creatively combine penalties to obtain more nice properties. For example, the elastic net method proposed by Zou and Hastie[55] is a successful attempt to combine the L_1 norm penalty and the L_2 norm penalty. In particular, they use $\lambda_1|\boldsymbol{\beta}|_1 + \lambda_2|\boldsymbol{\beta}|_2^2 = \sum_{j=1}^{p}\left\{\lambda_1|\beta_j| + \lambda_2\beta_j^2\right\}$ where λ_1 and λ_2 are both positive parameters. The elastic net outperforms the LASSO in certain situations, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model at the same time.

Liu and Wu[56] consider a combination of a L_0 norm-ish penalty and the L_1 norm penalty. In particular, they consider $p_0^{\epsilon}(\theta) = 1_{[|\theta|>\epsilon]} + |\theta|/\epsilon \cdot 1_{[|\theta|\le\epsilon]}$ as a continuous approximation to the noncontinuous $p_0(\theta) = 1_{[\theta\neq 0]}$, and then use the penalty term $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} p\left(\beta_j\right)$ where $p\left(\beta_j\right) = ap_0^{\epsilon}\left(\beta_j\right) + (1-a)\left|\beta_j\right|$, for $0 < a < 1$. The new combined penalty is expected to enjoy the strengths of both the L_0 and L_1 penalties. Lv and Fan[57] provide a unified approach to model selection and sparse recovery by considering a similar formulation: a smooth homotopy between the L_0 and L_1 penalties.

Inspired by the adaptive LASSO and the elastic net, Zou and Zhang[58] propose the adaptive elastic net (or AdaEnet). The AdaEnet method makes use of the penalty term, $p(\boldsymbol{\beta}) = \sum_{j=1}^{p}\left\{\lambda_1 w_j|\beta_j| + \lambda_2\beta_j^2\right\}$, where w_j is the adaptive weight. Clearly, when λ_1 is forced to be 0, the AdaEnet reduces to the ridge regression; when $\lambda_2 = 0$, the AdaEnet becomes the adaptive LASSO; when $\lambda_2 = 0$ and $w_j = 1$, it is ordinary LASSO; when $w_j = 1$, this is the same as the elastic net. While maintaining the oracle property shared by the adaptive LASSO, the AdaEnet method enjoys better finite sample performance especially when the collinearity problem exists, an advantage given by the elastic net component.

In the classification regime, Wang et al.[59] apply the combined L_1 and L_2 penalty to support vector machine which performs similarly to the elastic net in regression.

Study on such unified frameworks, which are originally motivated by combinations of multiple methods, turns out to be very fruitful, inspiring and helpful. Such combined formulations provide additional theoretical tools to explore the properties of the methods involved over a broader spectrum. Moreover, we often obtain additional computational gains from these methods.

CONCLUSION

We have briefly reviewed some of the major developments in variable selection using L_q penalties and their variants in recent decades. L_q regularizations and nonconvex regularizations often enjoy computational advantage, have nice properties such as the oracle property, and can be generalized to a boarder span of problems, from classification, linear mixed effect model, to quantile regression and nonparametric regression. For the more challenging ultrahigh-dimensional setting, independence learning methods use marginal information to reduce the dimension, while keeping all the important variables

with high probability. Fusions of different penalty functions nicely inherit the advantages of different methods.

There are more to be done in this active area of research. For a nonconvex penalty, despite many recent developments, the computational issue in the difficulty of finding global optimality has not been fully resolved. The choice of tuning parameters in a penalization method is a rather challenging issue in high and ultrahigh dimensions, especially because the size of the true signals relative to the sample size is typically unknown. In the ultrahigh-dimensional space, the independence learning approach has opened one door. However, effects beyond the marginal ones may be ignored because only marginal information is considered. This problem deserves further research efforts. On the other hand, high-dimensional data can be joined by the difficulty brought by big/massive data sets and/or complex data structures, and variable selection in these cases can be very involved. Overall, penalization is a very useful tool for variable selection. In light of these, more critical thinking and innovations are needed and there are many promising tracks to follow.

ACKNOWLEDGMENT

This work was partially supported by a grant from Simons Foundation (246649 to Xingye Qiao).

REFERENCES

1. Donoho DL. High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challenges Lect* 2000, 1–32.
2. Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In: *Proceedings of the International Congress of Mathematicians*, vol. 3, 2006, 595–622.
3. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space (invited review article). *Stat Sin* 2010, 20:101–148.
4. Faraway JJ. *Linear Models with R*. Boca Raton: CRC Press; 2004.
5. Ahn J, Marron J, Muller KM, Chi Y-Y. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* 2007, 940:760–766.
6. Hall P, Marron J, Neeman A. Geometric representation of high dimension, low sample size data. *J R Stat Soc [Ser B]* 2005, 670:427–444.
7. Bickel PJ, Li B, Tsybakov AB, van de Geer SA, Yu B, Valdés T, Rivero C, Fan J, van der Vaart A. Regularization in statistics. *Test* 2006, 150:271–344.
8. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970, 120:55–67.
9. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc [Ser B]* 1996, 58:267–288.
10. Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n. *Ann Stat* 2007, 35:2313–2351.
11. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993, 350:109–135.
12. Knight K, Fu W. Asymptotics for lasso-type estimators. *Ann Stat* 2000, 28:1356–1378.
13. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994, 810:425–455.
14. Donoho DL, Johnstone IM, Kerkyacharian G, Picard D. Wavelet shrinkage: asymptopia? *J R Stat Soc [Ser B]* 1995, 57:301–369.
15. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 1998, 200:33–61.
16. Zhao P, Yu B. On model selection consistency of lasso. *J Mach Learn Res* 2006, 7:2541–2563.
17. Candes EJ, Wakin MB, Boyd SP. Enhancing sparsity by reweighted 1 minimization. *J Fourier Anal Appl* 2008, 140:877–905.
18. Zhang C-H, Huang J. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann Stat* 2008, 360:1567–1594.
19. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of lasso and dantzig selector. *Ann Stat* 2009, 370:1705–1732.
20. Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery. *IEEE Trans Inf Theory* 2009, 550:2183–2202.
21. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001, 960:1348–1360.
22. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann Stat* 2004, 320:928–961.
23. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006, 1010:1418–1429.

24. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010, 380:894–942.
25. Efron B, Hastie T, Johnstone IM, Tibshirani R. Least angle regression. *Ann Stat* 2004, 320:407–499.
26. Rosset S, Zhu J. Piecewise linear regularized solution paths. *Ann Stat* 2007, 35:1012–1030.
27. Fu WJ. Penalized regressions: the bridge versus the lasso. *J Comput Graph Stat* 1998, 70:397–416.
28. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat* 2008, 2:224–244.
29. Fan J, Lv J. Nonconcave penalized likelihood with np-dimensionality. *IEEE Trans Inf Theory* 2011, 570:5467–5484.
30. Mazumder R, Friedman JH, Hastie T. Sparsenet: Coordinate descent with nonconvex penalties. *J Am Stat Assoc* 2011, 1060:1125–1138.
31. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann Stat* 2008, 360:1509–1533.
32. Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation. arXiv preprint arXiv:1210.5992, 2012.
33. Zhang C-H, Zhang T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Stat Sci* 2012, 270:576–593.
34. Wang L, Kim Y, Li R. Calibrating nonconvex penalized regression in ultra-high dimension. *Ann Stat* 2013, 410:2505–2536.
35. Wang Z, Liu H, Zhang T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. arXiv preprint arXiv:1306.4960, 2013.
36. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc [Ser B]* 2008, 700:849–911.
37. Wang H. Forward regression for ultra-high dimensional variable screening. *J Am Stat Assoc* 2009, 1040:1512–1524.
38. Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann Stat* 2008, 360:587–613.
39. Weng H, Feng Y, Qiao X. Regularization after retention in ultrahigh dimensional linear regression models. arXiv preprint arXiv:1311.5625, 2013.
40. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 1999.
41. Vapnik V. *Statistical Learning Theory*. New York: John Wiley & Sons; 1998.
42. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995, 200:273–297.
43. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press; 2000.
44. Bradley PS, Mangasarian OL. Feature selection via concave minimization and support vector machines. In: *International Conference on Machine Learning*, vol. 98, 1998, 82–90.
45. Zhu J, Rosset S, Hastie T, Tibshirani R. 1-norm support vector machines. *Adv Neural Inf Process Syst* 2004, 160:49–56.
46. Liu Y, Helen Zhang H, Park C, Ahn J. Support vector machines with adaptive L_q penalty. *Comput Stat Data Anal* 2007, 510:6380–6394.
47. Zou H, Yuan M. The f_∞-norm support vector machine. *Stat Sin* 2008, 18:379–398.
48. Zhang HH, Liu Y, Wu Y, Zhu J. Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electron J Stat* 2008, 2: 149–167.
49. Fan J, Fan Y. High dimensional classification using features annealed independence rules. *Ann Stat* 2008, 360:2605–2637.
50. Wu Y, Liu Y. Variable selection in quantile regression. *Stat Sin* 2009, 190:801–817.
51. Shen X, Tseng GC, Zhang X, Wong WH. On Ψ-learning. *J Am Stat Assoc* 2003, 980:724–734.
52. Liu Y, Shen X. Multicategory Ψ-learning. *J Am Stat Assoc* 2006, 1010:500–509.
53. Wu Y, Liu Y. Robust truncated hinge loss support vector machines. *J Am Stat Assoc* 2007, 1020: 974–983.
54. Fan Y, Li R. Variable selection in linear mixed effects models. *Ann Stat* 2012, 400:2043–2068.
55. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc [Ser B]* 2005. ISSN 1467-9868, 670:301–320.
56. Liu Y, Wu Y. Variable selection via a combination of the l0 and l1 penalties. *J Comput Graph Stat* 2007, 160:782–798.
57. Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. *Ann Stat* 2009, 370:3498–3528.
58. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat* 2009, 370:1733–1751.
59. Wang L, Zhu J, Zou H. The doubly regularized support vector machine. *Stat Sin* 2006, 160:589–615.

