Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 227-236

Variable Selection and Shrinkage via a Conditional Likelihood-based Penalty

Arpita Ghosh[1], Andrew B. Nobel[2, 3], Fei Zou[3] and Fred A. Wright[4]
[1]*Public Health Foundation of India, New Delhi, India*
[2]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, North Carolina, USA*
[3]*Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, USA*
[4]*Department of Statistics, North Carolina State University, North Carolina, USA*

Received 02 September 2013; Revised 03 January 2014; Accepted 19 April 2014

SUMMARY

The usefulness of penalized regression to analyze large datasets is increasingly recognized, with a growing role in genome-wide association scans and in the analysis of data from other -omics technologies. Penalized regression has been applied to data in fields as diverse as health sciences, economics, and finance. We investigate connections between procedures to address "significance bias" or "winner's curse" in genome-wide association studies and the shrinkage of coefficient estimates and variable selection that is applied in existing penalized regression procedures. We use a conditional likelihood approach that has been applied to correct for significance bias in order to propose a new penalized regression procedure. The approach has a natural interpretation when the number of predictors is smaller than the sample size. In addition, we describe an analogous procedure when the number of predictors is larger than the sample size. We demonstrate via data examples and simulations that the procedure performs favorably in terms of prediction error in both low-dimensional and high-dimensional settings in comparison to competing approaches, especially when the proportion of true nonzero coefficients is small.

Keywords: Variable selection, Shrinkage, Penalized regression, Conditional likelihood, Significance bias, Winner's curse.

1. INTRODUCTION

In applications of linear regression, the goal is to find a linear model that provides a concise description of how the measured predictors affect the response. The model selection problem entails selecting variables that might best describe that relationship, and estimating the coefficients corresponding to those variables. As a consequence of the simultaneous development of high-speed computing and high-throughput measurement technologies, many research problems involve data with a large number of predictors, and in many cases, more predictors than observations. For example, a typical gene expression data set has tens of thousands of genes (predictors) and only a few hundred arrays (observations). High-dimensional data arises in various fields of scientific research including computational biology, finance, biomedicaical imaging, satellite imagery, and many others. At the same time as they enhance the need for model selection in linear regression, high dimensional datasets provide a number of challenges to traditional model selection procedures.

It is common to judge the usefulness of a regression model on the basis of prediction accuracy and interpretability. The prediction accuracy of a model is typically measured by the expected prediction error of the regression fit, also known as test error or

Corresponding author: Arpita Ghosh
E-mail address: arpita.ghosh@phfi.org

generalization error (Hastie *et al*. 2001). Interpretability of a model is often more qualitative in nature, and involves discerning which variables play an important role in predicting the response.

Ordinary least squares (OLS), minimizing the residual sum of squares, is intuitively appealing but does not always provide a satisfactory model in terms of prediction accuracy and interpretability. It produces best linear unbiased estimates, but the variance of the predicted values is often high. The interpretability of the model is also seriously hampered since OLS retains all the predictors. With too many variables in the model, it is dificult to understand which variables are really important in predicting the response. Moreover, in the high-dimensional setting, it is not possible get an OLS solution, there being no unique solution to the system of linear equations involving the coefficients. Several penalized methods have been developed over the last two decades to remedy these problems and achieve better prediction accuracy. We briey review some of the most commonly used methods here.

Traditional approaches to model selection, such as best subset regression or stepwise regression, retain a subset of the candidate predictors, eliminate the rest, and use OLS to estimate the coefficients corresponding to the ones retained. Subset selection generally achieves better prediction accuracy than the full model by selecting only a subset of the candidate predictors. The selection of the subset of variables is based on either best subset regression or forward/backward stepwise selection. Among the sequence of models produced by each of the above procedures, it is common to select one that minimizes an estimate of the expected prediction error. Although it is conceptually simple and produces easily interpretable models subset selection suffers from the fact that it is a discrete process: it either makes a coefficient zero or inates it. This feature of the method makes it unstable with respect to small perturbations in the data.

Ridge regression (Hoerl and Kennard 1970), on the other hand, retains all the predictors in the model and modifies how the coefficients are estimated. Ridge regression achieves better performance than OLS through a bias-variance trade-off. It is a continuous process and the ridge estimates are stable: if we delete a single data point, the new ridge estimates, for the same tuning parameter will be close to the old ones. However, like OLS, ridge regression retains all the predictors in the fitted model, resulting in less interpretability. On the other hand, the use of regularization means that ridge regression can be used in the high-dimensional setting.

The nonnegative garrote, proposed by Breiman (1995), retains good features of both subset selection and ridge regression. The nonnegative garrote starts with the OLS estimates. As one increases the penalty (tightens the garrote), some of the coefficients are set to zero and the remaining ones are shrunk towards zero. Breiman showed via simulations that nonnegative garrote outperforms subset selection and is comparable to ridge regression unless the model has a large number of small effcts. In terms of stability, the nonnegative garrote is intermediate between subset selection and ridge regression. As it depends heavily on the OLS estimates, the nonnegative garrote estimates are expected to suffer in situations where the OLS estimates perform poorly, and cannot be used when there are more predictors than samples.

Motivated by the idea of nonnegative garrote, Tibshirani (1996) proposed a new technique called LASSO: least absolute shrinkage and selection operator. When there are a large number of candidate predictors parsimony is an important issue, LASSO can reduce coefficients to exactly zero and thus produce sparse solution. In other related work around the same time, Frank and Friedman (1993) introduced bridge regression, which includes subset selection, LASSO, and ridge regression as special cases.

LASSO can be implemented in the high-dimensional setting but it cannot select more variables than number of observations. Also, if there is a group of correlated variables among which the pairwise correlations are very high, then LASSO tends to choose any one variable from that group. In the usual regression setup, if the correlation between the predictors is high, ridge regression usually outperforms LASSO. In an attempt to retain good features of both ridge regression and LASSO, Zou and Hastie (2005) presented a new regularization and variable selection method, called the elastic net. The elastic net uses a penalty that is a convex combination of the lasso and ridge penalties. The elastic net penalty has a “grouping effect” property, in which highly correlated predictors have the same regression coefficients.

In 2001, Fan and Li proposed a non-convex penalty function, the smoothly clipped absolute deviation (SCAD) penalty. The authors argued that a "good" penalty function should be unbiased, sparse, and continuous in the data, and showed that the SCAD penalty possesses all three of these properties. Efron *et al*. (2004) introduced a new model selection algorithm on least angle regression (LARS). Simple modifications of LARS give LASSO and forward stagewise regression. The LARS procedure has affected implementation of the LASSO in large data sets.

All the methods described above can be viewed as applying different penalty functions to the OLS criterion, and can be regarded as penalized least squares procedures. The idea of applying penalty functions to the OLS criterion can be extended to a penalized likelihood framework, thereby encompassing likelihood-based models. Penalized likelihood estimators can also be interpreted from a Bayesian point of view, in which the penalty function plays the role of a log-prior density for the parameters. Thus the LASSO estimates can be viewed as the Bayes posterior mode when the parameters are a priori independent, each having a Laplacian prior distribution. Similarly, the ridge estimate can be interpreted as mode of the posterior distribution with independent Gaussian priors for the parameters. The SCAD penalty corresponds to an improper prior.

In some situations, LASSO is inconsistent for variable selection. Zou (2006) proposed a variant of the LASSO, called the adaptive LASSO, that incorporates data-dependent weights to reduce the bias of the ordinary LASSO. Yuan and Lin (2006) studied the problem of selecting grouped variables (factors) for achieving better prediction accuracy in regression problems where interest lies in finding important explanatory factors for the response variable. They extended LASSO, LARS, and nonnegative garrote to group LASSO, group LARS, and group nonnegative garrote for factor selection. Wang *et al*. (2007) developed group SCAD regression in the same spirit. Fan and Lv (2006) introduced sure independent screening (SIS) in high-dimensional problems based on a correlation learning. Wasserman *et al*. (2007) advocated a three-stage procedure for model selection and fitting: in the first stage, a set of candidate models, LASSO, marginal regression, and forward stepwise regression, are fit to the data; in the second stage one of the methods is selected by cross-validation; and in the third stage, hypothesis testing is used to eliminate some of the variables.

Candes and Tao (2007) proposed the Dantzig selector for linear regression models with a large number of predictors but a sparse set of non-zero coefficients. Dantzig selector minimizes the sum of the absolute values of the coefficients, the l1 norm, subject to a constraint on the correlation of the residuals with the predictors. James *et al*. (2009) explored the relationship between LASSO and Dantzig selector and described a new algorithm, DASSO, which uses a LARS-type algorithm to compute the entire solution path for the Dantzig selector. Radchenko and James (2008) described another modification to LASSO to prevent overshrinkage of LASSO by using two tuning parameters, one for selecting variables and the other to control the amount of shrinkage. Many other extensions and improvements of the methods described above have been proposed in the model selection literature for simultaneous variable selection and coefficient shrinkage. These include Octagonal Shrinkage and Clustering Algorithm for Regression (Bondell and Reich 2008), Bolasso (Bach 2008), Forward-Lasso Adaptive Shrinkage (Radchenko and James 2011), and the Bayesian LASSO (Park and Casella 2008, Yuan and Lin 2005, Kyung *et al*. 2010).

In this paper we present a method for that addresses the two major components of model selection, variable selection and estimation of coefficients, in a linked fashion. To select a variable, we test whether the regression coefficient corresponding to that variable is zero or not, based on the observed test coefficient. Then we estimate the regression coefficient based on a conditional likelihood that takes into account the result of the hypothesis test. In particular, we incorporate the information of whether the variable was found to be significant or not while constructing the conditional likelihood for estimation. This idea is an application of the conditional likelihood approach for overcoming the "winner's curse" (Lohmueller *et al*. 2003, Znollner and Pritchard 2007), or "significance bias" (Ghosh *et al*. 2008), in genomewide association studies. The conditional likelihood approach yields a non-convex penalty function that can be used in the penalized likelihood framework for coefficient shrinkage. Thus, our proposed penalty function has a natural motivation

based on the selection procedure involving the test coefficient. We call the resulting method Test Coefficient Shrinkage or TCS. We extend TCS to high-dimensional regression problems. We use real data examples and simulations to illustrate the performance of TCS and to compare it with other popular penalized regression methods.

3. THE TCS METHOD

Consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where the response $\mathbf{y}$ is a $n \times 1$ vector and the design matrix $\mathbf{X}$ is of order $n \times p$. So the data consists of $(y_i, \boldsymbol{x_i})$, $i = 1, ..., n$ where y_i is the response and $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})'$ is the vector of predictor values for the ith observation in the sample. Let ϵ_i, $i = 1, ..., n$ be independently and identically distributed as $N(0, \sigma^2)$. We assume, without loss of generality, that the predictors are standardized and the response is centered so that $\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = 1$, $j = 1, ..., p$, and $\sum_i y_i = 0$.

Our goal is to find the best linear fit to the data in terms of expected prediction error, $E_{\hat{\beta}}\left[(y - \mathbf{x}'\hat{\beta})^2\right]$. With this objective in mind, we propose a shrinkage method based on a penalized likelihood. In the linear regression setup, the penalized log-likelihood assumes the form

$$l_{penalized}(\beta, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) - \sum_j p_\lambda\left(|\beta_j|\right), \quad (1)$$

where $\lambda \geq 0$ is the complexity or tuning parameter and the amount of shrinkage is dictated by λ. We suggest a novel penalty function

$$p_\lambda(|\beta_j|) = \log[\Phi(-\lambda - \mu_j) + \Phi(-\lambda + \mu_j)],$$

where we define $\mu_j = \dfrac{\beta_j}{SE(\hat{\beta}_j^0)}$, with $\hat{\beta}_j^0$ being the ordinary least squares (OLS) estimate and $SE(\hat{\beta}_j^0)$ its standard error. Φ is the cumulative distribution function for a standard normal variate. To obtain penalized maximum likelihood estimate of β for a given λ, we maximize (1) with respect to (β, σ^2). We choose λ to minimize an estimate of the expected prediction error.

The motivation for this penalty function stems from the testing-based selection of regression coefficients. Let us consider the situation where we have only one predictor x,

$$y_i = \beta x_i + \epsilon_i,\ i = 1, ..., n,$$

with $\sum_i x_i = 0, \sum_i x_i^2 = 1$, and $\sum_i y_i = 0$. We assume that the error variance σ^2 is known. Our goal is to build a model for y. We first test whether x has any predictive ability, that is, we test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ on the basis of the test statistic $Z = \dfrac{\hat{\beta}^0}{\sqrt{Var(\hat{\beta}^0)}} = \dfrac{\hat{\beta}^0}{\sigma}$ where $\hat{\beta}^0$ is the OLS estimator. We reject H_0 if observed value of Z is greater in magnitude than some prespecified quantity λ. If we are unable to reject the null hypothesis based on our sample, then we predict y using $\bar{y}$. Whereas, if we are able to reject the null hypothesis, we need an estimate for β to be able to predict y. We construct a conditional likelihood for β accounting for the fact that the null hypothesis has been rejected.

We note that $Z \sim N(\mu, 1)$, where $\mu = \dfrac{\beta}{\sigma}$. Thus the conditional likelihood for μ is

$$L_c(\mu) = \frac{p_\mu(z)}{P(|Z| > \lambda)} = \frac{\phi(z - \mu)}{\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)}. \quad (2)$$

We maximize $L_c(\mu)$ with respect to μ to derive $\tilde{\mu}$, the conditional maximum likelihood estimate of μ. Note that $\tilde{\mu}$ may be regarded as a penalized likelihood estimator with penalty $\log [\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)]$, since

$$\begin{aligned} \tilde{\mu} &= \text{argmax}_\mu\, L_c(\mu) \\ &= \text{argmax}_\mu \log L_c(\mu) \\ &= \text{argmax}_\mu \{\log \phi(z - \mu) - \log[\Phi(-\lambda - \mu) \\ &\quad + \Phi(-\lambda + \mu)]\}. \end{aligned}$$

Thus we call $\log [\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)]$ the Test Coefficient Shrinkage (TCS) penalty function. From a Bayesian point of view $\tilde{\mu}$ can be interpreted as the Bayes posterior mode under the (improper) prior $p(\mu) \propto [\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)]^{-1}$.

The estimate $\tilde{\mu}$ can easily be converted into an estimate $\tilde{\beta} = \tilde{\mu}\sigma$ for β. Thus we define $\tilde{\beta}$ according to the outcome of the hypothesis test as

$$\tilde{\beta} = \begin{cases} 0, & \text{if } H_0 \text{ accepted} \\ \tilde{\mu}\sigma, & \text{if } H_0 \text{ rejected} \end{cases}. \quad (3)$$

Thus $\tilde{\beta}$ is a thresholding rule shrinking the OLS estimate $\hat{\beta}^0$ to zero if we fail to reject the null hypothesis; and shrinking it to some non-zero value if we reject the null. The larger the value of λ, the greater is the shrinkage.

We note that the estimate $\tilde{\beta}$ is not a continuous function of the OLS estimate $\hat{\beta}^0$. If we believe continuity of a penalty function to be a desirable property, as advocated by Fan and Li (2001), we can define a modified estimate

$$\tilde{\beta}^* = \tilde{\mu}\sigma. \quad (4)$$

$\tilde{\beta}^*$ is purely a shrinkage estimator being no longer subject to thresholding and variable selection. In the general $n > p$ setting, we apply the TCS penalty, $p_\lambda(|\beta|)$ = log $[\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)]$ with $\mu = \frac{\beta}{SE(\hat{\beta}^0)}$, to the penalized log-likelihood in (1) to derive the shrinkage estimate of β for a given λ.

3.1 High-dimensional Setup

The direct application of TCS in the $p > n$ regime is not possible since the penalty term involves standard errors of OLS estimates, which are not defined in the high-dimensional setup. To address this, we make use of an iterative procedure in which we obtain shrinkage estimate of one regression coefficient at a time using the univariate TCS procedure. In particular, we obtain estimates $\hat{\beta}^0$ and $SE(\hat{\beta}^0)$ from univariate regression, calculate $z = \frac{\hat{\beta}^0}{SE(\hat{\beta}^0)}$, then derive the shrinkage estimate for the regression coefficient β by applying either the thresholded or the non-thresholded TCS penalty. For a prespecified value of λ, the thresholded estimate of β is

$$\tilde{\beta} = \begin{cases} 0, & |z| \le \lambda \\ \tilde{\mu}SE(\hat{\beta}^0), & |z| > \lambda \end{cases}$$

and the non-thresholded estimate of β is $\tilde{\beta}^* = \tilde{\mu}SE(\hat{\beta}^0)$, where $\tilde{\mu}$ is as defined before.

We have developed an iterative procedure in which we apply the univariate regression with residuals as the response variable. This idea is similar to the coordinate-wise descent algorithm (Friedman *et al*. 2007) for convex optimization problems. At each step of the iterative procedure we start with an initial estimate of β, update it one regression coefficient at a time till we loop through all the predictors, then repeat this process with the updated estimate as the initial value for the next iteration cycle. Let $\tilde{\beta}^{(k-1)}$ be the initial estimate of β at the k^{th} step of the procedure. We use the subscript $-j$ to signify that the j^{th} column or component is left out. For the j^{th} predictor we regress $\mathbf{y} - \mathbf{X}_{-j}\tilde{\beta}_{-j}$ on $\mathbf{x}_j$, get $\tilde{\beta}$ by shrinking $\hat{\beta}^0$, and replace $\tilde{\beta}_j^{(k-1)}$ by $\tilde{\beta}$. We then move on to the next predictor. After we have cycled through all the predictors, one predictor at a time, we finally have $\tilde{\beta}^{(k)}$. We then start the $(k + 1)^{th}$ step with $\tilde{\beta}^{(k)}$ as the initial estimator.

Getting $\tilde{\beta}$ from $\hat{\beta}^0$ is very fast and the computation of the residuals is also quick because only one component gets updated each time, which makes the whole iterative procedure very efficient. For the first step we define $\tilde{\beta}^{(0)}$ as a vector of $\tilde{\beta}$'s obtained by shrinking the marginal regression coefficients. For a particular step of the iteration we have to loop through all the predictors, one predictor at a time, but we need to decide on an order in which to loop through the variables. We consider the predictors in decreasing order of magnitude of the initial estimator for that iteration. We decided on this order so as to eliminate randomness from the iterative process, and to ensure that we end up with the same estimate of β every time we run the procedure for a particular dataset. Also, we need a stopping rule for the iterative procedure. Tseng (1988) established that coordinate-wise algorithms fo convex optimization problems converge to their optimal solution under separability of the penalty function. For TCS penalty, the iterative procedure does not enjoy such convergence properties since the penalty function is not convex. Thus we continue the iteration for 50 steps and then choose the $\tilde{\beta}$ which gives the minimum training error in the last 10 steps. This strategy is based on the empirical observation that the training and the test errors have similar paths over the iteration steps,

which led us to believe that training error can serve as an adequate stopping rule criterion.

3.2 Data Analysis

We consider two different datasets and apply commonly-used penalization techniques and the proposed TCS method to study how each method performs in finding the best linear fit to the data. We first standardize the predictors to have 0 mean and unit variance. We randomly split the data into a training set and a test set. Every method, other than OLS, involves a tuning parameter which is chosen to minimize an estimate of the prediction error based on 10-fold cross-validation. We follow a "one-standard error" rule, in which the least complex model is chosen whose estimated prediction error is one standard deviation above the minimum estimated prediction error. This conservative approach follows from the thought that prediction error is estimated with some error. We fix a, tuning parameter in the SCAD function, at 3.7. We use a two-dimensional grid search for elastic net and adaptive lasso tuning parameters. The final chosen model is then applied to the test set to assess its prediction error.

3.2.1 Prostate Cancer Data

The data on prostrate cancer has been widely used in the variable selection literature. The data comes from a study conducted by Stamey *et al*. (1989). The dependent variable is *lpsa*: level of prostate specific antigen in blood serum. The relevant covariates are a number of clinical measures in men about to receive a radical prostatectomy: *lcavol* (log cancer volume), *lweight* (log prostate weight), *age*, *lbph* (log of the amount of benign prostatic hyperplasia), *svi* (seminal vesicle invasion), *lcp* (log of capsular penetration), *gleason* (Gleason score), and *pgg45* (percent of Gleason scores 4 or 5). The dataset has 97 observations. We randomly split the data into a training set of size 67 and a test set of size 30.

3.2.2 National Family Health Survey 3

The 2005-06 National Family Health Survey (NFHS)-3 is the third round in a series of national surveys to provide estimates on key indicators of maternal and child health in India at the state- and national-level. NFHS-3 is a household survey; individual interviews are conducted as well with women aged 15-49 and men aged 15-54. Details on the study design, sampling, questionnaires and data collection can be found elsewhere (http://www.rchiips.org/nfhs/nfhs3.shtml). For our analysis we used data on height-for-age of children between 0 and 35 months old and alive at the time of the interview. We further restricted our analysis to only the last two children who were born as singletons to ever-married women aged 15-49. We examined correlation between the proportion of children stunted, defined as more than two standard deviations below theWorld Health Organization-determined median scores by age and gender, in a region and the socio-economic and demographic characteristics of that region. We defined regions on the basis of state and rural/urban classifications. For our analysis, we considered predictors involving sanitation, infrastructure and general well-being of the region; women's nutritional status, their educational attainment, autonomy, employment and exposure to media. Specifically, we considered proportion of households in the region defecating in the open (*opendefecation*), proportions of households with electricity (*haselectricity*), improved drinking water (*hasdrinkingwater*) and television (*hastele*), average maternal height (*momheight*) and BMI (*momBMI*) in the region, proportions of women who have primary (*wprimaryeducated*), secondary (*wsecondaryeducated*) and higher (*whighereducated*) education, proportion of women who are illiterate (*williterate*), proportions of women who at least once a week read newspaper (*wreadnewspaper*), listen to radio (*wlistenradio*), and watch television (*wwatchtele*), proportion of women who have not worked in the past 12 months (*wnotwork*), proportions of women who have say in matters of one's own health care (*wsayinhealth*) and spending of money earned by husband (*wsayinmoney*) and the proportion of women who have not heard of oral rehydration salts (ORS) packets for treating children with diarrhea (*wORSnotknow*) as predictors in the linear regression model. We had 58 data points (rural and urban sectors for each of 29 states) and we split the data into a training set of size 40 and a test set of size 18.

3.3 Simulations

We use a simulation study to compare OLS, ridge, LASSO, SCAD, elastic net, and TCS in the usual $n > p$ situation where n and p are the number of

observations and predictors, respectively. We simulated 100 datasets consisting of n observations from the model

$$\mathbf{y}^{n \times 1} = \mathbf{X}^{n \times p}\boldsymbol{\beta}^{p \times 1} + \in^{n \times 1}, \in \sim N_n(0, \sigma^2 I_n),$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The columns of $\mathbf{X}$ are standard normal. The correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $\rho^{|i-j|}$ with $\rho = 0.5$. This numerical example has been used in several publications (Fan and Li 2001, Tibshirani 1996, Zou and Hastie 2005) to discuss relative merits of different variable selection and shrinkage procedures. First we choose $\sigma = 3$ and $n = 40$. Then we reduce σ to 1 and finally increase the sample size to 60. Following Zou and Hastie (2005), for each simulation we have a training set, an independent validation set, and an independent test set. We use the training set to fit the data, the validation set to choose the tuning parameter(s), and the test set to estimate mean squared error. We compute RMSE (relative mean squared error) for each procedure as the mean squared error of the procedure relative to OLS. We use the median of the relative mean squared errors over 100 datasets (MRMSE) to compare performance of different methods. We also compare the performance of an oracle estimator to OLS.

To judge the performance of our proposed method in the $p > n$ situation, we simulate data from the same linear model but with fewer observations than predictors. We set $n = 100$, $p = 1000$, and $\sigma = 1$. The non-zero β's constitute a random sample from normal distribution with mean zero and variance σ_β^2. We examine the performance of TCS, both the thresholded and the non-thresholded versions, and compare them with ridge, LASSO, and elastic net over a range of simulation setups, generated by varying the number of non-zero predictors from 5 to 1000 and σ_β from 0.1 to 2. In our simulations, we cover the two extreme situations of having very few big predictors and many small ones. Through our detailed numerical exercise we wish to verify the empirical observations about the performance of the various variable selection techniques, and to understand in which cases our method works best. For each simulation setup we compute the estimated test error for all three methods over 100 replications and judge their relative performance on the basis of the average test error. For each replication we have a training set of size n to fit the model over a range of values of the tuning parameter, a validation set of size n on the basis of which we decide on the value of the tuning parameter, and a test set of size 10,000 to estimate the test error of the fitted model. We standardize the covariates and center the response variable before analysis.

4. RESULTS

Table 1 shows the results for the prostate cancer data for different variable selection and shrinkage methods. We see that ridge regression reduces OLS test error only by a small margin, whereas LASSO offers substantial improvement over OLS. Test errors for SCAD, and elastic net are also similar. Our proposed penalty has the lowest test error estimate with a small standard error.

Table 1. Estimated coefficients and test error results for prostate data

Term	OLS	Ridge	LASSO	SCAD	Elastic net	TCS
Intercept	2.480	2.472	2.477	2.483	2.479	2.478
lcavol	0.680	0.337	0.550	0.810	0.607	0.800
lweight	0.305	0.234	0.205	0.109	0.255	0.044
age	–0.141	–0.016	0.000	0.000	0.000	0.000
lbph	0.210	0.145	0.059	0.006	0.120	0.024
svi	0.305	0.205	0.129	0.000	0.185	0.019
lcp	–0.288	0.052	0.000	0.000	0.000	–0.004
gleason	–0.021	0.051	0.000	0.000	0.000	0.005
pgg45	0.267	0.112	0.034	0.000	0.072	0.021
Test error	0.586	0.552	0.483	0.473	0.478	0.461
Std error	0.184	0.174	0.155	0.132	0.138	0.133

For the NFHS-3 data there were several correlated predictors, with pairwise correlations as high as 0.99. Table 2 presents the results for different variable selection and shrinkage methods applied to the NFHS-3 data on stunting in children under 3. OLS performs poorly, SCAD and LASSO result in very little improvement over OLS. Ridge regression outperforms the LASSO due to the presence of many highly correlated predictors. Elastic net offers marginal improvement over ridge regression. The TCS penalty has the smallest prediction error and standard error as well.

Table 2. Estimated coefficients and test error results for NFHS-3 data on stunting in children under 3

Term	OLS	Ridge	LASSO	SCAD	Elastic net	TCS
Intercept	0.377	0.381	0.380	0.379	0.382	0.382
opendefecation	0.046	0.005	0	0	0.012	0.001
haselectricity	0.076	–0.003	0	0	0	–0.001
hasdrinkingwater	0.030	0.001	0	0	0	0
hastele	–0.006	–0.004	0	0	0	–0.004
momheight	–0.063	–0.003	0	0	–0.007	–0.001
momBMI	–0.026	–0.005	–0.009	–0.008	–0.016	–0.001
wprimaryeducated	–0.067	–0.005	0	0	–0.019	–0.025
wsecondaryeducated	0.086	–0.004	0	0	0	–0.005
whighereducated	–0.018	–0.004	–0.002	–0.001	–0.009	–0.001
williterate	0.048	0.004	0	0	0	0.013
wreadnewspaper	0.004	–0.004	–0.011	–0.010	0	–0.001
wlistenradio	0.026	0	0	0	0	0
wwatchtele	–0.062	–0.004	0	0	0	–0.004
wnotwork	0.034	0	0	0	0	0
wsayinhealth	0.024	0	0	0	0	0
wsayinmoney	0.015	0.004	0	0	0.014	0.001
wORSnotknow	–0.022	0.002	0	0	0	0
Test error	0.515	0.314	0.471	0.488	0.304	0.285
Std error	0.133	0.090	0.112	0.116	0.097	0.075

Table 3 shows the results for the simulations in the $n > p$ setup. The median of the relative mean squared errors over 100 datasets (MRMSE) for different procedures are reported in the table. Ridge regression performs poorly in all situations. Mean squared error for LASSO is smaller than that of ridge. When the noise level is high and sample size is small, *i.e.*, $\sigma = 3$ and $n = 40$, LASSO performs reasonably well, but its performance deteriorates quickly as the signal to noise ratio increases, *i.e.*, as we decrease σ or increase n. For $n = 60$, $\sigma = 3$, adaptive LASSO performs the best, with TCS and SCAD performing similarly well. Table 3 suggests that the proposed penalty performs remarkably well and is indeed a worthy competitor.

Table 3. Results for simulated numerical example in $n > p$ scenario

	MRMSE(%)		
Method	$n = 40$, $\sigma = 3$	$n = 40$, $\sigma = 1$	$n = 60$, $\sigma = 1$
Oracle	29.42	29.42	34.93
Ridge	86.41	100.00	100.00
Lasso	67.59	68.17	72.15
Scad	63.77	37.93	41.66
Elastic net	58.01	49.72	52.80
TCS	64.02	35.56	41.45

Table 4. Simulation results for $p > n$

	Number of non-zero predictors					
σ_β	5	10	50	100	500	1000
2	1.11^{T}	1.19^{T}	117.11^{E}	319.62^{E}	1699.79^{R}	379.85^{R}
	1.12^{N}	1.22^{N}	117.50^{L}	335.24^{L}	1732.26^{E}	3446.90^{E}
	1.29^{E}	1.87^{E}	136.87^{N}	335.88^{R}	1843.06^{N}	3628.62^{N}
	1.50^{L}	2.24^{L}	147.90^{T}	352.74^{N}	1969.42^{L}	4019.68^{L}
	17.16^{R}	32.69^{R}	168.85^{R}	383.04^{T}	2014.58^{T}	4032.96^{T}
1	1.17^{T}	1.37^{T}	31.08^{E}	80.90^{E}	425.84^{R}	845.80^{R}
	1.19^{N}	1.41^{N}	31.23^{L}	84.32^{L}	433.96^{E}	861.47^{E}
	1.28^{E}	1.75^{E}	36.39^{N}	84.85^{R}	460.77^{N}	906.27^{N}
	1.44^{L}	2.02^{L}	39.64^{T}	88.17^{N}	494.16^{L}	1004.66^{T}
	5.14^{R}	9.03^{R}	43.06^{R}	96.15^{T}	511.40^{T}	1006.20^{L}
0.5	1.21^{T}	1.50^{T}	9.22^{E}	21.27^{E}	107.34^{R}	212.31^{R}
	1.22^{N}	1.52^{N}	9.24^{L}	22.09^{R}	109.18^{E}	215.93^{E}
	1.24^{E}	1.58^{E}	10.53^{N}	22.14^{L}	115.30^{N}	227.68^{N}
	1.33^{L}	1.70^{L}	11.12^{T}	22.85^{N}	123.88^{L}	252.74^{L}
	2.12^{R}	3.09^{R}	11.62^{R}	24.82^{T}	127.01^{T}	255.05^{T}
0.1	1.05^{T}	1.09^{T}	1.47^{N}	1.93^{N}	5.39^{R}	9.59^{R}
	1.05^{N}	1.10^{N}	1.49^{T}	1.96^{R}	5.53^{E}	9.83^{E}
	1.06^{E}	1.10^{E}	1.50^{L}	1.98^{L}	5.60^{N}	10.08^{N}
	1.06^{L}	1.11^{L}	1.51^{E}	1.99^{T}	6.00^{L}	11.06^{T}
	1.15^{R}	1.19^{R}	1.53^{R}	1.99^{E}	6.04^{T}	11.12^{L}

R: Ridge, L: Lasso, E: Elastic net, N: Non-thresholded TCS, T: Thresholded TCS

Table 4 shows the results for the simulations in the $p > n$ situation. The different columns of the table are for the number of non-zero predictors and the rows signify different values of σ_β. For each cell corresponding to a particular number of non-zero predictors and a value of σ_β we have recorded the average test error over 1000 datasets for TCS,

thresholded and nonthresholded versions of it, ridge, LASSO, and elastic net in increasing order of magnitude.

If we are interested in the performance of a particular method we can trace the method across the grid and its position in various cells describes its overall performance. For example, if we compare the LASSO and ridge paths across the grid, we observe that in situations where we have fewer non-zero predictors than samples LASSO does better than ridge, while ridge does better than LASSO when the number of non-zero predictors is greater than the sample size. In the latter situation, LASSO is at a disadvantage since it cannot choose more predictors than the number of observations. In the case where there are equal number of nonzero predictors and observations, LASSO and ridge are very close, but LASSO does better when the non-zero coefficients are big whereas ridge outperforms LASSO when the nonzero coefficients are small in magnitude. Elastic net performs remarkably well when there are more non-zero predictors than number of observations. The table illustrates the complementary nature of the methods under study, and suggests that no method outperforms all other methods all situations. It is important to understand the strengths and weaknesses of different methods when applying them in different situations.

The first two columns of the table show that thresholded TCS almost always has the smallest test error among all the methods when β is truly sparse. But it performs poorly in situations where we have more non-zero predictors than samples. Though nonthresholded TCS is less frequently the best choice, it performs similarly to thresholded TCS in sparse situations, and performs acceptably in many situations when thresholded TCS performs poorly, often dominating the performance of the LASSO.

5. DISCUSSION

We used a conditional likelihood approach to propose a new regression penalty function, and showed that the proposed method compares favorably with other penalized regression procedures on a number of real and simulated datasets. By implementing a significance threshold as a tuning parameter for individual predictors, our procedure can create a sparse set of predictors with non-zero coefficient estimates. In addition, the proposed method can be used to obtain estimates when the number of predictors is greater than the sample size. When combined with cross-validation, our procedure is an automatic variable selection and coefficient shrinkage approach.

In the high-dimensional setting, we have judged the performance of the estimators in terms of prediction error. But the number of β-coefficients wrongly predicted as non-zero, better known as false-positives, or the number of β-coefficients wrongly predicted as zero, known as false-negatives, can also serve as criteria for judging the performance of these methods. Ridge regression and the non-thresholded version of proposed TCS penalty select all the candidate predictors and these measures will not mean much in these cases. But this criteria might be informative for comparing the performance of LASSO with the thresholded-TCS penalty.

Several issues warrant further research. In the usual $n > p$ setting we have used the TCS penalty in the penalized likelihood framework and implemented coefficient shrinkage, but the penalty is not a thresholding one and as a result does not participate in variable selection. We can implement a thresholded version of this in the $n > p$ scenario by investigating each predictor individually as we have done in the high-dimensional case. But we would have to deviate from the penalized likelihood framework where we optimize a single objective function over all the coefficients simultaneously. In the high-dimensional case we describe an iterative procedure that examines each predictor separately. So, the method for the highdimensional model is not an automatic extension of the method in the $n > p$ situation. For future work in this area it would be important to build a unified algorithm that can be applied to both the situations. Also, additional comparisons to other approaches would help us understand the method better.

ACKNOWLEDGEMENTS

Portions of the work were conducted while Drs. Wright and Ghosh were at the University of North Carolina at Chapel Hill. Part of Dr. Ghosh's research was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. The work of Dr. Andrew Nobel was supported in part by NSF Grant DMS 0907177. The work of Dr. Fei Zou was supported in part by NIH grant R01 GM074175.

REFERENCES

Bach, F.R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 33-40.

Bondell, H.D. and Reich, B.J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, **64,** 115-123.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37,** 373-384.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when *p* is much larger than *n*. *Ann. Statist.*, **35,** 2313-2351.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.,* 407-451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

Fan, J. and Lv, J. (2006). Sure independence screening for ultra-high dimensional feature space. *Arxiv preprint math.ST/0612857*.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics,* 109-135.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1,** 302-332.

Ghosh, A., Zou, F. and Wright, F. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *The Amer. J. Human Genet.*, **82,** 1064-1074.

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

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12,** 55-67.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Anal.*, **5,** 369-411.

Lohmueller, K., Pearce, C., Pike, M., Lander, E. and Hirschhorn, J. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.,* **33,** 177-182.

Park, T. and Casella, G. (2008). The bayesian lasso. *J. Amer. Statist. Assoc.*, **103,** 681-686.

Radchenko, P. and James, G. (2008). Variable inclusion and shrinkage algorithms. *J. Amer. Statist. Assoc.,* **103,** 1304-1315.

Radchenko, P. and James, G.M. (2011). Improved variable selection with forward-lasso adaptive shrinkage. *Ann. Appl. Statist.*, **5,** 427-448.

Stamey, T., Kabalin, J. and Ferrari, M. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. III. Radiation treated patients. *J. Urol.,* **141,** 1084-1087.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Series B (Methodological),* 267-288.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics,* **23,** 1486.

Wasserman, L., Roeder, K. and Pittsburgh, P. (2007). Multi-Stage Variable Selection: Screen and Clean. *Arxiv preprint arXiv*:0704.1139.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.*, 100.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc., Series B*, **68,** 49-67.

Znollner, S. and Pritchard, J. (2007). Overcoming the winners curse: estimating penetrance parameters from case-control data. *The Amer. J. Human Genet.*, **80,** 605-615.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101,** 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc., Series B*, **67,** 301-320.