A Confidence Region Approach to Tuning for Variable Selection

Funda Gunes and Howard D. Bondell
Department of Statistics
North Carolina State University

Abstract

We develop an approach to tuning of penalized regression variable selection methods by calculating the sparsest estimator contained in a confidence region of a specified level. Because confidence intervals/regions are generally understood, tuning penalized regression methods in this way is intuitive and more easily understood by scientists and practitioners. More importantly, our work shows that tuning to a fixed confidence level often performs better than tuning via the common methods based on AIC, BIC, or cross-validation (CV) over a wide range of sample sizes and levels of sparsity. Additionally, we prove that by tuning with a sequence of confidence levels converging to one, asymptotic selection consistency is obtained; and with a simple two-stage procedure, an oracle property is achieved. The confidence region based tuning parameter is easily calculated using output from existing penalized regression computer packages.

Our work also shows how to map any penalty parameter to a corresponding confidence coefficient. This mapping facilitates comparisons of tuning parameter selection methods such as AIC, BIC and CV, and reveals that the resulting tuning parameters correspond to confidence levels that are extremely low, and can vary greatly across data sets. Supplemental materials for the article are available online.

Keywords: Adaptive LASSO, Confidence region, Penalized regression, Tuning parameter, Variable selection

1 Introduction

Recently, penalized regression methods for variable selection have become popular. These methods continuously shrink the model parameters and perform selection by setting coefficients to zero. They control the shrinkage of the model parameters by a non-negative regularization parameter and, as this parameter increases, the regression coefficients shrink continuously. Least absolute

shrinkage and selection operator, LASSO, (Tibshirani, 1996), smoothly clipped absolute deviation, SCAD, (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), and octagonal shrinkage and clustering algorithm, OSCAR, (Bondell and Reich, 2008) are some examples of penalized regression techniques.

An important issue in using a penalized regression method is choosing the value of the regularization parameter, or tuning the procedure. Asymptotic results for methods such as SCAD and adaptive LASSO are available by letting the tuning parameter change at an asymptotic rate. However, in finite samples, the value of the tuning parameter needs to be chosen, and which criterion to use in practice is a difficult question.

Consider a generalized linear model where $\mathbf{Y}_{(n\times 1)}$ is the response vector, $\mathbf{X}_{(n\times p)} = [\mathbf{X}_1, ..., \mathbf{X}_p]$ is the design matrix and $\boldsymbol{\beta}_{(p\times 1)} = (\beta_1, ..., \beta_p)^t$ is the vector of model coefficients. Suppose the data $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$ is collected independently. Conditioning on $\mathbf{x}_i$, y_i has a density function $f(g(\mathbf{x}_i^t\boldsymbol{\beta}), y_i, \phi)$, where g is the known link function and ϕ is a possible nuisance parameter. A general form of the penalized regression problem is given by

$$\tilde{\boldsymbol{\beta}} = arg\ min_\beta\{-2\sum_{i=1}^{n}\log f(g(\mathbf{x}_i^t\boldsymbol{\beta}), y_i, \phi) + p_\lambda(\boldsymbol{\beta})\}, \tag{1}$$

where $\lambda \geq 0$ is the regularization parameter, and $p_\lambda(\boldsymbol{\beta})$ specifies the penalization on the regression coefficients. For example, LASSO uses the L_1 penalty, $p_\lambda(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p} \mid \beta_j \mid$, and adaptive LASSO uses weighted L_1 penalty, $p_\lambda(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p} \mid \beta_j \mid / \mid \hat{\beta}_j \mid$ where $\hat{\boldsymbol{\beta}}$ is the MLE. Note that for the adaptive LASSO case $p_\lambda(\boldsymbol{\beta})$ also depends on the data via $\hat{\boldsymbol{\beta}}$. As typically the case, to simplify the notation we will omit the dependence on the data, and write $p_\lambda(\boldsymbol{\beta})$. Often it is the case that $p_\lambda(\boldsymbol{\beta}) = \lambda p(\boldsymbol{\beta})$, and this form is considered in the remainder of the paper.

Various criteria can be applied to select the tuning parameter, for example, the C_p statistic (Mallows, 1973), Akaike information criterion, AIC, (Akaike, 1973), Bayesian information criterion, BIC, (Schwarz, 1978), generalized information criterion, GIC, (Konishi and Kitagawa, 1996), or K-fold cross-validation (Breiman and Spector, 1992). For a review of some of these selection criteria see Hastie et al. (2001).

Given the multitude of options, it can be difficult to decide which to use. Often these methods

can yield different results for the same problem. Even when one of the existing methods is selected, it can be hard to justify the reason for this choice to scientists and practitioners. To this end, we propose a tuning selection method based on the standard notion of confidence sets, a well accepted construction in statistical inference.

Consider a simple situation with $n = 20$ observations from a linear regression model with $p = 5$ predictors, and normally distributed errors having variance 1. Suppose that none of the predictors were important, i.e. the null model was true. In this setting, one would hope to choose the null model with high probability. Letting $0 < \alpha < 1$ and constructing a $1 - \alpha$ level joint confidence region and not rejecting the null model if the point $\boldsymbol{\beta} = 0$ were contained in the region will result in a family-wise error rate of α. Suppose that we use a penalized regression. We would like our tuning method to perhaps match this Type I error under the null model, while also having good selection properties when there truly are relevant predictors. We simulated this null model scenario 1000 times and used the adaptive LASSO to perform variable selection. Tuning via BIC selected the null model only 73% of the time for a 27% Type I error, while tuning via AIC selected the null model only 48% of the time, for a 52% Type I error. In terms of hypothesis testing, selecting via AIC would correspond to using a level of 48%, which in practice, may make an applied scientist uncomfortable.

We propose to choose as the estimate the point within a confidence region having the sparsest representation, where sparsity is measured via some complexity measure, $p(\boldsymbol{\beta})$. It is assumed that $p(\boldsymbol{\beta}) \geq 0$, and equality holds if and only if $\boldsymbol{\beta} = 0$. Under this assumption, if the constructed confidence region contains the origin, then the point $\boldsymbol{\beta} = 0$ will always be the chosen solution. Note that all typical penalty functions satisfy the above. Thus defining the approach in this manner will automatically maintain the desired family-wise error rate under the null model. This proposed formulation has an equivalent representation as a penalized regression with tuning parameter completely determined by choice of confidence level. We show that tuning in this manner is not only intuitively appealing and maintains the desired family-wise error rate under the null model, but when used with adaptive LASSO it can achieve asymptotic selection consistency for a sequence of confidence levels chosen appropriately to converge to one. Moreover, the proposed method enjoys

excellent finite sample behavior.

There are several advantages of tuning based on confidence regions. First, it is intuitive and easy to interpret for scientists and practitioners. Although asymptotic properties of selection criteria such as AIC, BIC, and CV have been studied (Stone, 1977; Shibata, 1981; Nishii, 1984; Shao, 1997; Wang et al., 2007), the interpretation of the tuning parameter as a confidence level is more natural to a non-statistician. Second, it has the ability to be used with a large variety of statistical methods where confidence regions can be created for model coefficients. Third, a default value of the tuning parameter can be chosen in practice, such as an a priori choice of say 90%, 95%, or 99% confidence level. But, perhaps the most important advantage is that, although asymptotic selection consistency theory exists for other methods of tuning, such as BIC (Shao, 1997; Wang et al., 2007), the proposed tuning method has shown strong finite sample selection properties.

In addition, based on the idea of tuning via confidence regions, any tuning parameter for the penalized regression can be mapped to a corresponding confidence level. This implied confidence level corresponding to standard methods of tuning such as AIC, BIC, and cross-validation can be extremely low, and can vary greatly across data sets.

The remainder of this paper proceeds as follows. Section 2 describes automatic variable selection via confidence regions in greater detail, which includes linear models and generalized linear models as special cases. Section 3 shows that variable selection consistency can be achieved for the proposed tuning method. Section 4 reports the Monte Carlo simulation results and compares our method with existing methods. Section 5 introduces a 2-stage estimation procedure and shows that it obtains the oracle property, and Section 6 presents a discussion. All proofs are given in the Appendix included with the online supplementary materials.

2 Tuning via Confidence Regions

2.1 The approach

The proposed approach is to build a confidence region for model parameters using some standard technique, such as likelihood-based confidence regions or Wald-based confidence regions. Among

the points in the confidence region, the model with the most sparse representation is chosen, where the sparsity is measured by some criterion such as the L_1 norm.

Let $p(\boldsymbol{\beta})$ specify the form of the criterion to measure the sparsity of the model coefficients. For some function $H(s,t)$, let the set of all $\boldsymbol{\beta}$ such that $H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}) \leq C_\alpha$ determine a $1-\alpha$ confidence region for $\boldsymbol{\beta}$, where C_α is chosen to yield the correct coverage probability. Then, the estimated model coefficients, $\tilde{\boldsymbol{\beta}}$, can be obtained by solving the following minimization problem:

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= argmin_\beta \; p(\boldsymbol{\beta}) \\ \text{subject to } & H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}) \leq C_\alpha \end{aligned} \tag{2}$$

Assuming convexity of both $p(\boldsymbol{\beta})$ and $H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta})$ in $\boldsymbol{\beta}$, the constrained minimization problem in (2) is equivalent to the following Lagrange formulation of optimization:

$$\tilde{\boldsymbol{\beta}} = argmin_\beta \; \{p(\boldsymbol{\beta}) + \theta_\alpha H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta})\}, \tag{3}$$

where θ_α is the Lagrange multiplier corresponding to the constraint in (2).

The problem given by (3) can also be represented by the following more common way of representation of penalized regression problems:

$$\tilde{\boldsymbol{\beta}} = argmin_\beta \; \{H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}) + \lambda_\alpha p(\boldsymbol{\beta})\}, \tag{4}$$

where $\lambda_\alpha = 1/\theta_\alpha$. Notice that (2) is equivalent to (4) in the sense that there exists a one to one function between C_α and λ_α. Therefore, for a specific value of C_α, the solution of (2) can be obtained using methods designed to solve (4). Thus, the proposed approach yields a tuning method for typical penalized regression methods, as the tuning parameter is fully determined via the initial specification of the confidence level α. In practice, this is most often taken to be the 95% confidence region. Hence, the tuning parameter has a natural default value that is familiar to practitioners. Note that if $p_\lambda(\boldsymbol{\beta}) \neq \lambda p(\boldsymbol{\beta})$, this duality would not hold, and the tuning parameter would not be fully determined by the confidence region level.

One common way to construct confidence region is to invert the likelihood ratio statistic. Suppose we wish to test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$ vs $H_a : \boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ fully specifies the regression model. Let $L(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y})$ be the likelihood function and $\hat{\boldsymbol{\beta}}$ be the usual MLE of $\boldsymbol{\beta}$ with no restrictions on the regression coefficients. For simplicity, let the log of the likelihood ratio test statistic, $log(\frac{L(\mathbf{X}\boldsymbol{\beta^*},\mathbf{Y})}{L(\mathbf{X}\hat{\boldsymbol{\beta}},\mathbf{Y})})$, be denoted by $\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta^*})$.

The likelihood ratio test rejects the null hypothesis for large values of $-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta^*})$, say $-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta^*}) \geq C_\alpha$ so that $P_{H_0}(-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta^*}) \geq C_\alpha) = \alpha$. Inverting the test yields a $1 - \alpha$ confidence set for $\boldsymbol{\beta}$ with the following form:

$$\{\boldsymbol{\beta} : \ -2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}) \leq C_\alpha\} \tag{5}$$

For some models, $-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}^*)$ may have a form that can be expressed via a known probability distribution. In this case, C_α can be determined exactly. However, if $-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}^*)$ does not have a known distributional form or if the exact distributional form is difficult to work with, a chi-square distribution can be used as an asymptotic approximation.

2.2 Implied Confidence Level

Given the duality between choice of confidence level and the tuning parameter, any method of choosing λ can be viewed as yielding an implied confidence level. In particular, the solution for any given tuning parameter would be the sparsest solution among a confidence region of level $1 - \alpha$. Somewhat surprisingly, standard tuning methods often result in confidence levels that are much smaller than would be used in practice.

(*** Figure 1 goes here ***)

Figure 1 shows the complex relationship between λ, as in (4), and the corresponding confidence region $1 - \alpha$ for 4 different linear regression data sets with sample size, $n = 100$, total number of predictors, $p = 20$, and true number of non-zero coefficients, $q = 4$. The choice of $p(\boldsymbol{\beta})$ is given by the adaptive L_1 norm $\sum_{j=1}^{p} \hat{w}_j \mid \beta_j \mid$, for $\hat{w}_j = 1/ \mid \hat{\beta}_j \mid$. This yields the adaptive LASSO penalization approach (Zou, 2006). This simulation design is used in Section 4.1 as example 1.

The vertical lines represent the locations of the models selected by the proposed tuning method with 95% and 99% confidence regions, and other well known tuning parameter selection criteria AIC, BIC and 5-fold cross-validation (CV).

(*** Figure 2 goes here ***)

Figure 2 shows boxplots of the implied confidence region from the three tuning methods based on 500 simulation runs. The left panel is the setup described above, while the right panel has only $q = 2$ non-zero coefficients. These box plots suggest that for CV the variability of the confidence region level is very large, whereas for AIC the variability is much smaller and the level is extremely close to 0. For BIC, as expected, the median confidence region level is higher than that for AIC, with more variability. The implied confidence coefficients from using the typical criteria for tuning are much smaller than what we propose with common 95% and 99% confidence regions. Note that fixing a $1 - \alpha$ confidence region level does represent the familywise Type I error under the null model. Hence a low implied confidence level from using an alternative criterion seems that it may also carry over to spurious selection of irrelevant predictors at models close to the null.

2.3 Linear Models

Consider the normal linear regression set up:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \tag{6}$$

Suppose that the columns of $\mathbf{X}$ matrix are standardized, i.e. $\sum_{i=1}^n x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$, $\forall_j$ $j = 1, .., p$. Also, the response vector $\mathbf{Y}$ is centered, $\sum_{i=1}^n y_i = 0$, so that there is no intercept in the model. Let $\hat{\boldsymbol{\beta}}$ be the MLE of $\boldsymbol{\beta}$, which is also the OLS estimator for normal linear models. As will be shown later, the asymptotic properties of the proposed tuning procedure will hold without the normality of the errors.

Suppose we wish to test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. For simplicity, let $Q(\boldsymbol{\beta}) = (Y - X\boldsymbol{\beta})^t(Y -$

$X\boldsymbol{\beta}$). In the case where σ^2 is known, the confidence set given in (5) reduces to

$$\frac{Q(\boldsymbol{\beta^*}) - Q(\hat{\boldsymbol{\beta}})}{\sigma^2} \leq C_\alpha, \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Note that $Q(\boldsymbol{\beta^*}) - Q(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})$. Therefore, (7) can also be written as

$$\frac{((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}))}{\sigma^2} \leq C_\alpha, \tag{8}$$

which has an exact chi-square distribution with p degrees of freedom under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta^*}$. Hence, the critical value C_α can be determined using the $1 - \alpha$ percentile of this chi-square distribution, i.e. $C_\alpha = \chi^2_\alpha$, where χ^2_α denotes the $1 - \alpha$ quantile of the chi-square distribution.

In practice, it is typically the case that σ^2 is not known. In this situation, an F distribution can be used to determine the critical value C_α. Here, the confidence set given in (5) reduces to

$$n \log(\frac{Q(\boldsymbol{\beta^*})}{Q(\hat{\boldsymbol{\beta}})}) \leq C_\alpha. \tag{9}$$

Then, by some algebra one can show that (9) is equivalent to

$$\frac{(Q(\boldsymbol{\beta^*}) - Q(\hat{\boldsymbol{\beta}}))/p}{Q(\hat{\boldsymbol{\beta}})/(n-p)} \leq K_\alpha, \tag{10}$$

where $K_\alpha = \frac{n-p}{n}\{\exp(C_\alpha/n) - 1\}$. Then (10) can be rewritten as

$$\frac{((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^T X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}))/p}{s^2} \leq K_\alpha, \tag{11}$$

where $s^2 = \frac{Q(\hat{\boldsymbol{\beta}})}{(n-p)}$ is the mean square error using the OLS estimator. Under the null, the left side of (11) has an F distribution with numerator degrees of freedom p and dominator degrees of freedom $n - p$. Now the critical value K_α can be determined using the $1 - \alpha$ percentile of $F_{(p,n-p)}$ distribution, denoted by F_α.

Based on the above derivation, any $\boldsymbol{\beta}$ lying in the $1-\alpha$ confidence region should satisfy:

$$\frac{((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T X^T X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))/p}{s^2} \leq F_\alpha \tag{12}$$

So, for normal linear models, the optimization problem given the choice of confidence level can be represented by

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= argmin_\beta \; p(\boldsymbol{\beta}) \\ \text{subject to } &\frac{((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))/p}{s^2} \leq F_\alpha \quad , \end{aligned} \tag{13}$$

where the constraint in (13) fully specifies the optimization problem for the model coefficients.

Note that the minimization problem given in (13) can also be expressed by

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= argmin_\beta \; p(\boldsymbol{\beta}) \\ \text{subject to } &\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 \leq T_\alpha \quad , \end{aligned} \tag{14}$$

where $T_\alpha = s^2(pF_\alpha + n - p)$. Now, this is in the form of (2) with $H(\mathbf{Y}, \mathbf{X}\beta) = \|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2$. Then, for a given confidence region level $1-\alpha$, (14) can be written as a typical penalized regression problem with nonnegative regularization parameter λ via the arguments in (2), (3), and (4). Section 2.3.1 gives a detailed explanation of computation via the standard penalized regression algorithms.

Note that the Dantzig Selector (Candes and Tao, 2007; James et al., 2009) can be put into this confidence region framework as it is exactly in the form of (3) with

$$p(\boldsymbol{\beta}) = \sum_{j=1}^{p} \mid \beta_j \mid, \text{ and } H(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}) = sup_{1 \leq j \leq p} \mid \mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \mid .$$

Hence, the Dantzig Selector can be viewed as constructing a confidence region based on the L_∞ norm of the score function, and then minimizing the L_1 norm within the region. Although, this confidence region is a non-standard type as compared with, a likelihood or Wald based region, a threshold may be chosen based on the distribution of $H(Y, X\beta)$, see Candes and Tao (2007) and

the discussion therein.

2.3.1 Computation

Given the duality, there is a one-to-one correspondence between the solution for a given value of λ and a particular confidence level $1-\alpha$. However, the relationship between $1-\alpha$ and λ is complex and data-dependent. Therefore, in order to find the solution, we propose a simple computational approach using standard algorithms. For example, if the complexity measure is either L_1 norm, $(\hat{w}_j = 1,\ \forall j)$, or adaptive L_1 norm, the LARS algorithm efficiently gives the entire solution path as a function of λ.

To find the solution using available algorithms, the solution path for the penalized regression, as in (4), should be obtained starting from a large enough λ value which would assign 0 to all model coefficients. Suppose $\boldsymbol{\beta}^*_\lambda$ is the vector of estimated model coefficients for a given tuning parameter λ. Then, for each $\boldsymbol{\beta}^*_\lambda$, it is checked if $\boldsymbol{\beta}^*_\lambda$ lies in the confidence region via checking $((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\lambda)^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\lambda))/ps^2 \leq F_\alpha$ or equivalently $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*_\lambda\|^2 \leq T_\alpha$. Among the models in the confidence region, the one with the smallest $p(\boldsymbol{\beta}^*_\lambda)$ value is the chosen solution. In practice, this only requires following the solution path from the origin, $(\boldsymbol{\beta} = 0)$, until the first time we hit the confidence region. In other words, the solution is simply the first time the solution path reaches the boundary of the confidence ellipsoid. This only requires computing $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ along a grid of λ values and noting when it crosses T_α. This can be obtained, for example, via standard output from the LARS package in R if $p(\beta) = \sum_{j=1}^{p} w_j \mid \beta_j \mid$ for any set of weights.

2.3.2 Simple Example

A simple example can be used to explain the tuning via confidence regions. This example will compare the solutions to the proposed method coupled with L_1 norm and adaptive L_1 norm, which are the forms of penalizations for LASSO (Tibshirani, 1996) and adaptive LASSO (Zou, 2006), respectively. For LASSO, $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \mid \beta_j \mid$, and for adaptive LASSO, $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \hat{w}_j \mid \beta_j \mid$, where $\hat{\boldsymbol{w}} = 1/\mid \hat{\boldsymbol{\beta}} \mid^\gamma$. We use $\gamma = 1$. Here, the L_1 norm forces all the coefficients to be penalized equally, whereas weighted L_1 norm assigns different weights to different coefficients.

Note that the $1-\alpha$ confidence region for linear models can be can be defined by $\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 \leq T_\alpha$ where $T_\alpha = s^2(pF_\alpha + n - p)$, and is independent of the form of the penalization. That is, for both L_1 and adaptive L_1 norm, we search for solutions within the same set. The solution to the proposed method using the L_1 norm is the point in the confidence region which gives the smallest $\sum_{j=1}^{p} \mid \beta_j \mid$, whereas for adaptive L_1 norm it is the point which gives the smallest $\sum_{j=1}^{p} \left(\mid \beta_j \mid / \mid \hat{\beta}_j \mid \right)$. As implied earlier, in practice this only requires following the solution path from the origin until the first time we hit the confidence region.

(*** Figure 3 goes here ***)

For the proposed method with a 95% confidence region, Figure 3 compares the L_1 norm with the adaptive L_1 norm for a linear regression setup with 2 predictors, where the vector of true regression coefficients is given by $\boldsymbol{\beta} = (0, 1)^T$. The ellipse is the confidence region centered around the OLS estimator. The solution paths for the L_1 norm and adaptive L_1 norm are the dot-dashed line and solid line respectively. Both of the paths connect the OLS estimate to the origin. We seek the points at which the solution path hits the confidence region for the first time. These points are shown by the dots at the intersection of the lines and the confidence region. Note that for this example the solution to the proposed method using the adaptive LASSO correctly identifies the zero coefficient, whereas the solution using the LASSO fails.

2.4 Generalized Linear Models

2.4.1 Likelihood Based Confidence Regions

Unlike linear models, for other generalized linear models (GLMs), the likelihood ratio statistic does not yield a simple form for an exact distribution. However, under the null hypothesis, it is known that for large samples the log-likelihood ratio test statistic can be approximated by a chi-square distribution, and an asymptotic $1 - \alpha$ confidence region can be constructed using this distribution.

Suppose we wish to test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. Let $\hat{\boldsymbol{\beta}}$ be the usual MLE estimator of $\boldsymbol{\beta}$. Note that in this case, a column of $1's$ should be added to the design matrix to include the intercept. We do not penalize this intercept. Then, the log-likelihood ratio statistic for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$,

$-2\Lambda(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}^*)$, has an asymptotic chi-square distribution under the null hypothesis.

Based on the likelihood ratio test, H_0 is rejected at a significance level of α if

$$2(\log L(\mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{Y}) - \log L(\mathbf{X}\boldsymbol{\beta}^*, \mathbf{Y})) \geq \chi^2_{p+1,\ \alpha} \tag{15}$$

Then, any $\boldsymbol{\beta}$ lying in the $1-\alpha$ confidence set should satisfy

$$2(\log L(\mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{Y}) - \log L(\mathbf{X}\boldsymbol{\beta}^*, \mathbf{Y})) \leq \chi^2_{p+1,\ \alpha} \tag{16}$$

Therefore, after some algebra, the proposed approach can be represented by

$$\begin{gathered} \tilde{\beta} = argmin_{\beta}\ p(\boldsymbol{\beta}) \\ \text{subject to } -\log L(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y})) \leq \frac{\chi^2_{p+1,\ \alpha}}{2} - \log L(\mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{Y})) \end{gathered} \tag{17}$$

Similar to linear models, for GLMs using the likelihood leads directly to the dual problem of regular penalized likelihood methods with the form given in (1), whose solution path can be computed as in Park and Hastie (2007), for example.

2.4.2 Wald Based Confidence Regions

Another common way of constructing confidence regions is based on a Wald-type statistic that uses the asymptotic normality of the MLE, $\hat{\boldsymbol{\beta}}$. Suppose $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow^d N(\mathbf{0}, \boldsymbol{\Sigma})$. For a given hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$, the Wald statistic is

$$W(\boldsymbol{\beta}^*) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T [\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \tag{18}$$

where $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ is the variance matrix for $\hat{\boldsymbol{\beta}}$, based on the Fisher Information. The asymptotic normality of $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-square distribution for W under H_0 and the degrees of freedom of the chi-square distribution is equal to the rank of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$, which we will assume to be full rank.

This alternative confidence region can be used for the proposed method and is given by

$$\tilde{\boldsymbol{\beta}} = argmin_{\beta}\ p(\boldsymbol{\beta})$$
$$\text{subject to } W(\boldsymbol{\beta}) \leq \chi^2_{p+1,\alpha}. \tag{19}$$

As discussed previously, the above can be reexpressed in the equivalent form

$$\tilde{\boldsymbol{\beta}} = argmin_{\beta}\ \{W(\boldsymbol{\beta}) + \lambda_{\alpha} p(\boldsymbol{\beta})\}. \tag{20}$$

where the tuning parameter is again automatically determined with the confidence region level $1-\alpha$.

The proposed tuning method via the Wald based confidence region results in the least square approximation (LSA) objective function (Wang and Leng, 2007) in the sense that for LSA, $W(\beta)$ is used as a simple approximation to the negative log-likelihood. Hence, the proposed method yields a new tuning method for LSA.

3 Asymptotic Selection Theory

Let $A = \{j : \beta_j^0 \neq 0\}$, where $\boldsymbol{\beta}^0$ is the true value of $\boldsymbol{\beta}$, and $A_n = \{j : \tilde{\beta}_j \neq 0\}$. Then the penalized least squares given in (1) is consistent in variable selection if and only if $P(A_n = A) \to 1$.

Selection via LASSO, with $w_j = 1$ for all j, can be inconsistent in variable selection (Zou, 2006; Zhao and Yu, 2006; Meinshausen and Buhlmann, 2006; Wainwright, 2009). Furthermore, Zou (2006) presented a simple necessary condition for the variable selection consistency of LASSO. If the necessary condition fails, there is no sequence of tuning parameters, λ_n, which makes LASSO consistent in variable selection. However, with the proper choice of λ_n, the adaptive LASSO enjoys variable selection consistency. Hence, for selection consistency of the proposed tuning method we will concentrate on the adaptive LASSO.

For the proposed method, the shrinkage parameter is the confidence region level $1-\alpha$. Although, in practice, a fixed confidence level will be specified by the user, for asymptotic consistency, we

must have $\alpha = \alpha_n$ decrease with the sample size. We examine the properties of the approach as $1 - \alpha_n \to 1$ or $\alpha_n \to 0$. However, if $\alpha_n \to 0$ too quickly, the quantile of the distribution which determines C_{α_n} will diverge rapidly, and hence the resulting region will not actually shrink asymptotically. Therefore, for selection consistency, as n increases, we need an appropriate decay rate in α_n, so that confidence region continues to shrink around the OLS estimator.

The theorems will be stated here, regularity conditions and proofs can be found in Appendix.

Theorem 1. *Variable Selection Consistency for Wald-Based and Likelihood Based Confidence Regions: Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be i.i.d. each with density* $f(X, Y, \beta)$ *that satisfies the regularity conditions* $(A1)-(A3)$ *in the Appendix. Let* $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\mid \beta_j \mid / \mid \hat{\beta}_j \mid \right)$. *If* $\alpha_n \to 0$ *and* $\frac{1}{n} \log \alpha_n \to 0$*, then the proposed tuning method for Adaptive LASSO with the Chi-square threshold for either the Wald-based region (19), or the likelihood-based region (17) is consistent in variable selection, i.e.* $P(A_n = A) \to 1$.

Theorem 2. *Variable Selection Consistency for Linear Models: Let* $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ *where* ε_i $(i = 1, \ldots, n)$ *are independent with* $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, *and independent of* $\boldsymbol{X}$, *and let* $\boldsymbol{X}^T\boldsymbol{X} \to \boldsymbol{C} > 0$. *Let* $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\mid \beta_j \mid / \mid \hat{\beta}_j \mid \right)$. *If* $\alpha_n \to 0$ *and* $\frac{1}{n} \log \alpha_n \to 0$*, then using the F-distribution threshold, the proposed tuning method for Adaptive LASSO is consistent in variable selection, i.e.* $P(A_n = A) \to 1$ *for linear models (13).*

It is clear from the above theorems that if $\log \alpha_n$ has an asymptotic decay rate smaller than n, then tuning via the confidence region will achieve variable selection consistency for generalized linear models, given that the asymptotic normality of the maximum likelihood estimator is met. Note that the thresholds such as the Chi-square or the F will not necessarily yield exact $1 - \alpha$ level regions. However, all that is needed for selection consistency is that the value of the threshold diverges at the appropriate rate. In fact, the thresholds can be replaced by any value C such that $C \to \infty$ and $n^{-1}C \to 0$ and would yield selection consistency, although the interpretation of tuning via a $1 - \alpha$ level confidence region would then be lost.

4 Simulation Studies

The proposed method is investigated through Monte Carlo simulations for linear models and GLMs using the adaptive L_1 norm. All simulations were conducted in *R*. Numerical studies are reported based on 500 simulation replications for each setting.

4.1 Linear Models

In the simulation studies for linear models, the LARS package in R is used. We simulated data for sample sizes $n = 50, 100, 150$ using a linear regression model set up, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where ε_i i.i.d. with $\varepsilon_i \sim N(0, 1)$ and $\mathbf{x}_{ij}$ are standard normal with the correlation between x_{ij_1} and x_{ij_2} given by $\rho^{|j_1 - j_2|}$, for $\rho = 0, 0.5, 0.9$. Out of $p = 20, 40$ predictors, the true numbers of predictors (q) are determined for various levels of sparsities. The true $\boldsymbol{\beta}$ vectors are generated by assigning 1 to each nonzero coefficient. The $\mathbf{X}$ matrix is standardized and $\mathbf{Y}$ is centered as described in Section 2.3.

For a given method and scenario, each entry in the tables shows the proportion of times that the correct model is chosen along with the median number of nonzero estimated coefficients for the chosen models in parenthesis. Adaptive LASSO is tuned via the proposed method for 95% and 99% confidence region levels. These are compared with tuning the Adaptive LASSO via typical methods: AIC, BIC, and 5-fold cross-validation (CV).

Example 1: In this simulation study, $p = 20$ and $\rho = 0, 0.5, 0.9$. The simulation results are summarized in Table 1 for $q = 2, 4, 10$ and in Table 2 for $q = 12, 14, 16$.

(*** Table 1 goes here ***)

(*** Table 2 goes here ***)

From Table 1 and 2, the most striking result is that the proposed method with either 95% or 99% confidence regions perform better than the AIC, BIC, or CV methods in terms of identifying the correct sparse models for all sample sizes and the various degrees of sparsity for $\rho = 0, 0.5$. For example, in Table 1 for $\rho = 0$ even when $n = 50$, the performance of the proposed method in identifying the correct sparse model for all degrees of sparsity is between 80% and 91%, whereas for

AIC, BIC, and CV combined the highest performance is only 56%. In the case of high correlation, $\rho = 0.9$, the performance of BIC is now slightly better than the proposed method, particularly in the smaller samples. The high correlation leads to some instability in the confidence regions making them larger and more likely to contain a sparse model. This can be seen by the lower median model sizes selected via the proposed approach compared to the others. As the sample size increases, the proposed approach behaves similarly to BIC in this high correlation setting.

The asymptotic theory suggests that for larger samples, a larger confidence level should perform better. The simulation results supports this theory. Notice that for $n = 50$ the 99% CR level does not perform as well as the 95% CR level, but the 99% CR level performs the best for $n = 100$ and $n = 150$. However, the performance of either choice of 95% or 99% is strong throughout the set of scenarios. We have also tried 90% and the results were similar, and not shown.

Example 2: For this example, $p = 40$ and $\rho = 0, 0.5, 0.9$. The results are summarized in Table 3 for $q = 4, 8, 20$. The results are similar to the $p = 20$ case. However, for $n = 50$ all methods show degraded performance as expected.

(*** Table 3 goes here ***)

Furthermore by using the median number of estimated coefficients one can see that while the proposed method tends to underselect, AIC, BIC, and CV tend to overselect. The overselecting tendency of prediction accuracy based criteria has been studied for LASSO type selection by Leng et al. (2006). For the proposed method it is not entirely surprising to observe underfitting as it is controlling the family-wise error rate, which can be conservative, and hence lead to underselection.

4.2 Generalized Linear Models

For Generalized Linear Models, the *penalized* package (Goeman, 2007) in *R* is used for computation. For each setting, 500 data sets are simulated from a logistic regression model with

$$Y_i \sim Bernoulli(\frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}), \tag{21}$$

where x_{ij} $(i = 1, \ldots, n)$ are standard normal with the correlation between x_{ij_1} and x_{ij_2} given by $\rho^{|j_1 - j_2|}$, for $\rho = 0, 0.5$, with sample sizes $n = 150, 200$. For the logistic fit we have used larger sample sizes due to the less information in the binary responses. The simulations results are shown by Table 4.

(*** Table 4 goes here ***)

From Table 4, we see that the proposed method compares favorably to other tuning methods in this scenario as well.

5 Estimation Accuracy

5.1 Asymptotic Results

Definition For $A = \{j : \beta_j^0 \neq 0\}$, an estimation procedure, δ, has the optimal estimation rate, if $\sqrt{n}(\tilde{\boldsymbol{\beta}}_A(\delta) - \boldsymbol{\beta}_A^0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is the covariance matrix knowing the true subset model.

Although it has been shown for the Adaptive Lasso (Zou, 2006) that an appropriate choice of tuning parameter sequence, λ_n, leads to both selection consistency and optimal estimation, the same is not true of tuning via a fixed sequence α_n.

There is a unique one-to-one mapping between λ_n and α_n for any given data set. Hence there does exist a sequence α_n that will also yield both properties, due to the one-to-one correspondence between α_n and λ_n for any given dataset. However, this mapping is data dependent, and thus, this matching sequence is a random sequence, and it is not possible to establish conditions on any fixed sequence of α_n that would achieve both properties. If $\alpha_n \rightarrow 0$ for any fixed sequence, it is not possible to obtain $\sqrt{n}$ consistency. Intuitively consider the simple univariate case for a linear model. Then, the confidence region is simply $\hat{\beta} \pm t_{\alpha_n} s/\sqrt{n}$. Now the solution will be on the boundary of the region (unless zero is inside the region), but if $\alpha_n \rightarrow 0$, then $t_{\alpha_n} \rightarrow \infty$, hence it follows that $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \rightarrow \infty$. But since, $\sqrt{n}(\hat{\beta} - \beta^0) = O_p(1)$, it must be that $\sqrt{n}(\hat{\beta} - \beta^0) \rightarrow \infty$. Hence it is not $\sqrt{n}$ consistent for any fixed sequence $\alpha_n \rightarrow 0$.

If prediction is the goal, in order to obtain better estimates of the coefficients we propose a two-stage estimation procedure; first select the model via the proposed method, and then refit the model with the MLE for the important predictors chosen in the first stage. Let the estimated model coefficients of the refitted model be the second stage estimates, $\tilde{\boldsymbol{\beta}}_{Refit}$

The following theorem shows the optimal estimation rate of $\tilde{\boldsymbol{\beta}}_{Refit}$.

Theorem 3. *Optimal Estimation Rate for the Two-stage Procedure: If* $\alpha_n \to 0$ *and* $\frac{1}{n}\log\alpha_n \to 0$*, then* $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{Refit_A} - \boldsymbol{\beta}_A^0) \to_d N(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$*, where* $\boldsymbol{\Sigma}^*$ *is the covariance matrix knowing the true subset model.*

This two-stage procedure hence obtains the oracle property as given by Fan and Li (2001) and Zou (2006)

5.2 Simulation Study

Consider the linear model with $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Let (x^{new}, y^{new}) be a new independent draw from the joint distribution. The expected prediction error for any estimator $\boldsymbol{\beta^*}$ is $E_\beta(\|y^{new} - x^{new}\boldsymbol{\beta^*}\|^2)$. This can be decomposed into $\text{ME}(\beta^*) + \sigma^2$, where $\text{ME}(\beta^*)$ is the model error for $\boldsymbol{\beta^*}$,

$$ME(\boldsymbol{\beta^*}) = (\boldsymbol{\beta^*} - \boldsymbol{\beta^0})^T\mathbf{V}(\boldsymbol{\beta^*} - \boldsymbol{\beta^0}), \tag{22}$$

where $\boldsymbol{\beta^0}$ are the true model coefficients, and $\mathbf{V}$ is the covariance matrix of the predictors. Note that σ^2 is the irreducible error and cannot be avoided unless $\sigma^2 = 0$. Hence, $\text{ME}(\beta^*)$ differs among estimators and can thus be used to compare the performance of the estimators.

We can define two stage version for each tuning method. Let ME-I denote the model error for a tuning method. Let ME-II denote the model error for a method using the 2-stage version. For any method, the ME-II is given by using maximum likelihood on the set of predictors chosen by the particular method, such as AIC, BIC, CV, or the proposed method

Using the same simulation set up given in Section 4.1, Table 5 ($n = 50$) and Table 6 ($n = 100$) compare the accuracy of first stage and second stage estimators. The proposed method of tuning the adaptive LASSO for 99% and 95% confidence region levels are compared with tuning via AIC,

BIC and CV. We have also compared to the full OLS estimator and the oracle estimator, which performs OLS on the true set of active predictors.

(*** Table 5 goes here ***)

(*** Table 6 goes here ***)

Table 5 and Table 6 indicate that the model errors for the two-stage procedure improves upon that of the other methods for both correlated and uncorrelated cases and different degrees of sparsity. This is obviously a by-product of the improved selection properties. As expected, the estimation performance of the proposed method is worse than others in the first stage. However, due to its selection performance, the second stage gives a dramatic improvement. ME-II's for the proposed method for both confidence region levels are close to their oracle value. We also observe that for $n = 50$ the ME-II's for 95% and 99% confidence region levels are very close to each other, and for $n = 100$ they are almost the same.

For GLM's we used Kullback-Leibler (KL) divergence to compare first stage (KL-I) and second stage (KL-II) model errors. KL divergence gives the discrepancy from the probability distribution of the estimated model to the true model. Let $f_\beta(Y \mid X)$ and $f_{\beta^*}(Y \mid X)$ be the probability density functions under the true model and the estimated model respectively. Then the KL divergence from the estimated model to the true model is given by $E_\beta\left(\log \frac{f_\beta(Y|X)}{f_{\beta^*}(Y|X)}\right)$.

(*** Table 7 goes here ***)

Table 7 shows the simulation results for KL divergences using the same simulation set up described in Section 4.2. We see that the results are very similar to the linear case. We again see that the proposed 2-stage estimaton method with both of the confidence region levels are performing well. Therefore, with the two-stage procedure the proposed method has the dual benefit of selection and estimation accuracy.

6 Discussion

In this article we proposed a method to select the tuning parameter for penalized regression methods based on the confidence region level. The proposed method can be used for linear models, generalized linear models or for any other type of models where confidence regions can be constructed. Tuning penalized regression methods in this way is intuitive and easily understood by scientists and practitioners. Although under moderate correlation, the confidence region approach exhibits excellent performance, under high correlation it can sometimes underselect, particularly for smaller samples. This is not unexpected, as methods that control family-wise error rates can be conservative.

Comparisons of tuning parameter selection methods based on AIC, BIC and CV, reveal that the resulting tuning parameters correspond to confidence levels that are extremely low, and can vary greatly across data sets. We compared two-stage estimation methods. However, if interest focused greatly on prediction, it is also possible to include regularization in the 2nd stage (Meinshausen, 2007). This should improve the estimation performance for methods such as AIC/BIC/CV which tend to overselect.

Using Bayesian approaches to the LASSO (Park and Casella, 2008) can yield a posterior credible interval for the penalty parameter. Although this is more in line with conducting inference, rather than model selection, one can imagine choosing the largest value of the penalty parameter within the given credible interval. This would result in the most penalized model, and be a way to select the sparsest model within a posterior region.

Recently, a large focus has been on variable selection in the ultra-high dimensional case with $p >> n$. Although, for the confidence regions it is assumed that $p < n$, the proposed tuning method can be used in this situation as well. Recent methods for ultra-high dimensional screening (Fan and Lv, 2008; Wang, 2009), first screen to a moderate dimension, and then penalized regression is used following this screening to $p < n$. It has been shown that the initial screening to $p < n$ will contain all relevant predictors with probability tending to one, under the assumption of sufficient sparsity. Hence an improved tuning method for this penalized regression in the second step is an important addition to the ultra-high dimensional case as well.

Supplemental Materials

R Code: R code to perform selection based on the user specified confidence region level for linear models with adaptive lasso penalty is supplied. (codeCR.R)

Appendix: The supplemental files include the Appendix which gives all proofs. (appendixCR.pdf)

Acknowledgements

The authors are grateful to Len Stefanski for helpful discussions during the preparation of the manuscript. The authors would like to thank the editor, associate editor, and three referees for their helpful comments that improved the manuscript. This work was partially supported by NSF Grant DMS-0952826 and NIH Grants R01 MH-084022 and P01 CA-142538.

References

Akaike, H. (1973), *Information Theory and the an Extension of the Maximum Likelihood Principle*, In B. Petrov, F. Csaki (Eds.), Second international symposium on information theory, Budapest: Academiai Kiado.

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous Regression Shrinkage, Variable Selection and Clustering of Predictors with OSCAR," *Biometrics*, 64, 115–123.

Breiman, L. and Spector, P. (1992), "Submodel Selection and Evaluation in Regression. The X-Random Case," *International Statistical Review*, 60, 291–319.

Candes, E. and Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, 35, 2313–2351.

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 Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space (with discussion)," *Journal of the Royal Statistical Society B*, 70, 849–911.

Goeman, J. J. (2007), "Penalized: L1 (lasso) and L2 (ridge) Penalized Estimation in GLMs and in the Cox Model," R package.

Hastie, T., Tibshirani, R., and Freidman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.

James, G., Radchenko, P., and Dasso, J. L. (2009), "Connections Between the Dantzig Selector and Lasso," *JRSS(B)*, 12, 127–142.

Konishi, S. and Kitagawa, G. (1996), "Generalised Information Criteria in Model Selection," *Biometrika*, 83, 875–890.

Leng, C., Li, Y., and Wahba, G. (2006), "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284.

Mallows, C. L. (1973), "Some Comments on CP," *Technometrics*, 15, 661–675.

Meinshausen, N. (2007), "Relaxed lasso," *Computational Statistics and Data Analysis*, 52, 374–393.

Meinshausen, N. and Buhlmann, P. (2006), "High Dimensional Graphs and Variable Selection with Lasso," *Annals of Statistics*, 34, 1436–1462.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *Annals of Statistics*, 12, 758–765.

Park, M. Y. and Hastie, T. (2007), "L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model," R package.

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

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 451–464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264.

Shibata, R. (1981), "An optimal selection of regression variables," *Biometrika*, 68, 45–54.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, 30, 44–47.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wainwright, M. J. (2009), "Information-Theoretic Limitations on Sparsity Recovery in the High-Dimensional and Noisy Setting," *IEEE Transections and Information Theory*, 55, 5728–5741.

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Wang, H. and Leng, C. (2007), "Unified LASSO Estimation by Least Squares Approximation," *Journal of the American Statistical Association*, 102, 1039–1048.

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568.

Zhao, P. and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

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 Series B*, 67, 301–320.

Table 1: Linear Models: Automatic variable shrinkage via 95% and 99% confidence regions compared with AIC, BIC, and 5-fold CV methods using adaptive L_1 norm for $n = 50, 100, 150$. Each entry represents the proportion of times that the correct model is chosen and the median number of nonzero estimated coefficients in parenthesis, for $p = 20$, $q = 2, 4, 10$

		Sample Size								
		50			100			150		
	Method	2	4	10	2	4	10	2	4	10
$\rho = 0$	95% CR	91(2)	91(4)	85(10)	96(2)	96(4)	98(10)	97(2)	98(4)	99(10)
	99% CR	90(2)	91(4)	80(10)	99(2)	99(4)	99(10)	99(2)	100(4)	100(10)
	AIC	12(6)	13(8)	9(13)	20(5)	17(7)	18(12)	17(5)	21(6)	24(12)
	BIC	56(2)	46(5)	32(11)	71(2)	67(4)	59(10)	78(2)	75(4)	72(10)
	CV	54(2)	55(4)	43(10)	58(2)	66(4)	54(10)	60(2)	67(4)	60(10)
$\rho = 0.5$	95% CR	86(2)	79(4)	60(10)	96(2)	97(4)	99(10)	97(2)	99(4)	100(10)
	99% CR	81(2)	75(4)	56(10)	99(2)	99(4)	100(10)	100(2)	100(4)	100(10)
	AIC	15(6)	14(8)	11(13)	24(5)	20(6)	24(12)	23(5)	27(6)	35(11)
	BIC	57(2)	48(5)	33(11)	74(2)	74(4)	70(10)	79(2)	82(4)	81(10)
	CV	53(2)	45(4)	28(10)	59(2)	60(4)	46(10)	61(2)	61(4)	51(10)
$\rho = 0.9$	95% CR	18(1)	4(3)	0(6)	54(2)	24(3)	4(8)	77(2)	59(4)	19(9)
	99% CR	12(1)	2(2)	0(6)	48(2)	17(3)	3(8)	71(2)	47(4)	12(9)
	AIC	6(7)	2(8)	0(12)	15(6)	11(7)	7(12)	22(5)	18(7)	17(12)
	BIC	19(2)	7(4)	0(9)	57(2)	34(4)	11(10)	70(2)	59(4)	38(10)
	CV	10(2)	2(3)	0(6)	23(2)	8(3)	3(9)	34(2)	18(4)	9(10)

Table 2: Linear Models: Automatic variable shrinkage via 95% and 99% confidence regions compared with AIC, BIC, and 5-fold CV methods using adaptive L_1 norm for $n = 50, 100, 150$. Each entry represents the proportion of times that the correct model is chosen and the median number of nonzero estimated coefficients in parenthesis, for $p = 20$, $q = 12, 14, 16$

		Sample Size								
		50			100			150		
	Method	12	14	16	12	14	16	12	14	16
$\rho = 0$	95% CR	84(12)	85(14)	82(16)	97(12)	98(14)	100(16)	99(12)	100(14)	100(16)
	99% CR	83(12)	82(14)	77(16)	99(12)	99(14)	100(16)	99(12)	100(14)	100(16)
	AIC	10(15)	15(16)	19(18)	20(14)	27(15)	32(17)	29(13)	34(15)	42(17)
	BIC	31(13)	31(15)	41(17)	61(12)	63(14)	64(16)	73(12)	75(14)	75(16)
	CV	44(12)	44(14)	36(17)	54(12)	47(15)	37(17)	55(12)	54(13)	32(17)
$\rho = 0.5$	95% CR	56(12)	50(14)	42(15)	98(12)	100(14)	100(16)	100(12)	100(14)	100(16)
	99% CR	48(12)	41(13)	32(15)	99(12)	99(14)	99(16)	100(12)	100(14)	100(16)
	AIC	15(14)	18(16)	22(18)	24(14)	34(15)	45(17)	32(13)	40(15)	49(17)
	BIC	34(13)	35(15)	40(17)	68(12)	72(14)	80(16)	77(12)	81(14)	83(16)
	CV	25(12)	23(15)	29(17)	43(12)	36(15)	33(17)	48(12)	37(15)	30(17)
$\rho = 0.9$	95% CR	0(8)	0(9)	0(9)	2(10)	1(11)	1(12)	17(11)	12(12)	8(14)
	99% CR	0(7)	0(8)	0(9)	0(9)	0(11)	0(12)	12(10)	5(12)	5(14)
	AIC	0(13)	1(14)	0(15)	9(14)	11(15)	13(17)	19(14)	25(15)	33(17)
	BIC	0(10)	1(11)	0(13)	15(12)	12(13)	14(15)	42(12)	42(14)	43(16)
	CV	0(8)	0(9)	0(10)	4(11)	5(14)	7(16)	8(12)	11(15)	11(17)

Table 3: Linear Models: Automatic variable shrinkage via 95% and 99% confidence regions compared with AIC, BIC, and 5-fold CV methods using adaptive L_1 norm for $n = 50, 100, 150$. Each entry represents the proportion of times that the correct model is chosen and the median number of nonzero estimated coefficients in parenthesis, for $p = 40$, $q = 4, 8, 20$

		Sample Size								
		50			100			150		
	Method	4	8	20	4	8	20	4	8	20
$\rho = 0$	95%CR	48(4)	35(8)	12(19)	95(4)	96(8)	95(20)	96(4)	97(8)	99(20)
	99%CR	25(3)	21(7)	5(18)	99(4)	99(8)	96(20)	99(4)	99(8)	99(20)
	AIC	0(27)	0(29)	0(33)	7(13)	4(16)	3(26)	9(10)	8(14)	7(24)
	BIC	17(7)	8(14)	1(29)	56(4)	47(9)	28(21)	61(4)	62(8)	50(20)
	CV	38(4)	14(7)	1(14)	68(4)	74(8)	67(20)	70(4)	77(8)	77(20)
$\rho = 0.5$	95% CR	30(3)	16(7)	2(18)	96(4)	96(8)	93(20)	97(4)	99(8)	100(20)
	99% CR	19(3)	9(6)	1(16)	97(4)	97(8)	92(20)	99(4)	99(8)	100(20)
	AIC	1(28)	0(29)	0(33)	7(12)	6(16)	6(26)	9(10)	8(14)	13(24)
	BIC	17(6)	5(14)	1(29)	63(4)	52(8)	40(21)	70(4)	68(8)	60(20)
	CV	16(3)	4(5)	0(10)	55(4)	49(8)	39(20)	64(4)	57(8)	54(20)
$\rho = 0.9$	95% CR	0(2)	0(5)	0(11)	16(3)	2(6)	0(15)	37(4)	18(7)	3(17)
	99% CR	0(2)	0(4)	0(9)	12(3)	1(6)	0(15)	28(3)	11(7)	1(17)
	AIC	0(31)	0(31)	0(32)	2(14)	1(18)	0(27)	8(11)	6(15)	1(26)
	BIC	0(14)	0(14)	0(26)	20(4)	6(8)	0.2(20)	42(4)	29(8)	9(20)
	CV	0(2)	0(2)	0(3)	4(2)	1(4)	0(11)	11(3)	2(4)	1(13)

Table 4: GLM's: Automatic variable shrinkage via 95% and 99% confidence regions, constructed by using asymptotic chi-square distribution with degrees of freedom $p + 1$, compared with AIC, BIC, and CV methods using adaptive L_1 norm for $n = 150, 200$. Each entry represents the percent of correct models identified, and the median number of nonzero estimated coefficients in parenthesis, for $p = 20$, $q = 2, 4, 10$.

		Sample Size					
		150			200		
	Method	2	4	10	2	4	10
$\rho = 0$	95 % CR	92(2)	92(4)	69(10)	93(2)	94(4)	87(10)
	99 % CR	91(2)	93(4)	70(10)	98(2)	98(4)	89(10)
	AIC	46(3)	35(5)	11(13)	38(3)	37(5)	16(13)
	BIC	90(2)	84(4)	37(11)	83(2)	82(4)	55(10)
	CV	40(3)	25(5)	14(12)	25(3)	27(5)	13(12)
$\rho = 0.5$	95 % CR	91(2)	84(4)	16(11)	95(2)	94(4)	58(10)
	99 % CR	93(2)	86(4)	15(10)	99(2)	96(4)	53(10)
	AIC	41(3)	33(5)	1(15)	43(3)	37(5)	9(13)
	BIC	88(2)	77(4)	9(11)	91(2)	85(4)	36(11)
	CV	27(3)	18(6)	2(13)	30(3)	23(6)	13(12)

Table 5: Linear Models: Comparisons of model errors using adaptive L_1 norm for uncorrelated ($\rho = 0$) and correlated case ($\rho = 0.5$), for $p = 20$, $q = 2, 10$ and $n = 50$.

	$\rho = 0$				$\rho = 0.5$			
	2		10		2		10	
Method	ME-I	ME-II	ME-I	ME-II	ME-I	ME-II	ME-I	ME-II
95% CR	0.664	0.175	1.105	0.489	0.684	0.176	1.179	0.517
	(0.017)	(0.002)	(0.025)	(0.009)	(0.02)	(0.002)	(0.028)	(0.013)
99% CR	0.835	0.171	1.271	0.483	0.850	0.179	1.333	0.596
	(0.021)	(0.002)	(0.034)	(0.008)	(0.023)	(0.002)	(0.038)	(0.035)
AIC	0.436	0.566	0.663	0.711	0.433	0.571	0.670	0.714
	(0.009)	(0.012)	(0.011)	(0.016)	(0.01)	(0.015)	(0.012)	(0.016)
BIC	0.277	0.287	0.615	0.610	0.288	0.281	0.642	0.613
	(0.005)	(0.007)	(0.011)	(0.01)	(0.005)	(0.007)	(0.013)	(0.013)
CV	0.446	0.312	0.855	0.597	0.520	0.298	1.077	0.787
	(0.12)	(0.01)	(0.189)	(0.015)	(0.15)	(0.009)	(0.307)	(0.041)
OLS	0.791	-	0.806	-	0.795	-	0.811	-
	(0.016)		(0.017)		(0.016)		(0.012)	
Oracle	-	0.167	-	0.476	-	0.165	-	0.477
		(0.002)		(0.006)		(0.002)		(0.008)

Table 6: Linear Models: Comparisons of model errors using adaptive L_1 norm for uncorrelated ($\rho = 0$) and correlated case ($\rho = 0.5$), for $p = 20$, $q = 2, 10$ and $n = 100$.

	$\rho = 0$				$\rho = 0.5$			
	2		10		2		10	
Method	ME-I	ME-II	ME-I	ME-II	ME-I	ME-II	ME-I	ME-II
95% CR	0.428	0.121	0.650	0.315	0.418	0.124	0.690	0.315
	(0.007)	(0.001)	(0.009)	(0.003)	(0.009)	(0.001)	(0.008)	(0.003)
99% CR	0.521	0.118	0.726	0.315	0.511	0.122	0.771	0.315
	(0.007)	(0.001)	(0.009)	(0.003)	(0.01)	(0.001)	(0.01)	(0.003)
AIC	0.263	0.358	0.398	0.430	0.257	0.345	0.403	0.423
	(0.004)	(0.006)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.006)
BIC	0.168	0.161	0.372	0.364	0.163	0.149	0.373	0.352
	(0.001)	(0.003)	(0.004)	(0.004)	(0.001)	(0.002)	(0.004)	(0.004)
CV	0.332	0.187	0.518	0.369	0.368	0.184	0.579	0.395
	(0.098)	(0.004)	(0.125)	(0.004)	(0.118)	(0.003)	(0.165)	(0.007)
OLS	0.489	-	0.496	-	0.490	-	0.496	
	(0.004)		(0.006)		(0.005)		(0.006)	
Oracle	-	0.117	-	0.315	-	0.122	-	0.315
		(0.001)		(0.003)		(0.001)		(0.003)

Table 7: GLM: Comparisons of KL distances to true model using adaptive L_1 norm for uncorrelated ($\rho = 0$) and correlated case ($\rho = 0.5$), for $p = 20$, $q = 2, 10$ and $n = 150$. Each entry is the KL distance$\times$100

	$\rho = 0$				$\rho = 0.5$			
	2		10		2		10	
Method	KL-I	KL-II	KL-I	KL-II	KL-I	KL-II	KL-I	KL-II
95% CR	4.538	0.818	7.893	8.274	4.422	0.897	8.138	5.638
	(0.155)	(0.045)	(0.158)	(0.253)	(0.125)	(0.051)	(0.198)	(0.335)
99% CR	6.557	0.772	8.982	8.333	6.526	0.875	10.070	5.754
	(0.237)	(0.050)	(0.193)	(0.246)	(0.181)	(0.039)	(0.184)	(0.266)
AIC	1.451	2.423	8.808	16.658	1.419	2.256	5.669	9.100
	(0.073)	(0.173)	(0.195)	(0.372)	(0.063)	(0.124)	(0.486)	(0.729)
BIC	1.072	0.903	5.343	9.127	0.898	0.941	4.364	6.215
	(0.051)	(0.043)	(0.096)	(0.250)	(0.045)	(0.065)	(0.169)	(0.365)
CV	1.190	2.790	4.531	11.113	1.003	2.213	4.322	7.790
	(0.037)	(0.098)	(0.103)	(0.227)	(0.056)	(0.139)	(0.095)	(0.340)
GLM	10.535	-	23.516	-	10.468	-	13.267	-
	(0.239)		(0.380)		(0.245)		(1.284)	
Oracle	-	0.759	-	5.588	-	0.839	-	4.723
		(0.051)		(0.132)		(0.035)		(0.190)

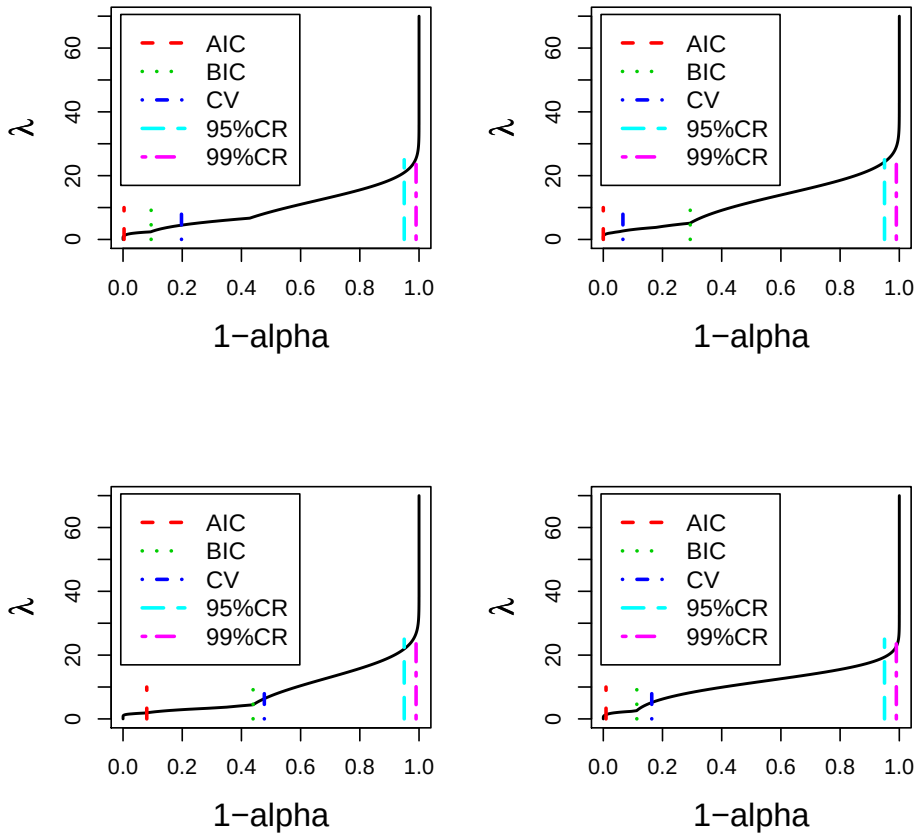
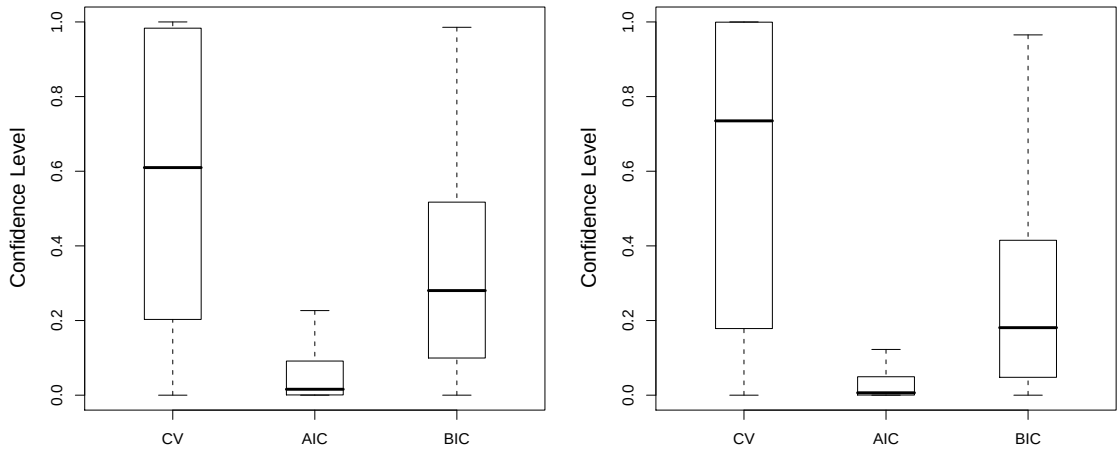


Figure 1: Relationship between λ and $1-\alpha$ for a sample of $n = 100,\ p = 20,\ q = 4$.



(a) (b)

Figure 2: Comparison of confidence region levels for models selected by CV, AIC and BIC. $n = 100$, $p = 20$ (a) q=4 (b) q=2.

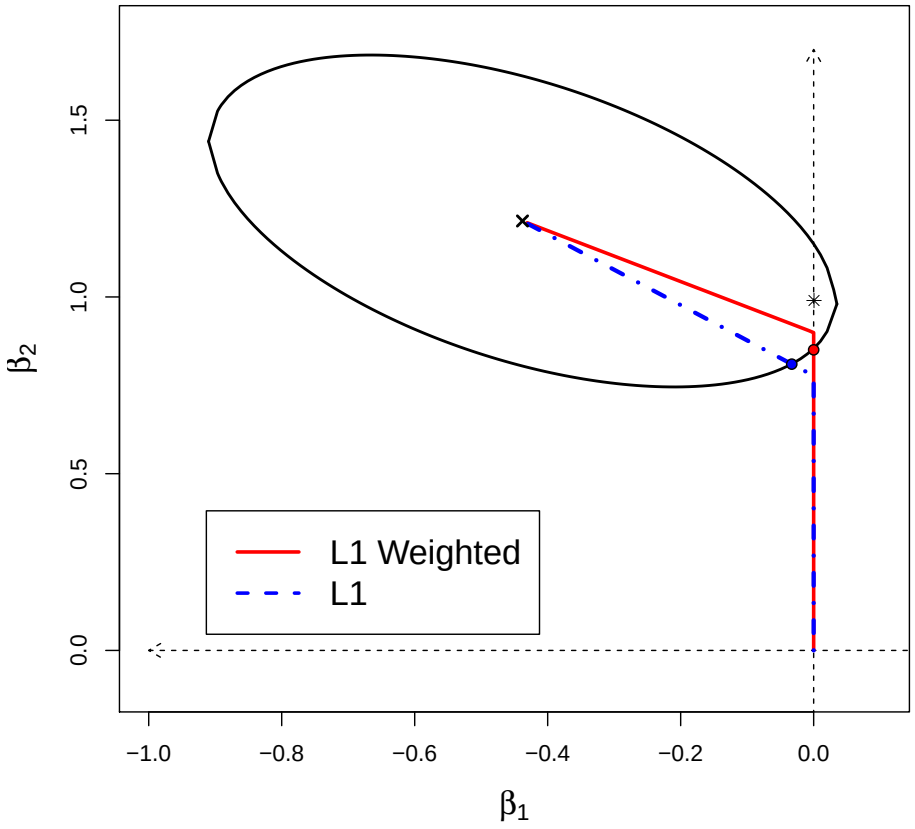


Figure 3: Schematic diagram of variable selection based on confidence region for L_1 norm and adaptive L_1 norm, for the case $n = 100$, $p = 2$ and true model coefficients, $\beta = (0, 1)$ (shown by a star on the y axes). The 'x' in the center of the ellipse represents the OLS estimate. The solid and dot-dashed lines are solution paths for adaptive L_1 norm and L_1 norm respectively. The point on each of these lines intersecting the confidence region is the solution to the proposed method.