Efficiency and Consistency for Regularization Parameter Selection in Penalized Regression: Asymptotics and Finite-Sample Corrections

Cheryl J. Flynn, Clifford M. Hurvich, and Jeffrey S. Simonoff

New York University

November 2, 2011

Extended Abstract

This paper studies the asymptotic and finite-sample performance of penalized regression methods when different selectors of the regularization parameter are used under the assumption that the true model is, or is not, included among the candidate model. In the latter setting, we relax assumptions in the existing theory to show that several classical information criteria are asymptotically efficient selectors of the regularization parameter. In both settings, we assess the finite-sample performance of these as well as other common selectors and demonstrate that their performance can suffer due to sensitivity to the number of variables that are included in the full model. As alternatives, we propose two corrected information criteria which are shown to outperform the existing procedures while still maintaining the desired asymptotic properties.

In the non-true model world, we relax the assumption made in the literature that the true error variance is known or that a consistent estimator is available to prove that Akaike's information criterion (AIC), C_p and Generalized cross-validation (GCV) themselves are asymptotically efficient selectors of the regularization parameter and

we study their performance in finite samples. In classical regression, AIC tends to select overly complex models when the dimension of the maximum candidate model is large relative to the sample size. Simulation studies suggest that AIC suffers from the same shortcomings when used in penalized regression. We therefore propose the use of the classical AIC_c as an alternative. In the true model world, a similar investigation into the finite sample properties of BIC reveals analogous overfitting tendencies and leads us to further propose the use of a corrected BIC (BIC_c). In their respective settings (whether the true model is, or is not, among the candidate models), BIC_c and AIC_c have the desired asymptotic properties and we use simulations to assess their performance, as well as that of other selectors, in finite samples for penalized regressions fit using the Smoothly clipped absolute deviation (SCAD) and Least absolute shrinkage and selection operator (Lasso) penalty functions. We find that AIC_c and 10-fold cross-validation outperform the other selectors in terms of squared error loss, and BIC_c avoids the tendency of BIC to select overly complex models when the dimension of the maximum candidate model is large relative to the sample size.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Least absolute shrinkage and selection operator; Model selection/ Variable Selection; Penalized regression; Smoothly clipped absolute deviation.

1 Introduction

Regularized (or penalized) regression methods have been widely used in recent years due to the increased availability of large data sets. In classical regression, variable selection (looking over all possible sets of predictors in a model) is commonly done using the Leaps and Bounds algorithm (Furnival and Wilson, 1974) but this method becomes infeasible when the number of predictors is much larger than 30 (Hastie et al., 2009). In contrast, in regularized regression increasing the amount of regularization increases the number of estimated coefficients that are set equal to zero thus performing "automatic" variable selection through the data-dependent choice of the regularization parameter, λ. For most penalty functions efficient algorithms

exist to compute the estimated models over a regularization path making it possible to do variable selection in high dimensions.

The performance of the estimated model heavily depends on the choice of the regularization parameter. In regularized regression several classical model selection procedures have been heuristically applied as selectors of this parameter including information criteria such as Akaike's information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), and Generalized cross-validation (GCV; Craven and Wahba, 1978) as well as data based selection procedures such as k-fold cross-validation (see, e.g., Fan and Li, 2001, Zou et al., 2007, Wang et al., 2007, and Zhang et al., 2010 for applications of these selectors to penalized regression estimators). The statistical properties of these model selection procedures have been widely studied in the context of classical regression and an ongoing research problem is to determine if these properties carry over to the context of penalized regression.

The asymptotic performance of model selection procedures can be studied under two important and distinct settings: (1) when the true model is not among the candidate models (the "non-true model world") and (2) when the true model is among the candidate models (the "true model world"). In the non-true model world a reasonable goal is *efficient* model selection, meaning that we would like to select the model that asymptotically performs the best amongst the candidate models. In contrast, in the true-model world most of the literature focuses on *consistent* model selection, meaning that the probability that the true model is chosen is asymptotically 1. Although the non-true model world has been extensively studied in classical regression (e.g., Shibata, 1981, Li, 1987, Hurvich and Tsai, 1989, 1991, Shao, 1997, and Burnham and Anderson, 2002) the majority of the research on model selection in penalized regression has focused on the true model world (e.g., Leng et al., 2006, Zou et al., 2007, and Wang et al., 2007). We feel that the non-true model world is more realistic in many situations since the data-generating process is likely to be too complex to know exactly. This setting should be of particular interest to researchers and data analysts in

areas such as social science and environmental health where a large number of predictors are expected to influence the dependent variable (too many to include in model fitting; Gelman, 2010) as well as machine learning where the goal is typically not to uncover the true data generating process but rather to find a model that can predict well.

To study model selection in regularized regression we consider the model

$$\mathbf{y}_n = \boldsymbol{\mu}_n + \boldsymbol{\varepsilon}_n$$

where $\mathbf{y}_n = (y_1, \ldots, y_n)^T$ is the $n \times 1$ response vector, $\boldsymbol{\mu}_n = (\mu_1, \ldots, \mu_n)^T$ is a $n \times 1$ unknown mean vector and the entries of the $n \times 1$ error vector ε_n are independent and identically distributed (iid) with mean 0 and variance σ^2. The mean vector is estimated by $\hat{\boldsymbol{\mu}}_n(\lambda) = \mathbf{X}_n\hat{\boldsymbol{\beta}}_n(\lambda)$ where $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is a $n \times d_n$ matrix of predictors and $\hat{\boldsymbol{\beta}}_n(\lambda)$ is the estimator which minimizes the penalized least squares function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \sum_{j=1}^{d_n} p_\lambda(|\beta_j|)$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{d_n}$ where d_n is the total number of predictors. This function consists of the residual sum of squares plus a penalty term which penalizes against model complexity and the size of the estimated coefficients, where the amount of penalization is controlled through the choice of λ. The minimum and maximum values that λ takes on are denoted by λ_{min} and λ_{max}, respectively.

Recently, Zhang et al. (2010) (hereafter ZLT) explored the use of the Generalized information criterion (Nishii, 1984),

$$GIC^{LS}_{\kappa_n}(\lambda) = \frac{1}{n}\left\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i\hat{\beta}_n(\lambda))^2 + \kappa_n\sigma^2 df_n(\lambda)\right\},$$

for choosing the regularization parameter λ for non-concave penalized estimators in both the non-true model world and the true-model world. Here $df_n(\lambda)$ is the effective degrees of

freedom. They showed that "AIC-type" versions of $GIC_{\kappa_n}^{LS}$ ($\kappa_n \to 2$) are efficient in the former case, while "BIC-type" versions of $GIC_{\kappa_n}^{LS}$ ($\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$) are consistent in the latter case. Unfortunately, the formula for $GIC_{\kappa_n}^{LS}$ includes the true error variance, σ^2, and their proofs operate under the assumption that this is known or that a consistent estimator is available. If the true model is not included in the set of candidate models then a consistent estimator of the true error variance may not be known (Shao, 1997) making the efficiency proofs of ZLT not applicable in practice.

This motivates us to extend the ZLT results in various ways. First, we show that the feasible version of GIC_2^{LS}, which corresponds to the well-known C_p measure (Mallows, 1973), is in fact efficient in the non-true model world. Second, we show that AIC and GCV, which do not require a consistent estimator of σ^2, are also efficient. Third, we show that although several model selection procedures may be asymptotically optimal, performance varies in finite samples. Specifically, we study performance when the number of predictors is allowed to be large relative to the sample size and show that AIC, BIC, C_p, and GCV all have a tendency to sometimes catastrophically overfit (lead to λ values approaching 0). In classical regression Hurvich and Tsai (1989) showed that AIC has a tendency to select overly complex models when the dimension of the maximum candidate model is large relative to the sample size and, recently, Chen and Chen (2008) showed that BIC suffers from the same issues. Hurvich and Tsai (1989) proposed a corrected version of AIC (AIC_c) and we further propose the corrected BIC (BIC_c) which is a simple analogue of AIC_c for the true model world. We show that these corrected versions preserve their respective asymptotic properties, but avoid the tendency of these methods to select overly complex models. We use Monte Carlo simulations to illustrate the properties of these methods in finite samples and compare their performance against the data-dependent method 10-fold CV. These results apply to a wide range of penalized regression estimators, including both non-concave penalized estimators and the well-known Least absolute shrinkage and selection operator (Lasso) estimator (Tibshirani, 1996).

K-fold CV is commonly used to select tuning parameters in both the statistical and machine learning literature. It operates by first randomly dividing the data set into k roughly equally sized subsets, then for each subset, the prediction error is computed based on the model fit using the data excluding that subset. The tuning parameter that minimizes the average square error computed across the subsets is then selected. In classical regression it has been shown that it should have the same asymptotic properties as $GIC^{LS}_{\kappa_n}$ with

$$\kappa_n = \frac{2k-1}{k-1}$$

(Shao, 1997). Applying this result, 10-fold CV should have the same asymptotic performance as $GIC^{LS}_{\kappa_n}$ with $\kappa_n = 2.\overline{11}$ implying that its behavior should be more closely related to the behavior of an efficient information criterion rather than a consistent one. Under the assumption of an orthonormal design matrix, Leng et al. (2006) showed that if the Lasso estimated model minimizes the prediction error then it will fail to select the true model with non-zero probability. The authors noted that this suggests that k-fold CV is inconsistent, but to our knowledge, the asymptotic properties of k-fold CV have not been fully established in the context of penalized regression. While a rigorous extension of the classical theory for k-fold CV is beyond the scope of this paper, the simulation results suggest that the same asymptotic properties apply.

The remainder of the paper is organized as follows. Section 2 briefly defines the model set-up and the model selection procedures that will be studied. Section 3 establishes the efficiency results for C_p, AIC, GCV and AIC_c without the assumption that the true population variance is known or that a consistent estimator exists, and explores the finite-sample behavior of the different selectors. Section 4 focuses on the true model world and studies with simulations the finite-sample performance of BIC and its corrected versions when the number of predictors is allowed to be large relative to the sample size. Concluding remarks are given in Section 5 and technical proofs are included in the supplementary material.

2 Model Set-up and Definition of Terms

Adopting the notation from ZLT, we let the index set $\mathcal{A}_n$ denote the class of all candidate models and we assume that $\bar{\alpha} = \{1, \ldots, d_n\}$ is the largest model in $\mathcal{A}_n$. For any $\alpha \in \mathcal{A}_n$, we define $d_n(\alpha)$ to be the number of predictor variables included in the candidate model. We further define the least squares estimated mean vector by $\hat{\boldsymbol{\mu}}_n^*(\alpha) = \mathbf{X}_n(\alpha)\hat{\boldsymbol{\beta}}_n^*(\alpha)$ where $\mathbf{X}_n(\alpha)$ is the matrix of predictors that are included in candidate model α and $\hat{\boldsymbol{\beta}}_n^*(\alpha)$ is the corresponding vector of the estimated least squares coefficients. The associated projection matrix is $\mathbf{H}_n(\alpha) = \mathbf{X}_n(\alpha)(\mathbf{X'}_n(\alpha)\mathbf{X}_n(\alpha))^{-1}\mathbf{X'}_n(\alpha)$. For a given λ, we define α_λ to be the model $\alpha \in \mathcal{A}_n$ whose predictors are those with non-zero coefficients in the penalized estimator $\hat{\boldsymbol{\beta}}_n(\lambda)$ and let $df_n(\lambda)$ denote the effective degrees of freedom. The least squares estimated mean vector based on the model α_λ is denoted by $\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) = \mathbf{X}_n(\alpha_\lambda)\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)$. In this equation, $\mathbf{X}_n(\alpha_\lambda)$ is the matrix of predictors whose coefficients are not shrunk to zero in the penalized estimator $\hat{\boldsymbol{\beta}}_n(\lambda)$ and $\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)$ are the estimated coefficients from the least squares model fit using these predictors. The associated projection matrix in this case is defined as $\mathbf{H}_n(\alpha_\lambda) = \mathbf{X}_n(\alpha_\lambda)(\mathbf{X'}_n(\alpha_\lambda)\mathbf{X}_n(\alpha_\lambda))^{-1}\mathbf{X'}_n(\alpha_\lambda)$.

If we assume that we are in the non-true model world, then a reasonable goal is efficient model selection. The L_2 loss is commonly used to assess the predictive performance of an estimator and is calculated as

$$L(\hat{\boldsymbol{\beta}}_n(\lambda)) = \frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n}.$$

For the efficiency proofs we further require the following notation. In classical regression the risk function is defined as

$$R(\hat{\boldsymbol{\beta}}_n^*(\alpha)) = E\left(\frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha)||^2}{n}\right) = \Delta_n(\alpha) + \frac{\sigma^2 d_n(\alpha)}{n}$$

where $\Delta_n(\alpha) = ||\boldsymbol{\mu}_n - \mathbf{H}_n(\alpha)\boldsymbol{\mu}_n||^2/n$. Letting $d_n(\alpha_\lambda)$ denote the number of predictors with

non-zero coefficients in the penalized estimator $\hat{\boldsymbol{\beta}}_n(\lambda)$, we further define the function

$$\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) = \Delta_n(\alpha_\lambda) + \frac{\sigma^2 d_n(\alpha_\lambda)}{n}$$

which is a random variable.

If we let $\hat{\lambda}_n$ denote the regularization parameter selected by a given selection procedure, then the procedure is defined to be *asymptotically loss efficient* if

$$\frac{L(\hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n))}{\inf_{\lambda\in[0,\lambda_{max}]} L(\hat{\boldsymbol{\beta}}_n(\lambda))} \to_p 1$$

and $\hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n)$ is said to be an *asymptotically loss efficient estimator*. If instead it is assumed that there exists a unique (minimal) true model, α_0, in the set of candidate models then the common goal in the literature is consistent model selection. If $\hat{\lambda}_n$ denotes the regularization parameter selected by a given selection procedure, then the procedure is defined to be *asymptotically consistent* if

$$P(\alpha_{\hat{\lambda}_n} = \alpha_0) \to 1$$

and $\hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n)$ is said to be an *asymptotically consistent estimator*.

2.1 Choice of Penalty Function

The theory and simulations presented here consider two penalized regression estimators. The Smoothly clipped absolute deviation (SCAD) penalty function was proposed by Fan and Li (2001). This penalty function is defined by

$$p'_\lambda(\beta) = \lambda\left\{I(\beta \le \lambda) + \frac{(a\lambda-\beta)_+}{(a-1)\lambda} I(\beta > \lambda)\right\}$$

for some $a > 2$ and $\beta > 0$. Fan and Li (2001) recommended setting the second tuning parameter in the SCAD penalty function, a, equal to 3.7 and this is commonly done in

practice; however, doing so will not necessarily guarantee that the SCAD objective function is convex and can result in convergence to local, but non-global, minima. As a result, in addition to studying the performance of SCAD with $a = 3.7$ (SCAD, 3.7), we study the performance of SCAD where $a = \max(3.7, 1 + 1/c^*)$ (SCAD) where c^* is the minimum eigenvalue of $n^{-1}\mathbf{X}_\mathbf{n}'\mathbf{X}_\mathbf{n}$. The latter choice will force the objective function to be convex (Breheny and Huang, 2011).

The wide use of SCAD is mainly due to the fact that it satisfies the "oracle property." This means that, assuming that the true model is in the set of candidate models and subject to certain regularity assumptions, there exists a sequence $\{\lambda_n\}$ such that if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ then with probability tending to one the SCAD-estimated regression based on the full model will correctly zero out any zero coefficients and have the same asymptotic distribution as the least squares regression based on the correct model. This result was proven originally for d_n fixed by Fan and Li (2001) and was extended to the case where $d_n < n$ but $d_n \to \infty$ by Huang and Xie (2007). These results are for an unknown deterministic sequence which needs to be estimated in practice.

Another common choice for the penalty function is the Lasso proposed by Tibshirani (1996). The Lasso penalty is the L_1-norm of the coefficients. Necessary and sufficient conditions have been established for the Lasso to perform consistent model selection (Zhao and Yu, 2006), but in general the Lasso produces biased estimates and does not satisfy the oracle property (Zou, 2006). However, in the non-true model world, the oracle property has no meaning, since there is no true model. Further, the oracle property is an asymptotic property. Therefore, we include the Lasso in the simulation studies.

In all examples, the Lasso regressions are fit using the R `lars` package (Hastie and Efron, 2011) and the SCAD regressions are fit using the R `ncvreg` package (Breheny and Huang, 2011). The `lars` package computes the entire regularization path for the Lasso and for SCAD the models are fit over a grid of 200 λ values from λ_{min} to λ_{max}, where the first 100 values of λ are fit on a log-scale and the last 100 values of λ are equally spaced. Breheny and Huang

(2011) considered a grid of 100 λ values in their simulation studies. We have chosen a grid that is twice as fine in order to remain closer to the theoretical assumption that all possible values of λ are considered. In all simulations, λ_{max} is selected so that all of the estimated coefficients are zero and λ_{min} is chosen to effectively produce the least squares estimate on the full model.

2.2 Model Selection Procedures

In addition to 10-fold CV, we study the performance of several information criteria. Specifically, we consider

$$AIC_\lambda = \log(\hat{\sigma}_n^2(\lambda)) + 2\frac{df_n(\lambda)}{n},$$

$$AIC_{c_\lambda} = \log(\hat{\sigma}_n^2(\lambda)) + 2\frac{df_n(\lambda)+1}{n-df_n(\lambda)-2},$$

$$BIC_\lambda = \log(\hat{\sigma}_n^2(\lambda)) + \log(n)\frac{df_n(\lambda)}{n},$$

$$GCV_\lambda = \frac{\hat{\sigma}_n^2(\lambda)}{(1-df_n(\lambda)/n)^2},$$

and

$$C_{p_\lambda} = \hat{\sigma}_n^2(\lambda) + 2\frac{df_n(\lambda)\tilde{\sigma}_n^2}{n}.$$

In the above we define

$$\hat{\sigma}_n^2(\lambda) = \frac{||\mathbf{y}_n - \mathbf{X}_n\hat{\boldsymbol{\beta}}_n(\lambda)||^2}{n}$$

and

$$\tilde{\sigma}_n^2 = \frac{||\mathbf{y}_n - \mathbf{X}_n\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha})||^2}{n-d_n-1}.$$

With the exception of 10-fold CV, all of the above model selection procedures require a definition of the effective degrees of freedom for the penalized regression method. In what follows, we use a heuristic definition and define the effective degrees of freedom to be the number of non-zero coefficients in $\hat{\boldsymbol{\beta}}_n(\lambda)$ and denote this by $d_n(\alpha_\lambda)$. Zou et al. (2007) proved

that the number of non-zero coefficients is an unbiased estimator of the degrees of freedom for the Lasso. For SCAD, Fan and Li (2001) proposed setting the degrees of freedom equal to the trace of the approximate linear projection matrix. Based on Proposition 1 from ZLT, our efficiency proofs would still hold if this alternate definition is used.

3 Non-True Model World

We show here that assuming that the true model is not in the set of candidate models C_{p_λ}, AIC_λ, GCV_λ, and AIC_{c_λ} are efficient selectors of the regularization parameter. The dimension of the full model, d_n, is allowed to tend to infinity with n but it is assumed that $d_n/n \to 0$. The efficiency proofs operate under the same assumptions as those of ZLT, which are presented here for completeness:

(A1) $(\frac{1}{n}\mathbf{X'}_n\mathbf{X}_n)^{-1}$ exists and its largest eigenvalue is bounded by a constant number C.

(A2) $E\varepsilon_1^{4q} < \infty$, for some positive integer q.

(A3) The risks of the least squares estimators $\hat{\boldsymbol{\beta}}_n^*(\alpha)$ satisfy

$$\sum_{\alpha \in \mathcal{A}_n} (nR(\hat{\boldsymbol{\beta}}_n^*(\alpha)))^{-q} \to 0.$$

(A4)

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{||\mathbf{b}_n||^2}{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \to_p 0,$$

where $\mathbf{b}_n$ is a $d_n \times 1$ vector where $b_{n,j} = p'_\lambda(|\hat{\beta}_{n_j}(\lambda)|)sgn(\hat{\beta}_{n_j}(\lambda))$ for all j such that $|\hat{\beta}_{n_j}(\lambda)| > 0$ and is equal to 0 otherwise.

The first three assumptions are common in the literature on model selection. Assumption (A4) is the only assumption that involves the penalty function and the authors provided three sufficient conditions for the assumption to be satisfied. It is important to note that although

ZLT only studied non-concave penalty functions, if the non-zero estimated coefficients, $\hat{\boldsymbol{\beta}}^1_n(\lambda)$, satisfy a relationship of the form

$$\hat{\boldsymbol{\beta}}^1_n(\lambda) = (\mathbf{X}'_n(\alpha_\lambda)\mathbf{X}'_n(\alpha_\lambda))^{-1}\mathbf{X}_n(\alpha_\lambda)\mathbf{y}_n + \left(\frac{1}{n}\mathbf{X}'_n(\alpha_\lambda)\mathbf{X}'_n(\alpha_\lambda)\right)^{-1}\mathbf{b}^1_n$$

with probability tending to 1 and (A4) is satisfied, then the efficiency proofs will hold for any penalty function. In particular, based on Lemma 2 of Zou et al. (2007), the Lasso satisfies this relationship and the same sufficient conditions provided by the ZLT for (A4) can be used. Therefore, the efficiency proofs will hold for the Lasso so it is interesting to compare the performance of the two penalty functions.

The asymptotic efficiency of C_{p_λ} is given by the following result.

Theorem 1. *Under the assumptions of ZLT and that $d_n/n \to 0$ as $n \to \infty$, the regularization parameter, $\hat{\lambda}_n$, selected by minimizing*

$$C_{p_\lambda} = \hat{\sigma}^2_n(\lambda) + \frac{2d_n(\alpha_\lambda)\tilde{\sigma}^2_n}{n}$$

yields an asymptotically loss efficient estimator, $\hat{\beta}_n(\hat{\lambda}_n)$.

To further establish the efficiency of AIC_λ, GCV_λ and AIC_{c_λ} we require the following two results.

Theorem 2. *Under the assumptions of ZLT and that $d_n/n \to 0$ as $n \to \infty$, the regularization parameter, $\hat{\lambda}_n$, selected by minimizing*

$$\Gamma_n(\lambda) = \hat{\sigma}^2_n(\lambda)\left(1 + \frac{2d_n(\alpha_\lambda)}{n}\right)$$

yields an asymptotically loss efficient estimator, $\hat{\beta}_n(\hat{\lambda}_n)$.

Theorem 3. *Under the assumptions of Theorem 2, any information criterion that can be*

written in the form

$$\tilde{\Gamma}_n(\lambda) = \hat{\sigma}_n^2(\lambda)\left(1 + \frac{2d_n(\alpha_\lambda)}{n} + \delta_n(\lambda)\right)$$

where

$$\sup_{\lambda\in[0,\lambda_{max}]} |\delta_n(\lambda)| \to_p 0 \tag{C1}$$

and

$$\sup_{\lambda\in[0,\lambda_{max}]} \frac{|\delta_n(\lambda)|}{L(\hat{\beta}_n(\lambda))} \to_p 0, \tag{C2}$$

is an asymptotically loss efficient procedure for selecting λ*.*

Condition (C2) in Theorem 3 is a stronger assumption than in the analogous result established by Theorem 4.2 in Shibata (1980) for selecting the optimal order of a linear process, but Theorem 3 is sufficient to show that AIC_λ, GCV_λ, and AIC_{c_λ} are asymptotically loss efficient model selection procedures for the regularization parameter. All three methods can be shown to satisfy (C1) and (C2) using Taylor series expansions. The details are provided in the supplementary material.

3.1 Finite Sample Performance

In this section we study the finite sample performance of the model selection procedures discussed in Section 2.2 when the true model is not included in the set of candidate models. The first set of simulations considers a trigonometric regression where the candidate models are in the neighborhood of the true model but never include the true model. This example is in line with the framework considered by Shibata (1980) and Hurvich and Tsai (1991). The second set of simulations look at an example where there is an omitted predictor. For example, the researcher may have access to some of the relevant predictors but may be missing others. This is the setting that was considered by ZLT.

In all of the examples, the results are based on 1000 realizations of samples with $n = 50, 100$, and 150, and the selection procedures are evaluated based on their L_2 loss efficiency,

L_2 loss, and the variability of the selected number of non-zero coefficients. For each realization, if we let $\hat{\lambda}$ denote the regularization parameter selected by a given selection procedure, then the loss efficiency is computed as

$$\frac{L(\hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n))}{\min_{\lambda\in[0,\lambda_{max}]} L(\hat{\boldsymbol{\beta}}_n(\lambda))}.$$

3.1.1 Trigonometric Regression

Here we consider a trigonometric example based on an example studied in Hurvich and Tsai (1991). The true model is the model described as

$$y_i = e^{4i/n} + \varepsilon_i$$

for $i = 1, \ldots, n$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The candidate models are SCAD and Lasso penalized regressions where the matrix of predictors, $\mathbf{X}_n = (\mathbf{x}_n^1, \mathbf{x}_n^2)$, is a $n \times d_n$ matrix with components defined by

$$x_{i,j}^1 = \sin\left(\frac{2\pi j}{n} i\right)$$

and,

$$x_{i,j}^2 = \cos\left(\frac{2\pi j}{n} i\right)$$

for $j = 1, \ldots, d_n/2$ and $i = 1, \ldots, n$. The maximum number of predictors is allowed to vary by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$. We consider values of c on the grid $(0.5, 0.7, 0.9, 0.98)$. Note that examining d_n close to n allows study of high-dimensional data problems, and is in the spirit of simulations performed in Tibshirani (1996) and Zou and Hastie (2005). Since the predictor variables are orthogonal in this example, setting $a = 3.7$ for SCAD satisfies the convexity constraint for all values of c.

We examine both $\sigma^2 = 50$ and $\sigma^2 = 100$, but the patterns for the two error variances are similar so only the results for $\sigma^2 = 100$ are reported. The average L_2 loss efficiency is presented in Table 1 for both SCAD and Lasso. For all values of c, the average loss efficiency

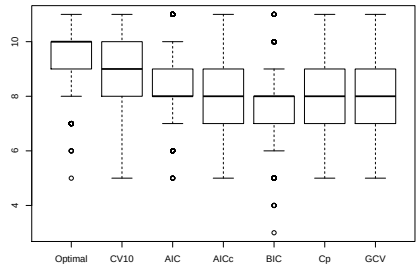
of AIC_{c_λ} and C_{p_λ} tend to one as the sample size increases, while the average loss efficiency of BIC_λ does not show signs of convergence. These patterns are consistent with the theoretical results. When the number of predictor variables is small relative to the sample size, the loss efficiency of AIC_λ also tends to one; however, as the number of predictors is increased, the performance of AIC_λ deteriorates. Figure 1 displays boxplots of the selected number of non-zero coefficients when $n = 100$ and $\sigma^2 = 100$. From this plot we see that AIC_λ often selects a model that is close to the full model when c is large. As the sample size is increased the full model becomes less desirable and AIC_λ suffers as a result. For SCAD, GCV_λ appears to suffer from a similar problem, but to a lesser extent than AIC_λ. The difference in performance for varying values of c suggests that the good asymptotic performance of AIC_λ and GCV_λ is strongly dependent on the fact that $d_n/n \to 0$ and these selectors may not perform well in finite samples when this ratio is close to 1.

Figure 2 presents boxplots of the L_2 loss for the 1000 realizations when $n = 100$. AIC_λ clearly suffers as c is increased and BIC_λ is outperformed by the remaining methods. From Figure 1 we see that BIC_λ generally selects a model that is more parsimonious than the optimal model, but when the number of parameters is large relative to the sample size it also has some tendency to pick models with dimension close to d_n. Furthermore, we see that the model dimension selected by C_{p_λ} and GCV_λ varies widely when the number of predictor variables is large, while AIC_{c_λ} and 10-fold CV are generally more stable in their choices.

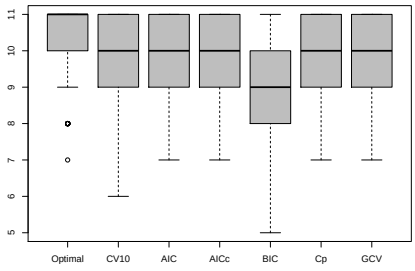
Table 1: Average L2 Loss Efficiency over 1000 simulations for the exponential model.

$\sigma^2 = 100$									
		SCAD				Lasso			
Info. Crit.	n	c=.5	c=.7	c=.9	c=.98	c=.5	c=.7	c=.9	c=.98
10-fold CV	50	1.08	1.25	1.31	1.33	1.03	1.14	1.25	1.32
	100	1.07	1.16	1.14	1.16	1.03	1.10	1.14	1.19
	150	1.06	1.15	1.11	1.13	1.03	1.08	1.11	1.16
AIC_λ	50	1.06	1.13	1.40	1.92	1.02	1.10	1.28	1.82
	100	1.06	1.11	1.60	2.44	1.03	1.08	1.34	2.15
	150	1.04	1.10	1.75	2.83	1.02	1.07	1.37	2.41
AIC_{c_λ}	50	1.08	1.20	1.34	1.38	1.04	1.24	1.47	1.49
	100	1.07	1.13	1.19	1.22	1.04	1.14	1.24	1.28
	150	1.05	1.11	1.16	1.18	1.03	1.11	1.19	1.22
BIC_λ	50	1.13	1.35	1.57	1.85	1.08	1.40	1.70	1.78
	100	1.13	1.39	1.56	1.74	1.11	1.52	1.69	1.69
	150	1.12	1.42	1.59	1.61	1.11	1.61	1.72	1.66
C_{p_λ}	50	1.07	1.15	1.24	1.48	1.02	1.12	1.19	1.38
	100	1.06	1.11	1.23	1.44	1.03	1.09	1.12	1.24
	150	1.05	1.10	1.22	1.43	1.03	1.08	1.10	1.22
GCV_λ	50	1.07	1.15	1.27	1.59	1.03	1.14	1.21	1.30
	100	1.06	1.11	1.28	1.73	1.03	1.10	1.14	1.23
	150	1.05	1.10	1.29	1.83	1.03	1.09	1.12	1.21

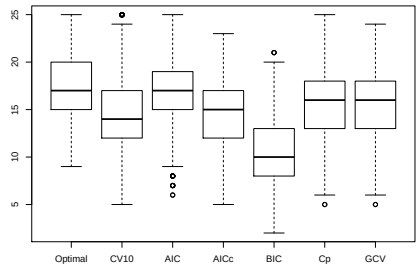
Figure 1: Comparison of model selection procedures based on the number of non-zero coefficients (includes intercept) in the selected model over 1000 simulations for the exponential model with $n = 100$ and $\sigma^2 = 100$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$.



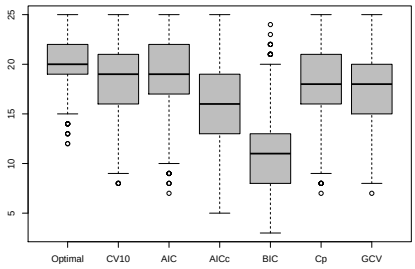
(a) SCAD, c=.5



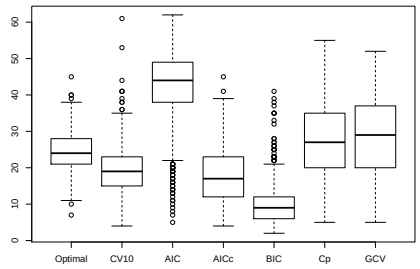
(b) Lasso, c=.5



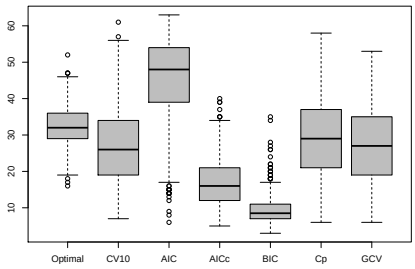
(c) SCAD, c=.7



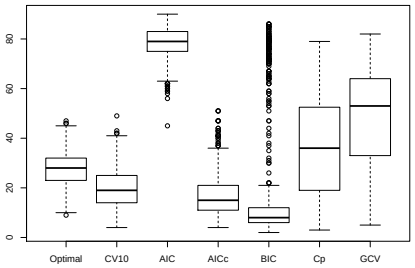
(d) Lasso, c=.7



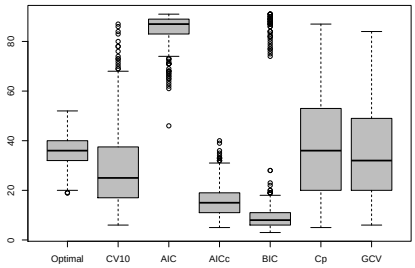
(e) SCAD, c=.9



(f) Lasso, c=.9

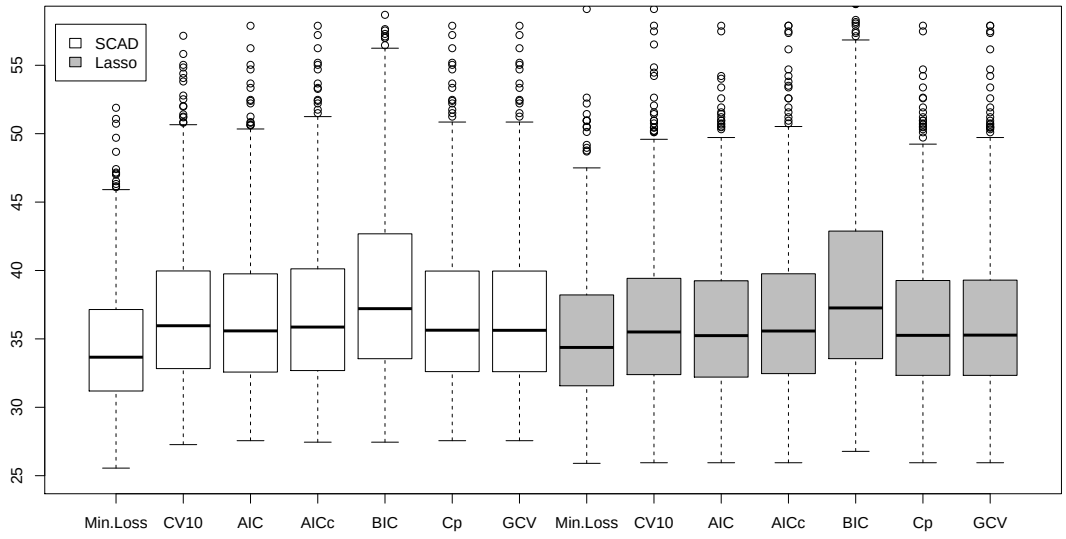


(g) SCAD, c=.98

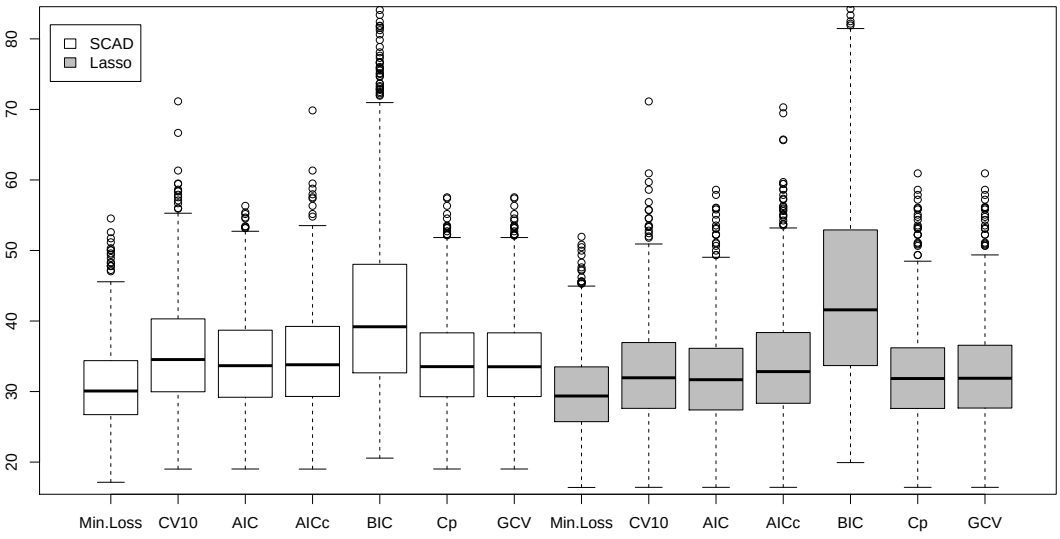


(h) Lasso, c=.98

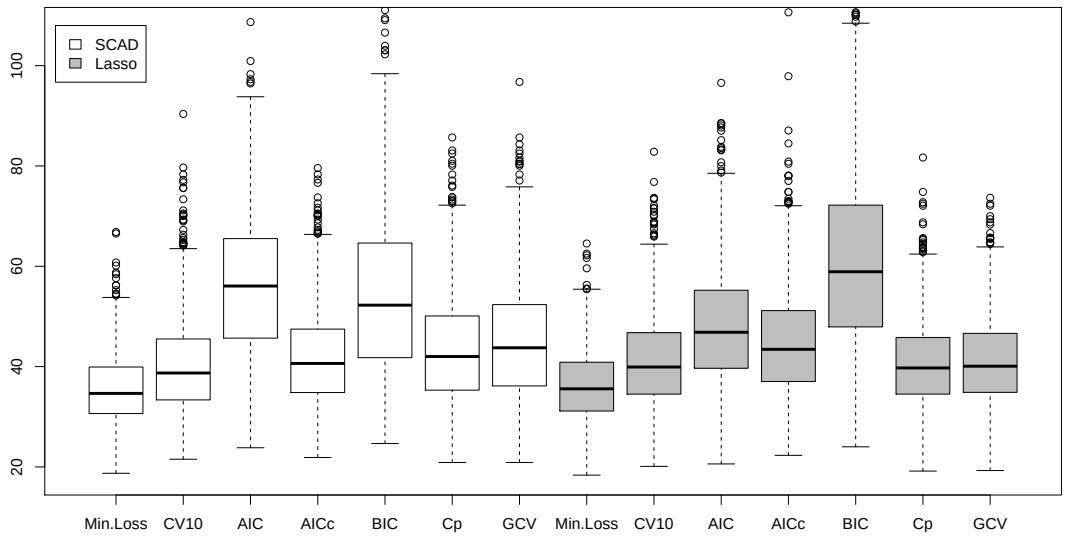
Figure 2: Comparison of model selection procedures based on L2 Loss over 1000 simulations for the exponential model with $n = 100$ and $\sigma^2 = 100$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$. In order to make it easier to compare the procedures, the limits of the vertical axis are specified so that all the boxes and whiskers appear but some of the outliers are not shown.



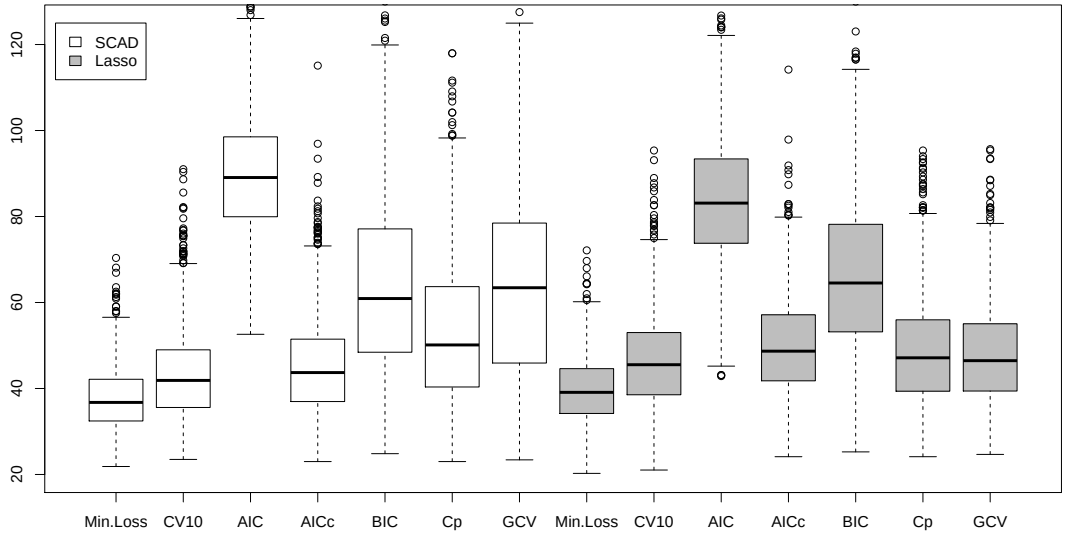
(a) c=.5



(b) c=.7



(c) c=.9



(d) c=.98

From Figure 2 we can also compare the performance of SCAD and the Lasso. Based on minimum loss, the difference between SCAD and Lasso is significant based on a signed rank test, though neither method is the clear winner, with SCAD outperforming the Lasso for $c = .5, .9$, and .98 and the Lasso outperforming SCAD when $c = .7$. Still, it is striking that

from a practical point of view the predictive accuracies of the two methods are very similar. Overall, the sensitivity to the value of c clearly hurts the performance of AIC_λ and can also negatively impact the performance of C_{p_λ} and GCV_λ. The impact on the latter two is more noticeable when looking at SCAD, but in both cases the extreme variability in the size of the selected model is undesirable. As a result, we recommend the use of AIC_c or 10-fold CV which are less sensitive to the closeness of d_n to n. 10-fold CV outperforms AIC_c for the Lasso, and for SCAD, AIC_{c_λ} outperforms 10-fold CV when $c = .7$, while the opposite is true when $c = .9$ and .98. The difference in performance in these cases is statistically significant based on a signed rank test.

3.1.2 Omitted Predictor

Here we study an example based on example 2 in ZLT where there is an omitted predictor. The true model is defined as

$$y_i = 3x_{i,1} + 1.5x_{i,2} + 2x_{i,10} + .2x_{i,13} + \varepsilon_i$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2 = 16)$. We let $\mathbf{X}_n$ be the $n \times (d_n + 1)$ matrix of predictors where the $\mathbf{x}_i's$ are simulated from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ where $\Sigma_{i,j} = \rho^{|i-j|}$ for $\rho = 0$ and 0.5. The candidate models are SCAD and Lasso penalized regressions based on $\mathbf{X}_n$ except with the 13^{th} column removed so that the true model is never included in the set of candidate models.

In both examples the number of superfluous variables included in the candidate models is allowed to vary by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$. For this example we consider values of c on the grid $(0.5, 0.7, 0.9, 0.98)$. In this example setting $a = 3.7$ will not satisfy the convexity constraint for all values of c. Therefore, we further compare the case where $a = 3.7$ (SCAD, $a = 3.7$) to the case where $a = \max(3.7, 1 + 1/c^*)$ (SCAD).

We first consider Figure 3 which presents boxplots comparing the three estimators based on loss when $n = 100$ and $\rho = 0.5$. From these plots it is immediately clear that all of the

information criteria perform better when a is allowed to be data-dependent, while 10-fold CV performs well regardless of the choice of a. One possible explanation for this is that all of the information criteria under consideration were derived for use in classical least squares regression so they should perform well assuming that the estimated models are close to the corresponding OLS models. When the second tuning parameter of SCAD is fixed at 3.7, the objective function is not necessarily convex so the SCAD-estimated models may be very far from the OLS models. On the other hand, 10-fold CV is a general model selection procedure which should work in a variety of settings. In general, we recommend using a data-dependent choice of a since it requires little additional cost and can greatly improve the performance of all of the information criteria.

Focusing only on the data-dependent choice of a, we see that the performance of the model selection procedures is similar for both SCAD and Lasso. Based on minimum loss, the difference between SCAD and Lasso is statistically significant based on a signed rank test, though again neither method is the clear winner with SCAD outperforming the Lasso for $c = .5, .7$, and .9 and the Lasso outperforming SCAD when $c = .98$. Figure 4 presents boxplots of the selected number of non-zero coefficients. As was the case with the exponential model, AIC_λ has a strong tendency to select models that contain almost all of the predictor variables and the dimension of the models selected by C_{p_λ} and GCV_λ become extremely variable as c is increased. In contrast, 10-fold CV and AIC_{c_λ} are less sensitive to the number of predictor variables included in the model. In Figure 3 it is clear that this sensitivity to the value of c impacts the performance of the model selection procedures, and as a result 10-fold CV and AIC_{c_λ} outperform the other procedures. We use a signed rank test to test the hypothesis that the performances of 10-fold CV and AIC_{c_λ} are equal. This test produces p-values equal to 0.0734, 0.0000, 0.6429, and 0.5286 for SCAD and 0.0153, 0.9736, 0.0144, 0.1880 for the Lasso for $c = .5, .7, .9$, and .98, respectively, suggesting that the performances of the two methods are comparable.

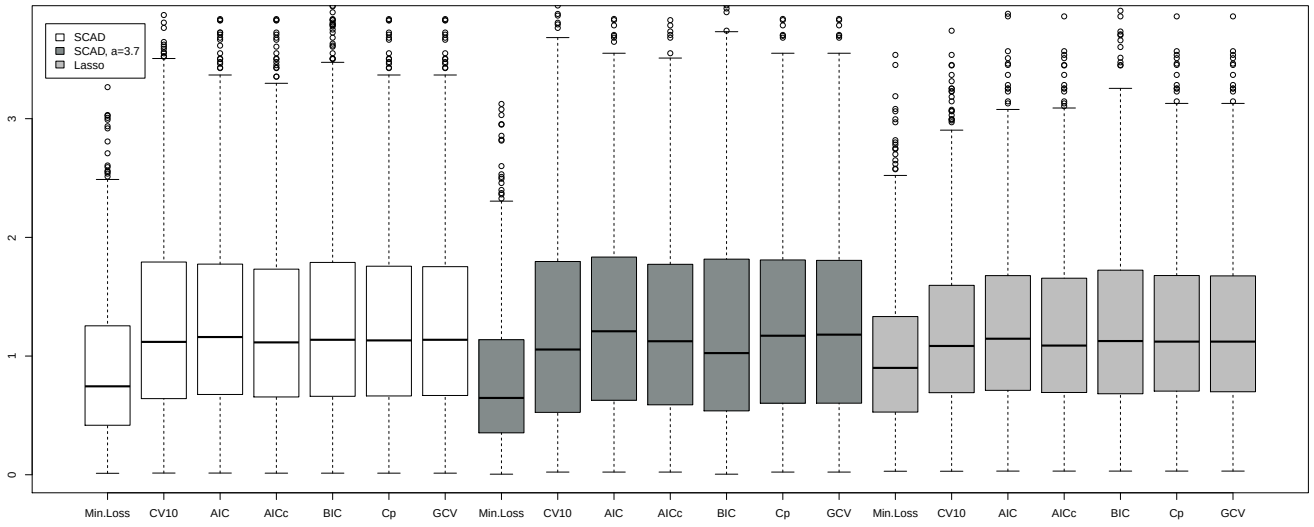
In order to study the asymptotic behavior of the selection procedures, Table 2 presents

the average loss efficiencies. The patterns are similar for both values of ρ so only the results for $\rho = 0.5$ are reported. In general, the loss efficiencies of AIC_{c_λ}, C_{p_λ}, and GCV_λ tend to 1, while the loss efficiency of BIC_λ does not show signs of convergence. Also, the results again show that AIC_λ performs poorly when the number of predictor variables is large relative to the sample size. Overall, the results corroborate the theoretical findings, but reinforce that the finite sample performance of asymptotically equivalent methods may vary greatly.

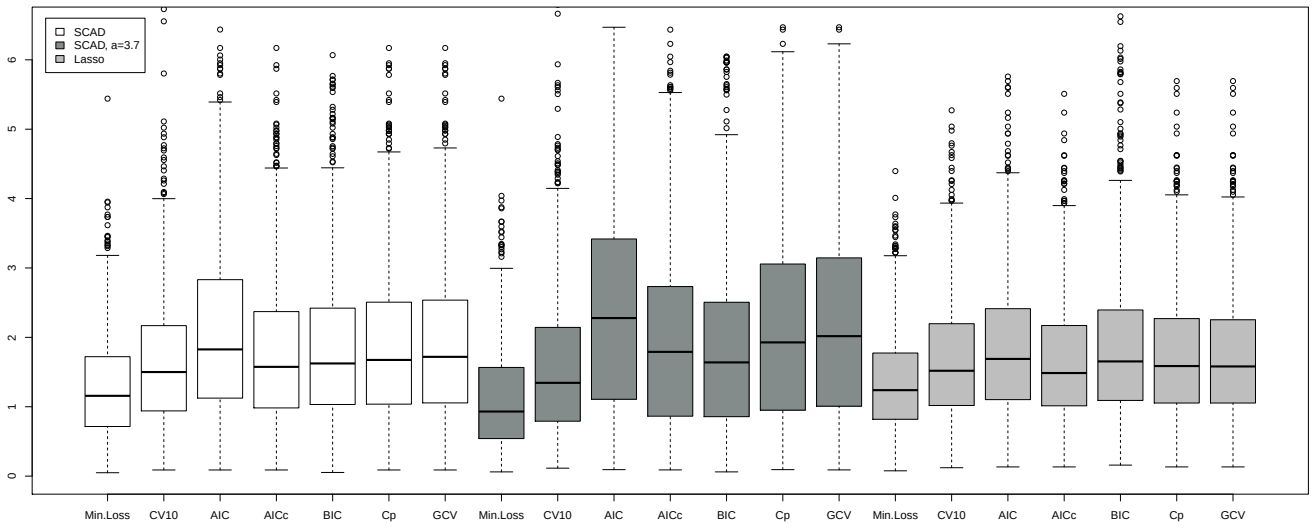
Table 2: Average L2 Loss Efficiency over 1000 simulations for the model with an omitted predictor.

$\rho = .5$													
		SCAD				SCAD, $a = 3.7$				Lasso			
Info. Crit.	n	c=.5	c=.7	c=.9	c=.98	c=.5	c=.7	c=.9	c=.98	c=.5	c=.7	c=.9	c=.98
10-fold CV	50	–	1.48	1.38	1.46	–	1.74	1.61	1.57	–	1.40	1.40	1.46
	100	1.89	1.42	1.27	1.28	2.14	1.66	1.54	1.48	1.34	1.31	1.28	1.30
	150	1.54	1.40	1.24	1.21	1.67	1.64	1.48	1.43	1.29	1.27	1.23	1.23
AIC_λ	50	–	1.90	3.42	6.17	–	2.16	4.59	6.94	–	1.57	2.84	6.09
	100	1.92	2.00	4.07	9.43	2.33	2.87	7.86	12.04	1.43	1.52	3.21	9.37
	150	1.73	2.03	4.13	12.40	2.05	3.24	11.91	18.91	1.36	1.48	2.95	2.95
AIC_{c_λ}	50	–	1.56	1.35	1.28	–	1.81	2.47	3.36	–	1.32	1.28	1.25
	100	1.83	1.59	1.25	1.24	2.21	2.12	3.52	5.54	1.37	1.29	1.22	1.21
	150	1.64	1.60	1.25	1.21	1.92	2.46	4.24	8.77	1.32	1.28	1.21	1.21
BIC_λ	50	–	1.60	1.62	3.77	–	1.71	2.39	5.46	–	1.42	1.45	3.58
	100	1.69	1.51	1.49	1.60	1.78	1.85	1.98	3.29	1.35	1.40	1.40	1.48
	150	1.59	1.60	1.53	1.51	1.70	1.87	1.95	2.04	1.38	1.41	1.42	1.42
C_{p_λ}	50	–	1.72	1.87	2.77	–	1.98	3.02	4.27	–	1.42	1.66	2.73
	100	1.87	1.74	1.62	2.48	2.26	2.41	4.33	6.04	1.41	1.40	1.48	2.46
	150	1.67	1.83	1.50	2.15	1.99	2.75	5.34	8.10	1.35	1.36	1.39	1.39
GCV_λ	50	–	1.73	1.92	2.28	–	2.01	3.61	5.76	–	1.42	1.52	2.08
	100	1.88	1.79	1.60	1.63	2.26	2.51	6.07	10.00	1.41	1.40	1.44	1.60
	150	1.67	1.81	1.52	1.41	2.00	2.91	8.31	16.01	1.34	1.36	1.36	1.36

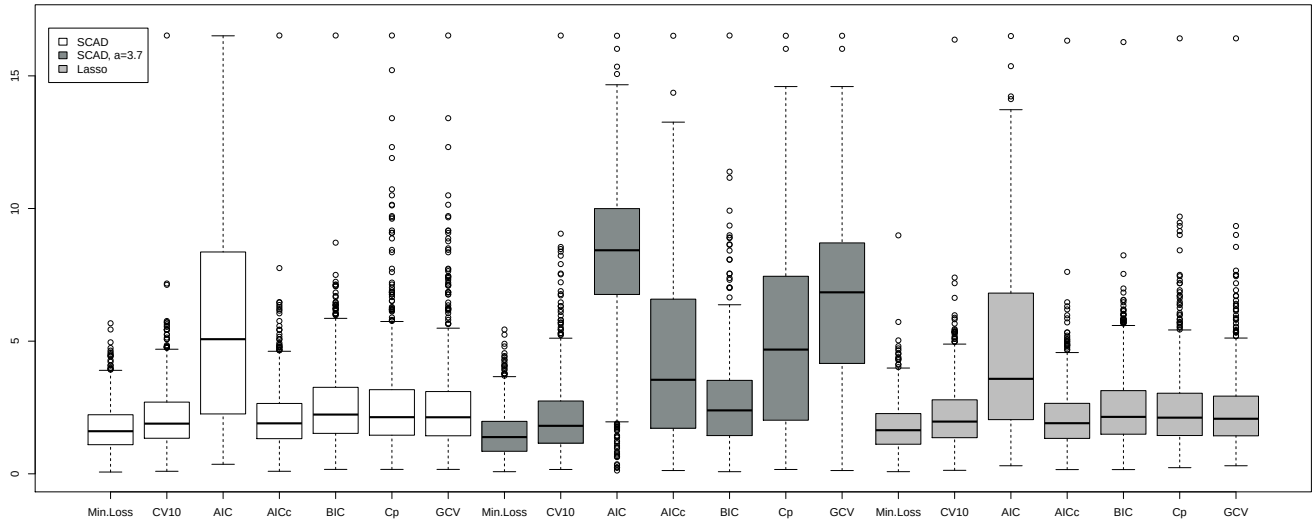
Figure 3: Comparison of model selection procedures based on L2 Loss over 1000 simulations for the model with an omitted predictor with $n = 100$ and $\rho = 0.5$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$. In order to make it easier to compare the procedures, the limits of the vertical axis are specified so that all the boxes and whiskers appear but some of the outliers are not shown.



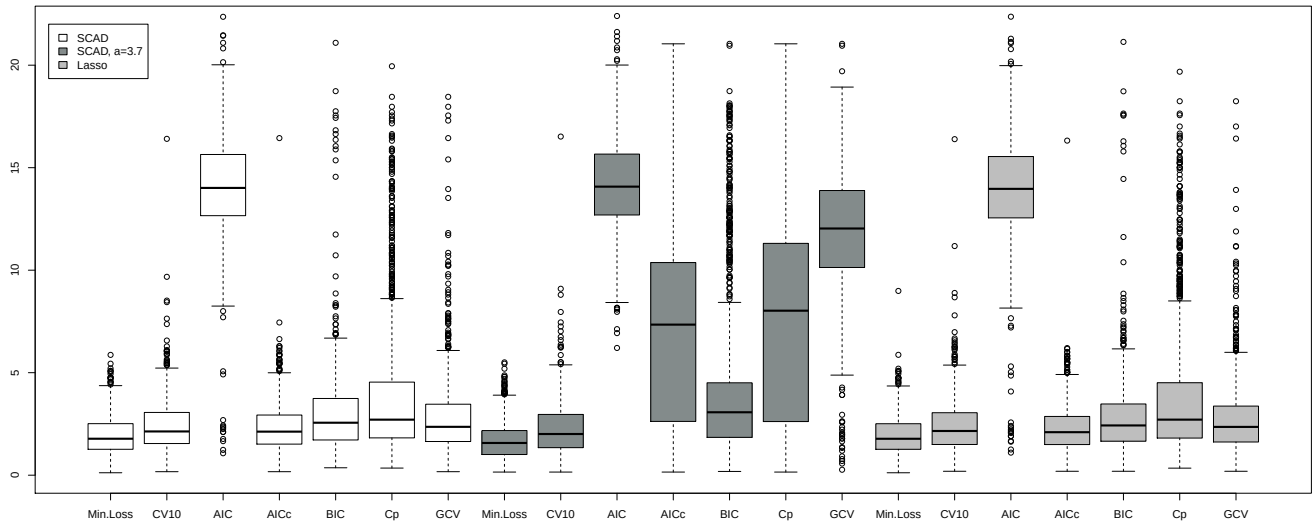
(a) c=.5



(b) c=.7

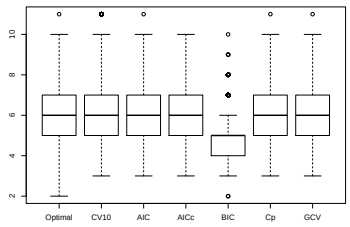

SCAD
SCAD, a=3.7
Lasso
0
5
10
15
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV

(c) c=.9

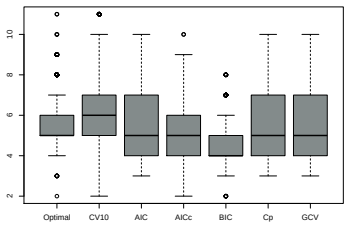

SCAD
SCAD, a=3.7
Lasso
0
5
10
15
20
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV
Min.Loss
CV10
AIC
AICc
BIC
Cp
GCV

(d) c=.98

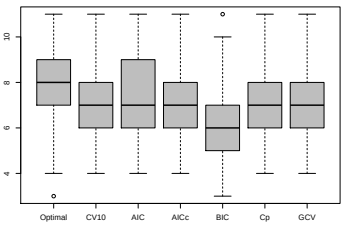
Figure 4: Comparison of model selection procedures based on the number of non-zero coefficients (includes intercept) in the selected model over 1000 simulations for the model with an omitted predictor with $n = 100$ and $\rho = 0.5$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$.



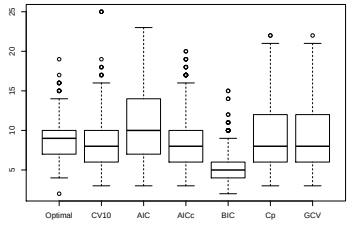
(a) SCAD, c=.5



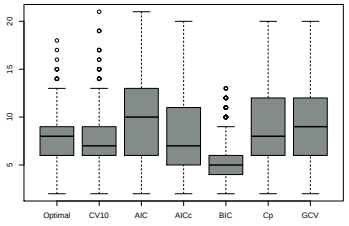
(b) SCAD, $a = 3.7$, c=.5



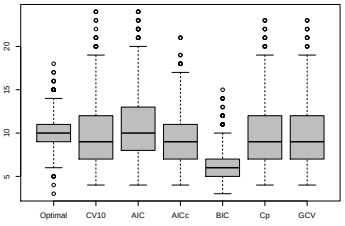
(c) Lasso, c=.5



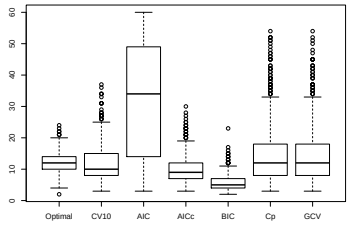
(d) SCAD, c=.7



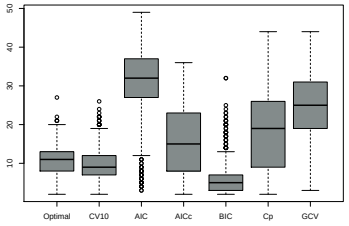
(e) SCAD, $a = 3.7$, c=.5



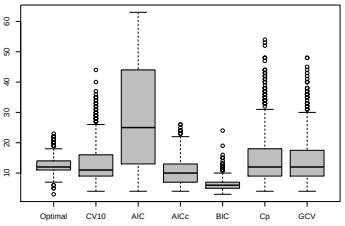
(f) Lasso, c=.7



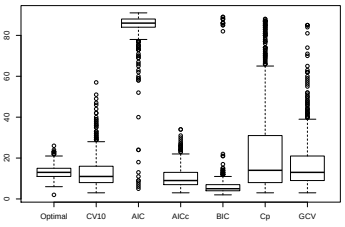
(g) SCAD, c=.9



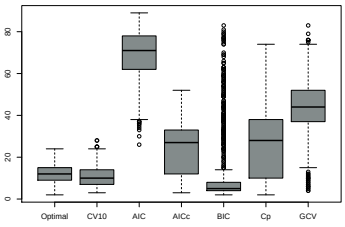
(h) SCAD, $a = 3.7$, c=.5



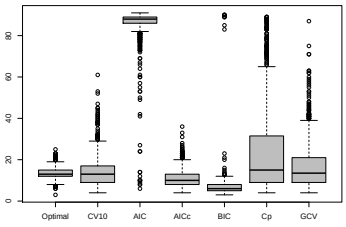
(i) Lasso, c=.9



(j) SCAD, c=.98



(k) SCAD, $a = 3.7$, c=.5



(l) Lasso, c=.98

4 True Model World

In this section we turn our attention to the finite-sample performance of the model selection procedures when the true model is included in the set of candidate models. Under certain regularity conditions and the assumption that d_n is fixed, Wang et al. (2007) proved that BIC_λ is a consistent selector of the regularization parameter for SCAD-penalized regression.

In classical regression it has been shown that BIC has a tendency to select overly complex models when the number of predictors is large relative to the sample size. Chen and Chen (2008) discussed the poor performance of BIC from a Bayesian perspective in the context of classical regression. In the supplementary material, Theorem 4 computes the probability that an information criterion will select the full model over the true model in classical linear regression. This can be used to further demonstrate a finite-sample overfitting property of BIC. For example, if $n = 50$, $d_n = 46$, and $d_0 = 3$, then the probability that BIC will select the full model over the true model is 0.1819 and if $n = 100$, $d_n = 90$, and $d_0 = 3$, then the probability that BIC will select the full model over the true model is 0.0017. This simple calculation clearly demonstrates that BIC has the potential to catastrophically overfit when the number of predictor variables is large relative to the sample size, particularly when the sample size is small.

Although the above calculation is done in the context of classical regression, simulations suggest that BIC_λ suffers from the same issues when used as a selector of the regularization parameter for SCAD and the Lasso. This motivates us to study the finite sample performance of two corrected versions of BIC_λ. The first is the corrected BIC_λ (BIC_{c_λ}),

$$BIC_{c_\lambda} = \log(\hat{\sigma}_n^2(\lambda)) + \log(n)\frac{(df_n(\lambda)+1)}{n - df_n(\lambda) - 2}, \tag{1}$$

which is a simple analogue of AIC_{c_λ} where the 2 has been replaced by $\log(n)$. The consistency proof of Wang et al. (2007) can be applied to BIC_{c_λ} to establish that the corrected version

preserves the desired asymptotic properties. The second is the Modified BIC_{c_λ} ($MBIC_\lambda$),

$$MBIC_\lambda = \log(\hat{\sigma}_n^2(\lambda)) + \log(n)\frac{df_n(\lambda)}{n}\log(\log(d_n)), \tag{2}$$

proposed by Wang et al. (2009). The authors proved that $MBIC_\lambda$ performs consistent model selection when $d_n \to \infty$ in SCAD-penalized regression where it is assumed that the $\lim_{n\to\infty}\sup(d_n/n^{\kappa^*}) < 1$ for some $\kappa^* < 1$. Again, in what follows, we define the $df_n(\lambda)$ to be $d_n(\alpha_\lambda)$.

If we consider these procedures in the context of classical regression then we can perform the same probability calculation. If $n = 50$, $d_n = 46$, and $d_0 = 3$, then the probability that BIC_c and $MBIC$ will select the full model over the true model is 0.0000 and 0.0249, respectively, and if $n = 100$, $d_n = 90$, and $d_0 = 3$, then the probability that BIC_c and $MBIC$ will select the full model over the true model is 0.0000 for both procedures. This suggests that $MBIC$ still has some tendency to overfit but only in the most extreme settings, while BIC_c does not suffer from this issue.

The following simulations study the finite-sample performance of these two methods as well as the methods described in Section 2.2 when the number of predictors is allowed to be large relative to the sample size. The set-up for the simulation is based on the example studied in Wang et al. (2007). We define $\sigma^2 = 9$ and

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$$

where the number of superfluous variables in $\boldsymbol{\beta}$ is allowed to vary by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$. We consider values of c on the grid $(0.5, 0.7, 0.9, 0.98)$. The $\mathbf{x}_i$'s are simulated from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ where $\Sigma_{i,j} = \rho^{|i-j|}$ for $\rho = 0$ and 0.5. The results are based on 1000 realizations of samples with n=50, 100, and 150. The selection procedures are evaluated based on the number of times the true model is selected, the L_2 Loss of the selected estimated models, and the selected

number of non-zero coefficients.

The patterns for the two values are ρ are similar so only the results for $\rho = 0.5$ are reported. The distribution of the selected models under each scenario is presented in Tables 3 and 4. Taking into account that our definition of degrees of freedom is different from Wang et al. (2007), the results appear to be fairly consistent with previous findings when $c = .5$. Furthermore, the results appear to be consistent with Wang et al. (2009) who found that sample sizes that are around 1600 are required before the percentage of times that the true model is selected is close to 100%.

Comparing the model selection procedures, the consistent criteria, BIC_λ, BIC_{c_λ}, and $MBIC_\lambda$, select the true model more frequently than the efficient criteria, AIC_λ, AIC_{c_λ}, C_{p_λ}, which is consistent with the theoretical results. Based on the simulations, 10-fold CV does not appear to behave like a consistent model selection procedure and has a strong tendency to overfit. This further supports the conjecture that its properties from classical regression still hold in the context of penalized regression. Focusing on the consistent selection procedures, both BIC_{c_λ} and $MBIC_\lambda$ select the true model more frequently than BIC_λ in all of the cases considered and from the tables it is clear that BIC_λ has a tendency to select models that are close to the full model when d_n is large relative to n. $MBIC_\lambda$ also seems to suffer from this behavior but only in the most extreme setting when $n = 50$ and $d_n = 46$. In contrast BIC_{c_λ} does not have this tendency to sometimes catastrophically overfit.

It is also interesting to compare the overall performance of the modeling procedures. Although asymptotically the Lasso is known to not satisfy the "oracle property," in the simulations the Lasso outperforms SCAD in terms of selecting the true model in some settings, specifically when the number of predictors is large relative to the sample size. Also, when the predictor variables are correlated and the number of predictors is large relative to the sample size the simulations suggest that a data dependent choice of the a can improve performance.

Although the true model is included in the set of candidate models, a data analyst may be more interested in predictive performance than in finding the true model. Figure 5 presents

boxplots comparing the procedures based on the L_2 loss of the selected models for $n = 100$ and $\rho = 0.5$. These plots also include the loss experienced when the oracle estimate (the least squares model fit using the three true predictors) is used. Using signed rank tests to compare loss performance, 10-fold CV performs as well as the consistent selection procedures when $c = .5$ and performs significantly better when $c = .7$, .9, and .98 for all three modeling procedures. AIC_{c_λ} also has good performance as a selector of the regularization parameter for SCAD with a data-dependent choice of a and the Lasso. The difference between 10-fold CV and AIC_{c_λ} is not statistically significant in these cases. Figure 6 presents boxplots comparing the selected number of non-zero coefficients. Although the true model includes only three predictors, these plots suggest that predictive performance can be improved by selecting and estimating based on a more complex model and that 10-fold CV and AIC_{c_λ} are more successful at selecting models with dimensions that are closer to the optimal model than are the consistent model selection procedures. It should be noted here that the good performance of an overly complex model is at least partly due to the bias of the penalized estimates since the oracle estimate outperforms all of the other methods, with some sort of bias/variance tradeoff also possibly involved.

Overall, the results indicate that a data analyst can benefit from using either BIC_{c_λ} or $MBIC_\lambda$ instead of BIC_λ since they are less sensitive to the number of superfluous predictors that are included in the model, and we would recommend using BIC_{c_λ} when the number of predictors is very close to the sample size. Furthermore, these simulations suggest that the cost of using 10-fold CV or AIC_{c_λ} in the true model world is less than the cost of using BIC_λ in a non-true model world. Therefore, when in doubt of which setting you are in, we recommend using 10-fold CV or AIC_{c_λ}.

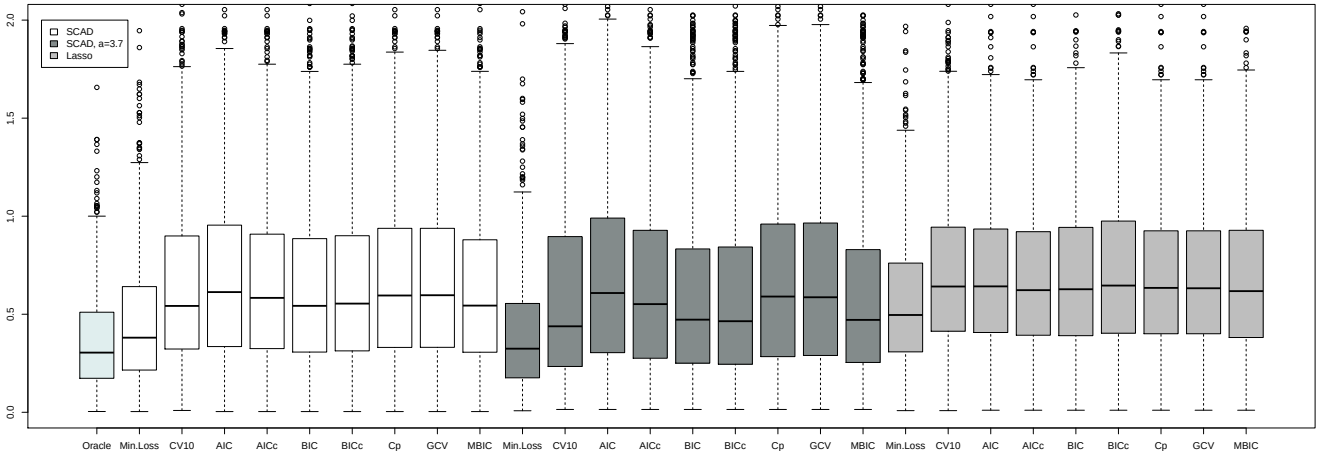
Table 3: Comparison of the distributions of the selected models over 1000 simulations for the true model world example with correlated predictors. A model is considered underfitted if it does not contain the true model and is considered correct if the true model is selected, and the maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$.

$c = .5$																					
	SCAD							SCAD, $a = 3.7$							Lasso						
			Number of Excess Predictors							Number of Excess Predictors								Number of Excess Predictors					
Info. Crit.	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$
$n = 50$																					
10-fold CV	108	285	607	0	0	0	0	143	292	565	0	0	0	0	21	10	969	0	0	0	0
AIC_λ	138	364	498	0	0	0	0	165	362	473	0	0	0	0	41	294	665	0	0	0	0
AIC_{c_λ}	161	421	418	0	0	0	0	195	411	394	0	0	0	0	49	375	576	0	0	0	0
BIC_λ	203	491	306	0	0	0	0	274	450	276	0	0	0	0	78	487	435	0	0	0	0
BIC_{c_λ}	221	537	242	0	0	0	0	303	480	217	0	0	0	0	101	559	340	0	0	0	0
C_{p_λ}	147	392	461	0	0	0	0	178	382	440	0	0	0	0	41	322	637	0	0	0	0
GCV_λ	145	392	463	0	0	0	0	177	381	442	0	0	0	0	41	321	638	0	0	0	0
$MBIC_\lambda$	164	418	418	0	0	0	0	199	407	394	0	0	0	0	49	376	575	0	0	0	0
$n = 100$																					
10-fold CV	5	175	788	32	0	0	0	15	281	687	17	0	0	0	0	0	947	53	0	0	0
AIC_λ	5	258	737	0	0	0	0	8	378	614	0	0	0	0	0	144	850	6	0	0	0
AIC_{c_λ}	7	289	704	0	0	0	0	9	425	566	0	0	0	0	0	185	812	3	0	0	0
BIC_λ	27	488	485	0	0	0	0	44	585	371	0	0	0	0	2	433	565	0	0	0	0
BIC_{c_λ}	37	527	436	0	0	0	0	58	611	331	0	0	0	0	2	500	498	0	0	0	0
C_{p_λ}	6	272	722	0	0	0	0	9	403	588	0	0	0	0	0	159	836	5	0	0	0
GCV_λ	6	269	725	0	0	0	0	9	397	594	0	0	0	0	0	153	842	5	0	0	0
$MBIC_\lambda$	20	456	524	0	0	0	0	38	554	408	0	0	0	0	1	364	635	0	0	0	0
$n = 150$																					
10-fold CV	1	211	763	25	0	0	0	2	330	644	24	0	0	0	0	0	895	105	0	0	0
AIC_λ	0	312	670	18	0	0	0	1	422	570	7	0	0	0	0	128	817	55	0	0	0
AIC_{c_λ}	0	340	651	9	0	0	0	1	464	533	2	0	0	0	0	150	820	30	0	0	0
BIC_λ	7	607	386	0	0	0	0	14	710	276	0	0	0	0	0	453	546	1	0	0	0
BIC_{c_λ}	7	631	362	0	0	0	0	15	732	253	0	0	0	0	0	483	517	0	0	0	0
C_{p_λ}	0	330	657	13	0	0	0	1	442	552	5	0	0	0	0	132	823	45	0	0	0
GCV_λ	0	323	666	11	0	0	0	1	439	555	5	0	0	0	0	135	823	42	0	0	0
$MBIC_\lambda$	5	591	404	0	0	0	0	13	699	288	0	0	0	0	0	436	563	1	0	0	0
$c = .7$																					
	SCAD							SCAD, $a = 3.7$							Lasso						
			Number of Excess Predictors							Number of Excess Predictors								Number of Excess Predictors					
Info. Crit.	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$
$n = 50$																					
10-fold CV	82	122	741	55	0	0	0	187	114	650	49	0	0	0	39	4	807	150	0	0	0
AIC_λ	75	98	657	170	0	0	0	167	100	641	92	0	0	0	54	92	678	176	0	0	0
AIC_{c_λ}	105	186	682	27	0	0	0	232	163	589	16	0	0	0	68	180	726	26	0	0	0
BIC_λ	132	298	558	12	0	0	0	290	256	449	5	0	0	0	114	315	560	11	0	0	0
BIC_{c_λ}	161	399	440	0	0	0	0	320	350	330	0	0	0	0	142	428	430	0	0	0	0
C_{p_λ}	87	124	697	92	0	0	0	211	128	611	50	0	0	0	56	130	723	91	0	0	0
GCV_λ	87	119	700	94	0	0	0	200	126	621	53	0	0	0	58	128	729	85	0	0	0
$MBIC_\lambda$	132	297	559	12	0	0	0	290	255	450	5	0	0	0	112	313	564	11	0	0	0
$n = 100$																					
10-fold CV	5	98	734	144	19	0	0	21	109	794	62	14	0	0	0	0	708	247	45	0	0
AIC_λ	2	90	521	261	126	0	0	8	135	468	293	96	0	0	1	72	583	248	96	0	0
AIC_{c_λ}	7	137	670	165	21	0	0	19	188	591	178	24	0	0	0	109	726	154	11	0	0
BIC_λ	23	392	578	7	0	0	0	80	364	546	9	1	0	0	3	392	601	4	0	0	0
BIC_{c_λ}	23	463	513	1	0	0	0	99	403	497	1	0	0	0	4	452	543	1	0	0	0
C_{p_λ}	5	112	600	217	66	0	0	15	165	554	213	53	0	0	0	91	648	212	49	0	0
GCV_λ	4	106	587	236	67	0	0	13	157	533	242	55	0	0	0	86	651	217	46	0	0
$MBIC_\lambda$	23	472	504	1	0	0	0	102	410	487	1	0	0	0	4	462	533	1	0	0	0
$n = 150$																					
10-fold CV	1	86	726	156	30	1	0	2	134	778	77	7	2	0	0	0	637	262	91	10	0
AIC_λ	1	93	527	177	196	6	0	3	194	401	202	200	0	0	0	69	546	239	134	12	0
AIC_{c_λ}	1	135	639	154	70	1	0	5	244	501	167	83	0	0	0	100	674	180	46	0	0
BIC_λ	5	505	489	1	0	0	0	26	522	449	3	0	0	0	0	436	563	1	0	0	0
BIC_{c_λ}	5	541	453	1	0	0	0	29	561	409	1	0	0	0	0	477	523	0	0	0	0
C_{p_λ}	1	117	598	164	118	2	0	3	225	471	182	119	0	0	0	79	602	225	91	3	0
GCV_λ	1	109	589	171	128	2	0	3	222	455	198	122	0	0	0	79	600	228	91	2	0
$MBIC_\lambda$	8	590	402	0	0	0	0	40	596	363	1	0	0	0	0	554	446	0	0	0	0

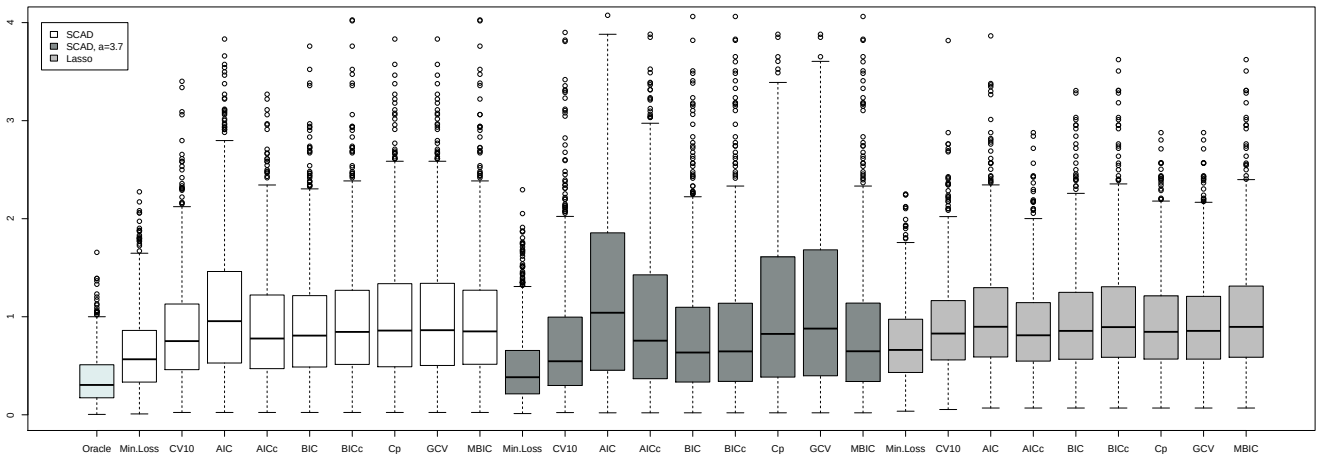
Table 4: Comparison of the distributions of the selected models over 1000 simulations for the true model world example with correlated predictors. A model is considered underfitted if it does not contain the true model and is considered correct if the true model is selected, and the maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$.

$c = .9$																					
	SCAD							SCAD, $a = 3.7$							Lasso						
			Number of Excess Predictors							Number of Excess Predictors								Number of Excess Predictors				
Info. Crit.	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$
$n = 50$																					
10-fold CV	68	66	584	193	85	4	0	251	62	580	99	6	2	0	60	5	568	248	107	12	0
AIC_λ	64	26	197	99	303	311	0	139	11	70	199	532	49	0	62	27	242	141	257	271	0
AIC_{c_λ}	102	126	656	102	14	0	0	250	93	349	236	72	0	0	100	132	664	96	8	0	0
BIC_λ	142	276	509	44	22	7	0	305	193	299	128	73	2	0	147	288	515	33	13	4	0
BIC_{c_λ}	195	400	404	1	0	0	0	350	309	330	11	0	0	0	197	411	391	1	0	0	0
C_{p_λ}	80	77	484	161	162	36	0	221	72	272	261	171	3	0	75	81	503	169	146	26	0
GCV_λ	75	67	476	182	167	33	0	184	25	194	296	296	5	0	66	65	533	189	132	15	0
$MBIC_\lambda$	186	372	422	14	4	2	0	344	278	301	61	16	0	0	188	373	425	9	4	1	0
$n = 100$																					
10-fold CV	1	45	522	251	152	27	2	39	44	581	270	66	0	0	0	0	458	303	202	32	5
AIC_λ	8	19	187	109	128	96	453	28	12	36	22	148	421	333	1	23	221	138	158	118	341
AIC_{c_λ}	2	69	591	229	103	6	0	38	58	373	104	294	129	4	0	69	613	224	92	2	0
BIC_λ	7	292	684	15	2	0	0	132	195	623	29	20	1	0	4	365	614	16	1	0	0
BIC_{c_λ}	7	388	601	4	0	0	0	158	237	596	6	3	0	0	6	443	549	2	0	0	0
C_{p_λ}	4	45	424	211	187	77	52	34	50	293	88	295	218	22	0	48	440	218	195	67	32
GCV_λ	4	37	421	236	201	63	38	31	30	144	60	286	378	71	0	41	440	242	195	65	17
$MBIC_\lambda$	10	524	465	1	0	0	0	205	294	499	2	0	0	0	10	561	428	1	0	0	0
$n = 150$																					
10-fold CV	0	46	474	260	175	37	8	8	47	563	294	83	5	0	0	0	398	303	229	60	10
AIC_λ	0	21	200	125	143	79	432	6	11	33	6	17	106	821	0	25	228	162	174	101	310
AIC_{c_λ}	0	58	497	267	153	24	1	8	64	403	122	110	183	110	0	58	503	265	153	21	0
BIC_λ	0	355	632	13	0	0	0	51	330	609	10	0	0	0	0	428	563	9	0	0	0
BIC_{c_λ}	0	393	605	2	0	0	0	55	375	566	4	0	0	0	1	472	526	1	0	0	0
C_{p_λ}	0	50	392	247	190	69	52	12	63	337	103	94	182	209	0	44	397	248	197	78	36
GCV_λ	0	46	383	246	213	70	42	7	34	144	52	68	201	494	0	40	385	242	232	71	30
$MBIC_\lambda$	0	596	404	0	0	0	0	92	493	415	0	0	0	0	1	623	376	0	0	0	0
$c = .98$																					
	SCAD							SCAD, $a = 3.7$							Lasso						
			Number of Excess Predictors							Number of Excess Predictors								Number of Excess Predictors				
Info. Crit.	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$	Underfit	Correct	$1-5$	$6-10$	$11-20$	$21-30$	$30+$
$n = 50$																					
10-fold CV	86	43	489	222	122	32	6	287	47	467	183	13	1	2	76	2	483	252	155	24	8
AIC_λ	104	1	10	2	10	17	856	172	1	0	4	42	156	625	84	3	9	3	10	20	871
AIC_{c_λ}	121	97	637	129	16	0	0	288	50	228	175	242	17	0	124	115	614	134	13	0	0
BIC_λ	135	131	272	20	5	7	430	259	53	110	35	99	129	315	131	157	278	21	5	11	397
BIC_{c_λ}	219	372	409	0	0	0	0	371	265	350	14	0	0	0	220	375	405	0	0	0	0
C_{p_λ}	106	92	289	99	115	88	211	255	73	161	76	235	129	71	109	104	279	99	115	90	204
GCV_λ	82	39	370	183	132	61	133	240	3	24	51	316	274	92	80	43	382	196	135	78	86
$MBIC_\lambda$	198	328	341	6	1	1	125	333	208	226	29	37	52	115	200	343	341	5	1	1	109
$n = 100$																					
10-fold CV	1	34	445	270	185	50	15	49	22	490	309	128	1	1	1	0	381	333	200	65	20
AIC_λ	20	2	6	1	2	2	967	86	1	0	0	0	3	910	12	1	11	1	5	2	968
AIC_{c_λ}	0	64	571	240	113	12	0	45	25	232	67	152	314	165	1	62	559	248	116	14	0
BIC_λ	9	283	675	18	2	2	11	140	157	528	23	24	29	99	6	351	617	16	2	1	7
BIC_{c_λ}	10	349	636	5	0	0	0	185	217	587	10	1	0	0	11	438	547	4	0	0	0
C_{p_λ}	2	53	323	165	158	78	221	64	39	248	55	96	221	277	1	59	334	164	160	72	210
GCV_λ	0	33	349	229	218	95	76	68	2	10	7	21	158	734	0	32	379	225	202	94	68
$MBIC_\lambda$	14	526	459	1	0	0	0	256	276	467	1	0	0	0	15	564	421	0	0	0	0
$n = 150$																					
10-fold CV	0	30	418	266	203	59	24	13	37	457	282	199	12	0	0	0	340	299	255	74	32
AIC_λ	23	1	2	1	6	2	965	51	1	0	0	0	0	948	13	0	5	4	3	2	973
AIC_{c_λ}	0	47	473	257	185	37	1	11	35	200	63	38	69	584	0	53	452	270	187	37	1
BIC_λ	0	344	645	9	2	0	0	66	267	646	12	1	0	8	0	431	556	10	3	0	0
BIC_{c_λ}	0	379	620	1	0	0	0	71	299	626	4	0	0	0	1	472	525	2	0	0	0
C_{p_λ}	0	44	355	151	161	80	209	11	60	278	55	25	46	525	0	59	338	155	178	72	198
GCV_λ	0	30	337	220	237	109	67	25	2	7	3	2	10	951	0	31	315	224	251	102	77
$MBIC_\lambda$	1	597	402	0	0	0	0	105	436	459	0	0	0	0	1	639	360	0	0	0	0

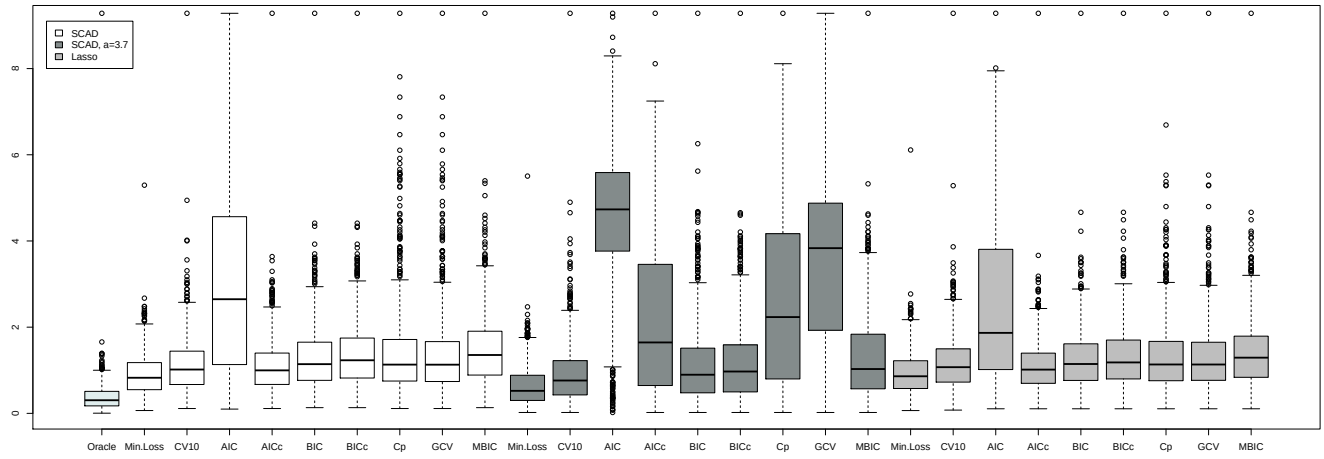
Figure 5: Comparison of model selection procedures based on L2 Loss over 1000 simulations for the true model world example with $n = 100$ and $\rho = 0.5$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$. In order to make it easier to compare the procedures, the limits of the vertical axis are specified so that all the boxes and whiskers appear but some of the outliers are not shown.



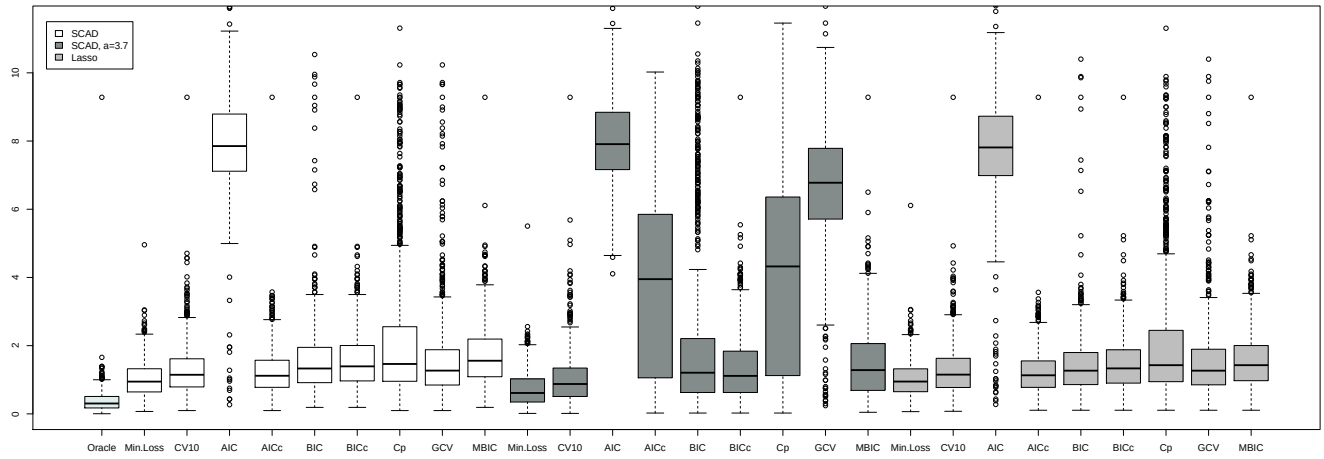
(a) c=.5



(b) c=.7

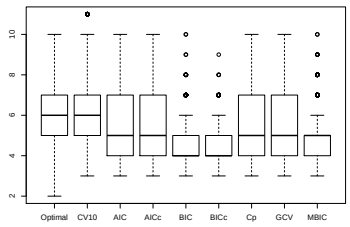

SCAD
SCAD, a=3.7
Lasso
0
2
4
6
8
Oracle
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC

(c) c=.9

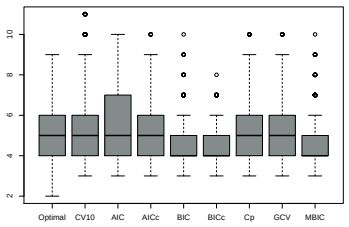

SCAD
SCAD, a=3.7
Lasso
0
2
4
6
8
10
Oracle
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC
Min.Loss
CV10
AIC
AICc
BIC
BICc
Cp
GCV
MBIC

(d) c=.98

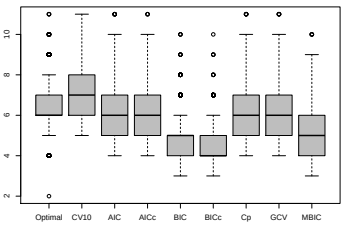
Figure 6: Comparison of model selection procedures based on the number of non-zero coefficients (includes intercept) in the selected model over 1000 simulations for the true model world example with $n = 100$ and $\rho = 0.5$. The maximum number of predictors is varied by letting $d_n = 2\lfloor n^c/2 \rfloor$.



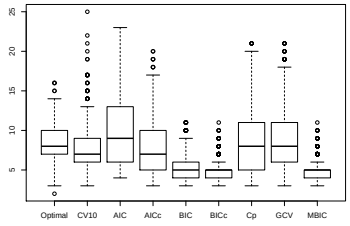
(a) SCAD, c=.5



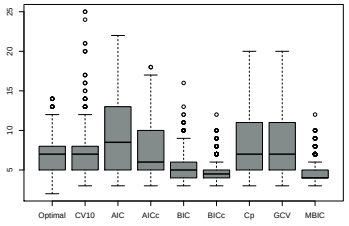
(b) SCAD, $a = 3.7$, c=.5



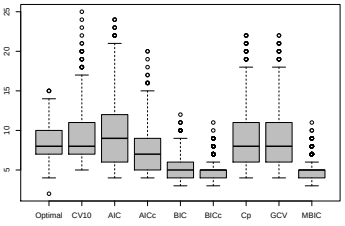
(c) Lasso, c=.5



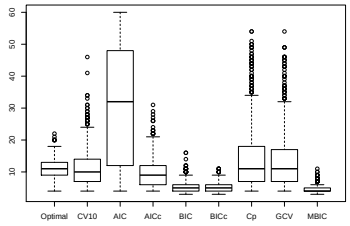
(d) SCAD, c=.7



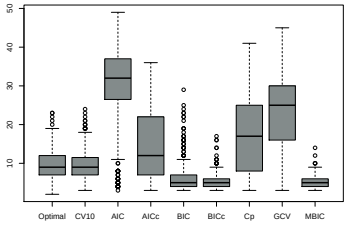
(e) SCAD, $a = 3.7$, c=.5



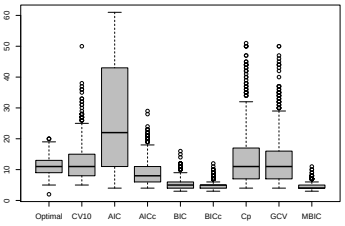
(f) Lasso, c=.7



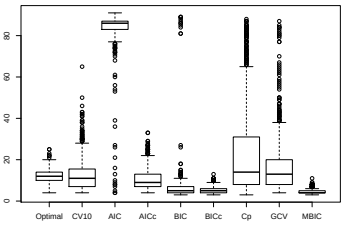
(g) SCAD, c=.9



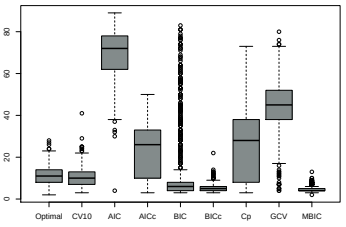
(h) SCAD, $a = 3.7$, c=.5



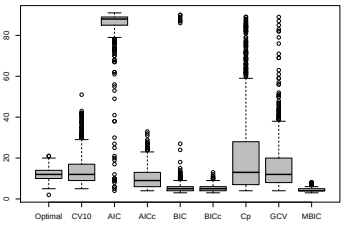
(i) Lasso, c=.9



(j) SCAD, c=.98



(k) SCAD, $a = 3.7$, c=.5



(l) Lasso, c=.98

5 Concluding Remarks

This paper studied the asymptotic and finite sample performance of classical model selection procedures in the context of regularized regression with and without the assumption that the true model is included amongst the candidate models. In the non-true model world we proved that AIC_λ, AIC_{c_λ}, C_{p_λ}, and GCV_λ are efficient selectors of the regularization parameter and the simulation studies yielded several interesting observations. As suspected, they showed that BIC_λ is outperformed by the efficient model selection procedures and demonstrated that AIC_λ, BIC_λ, C_{p_λ}, and GCV_λ are all sensitive to the number of predictor variables that are included in the full model and that their performance can suffer as a result. In light of this issue we recommend that researchers use a method that is insensitive to the number of variables included in the model. From the simulations, 10-fold CV has the best overall performance. However, it is 10 times more expensive to implement than using an information criterion, the asymptotic properties of 10-fold CV are not fully understood in this context, and the randomness involved in the procedure makes it difficult for data analysts to reproduce results. As an alternative, data analysts can consider using AIC_{c_λ}, which was shown here to be an efficient selection procedure for the tuning parameter, and which the simulations suggest has comparable performance to that of 10-fold CV. Lastly, the simulations suggest that there is no clear advantage to using SCAD in a world where the "oracle property" does not apply. Combining this with the facts that the Lasso also has the efficient 'Lars' algorithm available and does not involve a second tuning parameter that can greatly impact results, researchers may prefer to use the Lasso if they feel that they are in the non-true model world.

The simulations in the true model world demonstrated that BIC_λ can be outperformed by both $MBIC_\lambda$ and the proposed BIC_{c_λ} which are less sensitive to the number of the predictor variables that are included in the model. Furthermore, although 10-fold CV and AIC_{c_λ} have a tendency to select an overly complex model, the simulations suggest that their predictive performance is better than that of the consistent selection procedures. Therefore, if the data analyst is unsure about whether they are in the true model world or the non-true model

world or if predictive power or estimation of coefficients is of primary concern we recommend using one of these two procedures.

Although the focus of this paper was not on the second tuning parameter of SCAD, the simulations suggest that allowing this parameter to be data-dependent can improve the performance of the model selection procedures particularly when the goal is efficient model selection or when the predictor variables are correlated with each other. Further investigation into the selection of this parameter is an area for future research.

As a final remark, this paper dealt with the case when $d_n/n \to 0$ and the theoretical results cannot be directly extended to the case when d_n/n converges to something other than zero. The latter setting has received a great deal of attention in recent literature (in particular $d_n \gg n$) and is an area for future investigation.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, pages 267–281.

Breheny, P. and Huang, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, 5(1):232–253.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, 95(3):759–771.

Craven, P. and Wahba, G. (1978). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31(4):377–403.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Furnival, G. M. and Wilson, R. W. (1974). Regression by Leaps and Bounds. *Technometrics*, 16(4):499–511.

Gelman, A. (2010). Bayesian Statistics Then and Now. *Statistical Science*, 25(2):162–165.

Hastie, T. and Efron, B. (2011). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 0.9-8.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediciton*. Springer Series in Statistics. Springer, 2 edition.

Huang, J. and Xie, H. (2007). Asymptotic Oracle Properties of SCAD-Penalized Least Squares Estimators. *Lecture Notes-Monograph Series*, 55:149–166.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models. *Biometrika*, 78(3):499–509.

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

Li, K.-C. (1987). Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975.

Mallows, C. L. (1973). Some Comments on C_p. *Technometrics*, 15(4):661–675.

Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *The Annals of Statistics*, 12(2):758–765.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

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

Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *The Annals of Statistics*, 8(1):147–164.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*, 68(1):45–54.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters. *Journal of the Royal Statistical Society B*, 71:671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, 94(3):553–568.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization Parameter Selections via Generalized Information Criterion. *Journal of the American Statistical Association*, 105(489):312–323.

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(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "Degrees of Freedom" of the Lasso. *The Annals of Statistics*, 35(5):2173–2192.