Greedy and Relaxed Approximations to Model Selection: A simulation study

Guilherme V. Rocha and Bin Yu

April 6, 2008

Abstract

The Minimum Description Length (MDL) principle is an important tool for retrieving knowledge from data as it embodies the scientific strife for simplicity in describing the relationship among variables. As MDL and other model selection criteria penalize models on their dimensionality, the estimation problem involves a combinatorial search over subsets of predictors and quickly becomes computationally cumbersome.

Two approximation frameworks are: convex relaxation and greedy algorithms. In this article, we perform extensive simulations comparing two algorithms for generating candidate models that mimic the best subsets of predictors for given sizes (Forward Stepwise and the Least Absolute Shrinkage and Selection Operator - LASSO). From the list of models determined by each method, we consider estimates chosen by two different model selection criteria (AIC_C and the generalized MDL criterion - gMDL). The comparisons are made in terms of their selection and prediction performances.

In terms of variable selection, we consider two different metrics. For the number of selection errors, our results suggest that the combination Forward Stepwise+gMDL has a better performance over different sample sizes and sparsity regimes. For the second metric of rate of true positives among the selected variables, LASSO+gMDL seems more appropriate for very small sample sizes, while Forward Stepwise+gMDL has a better performance for sample sizes at least as large as the number of factors being screened. Moreover, we found that, asymptotically, Zhao and Yu's ((1)) irrepresentibility condition (index) has a larger impact on the selection performance of Lasso than on Forward Stepwise. In what refers to prediction performance, LASSO+AIC_C results in good predictive models over a wide range of sample sizes and sparsity regimes. Last but not least, these simulation results reveal that one method often can not serve for both selection and prediction purposes.

1 Introduction

The practice of statistics often refers to making efficient use of observed data to infer relationships among variables in order to either gain insight into an observed phenomenon (interpretation) or be able to make predictions based on partial information (prediction). In this paper, we focus on models designed to uncover how a dependent or response variable $Y \in \mathcal{Y}$ is affected by a set of p predictor variables $X \in \mathbb{R}^p$. Whether the goal is prediction or interpretation, the important task is to learn some "meaningful" or stable characteristics of the data across different samples of the data.

A traditional approach consists of postulating a class of models $\mathcal{F}$ indexed by a parameter β. An estimate $\hat{\beta}$ is often defined as:

$$\hat{\beta} \quad = \quad \arg\min_{\beta \in \mathcal{F}} \quad \sum_i L(Z_i, X_i^T \beta), \tag{1}$$

where $Z_i = (X_i, Y_i), i = 1, \ldots, n$ denotes then n observed data samples; $Y \in \mathbb{R}^n$ and is a vector containing the observed values for the dependent variable; $X \in \mathbb{R}^{n \times p}$ is a matrix containing the observed

values of the predictors in its rows; β is a model within the postulated class; and L is a loss function measuring the goodness of fit to the data Z for the model indexed by β. In this paper, we restrict attention to loss functions L defined by the *negative log-likelihood* (neg-loglikelihood) of probabilistic models. This framework is general enough to accommodate both regression and classification models and encompasses all Generalized Linear Models (2; 3). In this paper we will focus attention on the standard Gaussian linear regression model:

$$Y = X\beta + \varepsilon, \text{ with } \varepsilon \sim N(0, \sigma^2). \tag{2}$$

The minimization in (1) translates in this case to the L_2-loss:

$$\hat{\beta} \quad = \quad \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 \right\}. \tag{3}$$

Some of the information criteria we will be dealing with below also require an estimate of the variance σ^2. Unless otherwise stated, we use the likelihood estimate:

$$\hat{\sigma}^2 \quad = \quad \frac{\|Y - X\hat{\beta}\|^2}{n}. \tag{4}$$

However reasonable the estimate defined by (1) may be, this approach does not account for the fact that we often ignore what is an appropriate class of models in which to fit the data, that is, $\mathcal{F}$ should also be estimated. On the one hand, if we rely solely on the observable empirical neg-loglikelihood $L(Z, .)$ to decide between two classes of model $\mathcal{F}_1 \subset \mathcal{F}_2$, the larger class will be trivially preferred. On the other hand, the simpler model class $\mathcal{F}_1$ may be more representative of any structure contained the data and less sensitive to noise. The Minimum Description Length (MDL) principle introduced by Jorma Rissanen (4–6) addresses this problem by including the cost of coding the model itself into the picture. As more complex models are costlier to describe, parsimony is now rewarded.

The problem of identifying an adequate model class on which to search for an estimate $\hat{\beta}$ has been recognized since the seventies. It has since motivated developments of model (variable) selection criteria that penalize the neg-loglikelihood by measures of complexity of the model including the Minimum Message Length criterion (MML, 7), C_p (8), Akaike's Information Criterion (AIC, 9), Bayesian Information Criterion (BIC, 10), and various MDL methods (e.g. the generalized MDL criterion, gMDL, 11).

For linear models, many model selection criteria involve a penalty in the dimensionality of the model under evaluation (i.e., number of non-zero terms in $\hat{\beta}$), that is, the selected estimate is of the form:

$$\hat{\beta}(\lambda_n) \quad = \quad \arg\min_{\beta} \left\{ 2L(Z, \beta) + \lambda_n \|\beta\|_0 \right\}, \tag{5}$$

where $\|\beta\|_0 = \#\{j : \beta_j \neq 0\}$ and λ_n is a tuning parameter trading-off summarization performance and "complexity" of the model. We will refer to such penalties as ℓ_0-penalties in what follows. Perhaps the two most popular examples of model selection criteria within this family are AIC (9) and BIC (10) for which $\lambda_n = 2$ and $\lambda_n = \log(n)$ respectively. In this paper, we will be working with two criteria related to AIC and BIC: the AIC_C criterion (12) is a finite-sample corrected version of AIC; and the gMDL criterion (11) that tries to combine the virtues of AIC and BIC.

The strict computation of ℓ_0-penalized estimates leads to a costly combinatorial search over all subsets of the largest model: the best subset search problem is an NP hard problem in the number of predictors p as established by a recent formal proof (13). Exact solutions to this problem are computationally infeasible for modern massive data sets where the number of predictors p is in the orders of thousands as in gene expression data analysis and even millions as in text processing applications. Even if the

computation of models involving all subsets were feasible, it would still be wasteful. As the number of models to be compared is large, many of them are indistinguishable within the precision afforded by the number of samples practically available. As an example, a regression model involving 50 predictors (a modest size for many modern data sets) would require the comparison of $2^{50} \sim 10^{15}$ models.

Computationally feasible approximations to the ℓ_0-penalized estimates are currently a very active and exciting field of research in statistics. In this paper, we perform extensive simulations to compare the prediction and variable selection performance of models picked by AIC_C and gMDL from lists of candidate models generated by algorithms that identify subsets that are "approximately" the best for their dimension.

The first approximation we consider is the greedy Forward Stepwise regression for selecting variables. Forward Stepwise regression is an eminently algorithmic procedure. Starting from the null model, it selects the predictor whose coefficient corresponds to the largest sized term on the loss function gradient and refits the model involving the selected parameters at each step. For linear models and under the squared error loss (L_2-loss), that corresponds to picking the variables most correlated with the residuals at each step (see 14). The second approximation we study is the convex relaxation approach in which the ℓ_0-penalization is replaced by the ℓ_1-norm of the candidate vector of coefficients β. Early examples of the use of ℓ_1-norm as a penalization are the non-negative garrote (15), the Least Absolute Shrinkage and Selection Operator (LASSO, 16) and basis pursuit (17). The soft-thresholding rule used in VisuShrink (18) is also intimately related to the ℓ_1-penalty. This relaxation results in a convex penalization, easing the burden of computing estimates defined as the solution to an optimization problem (19).

The remainder of this paper is organized as follows. Section 2 reviews some of the theoretical results concerning exact ℓ_0 and ℓ_1 penalized estimates. Section 3 presents a brief overview of the greedy (Forward Stepwise) and relaxed (LASSO) regularization paths. There, we also present the selection criteria we will consider in our later simulation experiments. In Section 4, we present our simulation setup and results obtained for the squared error loss. Section 5 presents our conclusions.

2 Properties of ℓ_0 and ℓ_1-penalized estimates

The properties of ℓ_0-penalized estimates are well understood as various theoretical results have been obtained since their introduction in the seventies (e.g. 20–22). Two important examples of ℓ_0-penalized estimates are AIC (9) and BIC (10). These two criteria reflect a tension that exists between prediction performance and model selection accuracy. Well known results show that AIC-type criteria have the property of yielding the minimax-rate optimal of the regression function under the predictive L_2-loss (23–28), while BIC like criteria are consistent in terms of model selection (21).

Penalization by the ℓ_1-norm of β aims at solving an approximate solution to a convex relaxation of the optimization problem (5). The approximate estimate $\hat{\beta}_{LASSO}(\lambda_n)$ is defined as:

$$\hat{\beta}_{LASSO}(\lambda_n) \quad = \quad \arg\min_{\beta} L(Z, \beta) + \lambda_n \|\beta\|_1, \tag{6}$$

where $\|\beta\|_1 = \sum_j |\beta_j|$. Despite its relative youth, ℓ_1-penalized estimation has undergone intensive research in recent years and a series of theoretical results concerning its properties have been achieved (e.g. 29–38; 1; 39–41). Many of these results are either asymptotic in nature or concern the behavior of sparse approximation in the noiseless setting. In the deterministic setting, the use of convex relaxation of the ℓ_0-norm by the ℓ_1-norm was shown to recover the correct sparse representation under incoherence conditions (42; 31; 30; 32; 34).

In what concerns the predictive performance of ℓ_1-penalized estimates, results in (29) establish that, based on observed data, the actual out-of-sample prediction error can be estimated with greater pre-

cision for the non-negative garrote (closely related to ℓ_1-penalized estimates) than for subset selection procedures. As a result, non-negative garrote estimates can attain better predictive models than their ℓ_0 counterparts. As we will see in our experimental section, this seems to carry over to the LASSO.

In terms of model selection, asymptotic results by (38), (1) and (40) establish conditions for model selection consistency for ℓ_1-penalized estimates under the L_2-loss in the non-parametric setting (i.e., $p_n \to \infty$ as $n \to \infty$). Here, we define the *irrepresentability index* as:

$$II(\Sigma, \beta) \quad = \quad 1 - \|\Sigma_{21}\Sigma_{11}^{-1}\text{sign}(\beta)\|_\infty \tag{7}$$

where Σ_{11} is the covariance matrix of the covariates with non-zero coefficients and Σ_{21} is the partition of the covariance matrix of the covariates accounting for the correlation between the irrelevant and relevant covariates. Results from (1) show that a sufficient condition for the LASSO to be consistent in model selection for some sequence λ_n as $n \to \infty$ is that:

$$II(\Sigma, \beta) \quad > \quad 0 \tag{8}$$

Later in this paper, we will be investigating the effect of the irrepresentability index on the model selection performance in finite samples. Results in (39) refine the model selection consistency results for the LASSO by determining at what rates the number of relevant covariates q and the number of measured predictors p can increase as n grows for model selection consistency to be preserved.

3 Approximation algorithms and selection criteria

The strict implementation of model selection criteria of the form shown in (5) requires the computation of estimates for all possible subsets. As mentioned before this is both computationally infeasible and wasteful given the large number of candidates that must be compared. It does, however, suggest that two tasks are involved in the selection of a model: generating a series of candidate models and applying a criterion to pick the "best" among them.

We consider two algorithms (Forward Stepwise and LASSO) for generating candidate models based on approximations to the combinatorial problem (5). For selecting estimates out of the lists of candidates created by these two algorithms, we consider two different criteria: AIC_C (12) and gMDL (11).

Before we proceed, we point out that alternative algorithms for generating candidate models and alternative selection criteria exist. Boosting algorithms (43) are an important tool for generating list of candidate models. For an example of Boosting algorithms applied to model selection, see (44). Cross-validation (45–47) is an important tool for choosing among different models, especially in what refers to prediction. It is, however, limited by its computational cost and often inadequate for model selection purposes (24).

3.1 Description of the path-tracing algorithms

Although the exact solution to problem (5) is a combinatorial problem, a natural greedy approximation suggests itself. At first, initialize a set of active parameters $\mathcal{A}$ to be empty and set $\hat{\beta}_0 = 0$ – the sparsest possible solution. Then repeat the following process until no parameters are left out or a local optimal is attained. Pick the parameter corresponding to the entry in the gradient vector $\nabla_\beta L$ with the largest absolute value. Add the chosen parameter to the set $\mathcal{A}$ and refit the model adjusting the estimates of parameters contained in $\mathcal{A}$ (i.e., set the new estimate to be a vector such that the gradient of all variables in $\mathcal{A}$ are zero). We shall refer to this algorithm as the *Forward Stepwise algorithm* for the remainder of this paper. It has close connections to the orthogonal greedy algorithms from approximation theory (see,

for instance, 48; 49). [1]

The convex relaxation approximation takes a different route. As mentioned above, it replaces the exact solution of the problem (5) by an approximation based on convex relaxation as defined in (6). A series of candidate models is generated by letting λ_n vary over $[0, \infty)$. At a first glance, the convex relaxation approach seems radically different from the Forward Stepwise regression algorithm. However, the homotopy/LARS [2] algorithm introduced in (50; 51) to compute all LASSO candidates reveals a close connection between them. The homotopy/LARS algorithm also starts by setting an active set $\mathcal{A}$ of parameters to be empty and set $\hat{\beta}_0 = 0$. At each step, it then selects the parameters with the highest gradients, computes a direction preserving the gradient with respect to all active parameters equal in size and determines a step size in which one of two events happen. Either the gradient corresponding to an inactive term becomes as high as the ones in $\mathcal{A}$ in which case a new term is added to $\mathcal{A}$; or one of the parameter estimates in $\mathcal{A}$ hits zero in which case it is excluded from $\mathcal{A}$.

In the case of linear models fitted using an L_2-loss, an analysis in (51) gives the computational cost of the k-th interaction of these algorithms in terms of the current size of the active set a_k and the number of observed samples n. At the k-th step, the costlier operation to perform is determining the direction of the next step. To do so, it is necessary to invert the matrix $X'_{\mathcal{A}} X_{\mathcal{A}}$. This can be done efficiently by updating its Cholesky decomposition at each step of the algorithm at a cost in the order of $O(a_k^2 + a_k n)$.

For Forward Stepwise, the entire regularization path has exactly $r = \text{rank}(X) \leq \min\{p, n\}$ steps resulting in a cost of the order of $O(r^3 + r^2 n)$ for the entire Forward Stepwise path. The complete LASSO regularization path, on the other hand, allows variables to be dropped and re-added to the model along the way and hence has a random number of steps. Well behaved data will cause the computational cost of the LASSO and Forward Stepwise path to be roughly the same. In particular, if the positive cone condition in (51) is satisfied, the two paths are known to agree, thus involving approximately the same computational effort. On the other hand, the LASSO path is costlier when a lot of variable droppings take place. In our experience, we have observed more correlated designs to be associated with longer and consequently costlier paths for the LASSO.

3.2 Selection criteria for choosing an estimate from the regularization path

The Forward Stepwise and the LASSO algorithms above generate each a collection of models for us to choose from, which we call their *regularization paths*. We will focus our attention on two different criteria for picking models from the Forward Stepwise and LASSO regularization paths: the AIC_C (12) (corrected AIC) and the gMDL (11) criteria. We decide for these two criteria based on the good results reported in (11), (52; 44), and (53).

The AIC_C was proposed by Sugiura (12) as a finite sample correction for Akaike's AIC (9). The authors have previously used this criterion in the $n < p$ setting with good predictive performance (54). We use it here in place of cross-validation to reduce the computational cost of our experiments. For linear models based on the L_2-loss (Gaussian likelihood for residuals), the AIC_C estimate are defined as:

$$\hat{\beta}_{\text{AIC}_\text{C}} \quad = \quad \arg \min_{\beta \in \text{path}} \left\{ \frac{n}{2} \log \left(\sum_{i=1}^{n} \|Y_i - X_i \beta\|^2 \right) + \frac{1}{2} \cdot \frac{n \left(1 + \frac{K(\beta)}{n}\right)}{1 - \left(\frac{K(\beta)+2}{n}\right)} \right\}.$$

where $K(\beta)$ denotes an effective dimension of the model associated to β.

The second criterion we consider is the gMDL (11) criterion motivated as a data-driven bridging the AIC and BIC. We refer the reader to (11) for more details on the gMDL criterion. For a Gaussian (L_2-

[1] Boosting algorithms in their turn relate to the pure greedy algorithms in approximation theory.

[2] LARS standing for Least Angle Regression and Selection

loss) linear model and again letting $K(\beta)$ again denote an effective dimension of the model associated to β, the gMDL estimate is defined as:

$$\hat{\beta}_{\text{gMDL}} \quad = \quad \arg \min_{\beta \in \text{path}} \text{gMDL}(Z, \beta)$$

with:

$$\text{gMDL}(Z,\beta) = \begin{cases} \log\left(\frac{\|Y - X\beta\|^2}{n - K(\beta)}\right) + \frac{K(\beta)}{2} \log\left(\frac{\frac{\|X\beta\|^2}{K(\beta)}}{\frac{\|Y - X\beta\|^2}{n - K(\beta)}}\right) + \log(n), & \text{if } R^2 > \frac{K(\beta)}{n}, \\ \log\left(\frac{Y'Y}{n}\right) + \frac{1}{2}\log(n), & \text{otherwise.} \end{cases}$$

For both LASSO and the Forward Stepwise one effective dimensionality of the model $K(\beta)$ is given by the number of non-zero terms in β. For the LASSO, this is justified by the unbiased estimate for the degrees of freedom for LASSO estimates introduced in (55).

4 Simulation results

After reviewing the algorithms we will be using and some of the theoretical properties of ℓ_0 and ℓ_1 penalized estimates, we now present the results of our simulations. As seen in Section 3 above, LASSO and Forward Stepwise have some close connections and subtle differences. Natural questions regarding how their differences and similarities translate into selection accuracy and predictive performance arise. Our experiments below are geared to shed some lights on some of these questions.

Throughout this section, we work with the squared error loss (L_2-loss) and use the `lars` package implementation for both the LASSO and Forward Stepwise selection algorithms in `R`.

4.1 Simulation Set-up

The data in our simulations is generated according to:

$$Y \quad = \quad X\beta + \varepsilon,$$

where n observations are available and p predictors can be selected, that is, $Y, \varepsilon \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$. Throughout, $\varepsilon \sim N(0, \mathbb{I}_n)$. The predictors are also Gaussian with $X \sim N(0, \Sigma)$. To avoid systematic biases favoring one or another method, both the covariance matrix of the predictors $\Sigma \in \mathbb{R}^{p \times p}$ and the coefficients of the model $\beta \in \mathbb{R}^p$ are chosen randomly. Σ is given by $\Sigma = \frac{1}{p}W$, where W has a Wishart$(\mathbb{I}_p, \lceil dp \rceil)$ distribution. The higher the multiplier d in the degrees of freedom, the less correlation among the predictors as Σ concentrates around the orthogonal design with increasing d. For the coefficients, we fix the fraction of non-zero coefficients $s \in (0, 1)$ and randomly choose $q = \lfloor sp \rfloor$ coefficients to be non-zero. Conditional on the sparsity structure, the non-zero coefficients are sampled independently from a $N(0, 1)$ distribution and re-normalized to keep the signal to noise ratio fixed at 2.0.

Once the regularization paths are traced according to the Forward Stepwise and LASSO algorithms, model estimates are picked using the AIC_C and gMDL criteria. We also compare the selected models to models chosen from the path based on full information on the model. The *prediction oracle* is defined as the estimate in the path that minimizes the model error $(\hat{\beta} - \beta)'\mathbb{E}(X'X)(\hat{\beta} - \beta)$. The *selection oracle* estimate is the one model in the path minimizing the number of selection errors (i.e., the size of the symmetric difference between the selected set and the true set).

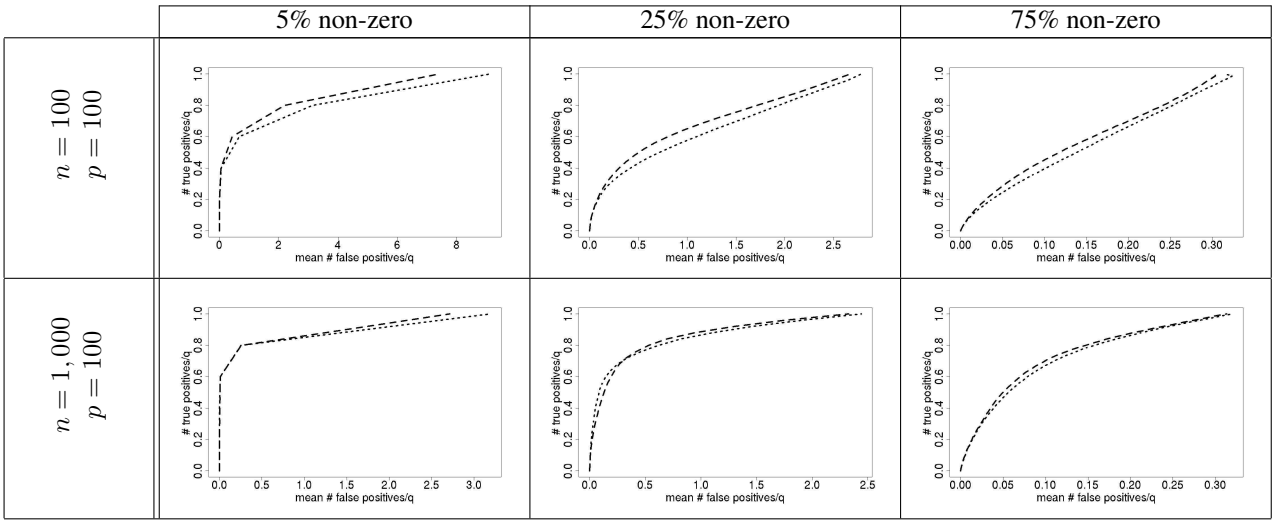

Figure 1: **'Mean ROC curves" for the LASSO (dashed) and Forward Stepwise (dotted):** Within each panel, the relative operating characteristic (ROC) curve shows the mean minimal number of false positives (horizontal axis) needed to achieve a given number of true positives (vertical axis) for both the LASSO (dashed lines) and Forward Stepwise (dotted lines). A selection procedure is better the more its curve approaches the upper left corner of the plot. As we can see here, both the LASSO and Forward Stepwise trade-off between false positives and true positives in a similar fashion for all sample sizes and sparsity levels.

4.2 Model selection results

We first take on the model selection aspects of LASSO and Forward Selection. We start by analyzing how the LASSO and Forward Stepwise trade-off between their ability of detecting true coefficients while keeping irrelevant predictors out of the model.

4.2.1 "Mean" ROC curves for the Forward Stepwise and LASSO

In this first step of our analysis, we compare the relative operating characteristic (ROC) curves for Forward Stepwise and the LASSO. An ROC curve will show the trade-off between the gain of adding a relevant variable and the loss of including an irrelevant variable as we move along the regularization path. By comparing the ROC curves of the LASSO and Forward Stepwise, we have a view of the model selection behavior of the two methods for all possible choices of the tuning parameter λ_n.

To estimate these curves, we fix a number of correctly selected variables and record the mean number of irrelevant variables included in the earliest model in the path containing that many true variables. The estimated curves are shown in Figure 1 for different sample sizes and sparsity levels.

Overall, we see a remarkable similarity in the ROC curves for the LASSO and Forward Stepwise. As a result, the ROC curves suggest that the LASSO and Forward Stepwise have similar behavior in terms of model selection accuracy over a wide range of settings.

4.2.2 The effect of the irrepresentability index and sample size

We now evaluate how much the irrepresentable index (7) affects the ability of the selection oracle to correctly select a model from the LASSO and Forward Stepwise. According to recent theoretical results (38; 1; 40), the presence of a model with all correct variables in the LASSO path is strongly related to

the irrepresentable index. Does the asymptotic results carry over to the small-n-large-p case? And how does the irrepresentable index affect the Forward Stepwise estimates if at all? The results presented in Figure 2 aim at answering these questions.

The two panels on Figure 2 show the minimum number of selection errors (ie, the number of errors committed by the selection oracle) plotted against the irrepresentability index for different sample sizes and sparsity levels as indicated. The results seem to imply that it takes large samples for the irrepresentability index to become a dominant effect on the model selection performance of the LASSO. Another interesting conclusion from Figure 2 is the relative insensitivity of Forward Stepwise to the irrepresentability index, especially in the sparser case. This was a somewhat surprising result given the similarity between the two algorithms. It also suggests that the coherence requirements in (35) as sufficient conditions for Forward Stepwise to recover the sparsest solution are overly restrictive.

4.2.3 Selection Oracle vs. AIC_C and gMDL

We now assess the model selection performance of the AIC_C and gMDL criteria for picking estimates from the LASSO and Forward Stepwise regularization paths. Figures 3 and 4 show the number of selection errors for gMDL and AIC_C and how they compare to the selection oracle for the LASSO and Forward Stepwise at different sparsity levels.

Throughout, gMDL outperforms AIC_C in keeping track of the minimal number of selection errors. Given the model selection consistency (alt. inconsistency) of BIC (alt. AIC) in the parametric case (21), that is not a surprising result: while gMDL strives to combine the virtues of BIC and AIC, the AIC_C simply adjusts the behavior of AIC for finite samples.

Also interesting is the fact that gMDL seems to approach the selection performance of the selection oracle as n increases for Forward Stepwise but not for the LASSO. That suggests that a selection criterion specifically designed for use with the LASSO regularization path can improve upon LASSO estimates picked by gMDL.

4.2.4 Specificity of Forward Stepwise and LASSO estimates

As an important application of variable selection consists of identifying potential relevant factors for further analysis, we now investigate how the Forward Stepwise and LASSO estimates fare in this respect. The important quantity in this case is the proportion of true positive effects among the selected effects. Given the imbalance between the proportion of true positives and true negatives in sparse models, a good performance in terms of number of selection errors does not necessarily translate into performance in terms of correct positive rate. Table 1 reports these results for Forward Stepwise and LASSO estimates picked by gMDL. The results for AIC_C were considerably worse and are not reported.

A high correct positive rate can be achieved by simply over-restricting the estimates. As a control for this, we also report the number of false positives among the selected predictors. A very low number of false positives serves as a warning of over-restriction. Overall, the number of false positives was about the same for Forward Stepwise and LASSO.

When an oracle is available, the correct positive rate for Forward Stepwise is significantly larger for all cases considered. However, when models are picked according to the feasible gMDL, an interesting effect occurs. For smaller samples ($n = 50, p = 100$), the LASSO estimates reach substantially higher correct positive rates than Forward Stepwise. As the sample sizes increase, Forward Stepwise gradually becomes better in the comparison to the LASSO and is preferable for large samples.

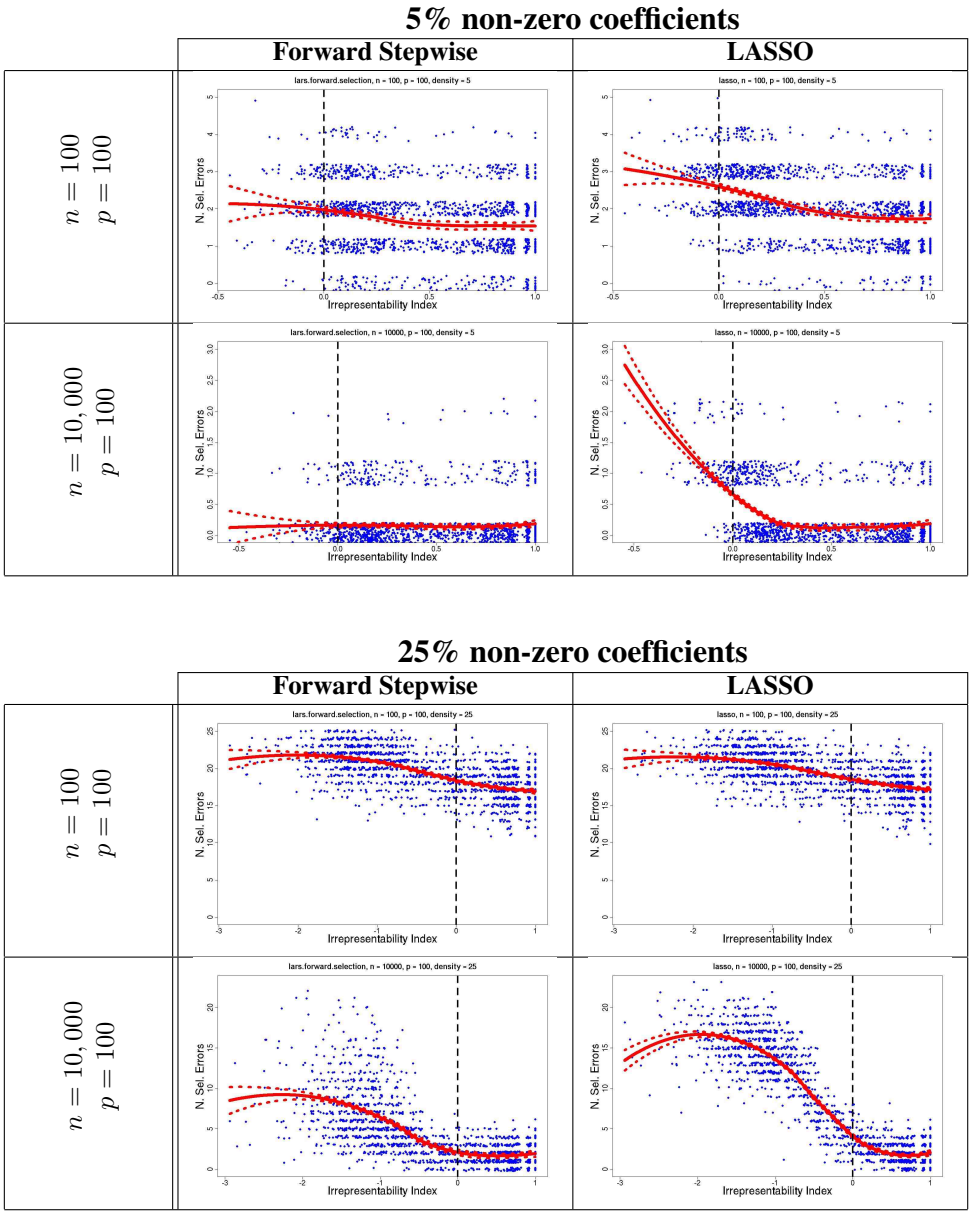

Figure 2: **Number of selection errors for LASSO and Forward Stepwise vs the irrepresentability index:** Each panel shows a plot of the (jittered) selection oracle number of selection errors vs. the irrepresentability index for the approximation and sample size indicated. In small samples, the irrepresentability index does not affect the model selection performance of neither the LASSO nor Forward Stepwise. Asymptotically, the irrepresentability index affects the LASSO more markedly than Forward Stepwise, particularly in the sparsest case.

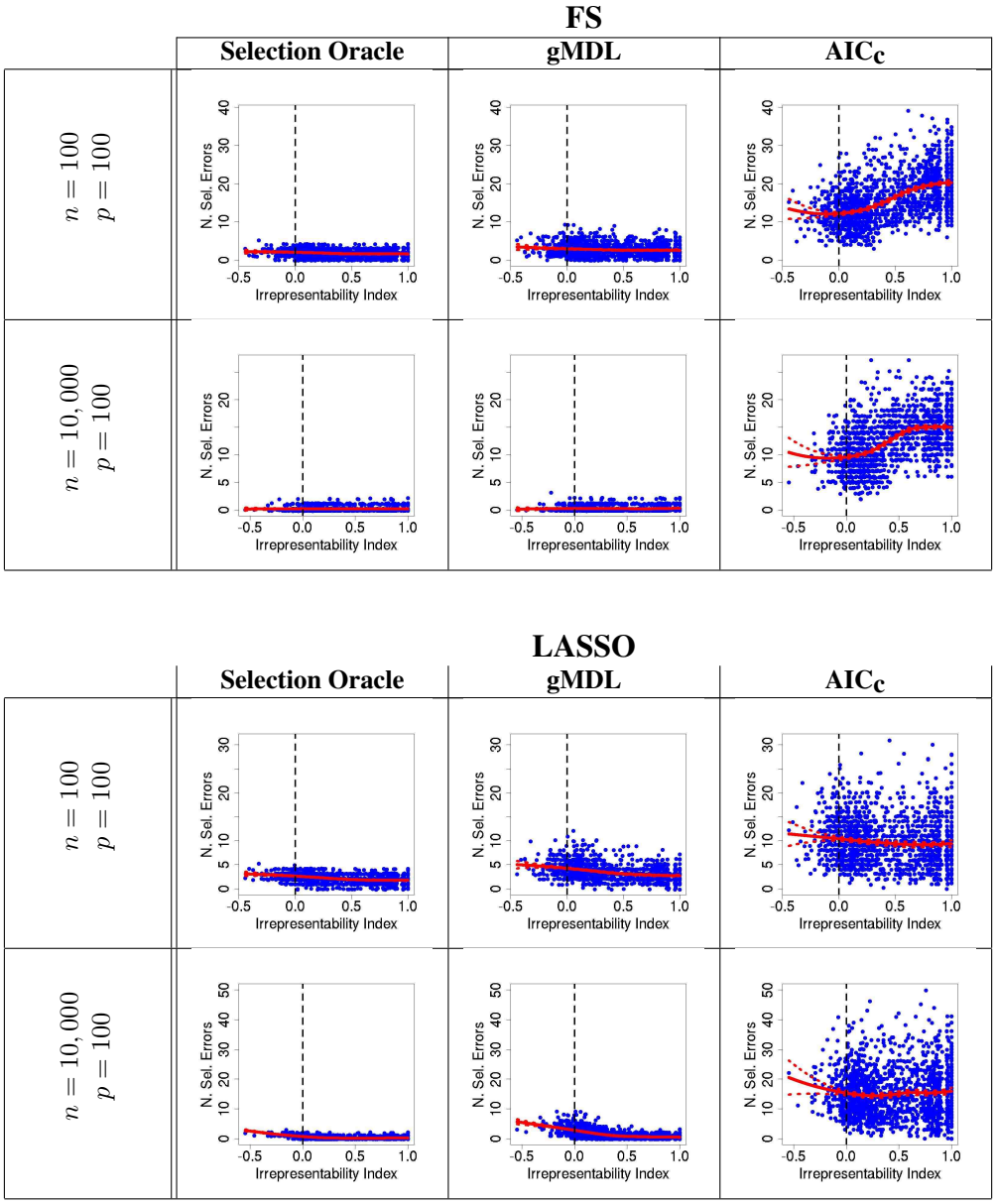

Figure 3: **Number of selection errors under** 5% **non-zero coefficients:** Each panel shows the (jittered) number of selection errors vs. the irrepresentability index for the indicated criterion and sample size. The gMDL criterion had a better performance than AIC_c in terms of number of selection errors for both LASSO and Forward Stepwise and all sample sizes considered. Using gMDL results in a slightly better selection performance for Forward Stepwise in comparison to LASSO.

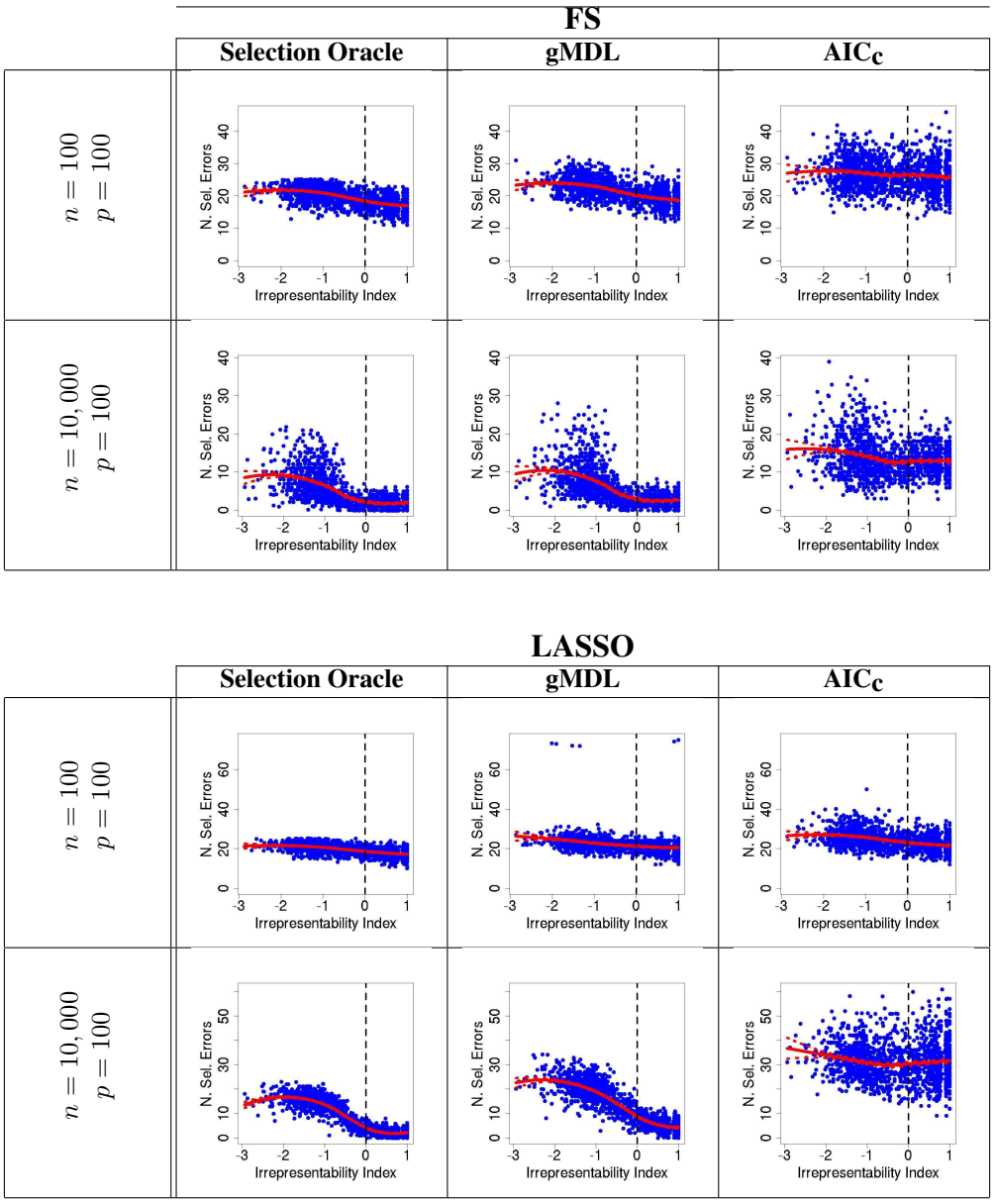

Figure 4: **Number of selection errors under** 25% **non-zero coefficients:** As in Figure 3, each panel shows the (jittered) number of selection errors vs. the irrepresentability index for the indicated criterion and sample size. The gMDL criterion still performs on par or slightly better than AIC_c in terms of number of selection errors for both LASSO and Forward Stepwise and all sample sizes considered. Again, using gMDL results in a slightly better selection performance for Forward Stepwise in comparison to LASSO.

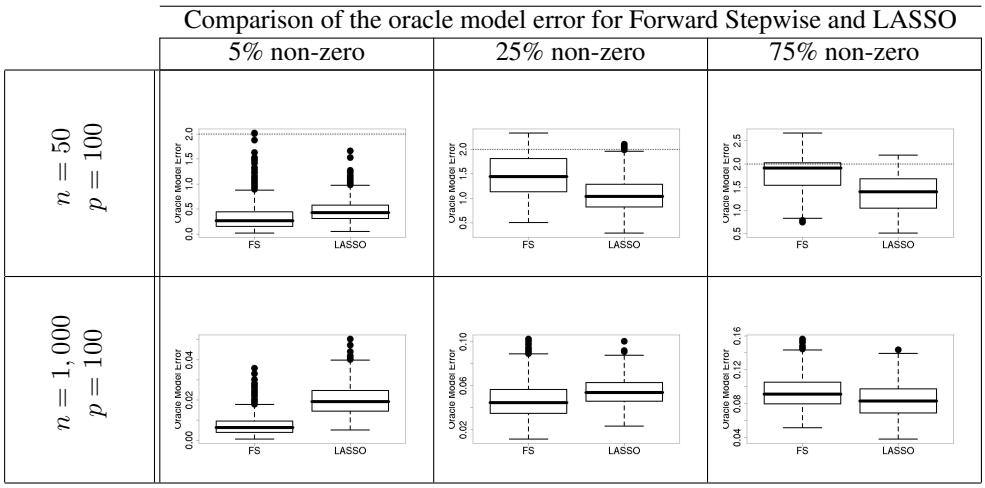

Figure 5: **Oracle model errors for Forward Stepwise and LASSO:** Each panel shows boxplots of the prediction oracle model errors for the Forward Stepwise and LASSO. The dotted lines in the upper panels indicate the model error of the null model (excluding the intercept error). In terms of the oracle model error, LASSO and Forward Stepwise perform similarly. LASSO has a slight advantage in small sample and less sparse settings, while Forward Stagewise seems better for sparser models and large sample sizes. The relative virtues of LASSO and Forward Stepwise for prediction change considerably when an oracle is no longer available (see Figure 6 below).

4.3 Prediction results

We end the exposition of our simulation results by evaluating how the LASSO and Forward Stepwise approximations compare in terms of predictive performance. Figure 5 shows boxplots comparing the model error associated to the LASSO and Forward Stepwise predictive oracles, that is, the models in the regularization path with the minimum model error. The best possible performance depends on the sparsity of the underlying model and the available sample size. In the sparsest case considered, the Forward Stepwise oracle had a better performance than its LASSO counterpart for all sample sizes considered. At less sparse regimes, the LASSO has an advantage for smaller samples, but Forward Stepwise catches up as the sample size increases.

When an oracle is not available and the sample size is small, the AIC_c estimate picked from the LASSO (LASSO+AIC_c estimate) is able to track the model error of the LASSO prediction oracle. The LASSO+AIC_c estimate had a competitive predictive performance across all simulated set-ups. This can be regarded as the LASSO version of earlier experimental and theoretical results (15; 29) for the non-negative garrote estimates. For large sample sizes and very sparse models, however, the Forward Stepwise+gMDL estimate can outperform the LASSO+AIC_c.

5 Discussion/Concluding Remarks

The MDL framework introduced by Jorma Rissanen is an instrumental tool in extracting knowledge from data. However, the high dimensional nature of many modern data sets poses computational challenges due to the combinatorial nature of the optimization problem defining many MDL estimates. A common approach to circumvent this problem consists in applying model selection criteria to a reduced list of candidates generated by algorithms that heuristically identify potentially good models.

In this paper, we present a series of experiments comparing models selected from the regularization

			Correct positive rate				# False positives			
			Sel. Oracle		gMDL		Sel. Oracle		gMDL	
n	p	q	FS	LASSO	FS	LASSO	FS	LASSO	FS	LASSO
50	100	5	98.7	97.4	66.3	75.6	2.49	2.74	2.34	2.73
			0.2	0.2	0.6	0.6	0.027	0.026	0.024	0.026
50	100	25	89.5	83.8	57.3	71.1	21.58	19.97	21.54	22.91
			0.4	0.4	0.6	0.8	0.064	0.088	0.040	0.039
100	100	5	99.1	98.0	80.3	75.2	1.67	2.01	1.63	1.82
			0.1	0.2	0.5	0.6	0.026	0.026	0.025	0.027
100	100	25	87.5	83.0	70.7	74.0	17.94	16.97	18.58	19.18
			0.3	0.3	0.5	0.5	0.092	0.099	0.060	0.082
1000	100	5	99.9	99.2	97.9	81.5	0.50	0.70	0.60	0.58
			0.0	0.1	0.2	0.5	0.017	0.019	0.018	0.018
1000	100	25	92.2	87.1	88.1	69.8	8.023	8.87	8.83	6.23
			0.3	0.3	0.4	0.4	0.109	0.118	0.085	0.072

Table 1: **Proportion of correct positives according to regression type and selection criterion:** If an oracle is available, Forward Stepwise can reach higher proportions of correctly selected variables than LASSO. Between gMDL and AIC_C, gMDL proved better for screening (hence, AIC_C is not shown). LASSO+gMDL is a better screener in small samples and Forward Stepwise+gMDL is a better screener for larger samples. Notice that the number of false positives is roughly the same for LASSO+gMDL and Forward Stepwise+gMDL within each experimental settings (n,p,q).

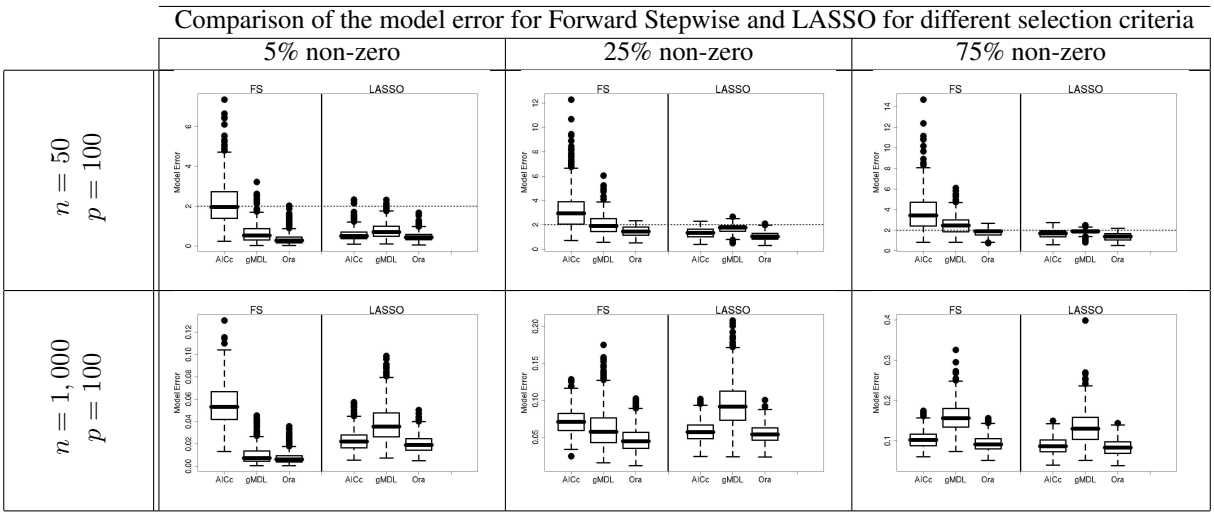

Figure 6: **Model errors for Forward Stepwise and LASSO for gMDL and AIC_C:** Each panel shows a boxplot of the model errors for Forward Stepwise and the LASSO and different selection criteria as indicated (Ora is the predictive oracle). The dotted line shows the model error of the null model. Throughout, the LASSO+AIC_C estimate managed to track the LASSO prediction oracle model error. The gMDL criterion can keep a good track of the oracle model error for Forward Stepwise in the sparsest case. Overall, LASSO+AIC_C have steadier predictive performance: it far exceeds Forward Stepwise+gMDL in the less sparse cases and it performs on par with Forward Stepwise+gMDL in the sparsest case.

path of either greedy (Forward Stepwise) or convex relaxation (LASSO) algorithms and selected by either AIC_C or the gMDL. We compare the selected models according to their prediction and variable selection performances.

In what concerns variable selection accuracy, the list of models generated by Forward Stepwise and the LASSO trade-off very similarly between false negatives and false positives, as evidenced by the experimental mean ROC curves (see Figure 1 for a definition). In terms of the number of variable selection errors, the Forward Stepwise+gMDL estimates seemed to have the best performance over the cases considered. For maximizing the correct positive rate among the selected variables, gMDL had the best results. For sample sizes smaller than the number of predictors being selected, the combination LASSO+gMDL had a better performance. As the sample sizes increased, the combination Forward Stepwise+gMDL achieved the best results.

Still regarding the selection performance of the two methods, our simulations suggest that, in small samples, the irrepresentability index (7) does not have a great influence on the oracle number of selection errors for neither the LASSO nor Forward Stepwise. Asymptotically, however, not even the selection oracle model picked from the LASSO path is model selection consistent for negative values of the irrepresentability index as postulated by theoretical results (38; 1; 40). The models picked from Forward Stepwise by the selection oracle for large samples were less affected by the irrepresentable index especially in the sparser cases. The incoherence conditions used in (35) provide sufficient conditions for the candidates recovered by Forward Stepwise to recover the best subsets, but our results suggest such conditions are overly restrictive.

In terms of prediction, the model error of models picked from the Forward Stepwise and LASSO paths by the prediction oracle performed very similarly. However, when an oracle was not available, the LASSO+AIC_C estimate had a good predictive performance across all settings tested. Such results reproduce for the LASSO, earlier simulation (15) and theoretical (29) findings for the non-negative garrote. They do provide compelling evidence to prefer the LASSO over Forward Stepwise in a reduced sample size situation. In that respect, we identify a minor theoretical gap: do Breiman's theoretical results (29) concerning the stability of the non-negative garrote carry over to the LASSO? Our simulation results seem to suggest so.

Finally, we observe an interesting parallel between the theoretical results for AIC and BIC for the all subsets case and our results. Regardless of the approximation used to obtain a list of candidate models, the AIC_C criterion was the best choice for prediction, whereas gMDL was the best performer for variable selection. Given that AIC_C and gMDL are "closer" to AIC and BIC respectively, it seems plausible that AIC-like (alt. BIC-like) criteria are more suitable for prediction (alt. variable selection) purposes when all subsets are substituted by a list of "approximately" best subsets.

References

[1] P. Zhao and B. Yu, "On model selection consistency of LASSO," Journal of Machine Learning Research, vol. 7, pp. 2541–2563, 2006. [Online]. Available: http://jmlr.csail.mit.edu/papers/volume7/zhao06a/zhao06a.pdf

[2] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," Journal of the Royal Statistical Society, Series A, vol. 135, no. 3, pp. 370–384, 1972.

[3] P. McCullagh and J. A. Nelder, Generalized Linear Models. London ; New York: Chapman & Hall, 1989.

[4] J. Rissanen, "Modeling by shortest data description," Automatica, vol. 14, pp. 465–471, 1978.

[5] ——, Stochastic Complexity in Statistical Inquiry, ser. World Scientific Series in Computer Science. Singapore: World Scientific, 1989, vol. 15.

[6] ——, Information and Complexity in Statistical Modeling, ser. Series: Information Science and Statistics. 233 Spring Street, New York, NY 10013, USA: Springer, 2007.

[7] C. S. Wallace and D. M. Boulton, "An information measure for classification," Computer Journal, vol. 11, no. 2, pp. 185–195, 1968. [Online]. Available: http://www.csse.monash.edu.au/~lloyd/tildeMML/Structured/1968-WB-CJ/

[8] C. L. Mallows, "Some comments on C_p," Technometrics, vol. 15, no. 4, pp. 661–675, 1973.

[9] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in 2nd International Symposium on Information Theory, B. N. Petrov and F. Csáki, Eds. Budapest: Akadémia Kiadó, 1973, pp. 267–281.

[10] G. Schwartz, "Estimating the dimension of a model," The Annals of Statistics, vol. 6, pp. 461–464, 1978.

[11] M. Hansen and B. Yu, "Model selection and the principle of minimum description length," Journal of the American Statistical Association, vol. 96, no. 454, pp. 746–774, 2001.

[12] N. Sugiura, "Further analysis of the data by Akaike's Information Criterion and finite corrections," Communications in Statistics, vol. A7, no. 1, pp. 13–26, 1978.

[13] X. Huo and X. S. Ni, "When do stepwise algorithms meet subset selection criteria?" Annals of Statistics, vol. 35, no. 2, pp. 870–887, 2007.

[14] S. Weisberg, Applied Linear Regression. New York: Wiley, 1980.

[15] L. Breiman, "Better subset regression using the nonnegative garrote," Technometrics, vol. 37, no. 4, pp. 373–384, 1995.

[16] R. Tibshirani, "Regression shrinkage and selection via the LASSO," Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267–288, 1996.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM Review, vol. 43, no. 1, pp. 129–159, 2001.

[18] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425–455, August 1994.

[19] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, UK ; New York: Cambridge University Press, 2004.

[20] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," Journal of the Royal Statistical Society, Series B, vol. 41, no. 2, pp. 190–195, 1979.

[21] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression," The Annals of Statistics, vol. 12, no. 2, pp. 758–765, 1984.

[22] R. Shibata, "An optimal selection of regression variables," Biometrika, vol. 68, no. 1, pp. 45–54, 1981.

[23] ——, "Asymptotic mean efficiency of a selection of regression variables," Annals of the Institute of Statistical Mathematics, vol. 35, no. 3, pp. 415–423, 1983.

[24] K.-C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," Annals of Statistics, vol. 15, no. 3, pp. 958–975, 1987.

[25] B. T. Polyak and A. Tsybakov, "Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression," Theory of Probability and its Applications, vol. 35, no. 2, pp. 293–, 1991.

[26] J. Shao, "An asymptotic theory for linear model selection (with discussions)," Statistica Sinica, pp. 221–264, 1997.

[27] Y. Yang and A. Barron, "Information theoretic determination of minimax rates of convergence," The Annals of Statistics, vol. 27, no. 5, pp. 1564–1599, 1999.

[28] Y. Yang, "Can the strengths of AIC and BIC be shared?" Biometrika, vol. 101, pp. 937–950, 2003.

[29] L. Breiman, "Heuristics of instability and stabilization in model selection," The Annals of Statistics, vol. 24, no. 6, pp. 2350–2383, 1996.

[30] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," Proceedings of the National Academy of Sciences, vol. 100, no. 5, pp. 2197–2202, 2003.

[31] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," IEEE Transactions on Information Theory, vol. 48, no. 9, p. 2558, 2002.

[32] R. Gribonval and M. Nielsen, "Sparse representation in unions of bases," IEEE Transactions on Information Theory, vol. 49, no. 12, p. 3320, December 2003.

[33] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," 2004, from the author's website. [Online]. Available: http://www-stat.stanford.edu/~donoho/Reports/2004/l1l0approx.pdf

[34] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," IEEE Transactions on Information Theory, vol. 50, no. 6, p. 1341, June 2004.

[35] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," IEEE Transactions on Information Theory, vol. 50, no. 10, pp. 2231 – 2242, October 2004.

[36] ——, "Recovery of short, complex linear combinations via ℓ_1-minimization," IEEE Transactions on Information Theory, vol. 51, no. 4, p. 1568, 2005.

[37] J. A. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," IEEE Transactions on Information Theory, 2007.

[38] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," The Annals of Statistics, vol. 34, no. 3, pp. 1436–1462, 2006.

[39] M. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," Department of Statistics, UC Berkeley, Tech. Rep., 2006. [Online]. Available: http://www.stat.berkeley.edu/tech-reports/709.pdf

[40] H. Zou, "The adaptive LASSO and its oracle properties," Journal of the American Statistical Association, vol. 101, pp. 1418–1429, 2006. [Online]. Available: http://www.stat.umn.edu/~zouhui/pub.htm

[41] E. Candes and T. Tao, "The Danzig Selector: Statistical estimation when p is much larger than n," The Annals of Statistics, 2007.

[42] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," IEEE Transactions on Information Theory, vol. 47, no. 7, p. 2845, 2001.

[43] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in Machine Learning: Proceedings of the Thirteenth International Conference, W. W. Cohen and A. Moore, Eds. Pittsburgh, Pennsylvania, USA: Morgan Kaufmann, 1996, pp. 148–156.

[44] P. Bühlmann, "Boosting for high dimensional linear models," The Annals of Statistics, vol. 34, no. 2, pp. 559–583, 2006.

[45] D. M. Allen, "The relationship between variable selection and data augmentation and a method for prediction," Technometrics, vol. 16, pp. 125–127, 1974.

[46] M. Stone, "Cross-validation choice and assessment of statistical predictions," Journal of the Royal Statistical Society, Series B, vol. 36, pp. 111–147, 1974.

[47] S. Geisser, "The predictive sample reuse method with applications," Journal of the American Statistical Association,, vol. 70, pp. 320–328, 1975.

[48] R. A. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," Advances in Computational Mathematics, vol. 5, pp. 173–187, December 1996.

[49] V. N. Temlyakov, "Weak greedy algorithms," Advances in Computational Mathematics, vol. 12, pp. 213–227, 2000.

[50] M. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its dual," Journal of Computational and Graphical Statistics, vol. 9, no. 2, pp. 319–337, June 2000. [Online]. Available: http://citeseer.ist.psu.edu/osborne99lasso.html

[51] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," The Annals of Statistics, vol. 35, pp. 407–499, 2004.

[52] P. Bühlmann and B. Yu, "Sparse boosting," Journal of Machine Learning Research, vol. 7, pp. 1001–1024, 2006.

[53] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai, "Smoothing parameter selection in nonparametric regression using an improved akaike information criterion," Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 60, no. 2, pp. 271–293, 1998. [Online]. Available: http://links.jstor.org/sici?sici=1369-7412%281998%2960%3A2%3C271%3ASPSINR%3E2.0.CO%3B2-6

[54] P. Zhao, G. Rocha, and B. Yu, "Grouped and hierarchical model selection through composite absolute penalties," Department of Statistics, UC Berkeley, Tech. Rep. 703, 2006. [Online]. Available: http://www.stat.berkeley.edu/users/gvrocha/papers/703.pdf

[55] H. Zou, T. Hastie, and R. Tibshirani, "On the "degrees of freedom" of the LASSO," Stanford University Department of Statistics, Tech. Rep., 2004. [Online]. Available: http://www-stat.stanford.edu/~hastie/Papers/dflasso.pdf