Controlling Variable Selection By the Addition of Pseudo Variables

Yujun Wu [1], Dennis D. Boos [2] and Leonard A. Stefanski [2]

Abstract

We propose a new approach to variable selection designed to control the *false selection rate* (FSR), defined as the proportion of uninformative variables included in selected models. The method works by adding a known number of pseudo variables to the real data set, running a variable selection procedure, and monitoring the proportion of pseudo variables falsely selected. Information obtained from bootstrap-like replications of this process is used to estimate the proportion of falsely-selected real variables and to tune the selection procedure to control the false selection rate.

KEY WORDS: False selection rate; forward selection; model error; model selection; subset selection.

[1]Division of Biometrics, The Cancer Institute of New Jersey, Department of Biostatistics, School of Public Health, University of Medicine and Dentistry, 195 Little Albany Street, New Brunswick, NJ 08901(E-mail: wuy5@umdnj.edu)

[2]Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: {boos, stefanski}@stat.ncsu.edu)

1 Introduction

Consider a regression model, $E(Y|X_1,\ldots,X_{k_T}) = g(\beta_1 X_1 + \cdots + \beta_{k_T} X_{k_T})$, and the problem of distinguishing the *informative* variables, defined as those with $\beta_j \neq 0$, from the *uninformative* variables, i.e., those with $\beta_j = 0$, using sample data $(\boldsymbol{Y}; \boldsymbol{X})$ ($\boldsymbol{Y}$ is $n \times 1$ and $\boldsymbol{X}$ is $n \times k_T$). Throughout we assume that $g()$ is known. Let k_I and k_U denote the number of informative and uninformative variables ($k_I + k_U = k_T$). A good variable selection procedure will on average include a high percentage of the k_I informative variables and a low but generally non-zero percentage of the uninformative variables. The non-zero, uninformative-variable inclusion rate is necessary to ensure a reasonable inclusion rate for weak informative variables, i.e., those with $|\beta_j| > 0$ but small. A method that tends to overfit includes an undesirable number of uninformative variables on average; whereas a method that tends to underfit systematically excludes weak informative variables and in doing so includes only a very small percentage of uninformative variables on average.

We propose a method based on the simple idea that the tendency of a variable selection method to overfit or underfit can be revealed by the use of pseudo explanatory variables. Suppose that $Z_1,\ldots,Z_{k_P}$ are pseudo explanatory variables, randomly generated to be independent of the response variable Y, so that

$$\begin{aligned} E(Y|\boldsymbol{X},\boldsymbol{Z}) &= g(\beta_1 X_1 + \cdots + \beta_{k_T} X_{k_T} + \beta_{k_T+1} Z_1 + \cdots + \beta_{k_T+k_P} Z_{k_P}) \\ &= g(\beta_1 X_1 + \cdots + \beta_{k_T} X_{k_T}). \end{aligned} \tag{1.1}$$

In the regression of Y on $(\boldsymbol{X},\boldsymbol{Z})$, $\beta_{k_T+1} = \cdots = \beta_{k_T+k_P} = 0$ and thus $Z_1,\ldots,Z_{k_P}$ are uninformative variables. Now consider applying a variable selection method repeatedly to B randomly generated, augmented data sets, $(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{Z}_b)$, where $\boldsymbol{Z}_b$ are $n \times k_P$ replicate pseudo-predictor matrices, $b = 1,\ldots,B$. Overfitting is indicated if the method

selects a high percentage of the k_P *phony* predictor variables on average; underfitting is indicated if the method selects very few or no phony variables on average. Once the underfitting or overfitting tendency of a selection method is revealed, the method can often be tuned to reduce undesirable fitting tendencies.

Others have noted the utility of adding noise variables to test or compare variable selection methods, (e.g., Miller, 2002, Chapter 4; Leonardo Auslender, personal communication); however, the systematic use of noise variables to tune variable selection methods is new. In related work, Luo et al. (2004) developed an approach to tune the entry significance level in forward selection based on adding additional noise to the response variable. Luo's method and our method both work by adding "noise" to the regression problem, although they differ with respect to the type of noise added. Another method in which noise is used advantageously is *stochastic resonance* wherein the noise is added to enhance the weak signals (Gammaitoni et al., 1998).

We develop and adapt the pseudo-variable method to forward selection in linear regression. We work with forward selection in this first exposition of the general approach because it is a simple, tunable selection method (the entry significance level controls overfitting/underfitting), and because forward selection is not limited to data with k_T small or $n > k_T$. These advantages of forward selection come at the expense of potentially bad performance in the presence of severe multicollinearity. However, the type and severity of multicollinearity that adversely affects forward selection is not all that common.

We show how to approximately tune forward selection to control the false selection of unimportant variables. For a given data set $(\boldsymbol{Y}; \boldsymbol{X})$ and selection method, let $U(\boldsymbol{Y}, \boldsymbol{X})$ and $I(\boldsymbol{Y}, \boldsymbol{X})$ denote the number of selected uninformative and informative variables,

respectively. The fraction of falsely-selected real predictors $\gamma(\boldsymbol{Y}, \boldsymbol{X})$ is defined as

$$\gamma(\boldsymbol{Y}, \boldsymbol{X}) = \frac{U(\boldsymbol{Y}, \boldsymbol{X})}{1 + I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})}, \tag{1.2}$$

or alternatively via the equation

$$U(\boldsymbol{Y}, \boldsymbol{X}) - \{1 + I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})\}\, \gamma(\boldsymbol{Y}, \boldsymbol{X}) = 0. \tag{1.3}$$

We add 1 to the sum $I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})$ primarily because most models have intercepts, but also because it avoids special cases needed to accommodate division by zero in (1.2). A false-selection-rate variable selection method is one that is designed to maintain, on average, $\gamma(\boldsymbol{Y}, \boldsymbol{X})$ near some predetermined target false selection rate γ_0, e.g., $\gamma_0 = .05$. Note that $(I(\boldsymbol{Y}, \boldsymbol{X}), U(\boldsymbol{Y}, \boldsymbol{X})) = (8, 1)$ and $(I(\boldsymbol{Y}, \boldsymbol{X}), U(\boldsymbol{Y}, \boldsymbol{X})) = (17, 2)$ both result in $\gamma(\boldsymbol{Y}, \boldsymbol{X}) = .1$ and thus the false selection fraction is not unique. In general we seek the largest model consistent with controlling false selection in order to include as many important variables as possible.

The alternative characterizations of $\gamma(\boldsymbol{Y}, \boldsymbol{X})$ in (1.2) and (1.3) give rise to different objective functions. If we replace $\gamma(\boldsymbol{Y}, \boldsymbol{X})$ by its target value γ_0 in (1.2) and take expectations, then we are led to seek a variable selection method having the property that

$$\gamma_0 = E\left\{\frac{U(\boldsymbol{Y}, \boldsymbol{X})}{1 + I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})}\right\}; \tag{1.4}$$

whereas starting with (1.3), substituting γ_0 for $\gamma(\boldsymbol{Y}, \boldsymbol{X})$ and taking expectations leads to seeking a variable selection method having the property that

$$\gamma_0 = \frac{E\{U(\boldsymbol{Y}, \boldsymbol{X})\}}{E\{1 + I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})\}}. \tag{1.5}$$

The expectations in (1.4) and (1.5) and in similar expressions throughout the paper are conditioned on $\boldsymbol{X}$ although we suppress the conditioning notation.

The key difference between the objective functions is that (1.4) is the expectation of a ratio, whereas (1.5) is the ratio of expectations. In Section 2 we present two false-selection rate methods corresponding to the two objective functions.

False selection rate (FSR) criteria are particulary relevant in applications where pursuing false leads is timely or costly, such as in drug discovery and certain genetic data research. Although such applications motivated our research, the new procedures are applicable more generally. More importantly, our simulation studies indicate that the new procedures are competitive with, and often superior to, existing methods in terms of other established performance criteria such as model error and model size.

Commonly used approaches for searching important subsets of variables are *all subsets*, *forward selection, backward elimination*, and *stepwise regression*. They generate a number of candidate subsets, and then an appropriate selection criterion is required to determine an optimal one. Selection criteria are often based on prediction accuracy (Breiman, 1992), as assessed by squared prediction error (PE),

$$\mathrm{PE}(\hat{\boldsymbol{g}}) = \frac{1}{n} E||\boldsymbol{Y}^{new} - \hat{\boldsymbol{g}}(\boldsymbol{X})||^2, \tag{1.6}$$

where the expectation is taken with respect to $\boldsymbol{Y}^{new}$ only, where $\boldsymbol{Y}^{new}$ is a new random vector with the same distribution as $\boldsymbol{Y}$, $\hat{\boldsymbol{g}}$ is the estimated prediction equation, and $||\cdot||$ is the Euclidean norm. For the linear model, (1.6) can be also written as

$$\mathrm{PE}(\hat{\boldsymbol{g}}) = \sigma^2 + \mathrm{ME}, \tag{1.7}$$

where σ^2 is the variance of $(Y|X_1, \ldots, X_{k_T})$ and

$$\mathrm{ME} = \frac{1}{n}||\boldsymbol{g} - \hat{\boldsymbol{g}}||^2. \tag{1.8}$$

Our definitions of PE and ME differ from those of Breiman (1992) by the factor $1/n$. The first term in PE (1.7) is not affected by the variable selection procedure, and

hence can be ignored. Many procedures are designed to minimize an estimate, $\widehat{\text{ME}}$, of ME. A widely used criterion is to minimize the C_p statistic (Mallows, 1973, 1995). However, Mallows's C_p uses a biased estimate of model error, and recently resampling methods have been used to reduce these biases, e.g., cross-validation (Breiman and Spector, 1992; Shao, 1993), the bootstrap (Shao, 1996), and the little bootstrap (Breiman, 1992). A second widely used family of criteria are those based on likelihood or information measures, such as the Akaike information criterion (AIC) (Akaike, 1973, 1977), the Bayesian information criterion (BIC) (Schwarz, 1978), and the recently developed deviance information criterion (DIC) (Spiegelhalter et al., 2002; Berg et al., 2004). They combine statistical measures of fit with penalties for increasing complexity (number of predictors). Good reviews of selection criteria can be found in Hocking (1976), Thompson (1978), Rao and Wu (2001), and Miller (2002). Bayesian approaches to model selection have been studied by George and McCulloch (1993), Chipman et al. (2001), Berger and Pericchi (2001), and Miller (2002, Chapter 7). In addition, Tibshirani (1996) proposed a subset selection method, the *LASSO*, that is similar to ridge regression, but can shrink some coefficients to 0, and thus implements variable selection. Recently, Efron et al. (2004) proposed a new model selection algorithm, *Least Angle Regression* (LARS), and show that it has good properties and that it implements the LASSO by a simple modification. The SCAD method of Fan and Li (2001) is similar to the LASSO in that it shrinks coefficient estimates to zero. Foster and Stine (2004) proposed a step-down testing variable selection procedure.

In Section 2, we describe the FSR methods in detail. Results of Monte Carlo simulations are reported comparing the new methods to established methods in Section 3. Section 4 illustrates the method with a real example, and we conclude in Section 5.

2 FSR Methods for Forward Selection

Let $\mathrm{FS}((\boldsymbol{Y};\boldsymbol{D}),\alpha)$ denote the set of columns of the design matrix $\boldsymbol{D}$ selected by running forward selection with entry level = α. Define $S(\alpha) = \mathrm{Card}\{\mathrm{FS}((\boldsymbol{Y};\boldsymbol{X}),\alpha)\}$. Then $S(\alpha) = I(\alpha) + U(\alpha)$ where $I(\alpha)$ and $U(\alpha)$ are the numbers of informative and uninformative variables selected (henceforth the dependence of I, U and S on $(\boldsymbol{Y};\boldsymbol{X})$ is suppressed). Define two FSR functions corresponding to (1.4) and (1.5) as

$$\gamma_{\mathrm{ER}}(\alpha) = E\left\{\frac{U(\alpha)}{1+S(\alpha)}\right\} = E\left\{\frac{U(\alpha)}{1+I(\alpha)+U(\alpha)}\right\}, \tag{2.1}$$

$$\gamma_{\mathrm{RE}}(\alpha) = \frac{E\{U(\alpha)\}}{E\{1+S(\alpha)\}} = \frac{E\{U(\alpha)\}}{E\{1+I(\alpha)+U(\alpha)\}}. \tag{2.2}$$

Our goal is to determine, at least approximately, α_* such that $\gamma_\bullet(\alpha_*) = \gamma_0$ for some specified γ_0, say $\gamma_0 = 0.05$ where $\gamma_\bullet$ denotes either γ_{ER} or γ_{RE}. Note that, $\gamma_\bullet(0) = 0$, $\gamma_\bullet(1) = k_U/(1+k_T)$ and $0 \le \gamma_\bullet(\alpha) \le k_U/(1+k_U)$ for all α. Thus if $k_U = 0$, then the only achievable FSR is 0; whereas for any $k_U > 0$ a solution to $\gamma_\bullet(\alpha_*) = \gamma_0$ exists for any $\gamma_0 < k_U/(1+k_T)$, assuming continuity of $\gamma_\bullet(\alpha)$. Formally we define

$$\alpha_* = \sup_\alpha\{\alpha : \gamma_\bullet(\alpha_*) \le \gamma_0\}$$

which avoids the issue of uniqueness and is consistent with identifying the largest possible model satisfying the false-selection rate bound.

Because $\gamma_\bullet(\cdot)$ is unknown, α_* can not be determined directly. We now show that it can be estimated approximately using Monte Carlo-estimated phony variables. We start considering methods based on $\gamma_{\mathrm{RE}}(\alpha)$

2.1 Method Based on Estimating $\gamma_{\mathrm{RE}}(\alpha)$

Suppose that $\boldsymbol{Z}_b$ is a set of random pseudo predictors satisfying (1.1). Let $S_{P,b}(\alpha) = \mathrm{Card}\{\mathrm{FS}((\boldsymbol{Y};\boldsymbol{X},\boldsymbol{Z}_b),\alpha)\}$. Then $S_{P,b}(\alpha) = I_{P,b}(\alpha) + U_{P,b}(\alpha) + U^*_{P,b}(\alpha)$, where $I_{P,b}(\alpha)$

and $U_{P,b}(\alpha)$ are the numbers of informative and uninformative real variables selected, and $U^*_{P,b}(\alpha)$ is the number of phony variables selected. Similar to (2.2) we define

$$\gamma_{\mathrm{RE,P}}(\alpha) = \frac{E\left\{U^*_{P,b}(\alpha)\right\}}{E\left\{1 + S_{P,b}(\alpha)\right\}} = \frac{E\left\{U^*_{P,b}(\alpha)\right\}}{E\left\{1 + I_{P,b}(\alpha) + U_{P,b}(\alpha) + U^*_{P,b}(\alpha)\right\}}. \tag{2.3}$$

Define

$$\overline{S}_P(\alpha) = B^{-1}\sum_{b=1}^{B} S_{P,b}(\alpha) \qquad \text{and} \qquad \overline{U}^*_P(\alpha) = B^{-1}\sum_{b=1}^{B} U^*_{P,b}(\alpha).$$

We estimate $\gamma_{\mathrm{RE,P}}(\alpha)$ with

$$\widehat{\gamma}_{\mathrm{RE,P}}(\alpha) = \frac{\overline{U}^*_P(\alpha)}{1 + \overline{S}_P(\alpha)}. \tag{2.4}$$

Note that $0 = \widehat{\gamma}_{\mathrm{RE,P}}(0)$ and $\widehat{\gamma}_{\mathrm{RE,P}}(1) = k_P/(1 + k_T + k_P)$. At a minimum we need $\widehat{\gamma}_{\mathrm{RE,P}}(1) > \gamma_0$, that is, $k_P > (1 + k_T)\gamma_0/(1 - \gamma_0)$, although Monte Carlo variance considerations suggest taking k_P much larger than the minimum. The left panel of Figure 2.1 displays plots of $\gamma_{\mathrm{RE}}(\alpha)$, $\gamma_{\mathrm{RE,P}}(\alpha)$ and $\widehat{\gamma}_{\mathrm{RE,P}}(\alpha)$ for one of the data generating models used in the simulation study in Section 3; $\gamma_{\mathrm{RE}}(\alpha)$ and $\gamma_{\mathrm{RE,P}}(\alpha)$ are population curves estimated by Monte Carlo methods using 100 replicate sample data sets, whereas $\widehat{\gamma}_{\mathrm{RE,P}}(\alpha)$ is the estimate defined in (2.4) for a single data set with $B = 500$.

Note that α_{**} satisfying $\gamma_{\mathrm{RE,P}}(\alpha_{**}) = \gamma_0$ does not equal α_* satisfying $\gamma_{\mathrm{RE}}(\alpha_*) = \gamma_0$. The former, α_{**}, controls the phony-variable selection rate to be γ_0 with data $(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{Z})$, whereas α_* controls the uninformative-variable selection rate to be γ_0 with data $(\boldsymbol{Y}; \boldsymbol{X})$. However, note that $\gamma_{\mathrm{RE,P}}(\alpha_*) = c$ for some $c > \gamma_0$. The right panel of Figure 2.1 illustrates the relationship between α_*, α_{**}, γ_0 and c for FSR curves $\gamma_{\mathrm{RE,P}}(\alpha)$ and $\gamma_{\mathrm{RE}}(\alpha)$. The critical values α_* and α_{**} can be very different depending on the average slope of $\gamma_{\mathrm{RE,P}}(\alpha)$ for small α which is an increasing function of k_P. Thus in order to use $\widehat{\gamma}_{\mathrm{RE,P}}(\alpha)$ to estimate α_* we need a method to adjust for the number of

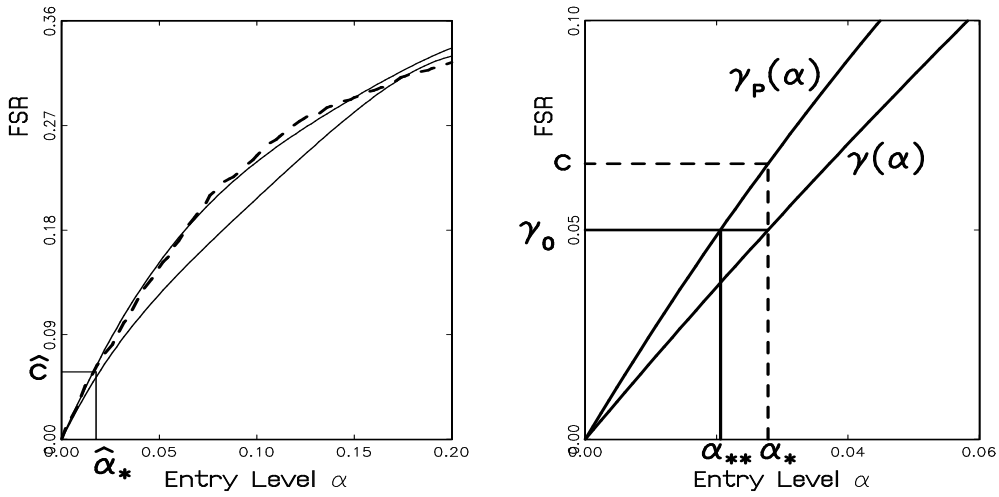


Figure 2.1: Left panel: $\widehat{\gamma}_{\mathrm{RE,P}}(\alpha)$ (dashed line, $\widehat{\alpha}_* = 0.018$ and $\widehat{c} = 0.058$) for a single simulated data set superimposed over Monte Carlo estimated versions of $\gamma_{\mathrm{RE}}(\alpha)$ and $\gamma_{\mathrm{RE,P}}(\alpha)$ (solid lines, $\gamma_{\mathrm{RE,P}}(\alpha)$ lies entirely above $\gamma_{\mathrm{RE}}(\alpha)$), for data with $n = 150$, $k_U = 15$, $k_I = 6$, $\rho = 0$ and $k_P = 21$. Right panel: Schematic diagram of theoretical FSR curves $\gamma_{\mathrm{RE}}(\alpha)$ and $\gamma_{\mathrm{RE,P}}(\alpha)$ showing the relationships between α_*, α_{**}, γ_0 and c.

phony variables used, or equivalently to estimate c in Figure 2.1. We derive a procedure under the assumptions:

(A1) $E\{U(\alpha)\} = E\{U_{P,b}(\alpha)\} = E\{U^*_{P,b}(\alpha)\}k_U/k_P$;

(A2) $E\{I_{P,b}(\alpha)\} = E\{I(\alpha)\}$.

Assumption (A1) states that real unimportant variables and phony unimportant variables have the same probability of being selected on average; (A2) states that the real important variables have the same probability of being selected whether phony variables are present or not. These assumptions are almost certainly to be violated, and in any case virtually unverifiable. Thus we recognize that our method is only approx-

imate, and we regard (A1) and (A2) more as guiding principles to the generation of phony variables, rather than as crucial mathematical conditions justifying the methods. In this vein we see that (A1) and (A2) indicate that the phony variables should behave as much like the real unimportant variables as possible (A1), and should have as little impact as possible on the selection rate of the real important variables (A2). We will revisit these issues when we discuss the generation of phony variables. For now we continue deriving the relationship between γ_0 and c.

Setting $\gamma_{\text{RE}}(\alpha) = \gamma_0$ and $\gamma_{\text{RE,P}}(\alpha) = c$ and solving for c using (A1) and (A2) shows that

$$c = \frac{\gamma_0 k_P}{\gamma_0 k_P + k_U},$$

which does not depend on any of the real- or phony-variable selection probabilities, but does depend on the unknown k_U. Thus apart from the dependence on k_U we are led to consider the equation $\widehat{\gamma}_{\text{RE,P}}(\alpha) = (\gamma_0 k_P)/(\gamma_0 k_P + k_U)$, i.e., with $\widehat{\gamma}_{\text{RE,P}}(\alpha)$ as defined in (2.4), we consider

$$\frac{\overline{U}^*_P(\alpha)}{1 + \overline{S}_P(\alpha)} = (\gamma_0 k_P)/(\gamma_0 k_P + k_U).$$

Algebraic rearrangement of this equation shows its equivalence to

$$\frac{k_U \overline{U}^*_P(\alpha)/k_P}{1 + \overline{S}_P(\alpha) - \overline{U}^*_P(\alpha)} = \gamma_0. \tag{2.5}$$

Note that under (A1) and (A2) the numerator above is estimating $E\{U(\alpha)\}$ and the denominator is estimating $E\{1 + I(\alpha) + U(\alpha)\}$. Define

$$\widehat{k}_U(\alpha) = k_T - \text{Card}\{\text{FS}((\boldsymbol{Y}; \boldsymbol{X}), \alpha)\}.$$

Then at the desired entry level of α we expect

$$\widehat{k}_U(\alpha) \approx k_U. \tag{2.6}$$

Combining (2.6) with (2.5) results in the estimator of $\gamma_{\text{RE}}(\alpha)$ given by

$$\widehat{\gamma}_{\text{RE}}(\alpha) = \frac{\widehat{k}_U(\alpha)\overline{U}^*_P(\alpha)/k_P}{1 + \overline{S}_P(\alpha) - \overline{U}^*_P(\alpha)}, \tag{2.7}$$

from which we obtain

$$\widehat{\alpha}_{\text{RE}} = \sup_{\alpha}\{\alpha : \widehat{\gamma}_{\text{RE}}(\alpha) \leq \gamma_0\}. \tag{2.8}$$

Then the final model selected variables are determined by applying forward selection to the real data with entry level = $\widehat{\alpha}_{\text{RE}}$. In application, the supremum in (2.8) is a maximum taken over a finite grid of α values. Restricting to a grid has little effect on the final model.

2.2 Method Based on Estimating $\gamma_{\text{ER}}(\alpha)$

Our method based on estimating $\gamma_{\text{ER}}(\alpha)$ is conceptually simpler and is less assumption dependent although it involves an approximation of a different nature. Note that $S(\alpha)$ in the definition of $\gamma_{\text{ER}}(\alpha)$ in (2.1) is known, as it is just the total number of variables, $\text{Card}\{\text{FS}((\boldsymbol{Y}; \boldsymbol{X}), \alpha)\}$, selected by forward selection with entry level = α. Thus if $U(\alpha)$ were also known, we could simply determine $\widehat{\alpha}_{\text{ER}}$ as

$$\widehat{\alpha}_{\text{ER}} = \sup_{\alpha}\{\alpha : U(\alpha)/\{1 + S(\alpha)\} \leq \gamma_0\}.$$

But $U(\alpha)$ is unknown. The suggested procedure is to replace the unknown random variable $U(\alpha)$ with a suitable predictor $\widehat{U}(\alpha)$. In light of the reasoning culminating in (2.7), we use $\widehat{U}(\alpha) = \widehat{k}_U(\alpha)\overline{U}^*_P(\alpha)/k_P$. Thus our second FSR method starts with

$$\widehat{\gamma}_{\text{ER}}(\alpha) = \frac{\widehat{k}_U(\alpha)\overline{U}^*_P(\alpha)/k_P}{1 + S(\alpha)}, \tag{2.9}$$

from which is calculated

$$\widehat{\alpha}_{\text{ER}} = \sup_{\alpha}\{\alpha : \widehat{\gamma}_{\text{ER}}(\alpha) \leq \gamma_0\}. \tag{2.10}$$

As in the first method, the final model selected variables are determined by running forward selection on the real data with entry level = $\widehat{\alpha}_{\text{ER}}$. In application, the supremum in (2.8) is again a maximum taken over a finite grid of α values.

The derivation of (2.9) used (A1) but not (A2) and is thus less dependent on assumptions. However, the joint distributions of $(\widehat{U}(\alpha), 1 + I(\alpha) + U(\alpha))$ and $(U(\alpha), 1 + I(\alpha) + U(\alpha))$ are clearly not equal and thus neither will the distributions of the ratios of their components. Thus while our second method does not depend on (A2), $\widehat{\gamma}_{\text{ER}}(\alpha)$ is not unbiased for $\gamma_{\text{ER}}(\alpha)$ in general.

2.3 Pseudo Variable Generation

As noted previously (A1) and (A2) provide guidance for generating pseudo variables. The objective is to generate pseudo variables so that the average inclusion probabilities of informative variables are approximately equal with data $(\boldsymbol{Y}; \boldsymbol{X})$ and $(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{Z})$; and the average inclusion probabilities of uninformative variables (real and pseudo) are approximately equal with data $(\boldsymbol{Y}; \boldsymbol{X})$ and $(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{Z})$.

We have studied four methods. In the first, entries in the $n \times k_P$ matrix $\boldsymbol{Z}$ are independent and identically distributed $\text{N}(0, 1)$, whereas in the second the n rows of $\boldsymbol{Z}$ are obtained by randomly permuting the rows of $\boldsymbol{X}$ (this restricts $k_P = k_T$). In both methods the pseudo predictors are stochastically uncorrelated with the true predictors, and in the second method the pseudo predictors have the same distribution as the real predictors. The third and fourth methods are variants of the first two in which $\boldsymbol{Z}$ is replaced by $(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{Z}$, where $\boldsymbol{H}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The variants are such that the pseudo predictors have sample means and sample correlations with the real predictors that are identically equal to zero. Thus the variants work like Monte Carlo swindles to reduce variation. Note that the variants are possible only for the case $n > k_T$. A

simulation study presented in the first author's PhD thesis (Wu, 2004) compared the four methods and found a slight advantage to the fourth method. Accordingly the results presented in this paper use the regression-residual, permutation method.

The advantage of the pseudo variables generated using the fourth method is not surprising. Permutation produces pseudo variables that when appended to the real data create what are essentially matched pairs. To each real variable there corresponds a pseudo variable with identical sample moments and also with preservation of correlations. In fact in the all-null model (no real important variables), the permutation method ensures that (A1) holds asymptotically. Projection via $(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{x}})$ helps reduce the effect of the pseudo variables on the selection probabilities of the real important variables, thus lessening the extent to which (A2) is violated.

3 Simulation Studies

We studied our two methods via Monte Carlo experiments and compared them with a variety of common selection procedures in the context of linear regression. In our FSR methods, we fixed $\gamma_0 = 0.05$ and repeatedly generated $B = 500$ sets of pseudo variables.

The two methods of estimating the entry significance level, $\widehat{\alpha}_{\text{RE}}$ and $\widehat{\alpha}_{\text{ER}}$ in (2.8) and (2.10) respectively, produced nearly identical results and thus we report only the results for $\widehat{\alpha}_{\text{ER}}$ as it is the least assumption-driven method. The similar performance of $\widehat{\alpha}_{\text{RE}}$ and $\widehat{\alpha}_{\text{ER}}$ is explained in part by the similarity of (2.1) and (2.2) for small α and uncorrelated predictors, although (2.1) and (2.2) differ more substantially for larger α and when there is high collinearity. Just as important is the fact that $\text{FS}((\boldsymbol{Y}; \boldsymbol{D}), \alpha)$ is a step function and thus $\widehat{\alpha}_{\text{ER}}$ and $\widehat{\alpha}_{\text{RE}}$ can differ yet still result in the same model.

3.1 Simulation Design

Luo et al. (2004) developed a comprehensive simulation design to study the performance of their *noise-addition model selection* procedure (NAMS). The design is based on one used initially by Tibshirani and Knight (1999) and related to that of Breiman (1992). Our design uses the same simulated data sets from Luo et al. (2004) in order to facilitate comparison with their results. Each data set contains 21 predictors generated from a multivariate normal distribution with mean 0 and an autoregressive covariance structure, where the covariance between any two predictors x_i and x_j is equal to $\rho^{|i-j|}$, with $\rho = 0$ and 0.7, respectively. The generated design matrix $\boldsymbol{X}$ is then held fixed for all runs. The response variable is generated from the linear regression model $y_i = \beta_0 + \mathbf{x_i}\boldsymbol{\beta} + \epsilon_i$, $i = 1, \cdots, n$, where $\mathbf{x_i}$ is the ith row of the design matrix $\boldsymbol{X}$, β_0 is the intercept (automatically fit for all selection procedures), $\boldsymbol{\beta}$ is the 21×1 regression coefficients vector, and the random errors ϵ_i independently follow a standard normal distribution $N(0, 1)$. The generation of the coefficients vector $\boldsymbol{\beta}$ are described in the following two stages.

1. The nonzero coefficients are clustered around two variables: x_7 and x_{14}, with their initial values given by

$$\beta_{7+j} = (h - j)^2, \quad \beta_{14+j} = (h - j)^2, \qquad |j| < h, \tag{3.1}$$

   where h has values 1, 2, 3, and 4. All other coefficients are zero.

2. The coefficients are then multiplied by a common constant to make the theoretical model R^2 equal to 0.75 or 0.35, where theoretical R^2 is defined as

$$R^2 = \frac{(\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{\beta}) + n\sigma^2}. \tag{3.2}$$

The results for $R^2 = 0.35$ and $R^2 = 0.75$ were qualitatively similar, and so we present only the $R^2 = 0.75$ case here. The four different values of h in (3.1) result in models with 2, 6, 10 and 14 nonzero coefficients. We designate these four different sets of coefficients by H1, H2, H3 and H4. In each model, the nonzero coefficients are specified with different values such that the associated variables make different contributions to the response variable. For model H1, there are two strong variables. At the other extreme, model H4 contains four weak variables. In addition, we also include the null model for which all coefficients are zero, denoted by H0. Each data set has either 50, 150 or 500 observations. For each combination of sample size n and ρ there are 100 repetitions for each model H0-H4.

For selection method evaluation, the average model error defined in equation (1.8) is used as the primary criterion. The average model size and false selection rate are also used to assess how well the selection procedures identify the correct size and control the selection of uninformative variables.

3.2 Operating Characteristics of the FSR Procedures

In this section, we study the dependence of the FSR methods on the number of pseudo variables generated (k_P) and the characteristics of estimated α_*.

3.2.1 The Number of Pseudo Variables

It is expected that the ratio $\overline{U}_P^*(\alpha)/k_P$ in (2.7) and (2.9) will change little for different k_P, and hence we anticipate that our methods are robust to the choice of k_P. We conducted a simulation study designed to investigate the sensitivity of the FSR procedures to choice of k_P. In the study the sample size was $n = 150$ and three values of k_P, 10, 21 and 42, were studied for each of the models H0-H4. The regression-residual, permutation method naturally generates $k_T = 21$ pseudo variables; so for $k_P = 10$ we

randomly selected 10 out of the 21 pseudo variables, and for $k_P = 42$ we combined two sets of regression-residual, permutation method phony variables. Average ME and average model size over the 100 Monte Carlo datasets were nearly identical for the range of models H0-H4 for three values of k_P, and so we use $k_P = k_T$ exclusively in the remainder of the simulation studies.

3.2.2 Estimated Significance Levels

Our methods estimate α_* designed to control the FSR. Intuitively, larger values of α_* should be obtained for larger models in order to avoid low power in identifying the informative variables, especially when some weak informative variables are present. Our simulation studies confirmed the relationship between α_* and model size. Boxplots of $\widehat{\alpha}_{\mathrm{ER}}$ for the range of models H0-H4 with sample size $n = 150$ are displayed in Figure 3.1. The trend with model size is evident. A similar trend was found for sample sizes $n = 50$ and $n = 500$.

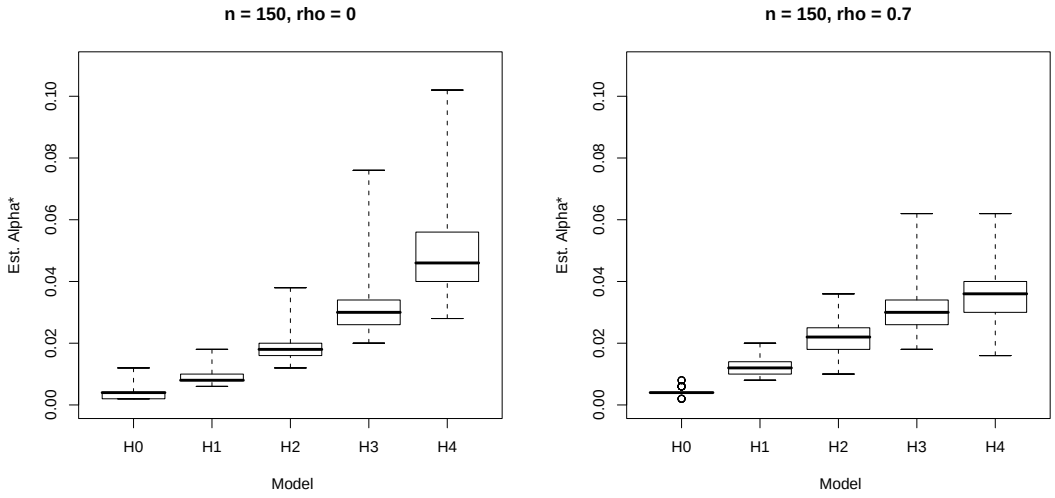


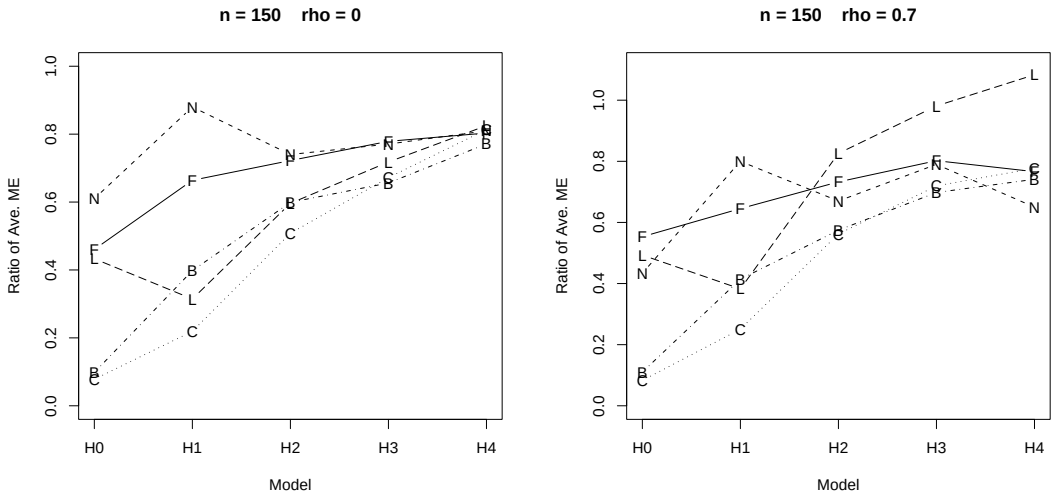
Figure 3.1: Boxplots of $\widehat{\alpha}_{\mathrm{ER}}$ found by the FSR method with $\gamma_0 = 0.05$ for $n = 150, k_T = 21$ and $k_P = 21$.

3.3 Comparison to Other Variable Selection Methods

To further study the performance of the FSR procedures, we compared them with NAMS (Luo et al., 2004), Minimum C_p (Mallows, 1973, 1995), the little bootstrap (LB) (Breiman, 1992), and the LASSO (Tibshirani, 1996). We also include in the study the *Best* selection method that selects the model having minimum model error from the forward selection candidate models based on inside knowledge of the true model. It plays the role of a gold standard and indicates the best that a forward selection procedure can do in prediction. Although Minimum C_p is based on all subsets searching, all other selection procedures (not including the LASSO) are based on the forward selection algorithm. In LB, we let $t = 0.6$ and B=40 as suggested by Breiman (1992). The results displayed and discussed in the following of the section were obtained by the FSR method based on estimating $\gamma_{\text{ER}}(\alpha)$.

We first concentrate on the results for $n = 150$. The ratios of average ME of Best selection to average ME of the FSR procedure, NAMS, Minimum C_p, LB, and LASSO are calculated and plotted versus model in Figure 3.2. Of course, the larger the ratio, the better is the selection method in terms of prediction. The most striking result is that the FSR procedure performs substantially better than Minimum C_p and LB. Compared to NAMS, the FSR procedure produces fairly close results when $\rho = 0$ except for models H0 and H1, but performs somewhat better when $\rho = 0.7$. The plots also indicate that the FSR procedure exhibits a significant advantage over the LASSO when $\rho = 0$. When $\rho = 0.7$, the FSR method does not do as well as the LASSO for models H3 and H4, but it still performs better in models H0 and H1.

The above results are also reflected in the comparison of model size. Figure 3.3 presents the deviations of average model size from Best selection for the five compared methods in the five H models. It is clear that Minimum C_p and LB always select

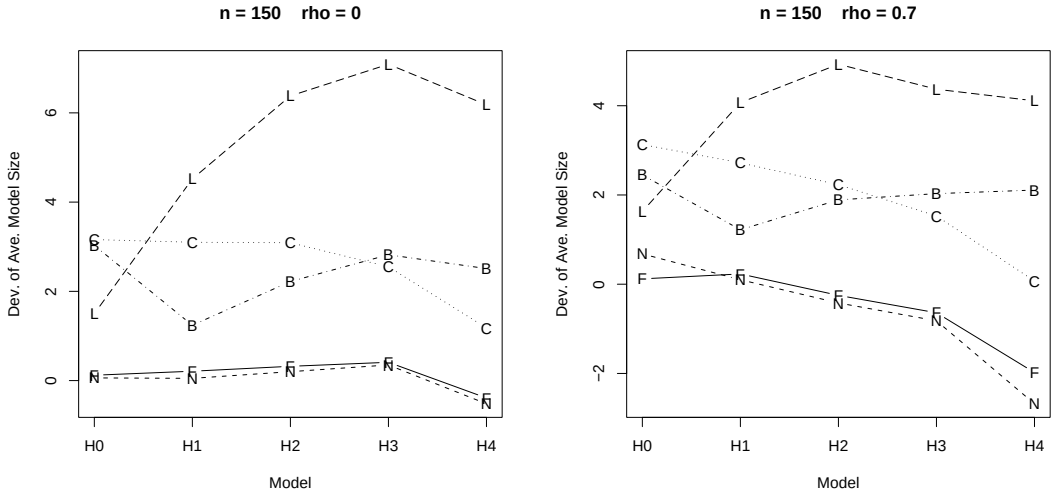


F : FSR N : NAMS C : C_p B : LB L : LASSO

Figure 3.2: The ratios of average ME of Best to average ME of the FSR method, NAMS, C_p, LB and LASSO for $n = 150, k_T = 21$ and $k_P = 21$. Standard errors of average ME range from 0.002 to 0.007.

larger models than the optimal selection by Best, indicating their tendency to overfit. The overfitting is considerable when the true model size is small, e.g., for models H0, H1 and H2, resulting in the model selected being penalized by a larger average model error. This overfitting becomes less damaging if there are many weak variables, because it will then tend to include them. For this reason, Minimum C_p and LB have good performance in model H4. For the LASSO, a substantial overfitting is also revealed in the picture, but the shrinkage of coefficients reduces the damage and results in good prediction for large size models, especially when the X-variables are correlated. In contrast, the FSR method and NAMS generally select parsimonious models and hence are the best in reducing complexity. Both methods achieve almost the same selection as Best except for H4 model where a mild underfitting is revealed.

We also compared false selection rates. Monte Carlo estimated false selection rates,

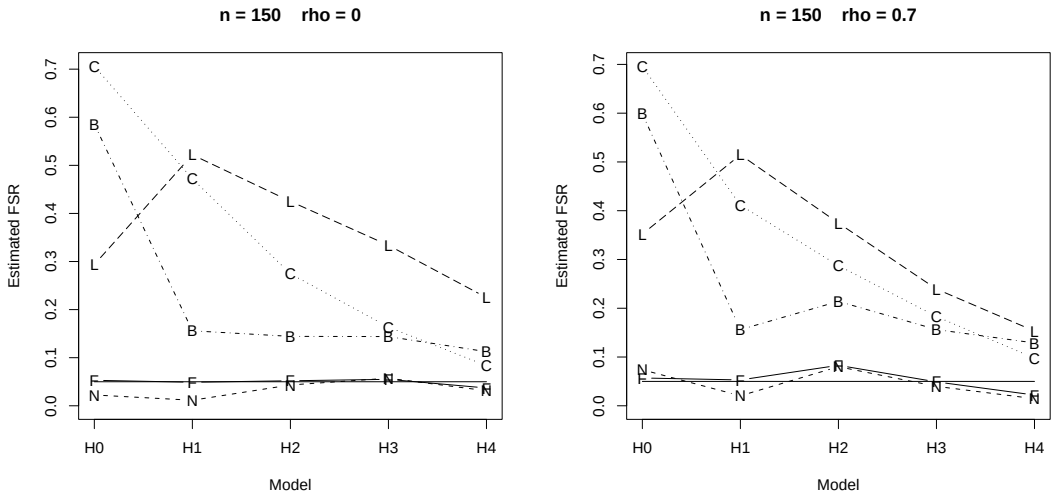


F : FSR N : NAMS C : C_p B : LB L : LASSO

Figure 3.3: Deviation of average model size for the FSR method, NAMS, C_p, LB, and LASSO from Best selection for $n = 150, k_T = 21$ and $k_P = 21$. Standard errors of average model size range from 0.04 to 0.15 for Best, the FSR method and NAMS, from 0.15 to 0.19 for C_p, from 0.29 to 0.50 for LB, and from 0.23 to 0.33 for LASSO.

computed by taking the average of proportions of uninformative variables in the final model over 100 replications, are plotted in Figure 3.4. The solid horizonal lines correspond to the value of 0.05, which is the desired level in our FSR method. It turns out that Minimum C_p, LB and LASSO tend to select many uninformative variables into the model as evidenced by their high false selection rates. As expected, the FSR method maintains the false selection rate near its target level 0.05, whereas NAMS tends to produce slightly more conservative (smaller) false selection rates.

Similar results are observed for $n = 50$ and 500 as well, for which the graphs are not shown. In particular, when $n = 50$, the FSR method shows an obvious advantage over NAMS for correlated X-variables in terms of average ME, and the selection by both the FSR method and NAMS tend to be more conservative (smaller selection). Compared to $n = 150$, the variation of estimated false selection rates for the FSR method across



F : FSR N : NAMS C : C_p B : LB L : LASSO

Figure 3.4: The Monte Carlo estimated false selection rates for the FSR method, NAMS, C_p, LB and LASSO for $n = 150, k_T = 21$ and k_P= 21. Standard errors range from 0.005 to 0.036.

the models is larger when $n = 50$, but becomes smaller for $n = 500$.

In addition, to study the effect in the different methods when more uninformative variables are included, we reran the sample size 150 cases, increasing the number of real predictors in the data sets to 42 ($k_T = 42$) by adding the squares of the original 21 variables. Because the true regression model is the same, each model now has 21 additional uninformative predictors. For average ME, average model size, and false selection rate, results similar to the $k_T = 21$ case were obtained. Table 3.1 presents the deviations of average ME from the average ME on the 21 original variables, i.e., $\text{Ave}(\text{ME}_{k_T=42}) - \text{Ave}(\text{ME}_{k_T=21})$, for different selection methods. An obvious result is that Minimum C_p and LB are adversely affected by the additional 21 predictors to a large extent as evidenced by the large deviations, whereas the FSR method and NAMS exhibit relatively stable performance with small loss in predictive activity. This

implies a large sensitivity of Minimum C_p and LB to the number of uninformative variables in the data. The adverse effect of the additional 21 predictors on the LASSO is considerable when $\rho = 0$, but is similar to the FSR method when $\rho = 0.7$.

Table 3.1: Deviation of average ME when 21 additional variables are added for sample size 150 cases (Ave. $\text{ME}_{k_T=42}$ − Ave. $\text{ME}_{k_T=21}$). Standard errors of the entries range from 0.002 to 0.007.

ρ	Method	H0	H1	H2	H3	H4
0	FSR	0.005	0.008	0.012	0.017	0.031
	NAMS	0.008	0.002	0.026	-0.001	0.023
	C_p	0.082	0.085	0.065	0.074	0.084
	LB	0.006	0.004	0.013	0.011	0.040
	LASSO	0.002	0.015	0.028	0.043	0.052
0.7	FSR	0.004	0.008	0.013	0.021	0.028
	NAMS	0.008	0.007	0.011	0.041	0.011
	C_p	0.086	0.076	0.075	0.076	0.085
	LB	0.006	-0.008	0.015	0.040	0.048
	LASSO	0.004	0.015	0.015	0.023	0.028

4 Application to Diabetes Data

We now demonstrate our methods with data used previously by Efron et al. (2004) to illustrate their model selection method, LARS. Ten baseline variables were obtained for each of $n = 442$ diabetes patients, including AGE, SEX, BMI (Body mass index), BP(average blood pressure), and six blood serum measurements, S1 - S6. The response is a quantitative measure of disease progression one year after baseline. A statistical model is desired to identify the important baseline factors in disease progression.

The data are first standardized as done by Efron et al. (2004) such that the baseline variables have mean 0 and unit length, and that the response has mean 0. Then, a linear regression model with baseline variables is fit by forward selection. With only

ten main effects considered, forward selection produced the following sequence.

Summary of Forward Selection

Step	Variable Entered	Model R-Square	Fvalue	Pr > F
1	BMI	0.3439	230.65	<.0001
2	S5	0.4595	93.86	<.0001
3	BP	0.4801	17.35	<.0001
4	S1	0.4920	10.27	0.0015
5	SEX	0.4999	6.84	0.0092
6	S2	0.5149	13.47	0.0003
7	S4	0.5163	1.26	0.2619

Our two FSR methods resulted in $\widehat{\alpha}_{\text{ER}} = 0.11$ and $\widehat{\alpha}_{\text{RE}} = 0.11$, and hence both methods stop forward selection in Step 6 producing identical six-variable models.

We also considered a quadratic model with ten main effects, 45 two-way interactions, and nine quadratic terms (each baseline variable except the dichotomous variable SEX). We did not force the hierarchy principle, and thus interaction effects can enter before main effects. Forward selection produces the sequence

Summary of Forward Selection

Step	Variable Entered	Model R-Square	Fvalue	Pr > F
1	BMI	0.3439	230.65	<.0001
2	S5	0.4595	93.86	<.0001
3	BP	0.4801	17.35	<.0001
4	AGE*SEX	0.4957	13.56	0.0003
5	BMI*BP	0.5066	9.60	0.0021
6	S3	0.5166	9.00	0.0029
7	SEX	0.5340	16.23	<.0001
8	S6*S6	0.5399	5.53	0.0192

Our methods resulted in $\widehat{\alpha}_{\text{ER}} = \widehat{\alpha}_{\text{RE}} = 0.01$, and each identifies the same seven-variable model.

For comparison with LARS, Table 4.1 shows the variables included in the model by the FSR methods and LARS. The variables are listed in the order of entering the model. LARS uses a C_p-type selection criterion for stopping. For the main-effects only modeling, the FSR methods and LARS produce very similar models with nearly identical R^2. For the quadratic modeling, the seven variables selected by the FSR methods are a subset of the sixteen variables included in the LARS model. The LARS model has a higher coefficient of determination ($R^2_{\text{LARS}} = .5493,\ R^2_{\text{FSR}} = .5340$), although the increase of .0153 comes at the expense of having nine additional variables in the model.

Table 4.1: Variables selected by the FSR method and LARS for the diabetes data

Model	Method	Variables in the model	Model R^2
Main Effects	FSR	BMI, S5, BP, S1, SEX, S2	0.5149
	LARS	S5, BMI, BP, S3, SEX, S6, S1	0.5146
Quadratic	FSR	BMI, S5, BP, AGE*SEX, BMI*BP, S3, SEX	0.5340
	LARS	BMI, S5, BP, S3, BMI*BP, AGE*SEX, $S6^2$, BMI^2, AGE*BP, AGE*S6, SEX, S6, AGE*S5, AGE^2, SEX*BP, BP*S3	0.5493

5 Conclusions

The false selection rate provides a meaningful performance measure for variable selection procedures that complements existing measures such as model error and model size. The key task is to control the false selection rate, keeping the selection procedure from including too many uninformative variables in the model while identifying

as many informative variables as possible. Our FSR methods accomplish this task with the aid of pseudo variables. Just as a physical scientist calibrates a measuring device by presenting to it known standards and recording measured amounts, we calibrate forward selection by presenting it with known phony variables and monitoring the proportion of those phony variables selected.

The proportion of phony variables selected at each entry level α is the key component of our approximate estimates of the false selection rate curves $\widehat{\gamma}_{\text{RE}}(\alpha)$ and $\widehat{\gamma}_{\text{ER}}(\alpha)$. The utility of the estimated false selection rate curves to control the actual false selection rates in replicate analyses was demonstrated in simulation results reported in Section 3 and in extensive simulation studies reported in the first author's PhD thesis (Wu, 2004). An unexpected but welcomed correlate was that controlling FSR at $\gamma_0 = .05$ produced parsimonious models with very favorable model error properties, comparable to and often better than established variable selection methods. In applications for which the false selection rate is the most meaningful performance criterion, γ_0 can be adjusted up or down as desired. We demonstrated the value of the FSR method in linear regression, but it applies to any regression model amenable to forward selection.

Because of the simulation required to compute $\widehat{\gamma}_{\text{RE}}(\alpha)$ and $\widehat{\gamma}_{\text{ER}}(\alpha)$, computation time is an issue. However, it is not a limiting factor except in large simulation studies. For the data sets of size $n = 150$ used in the simulation study, the FSR methods with $B = 500$ required an average of about 30 seconds on a desktop PC, compared to about 2 seconds for the LASSO. Thus while they are more time consuming than other methods, they are well within the feasible range for single data sets. In response to a reviewer's question about computation time, we reran many of the simulations reported in this paper taking $B = 100$. This lessened computation time considerably, with almost negligible degradation in performance in the FSR method.

ACKNOWLEDGEMENTS

We are grateful to the Editor and Referees for several helpful suggestions that substantially improved the paper, and to the National Science Foundation for supporting our research. Support for Y. Wu was also provided by the University of Medicine and Dentistry of New Jersey (UMDNJ) Foundation.

References

Akaike, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255-265.

—— (1977), "On Entropy Maximization Principle," *Applications of Statistics* (Krishnaiah, P. R., ed.), North Holland: Amsterdam, pp. 27-41.

Berg, A., Meyer, R., Yu, J., (2004), "Deviance Information Criterion for Comparing Stochastic Volatility Models," *Journal of Business and Economic Statistics*, 22, 107120.

Berger, J. O., and Pericchi, L. R. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison"(with discussion), *Institute of Mathematical Statistics Lecture Notes - Monograph Series* (Lahiri, P., ed.), vol. 38, 135-207.

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738-754.

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

Berk, Kenneth N. (1978), "Comparing subset regression procedures," *Technometrics*, 20, 1-6.

Chipman, H., George, E. I., and McCulloch, R. E. (2001), "The Practical Implementation of Bayesian Model Selection" (with discussion), *Institute of Mathematical Statistics Lecture Notes - Monograph Series*(Lahiri, P., ed.), vol. 38, 65-134.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, No. 2, 407-499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave penalized Likelihood and Its Oracle Property," *Journal of the American Statistical Association*, 96, 1348-1360.

Foster, D. P., and Stine, R. A. (2004), "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy," *Journal of the American Statistical Association*, 99, 303-313.

Gammaitoni, L., Hanggi, P., Jung, P., and Marchesoni, F. (1998), "Stochastic Resonance," *Reviews of Modern Physics*, Vol. 70, No. 1, 223-288.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881-889.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.

Luo, X., Stefanski, L. A., and Boos, D. D. (2004), "Tuning Variable Selection Procedures by Adding Noise," *Technometrics*, (to appear).

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

—— (1995), "More Comments on C_p," *Technometrics*, 37, 362-372.

Miller, A. J. (2002), *Subset Selection in Regression*, London: Chapman & Hall.

Rao, C. R., and Wu, Y. (2001), "On Model Selection (with discussion)," *Institute of Mathematical Statistics Lecture Notes - Monograph Series*(Lahiri, P., ed.), vol. 38, 1-64.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

Shao, J. (1996), "Bootstrap Model Selection," *Journal of the American Statistical Association*, 91, 655-665.

Spiegelhalter, D., Best, N. G., Carlin, B. P., and Linde, A. V. (2002), "Bayesian Measures of Model Complexity and Fit (with Discussion)," *Journal of the Royal Statistical Society, Series B*, 64, 583-639.

Thompson, M. L. (1978), "Selection of Variables in Multiple Regression," *International Statistical Review*, 46, 1-19 and 129-146.

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

Tibshirani, R., and Knight, K. (1999), "The Covariance Inflation Criterion for Adaptive Model Selection," *Journal of the Royal Statistical Society, Series B*, 61, 529-546.

Wu, Y. (2004), "Controlling Variable Selection By the Addition of Pseudo-Variables," *Ph.D. Thesis*, Statistics Department, North Carolina State University, Raleigh, NC.