Biometrics 65, 692–700
September 2009
DOI: 10.1111/j.1541-0420.2008.01127.x

Fast FSR Variable Selection with Applications to Clinical Trials

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

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

[2]Department of Biostatistics and Programming, Sanofis-Aventis Inc., Bridgewater, New Jersey 08807, U.S.A.

[*]*email:* boos@stat.ncsu.edu

Summary. A new version of the false selection rate variable selection method of Wu, Boos, and Stefanski (2007, *Journal of the American Statistical Association* **102,** 235–243) is developed that requires no simulation. This version allows the tuning parameter in forward selection to be estimated simply by hand calculation from a summary table of output even for situations where the number of explanatory variables is larger than the sample size. Because of the computational simplicity, the method can be used in permutation tests and inside bagging loops for improved prediction. Illustration is provided in clinical trials for linear regression, logistic regression, and Cox proportional hazards regression.

Key words: Bagging; False discovery rate; False selection rate; Forward selection; LASSO; Model error; Model selection; Regression.

1. Introduction

Variable selection methods are not used widely in the primary analyses of clinical trials. However, data collected in these trials are often used for secondary studies where variable selection plays a role. An example is the PURSUIT cardiovascular study that has generated many papers beyond the primary paper, Harrington et al. (1998).

Tsiatis et al. (2008) proposed model selection in a "principled" approach to covariance adjustment that allows variable selection to be used separately and independently in each arm of a clinical trial. That is the primary motivation for our use of variable selection in two other examples, one repeating their example in linear regression, and a second example using Cox regression.

The method we study is a variant of the false selection rate (FSR) method of Wu, Boos, and Stefanski (2007), henceforth WBS. In that paper, phony explanatory variables are generated, and the rate at which they enter a variable selection procedure is monitored as a function of a tuning parameter like α-to-enter of forward selection. This rate function is then used to estimate the appropriate tuning parameter so that the average rate that uninformative variables enter selected models is controlled to be γ_0, usually 0.05. The variant we develop requires no phony variable generation, but achieves the same net result. This allows us to estimate the tuning parameter from a summary table of the forward selection variable sequence and associated p-values. We show that if these p-values are monotone increasing, $p_1 \le p_2 \le \cdots \le p_{k_T}$, where k_T is the number of predictors, then the implied stopping rule is a version of an adaptive false discovery rate (FDR) method (Benjamini and Hochberg, 1995): choose the model of size k, where

$$k = \max\left\{ i : p_i \le \frac{\gamma_0(1+i)}{(k_T - i)} \text{ and } p_i \le \alpha_{\max} \right\},$$

and $\alpha_{\max}$ is defined later.

The scope of application is large, but we focus on linear, logistic, and Cox regression. Johnson (2008) illustrates that FSR can be used with other censored data regression techniques such as Buckley–James (Buckley and James, 1979). Because of the savings in computations, the method developed here also allows us to consider using the approach in permutation tests, and for improved prediction via bagging (Breiman, 1996b).

We first review the FSR method in Section 2 and then present the new "Fast" version in Section 3 and its connection to FDR. Section 4 illustrates the method in three clinical trial examples. Section 5 presents simulation results, and Section 6 illustrates using Fast FSR with bagging to improve predictions. Section 7 is a short conclusion.

2. The FSR Method in Regression Variable Selection

The variable selection method introduced by WBS is based on adding phony (noise) variables to the design matrix $\boldsymbol{X}$ and monitoring when they enter a forward selection sequence. The data consist of an $n \times 1$ vector response $\boldsymbol{Y}$ and an $n \times k_T$ matrix of explanatory variables $\boldsymbol{X}$. The variables are linked by a regression model where each column of $\boldsymbol{X}$ is associated with a parameter β_j. If $\beta_j \neq 0$, we say that X_j is "informative"; if

$\beta_j = 0$, we say that X_j is "uninformative." Our basic quantity of interest is the FSR given by

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

where $U(\boldsymbol{Y}, \boldsymbol{X})$ and $I(\boldsymbol{Y}, \boldsymbol{X})$ are the numbers of uninformative and informative variables in the selected model. The expectation is with respect to repeated sampling of the true model ($\boldsymbol{X}$ may be fixed or random). We include the 1 in the denominator because most models include intercepts and also because it avoids problems with dividing by 0. Our strategy is to specify a target FSR, $\gamma = \gamma_0$, and to adjust the selection method tuning parameters so that γ_0 is the achieved FSR rate. Typically, $\gamma_0 = 0.05$, although other values might be desired in specific problems.

Now consider a model selection method that depends on a tuning parameter α such that model size is monotone increasing in α. For forward selection, α is called the significance level for entry or α-to-enter, so that a new variable enters sequentially as long as its p-value to be included in the model (hereafter "p-to-enter") is $\leq \alpha$ (and is smaller than all other p-to-enter). For a given data set, let $U(\alpha) = U(\boldsymbol{Y}, \boldsymbol{X})$ when using the tuning parameter α, that is, $U(\alpha)$ is the number of uninformative variables in the model. Let $S(\alpha)$ denote the total number of variables (excluding the intercept) included in the model.

If we knew $U(\alpha)$, a simple estimator of the FSR in equation (1) as a function of α would be the empirical estimator $U(\alpha)/\{1 + S(\alpha)\}$. Then, setting this latter quantity equal to γ_0 and solving approximately for α would yield an estimated α as follows, $\hat{\alpha} = \sup_\alpha\{\alpha : U(\alpha)/[1 + S(\alpha)] \leq \gamma_0\}$. This $\hat{\alpha}$ is the largest α such that the FSR for the data $(\boldsymbol{Y}, \boldsymbol{X})$ is not more than γ_0. Thus $\gamma_0(\hat{\alpha}) \approx U(\hat{\alpha})/\{1 + S(\hat{\alpha})\}$. Next, we seek to mimic this approach to get a true estimator that satisfies this approximate equality on average.

Because $U(\alpha)$ is unknown, we estimate it with $\{k_T - S(\alpha)\}\hat{\theta}(\alpha)$, where $\{k_T - S(\alpha)\}$ estimates the total number k_U of uninformative variables in the data, and $\hat{\theta}(\alpha)$ is an estimate of the rate that uninformative variables enter the model using tuning parameter α. We define the target rate function as $\theta(\alpha) = E\{U(\alpha)/k_U\}$. The key idea in WBS for estimating $\theta(\alpha)$ is to generate k_P phony variables, append them to the original set of explanatory variables, and find the number of these that enter the model when using tuning parameter α, say $U_P(\alpha)$. This process is repeated B times, and $\hat{\theta}(\alpha) = \bar{U}_P(\alpha)/k_P$, where $\bar{U}_P(\alpha)$ is the average of $U_P(\alpha)$ over the B replications. This is a bootstrap type step, but note that only B sets of phony variables are generated; the original data $(\boldsymbol{Y}, \boldsymbol{X})$ are used in each replication.

Note also that $k_T - S(\alpha)$ overestimates the true number of uninformative variables k_U when α is small and underestimates it when α is large. But in the vicinity of an appropriate α, it is a reasonable estimate for relatively sparse models. Evidence of this claim is provided by the simulations of Section 5. In sparse models where k_I is small compared to k_U, $k_T - S(\hat{\alpha})$ is on average close to k_U, but in less sparse models it overestimates k_U. This is seen in Figure 4. For the smaller model sizes (M0, M5, M10), average model size $S(\hat{\alpha})$ is very close to k_I (0, 5, 10). However, for the larger models (M20, M40), average model size $S(\hat{\alpha})$ is less than k_I (20, 40).

Putting these pieces together yields

$$\hat{\alpha} = \sup_{\alpha \leq \alpha_{\max}}\{\alpha : \hat{\gamma}(\alpha) \leq \gamma_0\}, \quad \text{where}$$

$$\hat{\gamma}(\alpha) = \frac{\{k_T - S(\alpha)\}\hat{\theta}(\alpha)}{1 + S(\alpha)}. \tag{2}$$

Note that $\hat{\gamma}(\alpha) = 0$ at $\alpha = 1$ because $k_T - S(1) = 0$ when all the variables are in the model. Typically, $\hat{\gamma}(\alpha)$ descends to 0 for α in the range $[\alpha_{\max}, 1]$ for some $\alpha_{\max}$ because $\{k_T - S(\alpha)\}\hat{\theta}(\alpha)$ decreases with α for large α. So we do not consider α values beyond $\alpha_{\max} = 0.3$, suggested by extensive simulation results.

WBS studied a number of methods for generating phony variables but recommended permuting the rows of the original $\boldsymbol{X}$ leading to $k_P = k_T$, the number of phony variables equal to the number of original variables. A problem with this approach when k_T is large is that one then has to deal with selecting from $2k_T$ variables each of B times. The computational burden can be heavy, and some selection procedures may have trouble handling even one run with $2k_T$ predictors.

3. The Fast FSR Method in Forward Selection

3.1 *The Forward Selection Method*

In forward selection, the p-to-enter at step i is the smallest p-to-enter of all the variables currently not in the model. Although forward selection(α) is a variable selection method, we consider the full sequence of k_T p-values of the entered variables when $\alpha = 1$. Label these $p_1, \ldots, p_{k_T}$ and call the sequence of predictors in the order they enter the model the *forward addition sequence*.

When the p-to-enter are monotone increasing, $p_1 \leq p_2 \leq \cdots \leq p_{k_T}$, forward selection($\alpha$) chooses a model of size k, where $k = \max\{i : p_i \leq \alpha\}$. Additionally, for unique P_i, forward selection results in k_T nested models of sizes $S(0) = 0, S(p_1) = 1, S(p_2) = 2, \ldots, S(p_{k_T}) = k_T$. We need only consider α values equal to the p_i because those are where $S(\alpha)$ increases.

However, when the p-to-enter are not monotone increasing, we monotonize the original sequence by carrying the largest p forward and use the notation $\tilde{p}_1 \leq \tilde{p}_2 \leq \cdots \leq \tilde{p}_{k_T}$ for these monotonized p-values. Now forward selection(α) chooses a model of size k, where $k = \max\{i : \tilde{p}_i \leq \alpha\}$. Note that in this case at least two of these monotonized p-values must be equal and not all nested models of the forward addition sequence are chosen by forward selection(α) as Table 1 illustrates. The "Mono $\tilde{p}$" column has the monotonized p-values, and there are two equal values of 0.1168 at steps 8 and 9 where the original p-to-enter are not monotone. Thus, model 8 is not possible for forward selection(α).

3.2 *Fast FSR*

The key quantity in the FSR tuning method is the estimate $\hat{\theta}(\alpha) = \bar{U}_P(\alpha)/k_P$ in the numerator of $\hat{\gamma}(\alpha)$ of equation (2). The proposed new Fast FSR method simply uses $\theta(\alpha) = \alpha$ instead of estimating it by simulation. Then equation (2) becomes

$$\hat{\alpha}_F = \sup_{\alpha \leq \alpha_{\max}}\{\alpha : \hat{\gamma}_F(\alpha) \leq \gamma_0\} \quad \text{and} \quad \hat{\gamma}_F(\alpha) = \frac{\{k_T - S(\alpha)\}\alpha}{1 + S(\alpha)}, \tag{3}$$

Table 1
Summary from Cox regression in group 0 *of ACTG* 175 *data,* $k_T = 83$ *predictors*

Step	Variable	p-to-enter	Mono $\tilde{p}$	S	$\frac{0.05\{1+S(\tilde{p}_i)\}}{\{k_T - S(\tilde{p}_i)\}}$	$\hat{\gamma}_F(\tilde{p})$
1	CD40	$9e-08$	$9e-08$	1	0.0012	0.0000
2	CD80	$1e-05$	$1e-05$	2	0.0019	0.0003
3	Age^2	0.0083	0.0083	3	0.0025	0.1660
4	Anti	0.0093	0.0093	4	0.0032	0.1469
5	CD40^2	0.0517	0.0517	5	0.0038	0.6721
6	Wtkg ∗ Gender	0.0594	0.0594	6	0.0045	0.6534
7	CD40 ∗ HS	0.0715	0.0715	7	0.0053	0.6792
8	Drugs ∗ Race	0.1168	0.1168			0.8643
9	CD40 ∗ Drugs	0.0647	0.1168	9	0.0068	0.8643

Mono $\tilde{p} = p$-values to enter, monotonized to be increasing with step. S is the size of the model chosen by forward selection at $\alpha = \tilde{p}$. The model of size 8 is not possible with forward selection. Variables are defined in Example 2.

where here we define $\alpha_{\max}$ to be the α value where $\hat{\gamma}_F(\alpha)$ attains its maximum value. These definitions lead to the simple rule for model size selection,

$$k(\gamma_0) = \max\left\{ i : \tilde{p}_i \le \frac{\gamma_0[1+S(\tilde{p}_i)]}{k_T - S(\tilde{p}_i)} \text{ and } \tilde{p}_i \le \alpha_{\max} \right\}, \quad (4)$$

where note that the value of $\hat{\alpha}_F$ is not required. Table 1 illustrates with the results of the Cox regression discussed later in Example 3. Looking at the bounding column $0.05\{1 + S(\tilde{p}_i)\}/\{k_T - S(\tilde{p}_i)\}$, we see that $\tilde{p}_2 = 0.00001$ is less than the bound 0.0019, but $\tilde{p}$ values after that are not less than their bounds. So $\gamma_0 = 0.05$ leads to a model of size $k(0.05) = 2$.

Once $k(\gamma_0)$ is found from (4), then the linearity of $\hat{\gamma}_F(\alpha)$ between jumps gives $\hat{\alpha}_F = \gamma_0\{k_T - k(\gamma_0)\}/\{1 + k(\gamma_0)\}$. In other words, the estimated α is just the bound in (4) associated with the model size chosen. Figure 1 illustrates: $\gamma_0 = 0.05$ chooses $\hat{\alpha}_F = 0.002$, and $\gamma_0 = 0.20$ chooses $\hat{\alpha}_F = 0.013$. Note that $k_T = 83$ and $\alpha_{\max} = 0.12$. Although Table 1 does not have the bound for $\gamma_0 = 0.20$, we could find that the chosen model for $\gamma_0 = 0.20$ is of size $k(0.20) = 4$ because the last model with $\hat{\gamma}_F(\tilde{p})$ less than 0.20 is of size 4. So the last column of Table 1 can be used to choose a model for arbitrary γ_0.

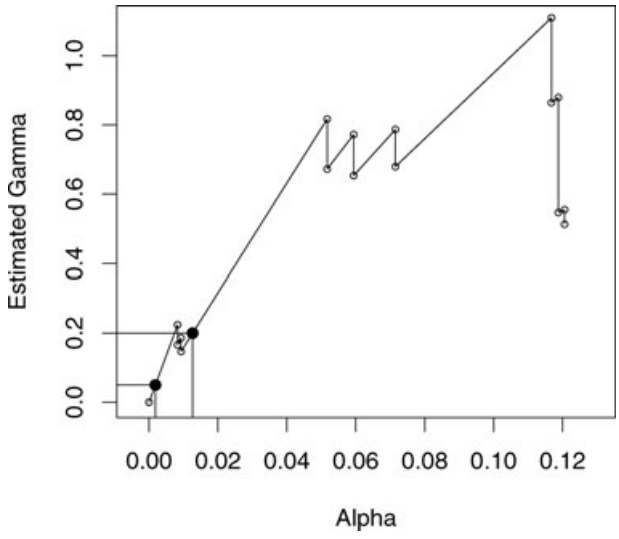

Figure 1. $\hat{\gamma}_F(\alpha)$ for the Table 1 data. Solid dots are where $\hat{\gamma}_F(\alpha)$ intersects $\gamma_0 = 0.05$, yielding $\hat{\alpha}_F = 0.002$, and $\gamma_0 = 0.20$, yielding $\hat{\alpha}_F = 0.013$. $\alpha_{\max} = 0.12$. Jumps in $\hat{\gamma}_F(\alpha)$ occur at $\tilde{p}$ where $\hat{\gamma}_F(\tilde{p})$ are the lower values.

3.3 *Relationship of Fast FSR to FDR*

The FDR method was introduced into the multiple comparisons literature by Benjamini and Hochberg (1995) for the case of k_T hypothesis tests. FDR is an improvement over familywise error-rate methods that tend to be fairly conservative. Let the ordered p-values be denoted $p_{(1)} \le p_{(2)} \le \cdots \le p_{(k_T)}$ and the associated null hypotheses $H^0_{(1)}, \ldots, H^0_{(k_T)}$. The FDR rule is to reject $H^0_{(1)}, \ldots H^0_{(k)}$, where

$$k = \max\left\{ i : p_{(i)} \le \frac{i}{k_T}\gamma_0 \right\}. \quad (5)$$

Using our notation, suppose that there are k_U true null hypotheses and $k_I = k_T - k_U$ false hypotheses, and let U be the number of true null hypotheses that are rejected out of a total of S hypotheses that are rejected using equation (5). Benjamini and Hochberg (1995) proved that if the tests are independent then $\text{FDR} = \text{E}\{U/S \mid R > 0\}P(R > 0) \le \gamma_0 k_U/k_T$. That is, equation (5) controls the FDR to be less than or equal to γ_0 because $k_U/k_T \le 1$. Benjamini and Yekutieli (2001) extended the method to certain types of dependencies in the test statistics and noted that the bound applies to any type of dependency if γ_0 in equation (5) is replaced by $\gamma_0/(\sum_{i=1}^{k_T} i^{-1})$. Benjamini and Hochberg (2000) and Benjamini, Krieger, and Yekutieli (2006) give adaptive versions of equation (5) whereby an estimate $\hat{k}_U$ of k_U is used, replacing γ_0 by $\gamma_0 k_T/\hat{k}_U$ and $\gamma_0 k_T/\{\hat{k}_U(1+\gamma_0)\}$, respectively. Thus, the first of these adaptive procedures uses the bound

$$p_{(i)} \le \frac{i}{k_T}\gamma_0\left(\frac{k_T}{\hat{k}_U}\right) = \frac{i\gamma_0}{\hat{k}_U}, \quad (6)$$

whereas the Fast FSR procedure uses

$$\tilde{p}_i \le \frac{\{1+S(\tilde{p}_i)\}\gamma_0}{k_T - S(\tilde{p}_i)}. \quad (7)$$

Clearly, then, the Fast FSR procedure is a type of adaptive FDR applied to the monotonized forward selection p-values $\tilde{p}_1 \le \tilde{p}_2 \le \cdots \le \tilde{p}_{k_T}$. When the p-values to enter are already strictly monotone, then $S(\tilde{p}_i) = i$, and there are two differences between equations (6) and (7): (i) the addition of 1 to $S(\tilde{p}_i)$ in the numerator of equation (7) arises from the definition of FSR with $1 + S$ in the denominator instead of S, (ii) the use of $k_T - S(\tilde{p}_i)$ in place of a fixed estimate $\hat{k}_U$ for the number of uninformatives. We originally considered using a fixed estimate and then iterating but found that it typically led to the same chosen model as equation (7).

The only use of FDR in regression that we are aware of is due to Bunea, Wegkamp, and Auguste (2006). They proposed using the conservative bound $i\gamma_0/(k_T \sum_{i=1}^{k_T} i^{-1})$ with full model p-values and show that under certain regularity conditions, the method results in consistent variable selection. However, we tried this approach using the less conservative bound $i\gamma_0/k_T$ with our simulations, and found it not competitive unless the predictors are uncorrelated.

3.4 *Justification of* $\theta(\alpha) = \alpha$

First consider Fast FSR in the case of normal linear regression, known error variance σ^2, and orthogonal design matrix $\boldsymbol{X}$. If the jth variable is uninformative, then its p-value is uniformly distributed on (0, 1). Also, all the p-values are independent. Thus the number of uninformative variables included in the model when using forward selection with tuning parameter α is binomial(k_U, α). In this case, all conditions for using FDR are met, and we could appeal to FDR theorems to justify the Fast FSR procedure because of the similarity of Fast FSR to adaptive FDR. In general, though, the forward selection p-values to enter do not have the properties required for formal FDR theorems. For the orthogonal case, estimating σ^2 in the usual way at each step of forward selection alters the uniform distribution and independence slightly, but the expected number of uninformative variables in the model should still be close to $k_U\,\alpha$.

Simulations for the case of $\boldsymbol{X}$ generated from independent normal variables show that $\theta(\alpha)$ is very close to α, but that $\hat{\gamma}_F$ from equation (3) is larger than the true $\gamma(\alpha) = E[U(\alpha)/\{1 + S(\alpha)\}]$ because $k_T - S(\alpha)$ is not a perfect estimator of k_U. To illustrate a situation that is far away from the orthogonal $\boldsymbol{X}$ case, the left panel of Figure 2 plots an "oracle" estimate of $\theta(\alpha)$ (solid irregular line) given by the average of $U(\alpha)/k_U$ from 1000 Monte Carlo replications of a model where the matrices $\boldsymbol{X}$ were generated with autocorrelated $N(0, 1)$ variables ($\rho = 0.7$). The situation is similar to model H3 in WBS: $n = 150, k_T = 21$ total variables, $k_I = 10$ variables with nonzero coefficients at variables 5–9 and 12–16 with values (25, 16, 9, 4, 1), respectively, and then multiplied by a constant to have R^2 near 0.35. The solid straight line in the left panel of Figure 2 is the simple $\theta(\alpha) = \alpha$ of Fast FSR. The dashed line is the average of $\hat{\theta}$ from using the phony variable method of WBS. The reason the latter is less noisy than the solid oracle line is because each $\hat{\theta}$ is based on 500 bootstrap averages. We have plotted for $\alpha \in (0, 0.05)$, but the estimated α chosen by the FSR method in this situation with $\gamma_0 = 0.05$ is ≈ 0.01. Thus, most of the "action" occurs on the extreme left side of both plots. On that portion of the left graph, the average of $\hat{\theta}(\alpha)$ from regular FSR with phony variables is quite close to $\theta(\alpha) = \alpha$ of Fast FSR.

As a second illustration, we consider a case with $n = 200, k_T = 80, k_I = 20$ nonzero βs all equal, $R^2 = 0.5$, and the $\boldsymbol{X}$ matrices generated from an autoregressive (AR) process with $\rho = 0.6$.In addition, we randomly permute the columns of $\boldsymbol{X}$ for each Monte Carlo data set. The right panel of Figure 2 gives plots analogous to the left panel of Figure 2. Here we see that the regular phony-generated FSR method and Fast FSR are very close to one another, justifying the replacement of regular FSR by Fast FSR. However, the true $\theta(\alpha)$ curve is a bit far from the average of the estimates, leading to inflated FSR rates (see later simulation results). On the other hand, the estimate of the true γ curve (not shown) rises steeply, and a good estimator of it would lead to very small models chosen and poor prediction. The poorer correspondence between $\theta(\alpha)$ and $\theta(\alpha) = \alpha$ in the right panel is due to high correlation between the informative and noninformative variables.

4. Examples

Example 1. *Logistic Regression in the PURSUIT Study.* The **P**latelet Glycoprotein IIb/IIIa in **U**nstable Angina: **R**eceptor **S**uppression **U**sing **I**ntegrilin **T**herapy (PURSUIT) study was a multicountry, multicenter, double-blind, randomized,

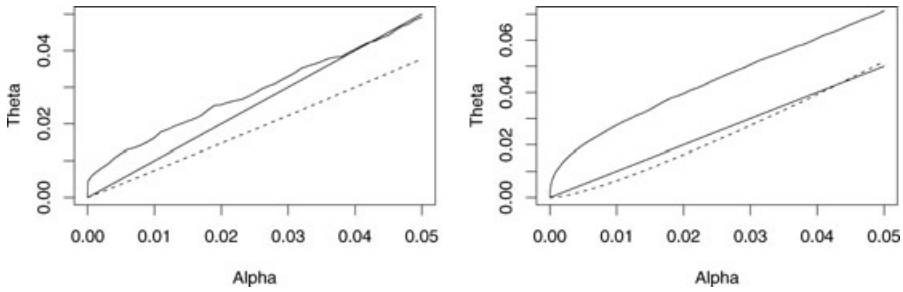

Figure 2. $\theta(\alpha)$ curves. Left panel: the Fast FSR proposal $\theta(\alpha) = \alpha$ (solid straight line) and averages of $U(\alpha)/k_U$ (solid line, S.E.s ≤ 0.002) that approximate $\theta(\alpha)$ and averages of $\hat{\theta}$ (dashed line, S.E.s ≤ 0.0002) from 1000 Monte Carlo replications of the H3 model: $n = 150, R^2 = 0.35, k_T = 21, k_I = 10, k_U = 11$. $\boldsymbol{X}$ matrices were generated from an AR(1) standard normal process with $\rho = 0.7$. The right panel is from 1000 replications of the M20 model: $n = 200, R^2 = 0.50, k_T = 80, k_I = 20, k_U = 60$. $\boldsymbol{X}$ matrices were generated from an AR(1) standard normal process with $\rho = 0.6$.

placebo-controlled study comparing a regimen of integrilin or placebo added to aspirin and heparin in 10,948 patients (Harrington et al., 1998). We consider a subset of patients from North America and Western Europe. The primary endpoint of death or heart attack in the first 30 days was significantly reduced in patients randomized to integrilin therapy, but further analysis revealed that the effect was strongest in males and in patients treated with percutaneous coronary intervention (PCI). We investigate whether other baseline characteristics are associated with the primary endpoint.

We used forward selection in logistic regression after forcing in five variables: treatment, gender, and PCI indicators, and interactions of the treatment indicator with the gender and PCI indicators. In a first run of forward selection on 34 variables and their interactions with gender ($k_T = 68$ variables not counting the included variables), we selected the top 14 main effects using a generous $\alpha = 0.15$ in the forward selection. To these 14 we added their interactions with gender and the squares after centering of the five continuous main effects in the top 14. Thus, in the second run we selected on 14 + 14 + 5 = 33 variables. However, we use $k_T = 68 + 12 = 80$ in the Fast FSR analysis because that would be the number of variables used for selection if we had added all 12 continuous quadratic terms. The main reason to do variable selection in two stages is that the sample size of 4888 complete cases increases to 5360 when we drop 20 of the original variables.

Fast FSR chose 12 additional terms of potential interest, including two quadratic terms and one interaction with gender. For the interactions and quadratic terms, we enforced the hierarchy principle that interactions can enter only after the associated main effects have entered, but only minor differences are introduced by not enforcing it. Having chosen $S = 12$ terms, the estimated α-to-enter is just the associated bound, $\hat{\alpha}_F = 0.0096$.

Example 2. *Covariate Adjustment with Linear Regression.* Tsiatis et al. (2008) introduce a new approach to covariance adjustment in randomized studies that allows one to separately model mean responses in the different treatment groups using model selection methods. For two treatment groups, labeled 0 and 1, the adjusted estimate of the mean difference is given by

$$\frac{n_1}{n}\bar{Y}^{(1)} + \frac{n_0}{n}\hat{Y}_0^{(1)} - \left\{\frac{n_0}{n}\bar{Y}^{(0)} + \frac{n_1}{n}\hat{Y}_1^{(0)}\right\}, \tag{8}$$

where n_0 and n_1 are the sample sizes in groups 0 and 1, $n = n_0 + n_1$, $\bar{Y}^{(1)}$ and $\bar{Y}^{(0)}$ are the sample means. Furthermore, $\hat{Y}_0^{(1)}$ is the predicted treatment 1 mean based on the explanatory variables from group 0 using a model developed from the group 1 data, and $\hat{Y}_1^{(0)}$ is similarly the predicted treatment 0 mean from the group 1 data. Thus, the estimate is an intuitive difference of weighted means. Each weighted mean is essentially an estimate of the appropriate estimate if each group could receive both treatments. In their approach, Tsiatis et al. (2008) suggest that the modeling in each group be done separately, ideally by independent statisticians, each using a model selection method of their choosing. An alternative approach is to have a very carefully specified protocol for doing the model selection in each group, thus avoiding the possibility of biasing the estimated treatment mean difference by choice of models.

We use the example from Tsiatis et al. (2008) to illustrate Fast FSR in this context. The data are from AIDS Clinical Trials Group Protocol 175 (ACTG 175; Hammer et al., 1996) with $n_0 = 532$ in the zidovudine monotherapy group (group 0), and three other treatments groups combined to yield $n_1 = 1607$ subjects in group 1. Following Tsiatis et al. (2008), we look for a mean difference in CD4 counts at approximately week 20 (±5 weeks) using 12 possible baseline covariates: CD4 counts (CD40), CD8 counts (CD80), age in years (Age), weight in kilograms (Wtkg), Karnofsky scores (Karn, 0–100), hemophilia indicator (Hemo, 1 = yes), homosexual indicator (HS, 1 = yes), history of intravenous drug use indicator (Drugs, 1 = yes), race (Race, 0 = white, 1 = nonwhite), gender (Gender, 1 = male), history of antiretroviral use (Anti, 0 = naive, 1 = experienced), and symptomatic status (Symp, 1 = present). The first five are continuous and the rest are binary. Tsiatis et al. (2008) used forward selection with fixed entrance level $\alpha = 0.05$ along with the 12 explanatory variables (forward-1 in their notation) and then with the full quadratic model (forward-2) having 83 possible explanatory variables (12 + 5 quadratic + 66 interactions). The forward-1 models had 4 and 7 variables selected, respectively, for groups 0 and 1, and forward-2 had 4 and 10 variables selected.

For the 12-variable linear case, Fast FSR chose three variables in group 0 ($\hat{\alpha} = 0.022$) and seven in group 1 ($\hat{\alpha} = 0.080$). For the quadratic model we centered the five continuous variables in both groups by subtracting overall means before creating squared and interaction terms. For the 83 variables of the full quadratic case, Fast FSR chose four variables in group 0 ($\hat{\alpha} = 0.003$) and eight in group 1 ($\hat{\alpha} = 0.006$). The estimates from equation (8) and associated tests were all similar for these different modeling approaches with highly significant approximately normal test statistics ≈ 10.

For a more inferentially challenging example, we randomly sampled $n_0' = 100$ from the 532 group 0 cases and $n_1' = 100$ from the 1607 group 1 cases. Then we repeated this sampling for $n_0'' = 200$ and $n_1'' = 200$. Because there might be concern that the standard errors given in Tsiatis et al. (2008) for equation (8) (see their equations 18 and 19) could be affected by model selection, we used approximate permutation p-values based on 100,000 random permutations. Each permutation employed all data manipulations such as centering and model selection within each group used to calculate the test statistic. The normal approximations are quite good even for sample sizes of 100: the two-sided permutation p-values are 0.010 and 0.013 for linear and quadratic covariate adjustments, respectively, compared to 0.011 and 0.014 for the normal approximations.

Fast FSR facilitated computation of the permutation p-values in reasonable time. Fast FSR, however, did not select many variables; 2 and 1 for the linear case ($k_T = 12$); 1 and 1 for the quadratic case ($k_T = 83$) for $n_0' = n_1' = 100$; and 3 and 1 (linear), and 3 and 1 (quadratic) for $n_0'' = n_1'' = 200$. Therefore, because little selection actually occurred, it is not surprising that the standard error and normal approximation work well here. There is also a hint of a practical suggestion here: for small samples, it may be wise to just select from the linear terms, leaving quadratic models for larger sample sizes.

Example 3. *Variable Selection in Cox Regression.* Lu and Tsiatis (2008) used the ACTG 175 data from the previous example to illustrate covariate adjustment with a composite survival endpoint defined as the first time a subject's CD4 count decreased to 50% of baseline or they developed an AIDS-defining event or they died. In the original paper, Hammer et al. (1996), this endpoint was used to show the value of combined therapies compared to the use of zidovudine alone (group 0). Here we work with only the 532 cases in group 0 and use the full quadratic set ($k_T = 83$) of variables to select a proportional hazards model. `PROC PHREG` in `SAS` was used to get the forward sequence of variables and p-values based on score statistics displayed in Table 1, used previously in Section 3.2 to illustrate Fast FSR. As mentioned previously, for $\gamma_0 = 0.05$, we choose a two variable model, with terms CD40 and CD80. The associated estimated α is $\hat{\alpha}_F = 0.05(1+2)/(83-2) = 0.00185$.

5. Simulation Results

Here we report on two sets of simulations for linear regression. We also have done simulations with logistic regression similar to those found in Wu (2004), but we do not report on these here except to mention that Fast FSR gives results similar to the original FSR method.

5.1 *Linear Regression with Fixed* $\boldsymbol{X}$

Two 150 × 21 design matrices were generated from $N(0,1)$ distributions, independent in the first case ($\rho = 0$) and autocorrelated, AR(1) with $\rho = 0.7$, in the second case. In addition, we added squares of the 21 original variables to make a $k_T = 42$ case and then added all pairwise interactions to make a full quadratic case with $k_T = 252$. In all cases, the added variables have true coefficients equal to 0. We ran simulations for $k_T = 21, k_T = 42, k_T = 100$, and $k_T = 252$, but report below only on the $k_T = 42$ and $k_T = 252$ (not displayed) because results for the other two cases can be anticipated from the ones reported.

The five mean models considered are the same as in WBS: H0 = all βs = 0; H1 = 2 equal nonzero βs for variables 7 and 14; H2 = 6 nonzero βs at variables 6–8 and 13–15 with values (9,4,1); H3 = 10 nonzero βs at variables 5–9 and 12–16 with values (25,16,9,4,1), respectively; and H4 = 14 nonzero βs at variables 4–10 and 11–17 with values (49, 36, 25, 16, 9, 4, 1), respectively. These nonzero βs were multiplied by a constant to make the theoretical $R^2 = 0.35$, where theoretical $R^2 = \boldsymbol{\mu}^T\boldsymbol{\mu}/\{\boldsymbol{\mu}^T\boldsymbol{\mu} + n\sigma^2\}, \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, and σ^2 is the error variance.

For simplicity we do not enforce the hierarchy (requiring linear terms to come in the model before related quadratic terms). In fact, here the addition of the quadratic and interaction variables is just a simple way to increase the number of explanatory variables.

Figure 3 gives average model errors and FSR rates for the Fast FSR and minimum Bayesian information criterion (BIC) methods based on the forward addition sequence, and for the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) using 5-fold cross-validation averaged 10 times.The model error for one data set is $n^{-1}\sum_{i=1}^{n}\{\hat{f}(\boldsymbol{x}_i) - \mu_i\}^2$, where $\hat{f}(\boldsymbol{x}_i)$ is the prediction for the ith case and μ_i is the true mean for that case. The FSR for one data set is $U(\boldsymbol{Y}, \boldsymbol{X})/\{1 + I(\boldsymbol{Y}, \boldsymbol{X}) + U(\boldsymbol{Y}, \boldsymbol{X})\}$, where $I(\boldsymbol{Y}, \boldsymbol{X})$ and $U(\boldsymbol{Y}, \boldsymbol{X})$ are the numbers of informative and uninformative variables in the selected model. These measures are averaged over the 100 generated data sets. The Fast FSR and minimum BIC methods use the `leaps` package in `R`, and the LASSO was computed with the `lars` package in `R`. Figure 3 displays model error ratios obtained by dividing the minimum model error of all models in the forward addition sequence by the method model error. Thus large values indicate good performance. The minimum BIC method is based on choosing the model from the forward addition sequence that has the lowest BIC value.

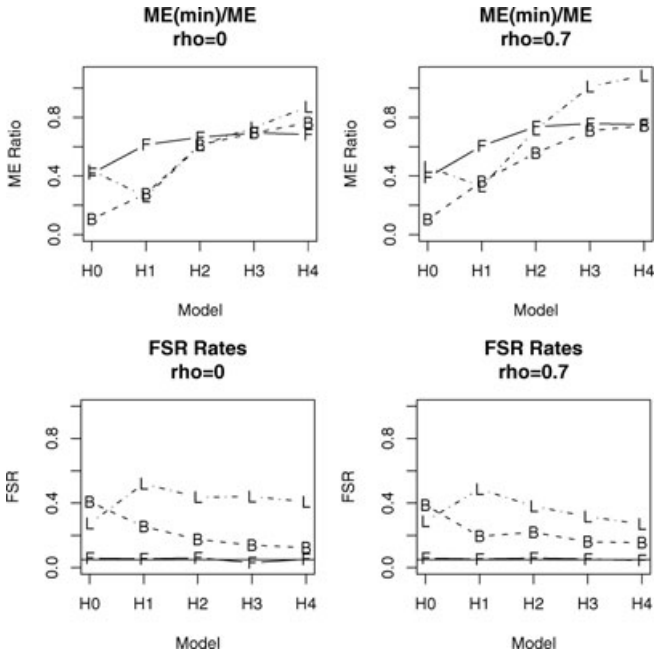

Figure 3. Model errors divided into the minimum model error possible for forward selection and FSR rates for Fast FSR (F), BIC (B), and LASSO (L). Based on 100 Monte Carlo replications for a situation with $n = 150, R^2 = 0.35$, and $k_T = 42$. Standard errors for the model error ratios are bounded by 0.06 except for model H0 and range from 0.01 to 0.04 for the FSR rates.

In terms of model error, BIC does fine for $k_T = 42$ in models H2–H4 (Figure 3), but not well for models with $k_T = 252$ (not displayed). In fact we had to limit the models searched to the first 60 of the forward addition sequence to even get these BIC results. The BIC curve keeps decreasing when k_T is too large. The LASSO does well in terms of model error and even beats the best possible model error of the forward addition sequence in H4 for $k_T = 42$.

Basically, the LASSO is better than Fast FSR for all the H4 cases and one of the H3 cases. However, the LASSO has much higher FSR rates because it admits many variables. In the $k_T = 252$ case, the average number of terms chosen by the LASSO for H4 was 32 ($\rho = 0$) and 28 ($\rho = 0.7$). In comparison, Fast FSR had average model sizes of 2.9 and 2.3, seemingly too small because the true model for H4 has 14 nonzero coefficients. The minimum model error "oracle" method had average model sizes of 5.1 and 2.9 for H4 at $k_T = 252$. So

Fast FSR has low model size apparently to achieve the FSR rate near 0.05, but in that H4 case it is not too far from the average optimal size.

In general, these simulations suggest that Fast FSR has reasonable model error and close to the advertised 0.05 FSR rate. BIC fails miserably for $k_T = 252$. The LASSO has good model error for larger models, but at the cost of a high FSR.

5.2 *Linear Regression with Random X*

The previous section used models that were relatively sparse, especially when adding uninformative interaction terms. Thus, we wanted to provide a more challenging situation for forward selection and our Fast FSR method.

The data were generated from a mean zero normal AR(1) process, this time with $\rho = 0.6$ and $k_T = 80$ variables. However, we transformed 20 of the variables by taking their absolute value and dichotomized another 20 variables as $I(X_{ij} > 0)$. Finally, the 80 variables were randomly permuted and rescaled to have sample mean equal to 0 and sample variance equal to 1 before creating the responses. This process was repeated for each of the 100 design matrices $\boldsymbol{X}$ in the simulation. Thus, this is a case with random design matrices, in contrast to the previous section. The models used were M0 = no informative variables, M5 = 5 equal nonzero βs, M10 = 10 equal nonzero βs, M20 = 20 equal nonzero βs, and M40 = 40 equal nonzero βs. The βs were multiplied by a constant in each case to have theoretical $R^2 = 0.5$.

The top left panel of Figure 4 gives model error ratios defined in Section 5.1. We see that similar to previous results,

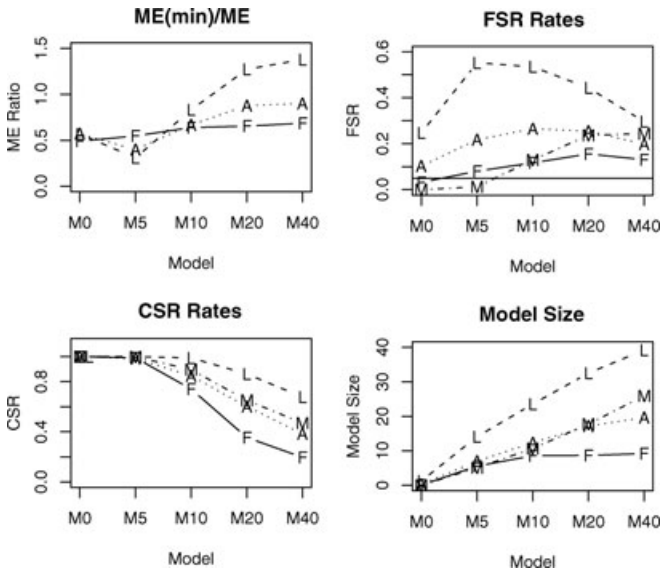

Figure 4. Model errors divided into the minimum model error possible for forward selection, FSR rates, CSR rates, and model sizes for Fast FSR (F), LASSO (L), LSA (A), and minimum model error (M). Based on 100 Monte Carlo replications for a situation with $n = 200, R^2 = 0.5, \rho = 0.6$, and $k_T = 80$. Standard errors for model error ratios are $\leq$0.03 except for model M0; for FSR rates are $\leq$0.04; for CSR rates are $\leq$0.02; for model size are $\leq$0.89.

Fast FSR performs well for sparse models with few informative predictors, but not as well for the larger models. Here, that tendency is accentuated because there are even larger models, and all the βs have the same value. This is not a defect of Fast FSR but rather of forward selection in this type situation—note that the LASSO is much better at M20 and M40 than the oracle minimum model error for the forward addition sequence. We also added another method, the least squares approximation (LSA) for LASSO estimation of Wang and Leng (2007) using a BIC stopping rule (see their equation 3.5). Marked with an "A" in Figure 4, the LSA performs intermediate to the LASSO and Fast FSR, perhaps not a surprising result because its BIC stopping rule tends to choose smaller models than regular LASSO. We also give average FSR rates, Correct Selection Rate (CSR) equal to the proportion of informative predictors selected, and model size.CSR is complementary to FSR—together with average model size they give a fuller picture of a model selection procedure's characteristics. Here the LASSO chooses large model sizes and therefore has high FSR rates and high CSR rates. In contrast, Fast FSR chose much smaller models and has low FSR rates and low CSR rates. The LSA is in between those two.

The FSR rate of Fast FSR exceeds the target 0.05 rate in models M10–M40 (cf, Figure 3). The problem lies in the higher correlation between informative predictors and noninformative predictors and the larger models. Although we have not displayed the phony variable version of FSR, it performs similar to Fast FSR. We repeated some simulations for the situation of Figure 3 using randomly permuted X matrices, and we see elevated FSR rates. Also, we repeated simulations like those in Figure 4 but without the random permuting of columns, and the FSR rates were as advertised (near 0.05). In all cases, the informative variables are in the first columns so that without permuting columns, the informative variables are not very correlated with the uninformative ones. Thus we believe the problem is with the increased correlation between informative predictors and noninformative predictors induced by permuting the columns. Both Fast FSR and regular phony-generated FSR are based on independence between these two sets of predictors.

We could develop an improved FSR method to handle this correlation problem, but it is not clear that we want to. In Figure 4 we have also given results for the oracle minimum model error method, marked "M" on the graph. Note that its FSR rates and model sizes are considerably higher than Fast FSR. Thus, to achieve FSR rates around 0.05 as advertised, the model error and CSR performance of such an improved FSR method would be much worse. So, the quandary is that for large models with high correlation among the explanatory variables, any model selection procedure based on forward selection cannot have both low FSR rates and good model error. In the next section, we show that bagging Fast FSR can recover good model error performance in these situations.

6. Bagging

Breiman (1996a,b) showed that some model selection procedures are unstable in the sense that perturbing the data can result in selection of very different models. One of his proposals for improving model selection stability is bagging, essentially averaging selected models over bootstrap data sets.

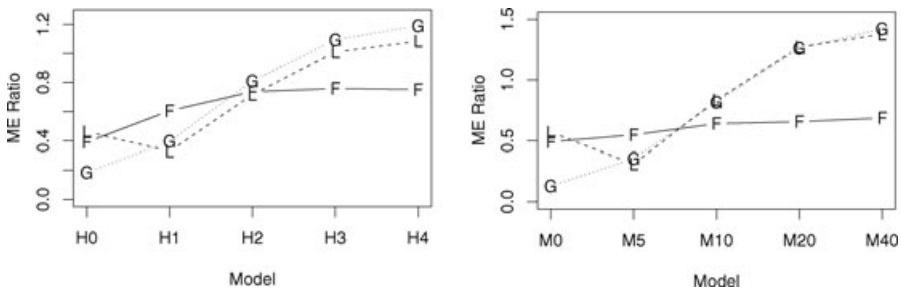

Figure 5. Model errors divided into the minimum model error possible for forward selection, Fast FSR (F), LASSO (L), and Bagged Fast FSR (G). Based on 100 Monte Carlo replications for the situation in the right side of Figure 3 with $n = 150, R^2 = 0.35, \rho = 0.7$, and $k_T = 42$ (left panel) and for the situation of Figure 4 with $n = 200, R^2 = 0.5, \rho = 0.6$, and $k_T = 80$ (right panel). Standard errors for the model error ratios are bounded by 0.03 except for model M0.

The simplicity of Fast FSR enables its use in bagging as follows: Randomly draw with replacement from the pairs $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$; run Fast FSR on the bootstrap sample and obtain $\hat{\boldsymbol{\beta}}^*$; repeat B times. For linear regression, average the bootstrap $\hat{\boldsymbol{\beta}}^*$, say $\bar{\boldsymbol{\beta}}^*$, and predict from $\bar{\boldsymbol{\beta}}^*$.

Note that with this model averaging $\bar{\boldsymbol{\beta}}^*$ typically has no zeroes even though each $\hat{\boldsymbol{\beta}}^*$ has many zeroes—so there is no variable selection in the averaged model. In our simulations, bagging Fast FSR had a large improvement over Fast FSR in terms of model error for the less-sparse sampling situations. The left panel of Figure 5 repeats simulations from the right panel of Figure 3, and the right panel repeats the upper left panel of Figure 4. In the left panel of Figure 5, bagging Fast FSR is better in terms of model error than Fast FSR for models H2–H4 and better than the LASSO in models H1–H4. Similar results are found in the right panel of Figure 5 except that bagging Fast FSR and the LASSO are nearly identical for models M5–M40. Bagging Fast FSR is better than the oracle minimum model error for the forward addition sequence in 4 of the 10 models displayed in Figure 5 (where the model error ratios are >1).

A feature of the bagged Fast FSR method is that the $\hat{\alpha}$ chosen by Fast FSR on the bootstrap data sets is larger on average than those chosen by Fast FSR on the parent data sets. Our explanation for this phenomenon is that in the bootstrap world created by randomly resampling pairs, the "true" model is effectively the full least squares solution, and thus there are no nonzero βs in that world. Thus, it makes sense for Fast FSR to try and pick larger models on the bootstrap data sets. This appears somewhat fortuitous because bagging likely does better for small overfitting compared to small underfitting.

Clearly, bagging seems to be useful for prediction (as measured by model error) in the larger models used in Figure 5. Can we anticipate when it will be useful to use bagging of Fast FSR instead of merely Fast FSR? Yuan and Yang (2005) suggested a measure of instability given by the derivative of $I(c)$ at $c = 0$ where

$$I(c) = \frac{1}{M\hat{\sigma}} \sum_{j=1}^{M} \left[\frac{1}{n} \sum_{i=1}^{n} \{\hat{f}_j(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i)\}^2 \right]^{1/2},$$

$\hat{f}(\boldsymbol{x}_i)$ is the prediction from a selected model, $\hat{\sigma}$ is an estimate of the error standard deviation from the selected model, and $\hat{f}_j(\boldsymbol{x}_i)$ is the prediction from the selected model from the jth of M bootstrap data sets of the form $(\tilde{Y}_1, \boldsymbol{x}_1), \ldots, (\tilde{Y}_n, \boldsymbol{x}_n)$, where $\tilde{Y}_i = Y_i + W_i$ and $W_1, \ldots, W_n$ are i.i.d. from a $N(0, c^2\hat{\sigma}^2)$ distribution. The derivative at $c = 0$ of $I(c)$ is called the "perturbation instability in estimation" (PIE) and is estimated by regression through the origin using a grid of c values and the corresponding $I(c)$ values. In our implementation, we used $M = 100$ and $c = (0.1, 0.3, 0.5, 0.7, 0.9)$.Yuan and Yang (2005) recommend that some type of model averaging approach be used when the PIE values are higher than 0.4 to 0.5.

For the simulation data sets used to make the right panel of Figure 5, the average values of PIE with standard deviations in parentheses were PIE(M0) = 0.09 (0.06), PIE(M5) = 0.35 (0.10), PIE(M10) = 0.66 (0.09), PIE(M20) = 0.78 (0.11), PIE(M40) = 0.84 (0.10). Thus, from these averages and standard deviations, the Yuan and Yang (2005) rule of thumb suggests bagging most of the time when sampling from models M10, M20, and M40, precisely the models in Figure 5 where bagging improves over Fast FSR in terms of model error. So the good news is that situations where bagging is needed to obtain good predictions can be identified.

Of course, bagging is at odds with variable selection. However, it is not clear that good prediction and low FSRs are compatible in certain situations, at least when using forward selection. Note in Figure 4 that the FSRs of the minimum model error forward sequence model are high for (M10, M20, M40).

7. Conclusion

Fast FSR in forward selection provides an intuitive choice of model size (and associated $\hat{\alpha}_F$) that is a type of adaptive FDR applied to the monotonized forward selection p-values. It can be used with essentially any regression method; we have used it with linear, logistic, and Cox regression.

Forward selection's prediction performance can degrade when there are a large number of highly correlated predictors with a large number of informative predictors relative to sample size. Although we never know the true number of informative predictors, if such a situation is suspected, then we suggest bagging Fast FSR for predictions. Computing the

PIE measure aids in making the choice between regular Fast FSR and bagged Fast FSR.

Acknowledgements

We thank Duke Clinical Research Institute for the PURSUIT data and Marie Davidian and Butch Tsiatis for providing the ACTG 175 data and preprints of papers. FSR programs in `SAS` and `R` are available at `http://www4.stat.ncsu.edu/~boos/var.select`. This work was supported by NSF grant DMS-0504283.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25,** 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29,** 1165–1188.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93,** 491–507.

Breiman, L. (1996a). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24,** 2350–2383.

Breiman, L. (1996b). Bagging predictors. *Machine Learning* **24,** 123–140.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Bunea, F., Wegkamp, M. H., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* **136,** 4349–4364.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C., for the Aids Clinical Trials Group Study 175 Study Team. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine* **333,** 1081–1089.

Harrington, R. A. for the PURSUIT Trial Investigators. (1998). Inhibition of platelet glycoprotein with eptifibatide in patients with acute coronary syndromes. *The New England Journal of Medicine* **339,** 436–443.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistic Society, Series B.* **70,** 351–370.

Lu, X. and Tsiatis, A. A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* **95,** 679–694.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistic Society, Series B* **58,** 267–288.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **29,** 4658–4677.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *Journal of the American Statistical Association* **102,** 1039–1048.

Wu, Y. (2004). Controlling variable selection by the addition of pseudovariables. Unpublished doctoral thesis. Statistics Department, North Carolina State University.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102,** 235–243.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* **100,** 1202–1214.

Received October 2007. *Revised May* 2008.
Accepted June 2008.