

# FSR Methods for Second-Order Regression Models

Hugh B. Crews

Center for Marine Science

University of North Carolina Wilmington

and

Dennis D. Boos, Leonard A. Stefanski

Department of Statistics

North Carolina State University

March 22, 2009

## Summary

Most variable selection techniques focus on first-order linear regression models. Often, interaction and quadratic terms are also of interest, but the number of candidate predictors grows very fast with the number of original predictors. Thus, we develop forward selection algorithms that enforce natural hierarchies in second-order models to control the entry rate of uninformative effects. Also, a general method of controlling false selection rates for two groups of predictors is proposed that results in equal contributions to the false selection rates from first-order and second-order terms. Method performance is compared through Monte Carlo simulation, and an illustration is provided using a response surface experiment.

*Key words and phrases:* Bagging; Model selection; Response optimization; Variable selection.

# 1 Introduction

Variable selection techniques are used in a variety of settings with most attention focused on selecting a subset of the measured variables. In some applications such as response surface optimization, selecting interaction and quadratic terms is important. In such applications, second-order terms can increase the model’s predictive accuracy and reveal patterns that would be missed when only considering the measured variables (see, for example, Hamada and Wu, 1992). However, the number of possible second-order terms grows exponentially with the number of predictors, making the problem of selecting the best subset difficult.

Several methods have been proposed to limit the number of terms under consideration. One natural approach is to restrict to models that are invariant to changes in measurement scale (Peixoto, 1990). In stepwise search this leads to including main effects before interactions, called *strong heredity* or *strong hierarchy*. *Weak heredity* or *weak hierarchy* refers to requiring only one main effect to be in the model before allowing in an associated interaction. Chipman (1996) and Chipman, Hamada, and Wu (1997) use Bayesian methods to enforce these hierarchies in constructing second-order models. Yuan, Joseph, and Lin (2007) proposed LARS algorithms that enforce a hierarchy for analyzing experimental designs.

In the context of first-order regression models, Wu, Boos, and Stefanski (2007) developed a general simulation-based method for estimating the tuning parameter of variable selection techniques to control the False Selection Rate (FSR) of variables. More recently, Boos, Stefanski, and Wu (2009), henceforth BSW, proposed a “Fast” FSR approach that requires no simulation to estimate  $\alpha$ -to-enter to use with forward selection. In this paper, we apply Fast FSR methodology to forward selection algorithms that enforce either the strong or weak hierarchies in second-order regression models. We also propose a new approach to forward selection by using different  $\alpha$ -to-enter values for first-order and second-order terms. Then, by estimating the entry levels appropriately, we attain approximately equal contributions to the FSR from both first and second-order effects.

The remainder of the paper is organized as follows. Section 2 reviews the basic Fast FSR approach to variable selection and generalizes the method to handle hierarchy-based algorithms. Section 3 presents a new forward selection algorithm that uses separate entry levels for first-order and second-order terms and approximately equalizes their contributions to the FSR. Section 5 compares the methods via Monte Carlo simulation, and Sections 4 and 6 illustrate the methods

with examples. Section 7 gives a short conclusion.

## 2 Fast FSR Methods

### 2.1 Fast FSR

When using a variable selection procedure on a data set with an  $n \times 1$  response vector,  $\mathbf{Y}$ , and an  $n \times k_T$  matrix of explanatory variables,  $\mathbf{X}$ , the False Selection Rate (FSR) is defined as

$$\gamma = E \left\{ \frac{U(\mathbf{Y}, \mathbf{X})}{1 + I(\mathbf{Y}, \mathbf{X}) + U(\mathbf{Y}, \mathbf{X})} \right\}, \quad (1)$$

where  $I(\mathbf{Y}, \mathbf{X})$  and  $U(\mathbf{Y}, \mathbf{X})$  are the number of informative and uninformative variables in the selected model. Informative variables are defined as those whose regression coefficients are nonzero. The goal of FSR variable selection is to tune the variable selection procedure such that the FSR is some desired level,  $\gamma_0$ . Typically,  $\gamma_0 = 0.05$ , but other choices may be appropriate.

When using forward selection with  $\alpha$ -to-enter value  $\alpha$ , let  $U(\alpha) = U(\mathbf{Y}, \mathbf{X})$  be the number of uninformative variables selected and let  $S(\alpha)$  be the total number of variables selected. If  $U(\alpha)$  were known, then a simple estimator for the FSR would be  $U(\alpha)/\{1 + S(\alpha)\}$ . Although  $U(\alpha)$  is unknown, it can be estimated by  $\hat{N}(\alpha)\theta(\alpha)$ , where  $\theta(\alpha)$  is the rate that uninformative variables enter the model, and  $\hat{N}(\alpha) = k_T - S(\alpha)$  is an estimate of the total number of uninformative variables available for selection. In order to estimate  $\theta(\alpha)$ , Wu, Boos, and Stefanski (2007) generated phony explanatory variables and monitored their rate of entry over a grid of  $\alpha$  values.

In the Fast FSR approach, BSW use  $\theta(\alpha) = \alpha$ , and therefore no phony variable simulation is required. This leads to the Fast FSR estimate,

$$\hat{\gamma}_F(\alpha) = \frac{\hat{N}(\alpha)\alpha}{1 + S(\alpha)} = \frac{\{k_T - S(\alpha)\}\alpha}{1 + S(\alpha)}. \quad (2)$$

The goal is to use the largest  $\alpha$  such that  $\hat{\gamma}_F(\alpha)$  is no greater than  $\gamma_0$ . However, because  $S(\alpha) \rightarrow k_T$  as  $\alpha \rightarrow 1$ ,  $\hat{\gamma}_F(\alpha)$  typically underestimates the FSR over the range  $[\alpha_{max}, 1]$ , where  $\alpha_{max}$  is the  $\alpha$  value such that  $\hat{\gamma}_F(\alpha)$  is at its maximum. Therefore,  $\alpha$  is estimated using

$$\hat{\alpha} = \sup_{\alpha \leq \alpha_{max}} \{\alpha : \hat{\gamma}_F(\alpha) \leq \gamma_0\}. \quad (3)$$

If the observed  $p$ -to-enter values are monotone increasing,  $p_1 \leq p_2 \leq \dots \leq p_{k_T}$ , then using forward selection with  $\hat{\alpha}$  chooses a model of size  $k$ , where  $k = \max\{i : p_i \leq \hat{\alpha} \text{ and } p_i \leq \alpha_{max}\}$ .

Therefore, one may restrict attention to these observed  $p$ -to-enter values as possible values for  $\hat{\alpha}$  because  $S(\alpha)$  only increases at these values. However, if the  $p$ -to-enter values are not monotone increasing, then they must be monotonized in order to use the function  $forward-selection(\alpha)$  that defines nested models by inclusion if a  $p$ -value is less than or equal to  $\alpha$ . Given a sequence of observed  $p$ -to-enter values,  $\{p_1, \dots, p_{k_T}\}$ , the monotonized sequence of  $p$ -to-enter values is  $\{\tilde{p}_1, \dots, \tilde{p}_{k_T}\}$ , where  $\tilde{p}_i = \max\{p_1, \dots, p_i\}$ . Using these definitions leads to the Fast FSR rule for model size,

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

Using  $k(\gamma_0)$  from (4), the solution to (3) is

$$\hat{\alpha} = \min \left\{ \frac{\gamma_0\{1 + k(\gamma_0)\}}{k_T - k(\gamma_0)}, \alpha_{max} \right\}.$$

BSW show that (4) can be viewed as a type of adaptive false discovery rate (FDR) method applied to the monotonized  $p$ -to-enter values.

We call the sequence of models corresponding to the original, possibly nonmonotone  $p$ -values, the *forward addition sequence*. It has  $k_T$  steps and model sizes  $S(i) = i$ ,  $i = 1, \dots, k_T$ . However,  $forward-selection(\alpha)$  denotes the variable selection method as a function of the  $\alpha$ -to-enter value  $\alpha$  that has model size  $S(\alpha)$  changing only at the monotonized  $p$ -values. Thus,  $S$  has a dual notation for model size, one for the steps of the forward addition sequence and one as a function of  $\alpha$  in  $forward-selection(\alpha)$ . In general,  $S(i) = S(\tilde{p}_i)$  only when observed  $p$ -to-enter values are strictly increasing.

## 2.2 Fast FSR for Second-Order Models

The observed data are  $n$  pairs  $(Y_1, \mathbf{d}_1), \dots, (Y_n, \mathbf{d}_n)$ , where  $\mathbf{d}_i$  is a  $p \times 1$  vector of design constants. We refer to these predictor variables as main effects. When estimating response surfaces, it is typical to also use the squares and products of the predictor variables. In other situations, it often makes sense to check for interactions and nonlinearities. Correlation among the predictor variables, however, generally makes variable selection more difficult. Thus, before adding quadratic terms, we first center the main effects to reduce correlation between second-order effects and parent main effects. For example, the sample correlation of  $X$  and  $X^2$  when  $X$  consists of the integers 1 to 10 is .97, whereas the sample correlation of  $X - 5.5$  and  $(X - 5.5)^2$  is 0. One may also rescale the variables although this has no effect on our forward selection approach. Then we relabel all  $k_T = 2p + \binom{p}{2}$  variables  $\mathbf{x}^T = (x_1, \dots, x_{k_T})$  so that the first  $p$  of these are the centered and rescaled

main effects, the second  $p$  are the squares of the first  $p$ , and the remaining  $\binom{p}{2}$  are the main effect cross products. If some of the variables are binary, then the number of squared terms is less than  $p$ . The full  $n \times k_T$  design matrix with rows  $\mathbf{x}_i^T$ ,  $i = 1, \dots, n$ , is  $\mathbf{X}$ .

A simple approach for selecting a second-order model is to ignore the hierarchy between main effects and higher-order terms and treat each effect as a separate variable. If we run forward selection with this *No Hierarchy* approach, then each effect is a candidate for entry at the beginning of the forward selection process. Therefore, Fast FSR with No Hierarchy works exactly as described in Section 2.1 with  $k_T = 2p + \binom{p}{2}$ .

A standard approach with some philosophical appeal is to enforce a hierarchy throughout variable selection. When running forward selection with *Strong Hierarchy* (or strong heredity), an interaction cannot enter the model until both of its parent main effects are in the model. Similarly, a quadratic term cannot enter the model until its parent main effect is in the model. A less restrictive alternative, called *Weak Hierarchy* (or weak heredity), allows an interaction to enter the model provided at least one of its parent main effects is in the model. Thus, we consider three hierarchy principles to use in building second-order models via forward selection: No Hierarchy, Strong Hierarchy, and Weak Hierarchy.

Adjustment of the Fast FSR formulas under hierarchy restrictions takes some care. In (2)  $\widehat{N}(\alpha) = k_T - S(\alpha)$  estimates the total number of uninformative variables available to enter. For the Strong and Weak Hierarchy approaches, the number of candidate variables depends on  $\alpha$  as well as on which variables are already in the model. Thus  $\widehat{N}(\alpha) = k_T - S(\alpha)$  is no longer an appropriate estimate of the number of uninformative variables available for selection. In these cases  $\widehat{N}(\alpha)$  is defined as follows.

1. If  $\alpha = \widetilde{p}_i$  for  $i$  such that a single variable enters,  $\widehat{N}(\alpha)$  equals one less than the number of variables available to enter at  $\alpha = \widetilde{p}_i - \epsilon$ , for  $\epsilon > 0$  suitably small. The reduction by one is for the variable entering at  $\alpha = \widetilde{p}_i$ .
2. Now consider the case where  $\alpha = \widetilde{p}_i$  and  $i$  is such that  $\widetilde{p}_i$  appears  $k$  times in the monotonized sequence. Then all  $k$  variables enter forward-selection( $\alpha$ ) at  $\alpha = \widetilde{p}_i$ . However, using the forward addition sequence defined at the end of Section 2.1, we define  $\widehat{N}(\alpha)$  to be one less than the the number of candidate predictors available right before the last of the  $k$  steps. The reduction by one is for the last of the  $k$  variables that enter at  $\alpha = \widetilde{p}_i$ .

For example, consider a Weak Hierarchy case with  $p = 4$  main effects and the sequence  $(p_1, p_2, p_3) = (.001, .0005, .0003)$  for entering terms  $(X_2, X_4, X_4^2)$  in the first three steps of forward selection. The corresponding number of available predictors before each step is  $(4, 7, 9)$ . Then, monotonicizing gives  $(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3) = (.001, .001, .001)$  and  $\hat{N}(.001) = 9 - 1 = 8$ . Notice that in terms of monotonicized  $p$ -to-enter values, all three terms come in at  $\alpha = .001$  so that  $S(.001) = 3$ , but we need step notation to keep track of the terms available sequentially.

This definition of  $\hat{N}(\alpha)$  is consistent with  $k_T - S(\alpha)$  when No Hierarchy is used and allows us to replace  $\{k_T - S(\alpha)\}\alpha$  in (2) with  $\hat{N}(\alpha)\alpha$ , leading to the more general Fast FSR formula

$$\hat{\gamma}_F(\alpha) = \frac{\hat{N}(\alpha)\alpha}{1 + S(\alpha)}. \quad (5)$$

The estimated  $\alpha$  remains defined by (3), however, the rule for model size is now

$$k(\gamma_0) = \max \left\{ i : \tilde{p}_i \leq \frac{\gamma_0 \{1 + S(\tilde{p}_i)\}}{\hat{N}(\tilde{p}_i)} \text{ and } \tilde{p}_i \leq \alpha_{max} \right\}, \quad (6)$$

where  $\alpha_{max}$  is again defined as the entry level when  $\hat{\gamma}(\alpha)$  is at its maximum.

### 3 Fast FSR Adjustment Methods for No Hierarchy and Weak Hierarchy Approaches

With  $p = 10$  main effects, there are  $10+45=55$  second-order terms. With  $p = 20$  there are  $20+190=210$  second-order terms. If all the terms are treated equally as in the No Hierarchy approach, the number of noninformative second-order terms entering the model by chance will be much larger than the number of noninformative main effects. Thus we now develop methods that allow the rate that uninformative variables enter the model to be the same for first-order and second-order terms when using the No Hierarchy or Weak Hierarchy approaches. That is, we want the FSR to be  $\gamma_0/2$  for each set of terms. Basically, the solution is a type of Bonferroni adjustment so that the number of noninformative main effects in the model is on average equal to the number of noninformative second-order effects in the model. To motivate the adjustment, we first define a forward addition sequence that allows different  $\alpha$ -to-enter values for first-order and second-order terms. Then we adapt the Fast FSR methods so that the false selection rates for the two groups of explanatory variables are the same.

### 3.1 Forward Selection with Effect-Specific Entry-Levels

Suppose we want to run forward selection using  $\alpha$ -to-enter =  $\alpha_1$  for main effects and  $\alpha$ -to-enter =  $\alpha_2$  for second-order effects, where  $\alpha_2 < \alpha_1$  to limit the number of second-order effects in our model. A simple way to do this is to multiply the  $p$ -to-enter values of the second-order terms by  $c = \alpha_1/\alpha_2$ , thereby creating a set of adjusted  $p$ -to-enter values. Then one enters the term with the smallest adjusted  $p$ -value at each step of forward selection. We formalize this method for No Hierarchy as *Algorithm 1* because the notation is easiest, but the algorithm extends easily to Weak Hierarchy.

#### Algorithm 1: Forward Selection with Adjusted $P$ -Values for No Hierarchy

1. Starting with an intercept term in the model, calculate the  $p$ -to-enter values for adding any single effect from the candidate set of all main effects, interactions, and quadratic terms. Call these  $p$ -to-enter values  $\{p_{1,1}, \dots, p_{1,k_T}\}$ . For Step 1, define

$$\text{adjusted } p\text{-to-enter values} = \begin{cases} p_{1,j} & \text{if } j \in \mathcal{M}, \\ cp_{1,j} & \text{otherwise,} \end{cases} \quad (7)$$

where  $c = \alpha_1/\alpha_2$  and  $\mathcal{M}$  is the set of main effect indices. Select the effect,  $X_{(1)}$ , corresponding to  $p_1 = p_{1,(1)}$ , the smallest adjusted  $p$ -to-enter value for the first step.

2. With  $X_{(1)}$  and an intercept term in the model, calculate the  $p$ -to-enter values for adding any single effect remaining in the candidate set. Next, calculate the adjusted  $p$ -to-enter values using (7) and select the effect,  $X_{(2)}$ , corresponding to  $p_2 = p_{2,(1)}$ , the smallest adjusted  $p$ -to-enter value for the second step.
3. Repeat this process until no adjusted  $p$ -to-enter value is less than or equal to  $\alpha_1$ .

**Example for No Hierarchy.** Suppose there are three predictor variables,  $X_1, X_2, X_3$ , and we run forward selection using  $\alpha_1 = 0.05$  and  $\alpha_2 = 0.01$ . Thus  $c = \alpha_1/\alpha_2 = 5$ . Table 1 contains the  $p$ -to-enter values and adjusted  $p$ -to-enter values for adding any one of the nine total effects into the model for the first three steps of forward selection. Because  $X_1^2$  has the smallest adjusted  $p$ -to-enter value at Step 1, 0.020, and this value is less than 0.05,  $X_1^2$  enters. The second part of Table 1 contains the  $p$ -to-enter values and adjusted  $p$ -to-enter values for adding any one of the eight remaining effects.  $X_1$  has the smallest adjusted  $p$ -to-enter value, 0.005. Because this value is less than 0.05,  $X_1$  enters. The last part of Table 1 contains the  $p$ -to-enter values and adjusted

Table 1: Example of Forward Sequence Using Adjusted  $P$ -to-enter Values

Variable	Step 1		Step 2		Step 3	
	$p$ -to-enter	Adj. $p$	$p$ -to-enter	Adj. $p$	$p$ -to-enter	Adj. $p$
$X_1$	0.031	0.031	0.005	0.005*		
$X_2$	0.064	0.064	0.032	0.032	0.022	0.022*
$X_3$	0.279	0.279	0.393	0.393	0.410	0.410
$X_1^2$	0.004	0.020*				
$X_2^2$	0.732	3.660	0.549	2.745	0.592	2.960
$X_3^2$	0.581	2.905	0.612	3.060	0.391	1.955
$X_1X_2$	0.008	0.040	0.011	0.055	0.010	0.050
$X_1X_3$	0.232	1.160	0.472	2.360	0.317	1.585
$X_2X_3$	0.109	0.545	0.184	0.920	0.166	0.830

\* Indicates the smallest adjusted  $p$ -to-enter at each step.

$p$ -to-enter values for adding any one of the seven remaining effects into the model. Now  $X_2$  has the smallest adjusted  $p$ -to-enter value, 0.022, and enters the model. Assuming that no more terms are added, the final model using  $\alpha_1 = 0.05$  and  $\alpha_2 = 0.01$  contains an intercept,  $X_1$ ,  $X_1^2$ , and  $X_2$ .

### 3.2 Controlling Two False Selection Rates

In this section the goal is to choose  $\alpha_1$  and  $\alpha_2$  so that the contribution to the false selection rate from first-order effects ( $\text{FSR}_m$ ) is equal to the contribution to the false selection rate from second-order effects ( $\text{FSR}_q$ ), in other words,

$$E[\text{FSR}_m] = E[\text{FSR}_q] = \gamma_0/2. \quad (8)$$

In order to get a forward sequence of effects using Algorithm 1, the multiplicative factor  $c$  of (7) is required. Assuming that there are  $N_m$  uninformative main effects and  $N_q$  uninformative second-order effects in the candidate set, Algorithm 1 with fixed  $\alpha_1$  and  $\alpha_2$  leads to  $N_m\alpha_1$  and  $N_q\alpha_2$  as approximately the expected number of falsely selected main effects and second-order effects, respectively. Setting  $N_m\alpha_1 = N_q\alpha_2$  and substituting  $\alpha_1 = c\alpha_2$  (as specified by Algorithm 1) gives  $c = N_q/N_m$ . Therefore,  $c$  should be the ratio of uninformative second-order effects to uninformative main effects. In order to avoid division by zero, we add 1 to the numerator and denominator, leading

to the target value of  $c$  to be used in Algorithm 1 to achieve (8),

$$c = \frac{1 + N_q}{1 + N_m}. \quad (9)$$

We now present a method for estimating  $c$  in the No Hierarchy case and then explain how it is used with Weak Hierarchy.

### 3.3 Fast FSR Sequential Adjustment for No Hierarchy

When estimating  $\hat{\gamma}_F(\alpha)$  for a given  $\alpha$ , we assume all effects in the model are informative and all effects in the candidate set are uninformative. Therefore, with only an intercept in the model there are  $p$  uninformative main effects and  $k_T - p$  uninformative second-order effects. Using these values, the initial estimate for  $c$  is

$$c^{(1)} = \frac{1 + k_T - p}{1 + p}. \quad (10)$$

Now use the first step of Algorithm 1 with  $c^{(1)}$  to obtain a variable  $X_{(1)}$  that has the smallest adjusted  $p$ -to-enter value. To update  $c$  for the next step, assume that  $X_{(1)}$  is an informative variable. Thus, if  $X_{(1)}$  is a main effect, then  $c^{(2)} = (1 + k_T - p)/p$ . Alternatively, if  $X_{(1)}$  is a second-order effect, then  $c^{(2)} = (k_T - p)/(1 + p)$ . After updating  $c$ , we run another step of Algorithm 1 to get  $X_{(2)}$  and the adjusted  $p$ -to-enter value. To describe the  $i$ th step, we need to use the dual step notation mentioned at the end of Section 2.1:  $S_m(i - 1)$  and  $S_q(i - 1)$  are the number of main effect and second-order effects in the model after  $i - 1$  steps;  $\hat{N}_m(i - 1) = p - S_m(i - 1)$  is the estimated number of uninformative after  $i - 1$  steps and similarly  $\hat{N}_q(i - 1) = k_T - p - S_q(i - 1)$  is the estimated number of uninformative second-order effects. Then at Step  $i$  of Algorithm 1,

$$c^{(i)} = \frac{1 + \hat{N}_q(i - 1)}{1 + \hat{N}_m(i - 1)}. \quad (11)$$

After running forward selection as described, we have a sequence of estimates for  $c$ , effects entered, adjusted  $p$ -to-enter values, and monotonized adjusted  $p$ -to-enter values  $\tilde{p}_1, \dots, \tilde{p}_{k_T}$ . To define  $\hat{\gamma}_F(\alpha)$ , we define  $S_m(\alpha) = S_m(i_\alpha)$ ,  $S_q(\alpha) = S_q(i_\alpha)$ , and  $c(\alpha) = c^{(i_\alpha)}$  at  $\alpha = \tilde{p}_i$  and constant elsewhere, where  $i_\alpha$  is the largest  $i$  associated with all  $\tilde{p}_i$  equal to  $\alpha$ . Because this is a No Hierarchy case,  $\hat{N}_m(\alpha) = p - S_m(\alpha)$  and  $\hat{N}_q(\alpha) = k_T - p - S_q(\alpha)$ . Then

$$\hat{\gamma}_F(\alpha) = \frac{\hat{N}_m(\alpha)\alpha + \hat{N}_q(\alpha)\alpha/c(\alpha)}{1 + S_m(\alpha) + S_q(\alpha)}. \quad (12)$$

After calculating  $\hat{\gamma}_F(\alpha)$  for each  $\tilde{p}_i$ , we choose the model of size

$$k(\gamma_0) = \max \left\{ i : \tilde{p}_i \leq \frac{\gamma_0[1 + S_m(\tilde{p}_i) + S_q(\tilde{p}_i)]}{\hat{N}_m(\tilde{p}_i) + \hat{N}_q(\tilde{p}_i)/c(\tilde{p}_i)} \text{ and } \tilde{p}_i \leq \alpha_{max} \right\}, \quad (13)$$

and let  $\hat{\alpha}_1 = \sup_{\alpha \leq \alpha_{max}} \{\alpha : \hat{\gamma}_F(\alpha) \leq \gamma_0\}$ .

### 3.4 Fast FSR Sequential Adjustment for Weak Hierarchy

In the Sequential Adjustment Method,  $c$  changes at each step of the forward selection process, and under Weak Hierarchy, the candidate set is dynamic, often changing by more than one term at each step. Here we explain how to combine these two approaches.

Under the Weak Hierarchy principle, only the  $p$  main effects are initially in the candidate set. With an intercept in the model, we calculate the  $p$ -to-enter values for the main effects and select the variable,  $X_{(1)}$ , with the smallest  $p$ -to-enter value. We then update the candidate set by adding the  $p$  second-order terms that involve  $X_{(1)}$  ( $p - 1$  interactions and  $X_{(1)}^2$ ). For Step 2, we calculate the  $p$ -to-enter values, adjust the  $p$ -values for the second-order terms using  $c^{(2)} = (1 + p)/p$ , and select the term,  $X_{(2)}$ , with the smallest adjusted  $p$ -to-enter value. If  $X_{(2)}$  is a main effect, then we add the  $p$  second-order terms that involve  $X_{(2)}$  to the candidate set. However, if  $X_{(2)}$  is a second-order effect, then no additional variables are added to the candidate set. In general, for Step  $i$  we update  $c$  using (11), where  $\hat{N}_q(i - 1)$  and  $\hat{N}_m(i - 1)$  are the number of second-order and first-order terms in the candidate set, respectively, at the beginning of the  $i^{th}$  step, that is, computed after entering the  $(i - 1)$ th variable. The rest of the method is similar to (12) and (13) of the the previous section except that the definitions of  $\hat{N}_m(\alpha)$  and  $\hat{N}_q(\alpha)$  follow the general prescription from Section 2.2.

## 4 Example with Cox Regression

Here we briefly illustrate the flexibility and simplicity of using the methods described in Sections 2 and 3. This analysis is made very easy due to a set of SAS macros found at <http://www4.stat.ncsu.edu/~boos/var.select>. We use the primary biliary cirrhosis data analyzed in Fleming and Harrington (1991, p. 153-162) and given in their Appendix D. This data set has been analyzed often and is easily available on the web and in the above website with full description of variables. There are 276 complete cases with  $p = 17$  predictors and  $k_T = 165$  total

quadratic terms (five of the predictors are binary). The response variable is survival time and there are 187 (60%) censored cases. We first used forward selection with the Cox proportional hazards model and just the 17 main effects. The first column of Table 2 shows the order in which the main effects entered the model. Fast FSR chose a model of size 9 with  $\hat{\alpha} = .0625$ . Zhang and Lu (2007, Table 5) show that the LASSO identifies exactly the same 9 variables, and the Adaptive LASSO chooses the same variables except for `ascites`. However, we may want to see if interactions can improve the fit. In column 2 of Table 2 we use Fast FSR with the Strong Hierarchy and obtain the same size model but two interactions, `ascites*copper` and `ascites*edema`, replace the main effects `albumin` and `prottime`. Notice that  $\hat{\alpha} = 0.012$  (lower than 0.0625 because more terms had to be considered) and that the BIC is somewhat lower. Next, we ran Fast FSR with No Hierarchy because we wanted to see if there are important interactions that could not enter the model due to weak parent effects. In column 3 of Table 2 we see that four interactions are in this model, three that did not overlap with those chosen by the Strong Hierarchy. Note that  $\hat{\alpha} = 0.0032$  is much smaller in order to control the FSR in the face of 165 possible predictors. To further analyze within the No Hierarchy setup, we used the adjusted  $p$ -values to limit the entry of second-order terms. The fourth column shows that this adjusted forward sequence gives a model not so different from the second column with the Strong Hierarchy. We feel that either model might be of interest for clinicians to evaluate. We also used the Weak Hierarchy approaches, but they resulted in models with different interactions and possibly harder interpretation.

## 5 Simulation Studies

In this section we summarize two simulation studies designed to assess the performance of the Fast FSR methods. In the first, performance criteria are related to prediction and interpretation. In the second, performance criteria are related to response surface optimization. Crews (2008) contains additional details and results for each study.

### 5.1 Simulation for Prediction and Interpretation

We compare the Fast FSR methods with the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) and with Bayesian Additive Regression Trees (BART) (Chipman et al., 2006). For LASSO we used 5-fold crossvalidation to determine a model, whereas for BART we

Table 2: Models Selected for the Primary Biliary Cirrhosis Data

Variable	Main			Adjusted $p$ -val No Hierarchy
	Effects Only	Strong Hierarchy	No Hierarchy	
bili	1	1	2	1
ascites	2	2		
stage	3	3	3	3
copper	4	4	5	5
albumin	5		7	9
prottime	6			
age	7	6		6
sgot	8	8		8
edema	9	7		
gender			8	
ascites*copper		5		7
ascites*edema		9	1	2
gender*edema			4	4
ascites*hepato			6	
spiders*prottime			9	
$\hat{\alpha}$	0.063	0.012	0.0032	.026
BIC	979.1	961.3	952.3	946.0

Entries are the order than terms entered the model.

used the default settings, i.e., BART-default as described in Chipman et al. (2006). Recall that all predictor variables are first centered by subtracting the mean and then are divided by the sample standard deviation. Forward selection with  $p$ -values determined from the usual least squares  $F$ -tests are used in all the Fast FSR methods. Terms above second-order are not considered. The following Fast FSR methods are studied.

**Fast FSR with No Hierarchy (FFSR-NH):** All  $k_T = 2p + \binom{p}{2}$  terms are available at all steps, and model size is chosen by (4).

**Fast FSR with Strong Hierarchy (FFSR-SH):** Interactions  $X_i X_j$  are available only after both  $X_i$  and  $X_j$  are in the model, whereas  $X_i^2$  is available after  $X_i$  is in. Model size is chosen by (6).

**Fast FSR with Weak Hierarchy (FFSR-WH):** Interactions  $X_i X_j$  are available only after  $X_i$  or  $X_j$  are in the model, whereas  $X_i^2$  is available after  $X_i$  is in. Model size is chosen by (6).

**Fast FSR with Sequential Adjustment for No Hierarchy (FFSR-NH<sub>adj</sub>):** Same as FFSR-

NH except that the second-order  $p$ -to-enter values are multiplied by  $c^{(i)}$  of (11) at step  $i$ . Model size is chosen by (13).

**Fast FSR with Sequential Adjustment for Weak Hierarchy (FFSR-WH<sub>adj</sub>):** Same as FFSR-WH except that the second-order  $p$ -to-enter values are multiplied by  $c^{(i)}$  of (11) at step  $i$ . Model size is chosen by (13).

We studied models with  $p = 20$  original predictors, and so there were  $p_q = 230$  total predictors. The original predictors were generated as either  $N(10, 20)$  or  $\chi_{10}^2$  random variables with both correlated and uncorrelated cases and sample sizes  $n = 100$  and  $n = 500$ .  $N = 100$  independent data sets were generated for each situation. Correlated predictors had the following correlation structure:

$$\text{Corr}(X_i, X_j) = \begin{cases} 0.7 - 0.1(|i - j| - 1) & \text{if } 1 \leq |i - j| < 8, \\ 0 & \text{if } 8 \leq |i - j| < 13, \\ 0.7 - 0.1(19 - |i - j|) & \text{if } 13 \leq |i - j| \leq 19. \end{cases}$$

Note that this correlation initially decays linearly, is zero for lags 8 through 12, and then rises linearly again. However, because we randomly permute the columns of each data set, the only important fact is that there are 20 pairs of  $\mathbf{X}$  columns with correlations 0.1 to 0.7, respectively, and 50 pairs of columns with no correlation.

The models are:

1.  $Y = -100 + 25X_1 + 15X_{13} - 20X_{17} + X_1^2 - 3X_1X_9 + \epsilon;$
2.  $Y = -3 + X_1 - X_4 + 2X_9 - 1.2X_{13} + 1.6X_{17} + \epsilon;$
3.  $Y = 50 + 15X_1 - 25X_9 + 1.2X_1^2 - 1.6X_9^2 + 3X_1X_9 + \epsilon.$

For each model,  $\epsilon \sim N(0, \sigma^2)$ , where  $\sigma$  was chosen to achieve *theoretical*  $R^2$  values 0.25 and 0.50, where

$$\text{theoretical } R^2 = \text{Var} \left( \sum_{j=1}^p \beta_j X_j \right) / \text{Var} \left( \sum_{j=1}^p \beta_j X_j + \epsilon \right).$$

The key measure of performance used was average model error (AME),

$$\text{AME} = (nN)^{-1} \sum_{i=1}^N \sum_{j=1}^n (\hat{Y}_{ij} - \mu_{ij})^2.$$

Although this definition of model error corresponds to a fixed design, we use it here because of the random permutation of the design matrices after generation. To maintain similar scales, results are given in terms of the ratio of the AME of the true model to the AME of a particular method. We call this measure the AME Ratio and note that methods with a high AME Ratio are preferred.

Treating the simulation results as repeated measures ANOVA with 7 methods and a  $2^4 \times 3$  factorial treatment structure, we fit a linear model in SAS `proc mixed` with AME Ratio as our response and with the following factors:  $\mathbf{X}$  distribution ( $N(10, 20)$  or  $\chi_{10}^2$ ), predictor correlation (presence or absence), theoretical  $R^2$  (0.25 or 0.50), sample size (100 or 500), and model (1-3). Tables 3 and 4 present part of the ANOVA results.

Table 3 shows that FFSR-SH was the best overall performer. Statistical significance at the experiment .05 level was assessed using Tukey’s range test. Among the FSR methods, No Hierarchy fared the worst, and FFSR-WH and FFSR-NH<sub>adj</sub> were roughly equivalent, with FFSR-WH<sub>adj</sub> slightly better than those two. Clearly,  $p$ -value adjustment made a major improvement in the No Hierarchy method and a minor improvement in the Weak Hierarchy approach. Overall, BART and the LASSO were not competitive except when  $n = 100$  and  $R^2 = 0.25$ . The LASSO generally captured a large proportion of the informative effects, but because it tends to include a large number of effects, it also had large AFSR. Neither the LASSO nor BART used any hierarchy structure, and therefore suffered from overfitting interactions, in addition to lessening interpretability. Yuan, Joseph, and Lin (2007) show how to enforce hierarchy restrictions with LARS, a close relative of the LASSO.

Analysis among only the FSR methods reveals that Method does not interact strongly with the other factors. Among the five factors, model and sample size had large main effects as well as interactions with each other and with sample size.  $R^2$  did not have a strong main effect, but it did have a strong interaction with model. Distribution type for the  $X$  matrix (means 0.39 for  $\chi^2$  and 0.38 for normal) had no significant difference, and correlation within the  $X$  matrix (means 0.40 for no correlation and 0.37 for correlation) had only small effects. Table 4 shows the means for the important effects.

Model 1 was a combination of main effects and second-order terms, whereas Model 2 contained all main effects. Model 3 was a full quadratic in two variables and the toughest model to fit. For Models 1 and 2, as the sample size and  $R^2$  increased, the methods performed better relative to the

Table 3: Comparison of AME Ratio Means Using Tukey Range Test

Method	Mean	Grouping
FFSR-SH	0.44	A
FFSR-WH <sub>adj</sub>	0.40	B
FFSR-WH	0.38	C
FFSR-NH <sub>adj</sub>	0.38	C
FFSR-NH	0.32	D
LASSO	0.21	E
BART	0.18	E

Methods with the same letter are not significantly different.

Standard errors for entries and differences are 0.01 – 0.02.

Table 4: AME Ratio Means for Assessing Factor Interactions

	$R^2 = .25$	$R^2 = .50$	Model 1	Model 2	Model 3
$n = 100$	0.38	0.32	0.24	0.28	0.52
$n = 500$	0.37	0.46	0.36	0.53	0.36
$R^2 = .25$			0.28	0.35	0.49
$R^2 = .50$			0.32	0.47	0.39

Standard errors of entries are at most .02.

true model. However, for Model 3 this was not the case.

The average false selection rate (AFSR) is defined as the average over the Monte Carlo data sets of the number of uninformative effects selected divided by 1 plus the total number of effects selected. AFSR can be partitioned into contributions from main effects,  $\text{AFSR}_m$ , and second-order effects,  $\text{AFSR}_q$ .

Figure 1 illustrates that for uncorrelated predictors, the Fast FSR methods performed as expected choosing models whose AFSR were close to  $\gamma_0 = 0.05$ . For correlated predictors and  $n = 100$ , however, AFSR rates were on average around .10 although they generally improved for  $n = 500$ . The adjusted Fast FSR methods were designed to ensure that  $E(\text{FSR}_m) = E(\text{FSR}_q) \approx E(\text{FSR})/2$ .

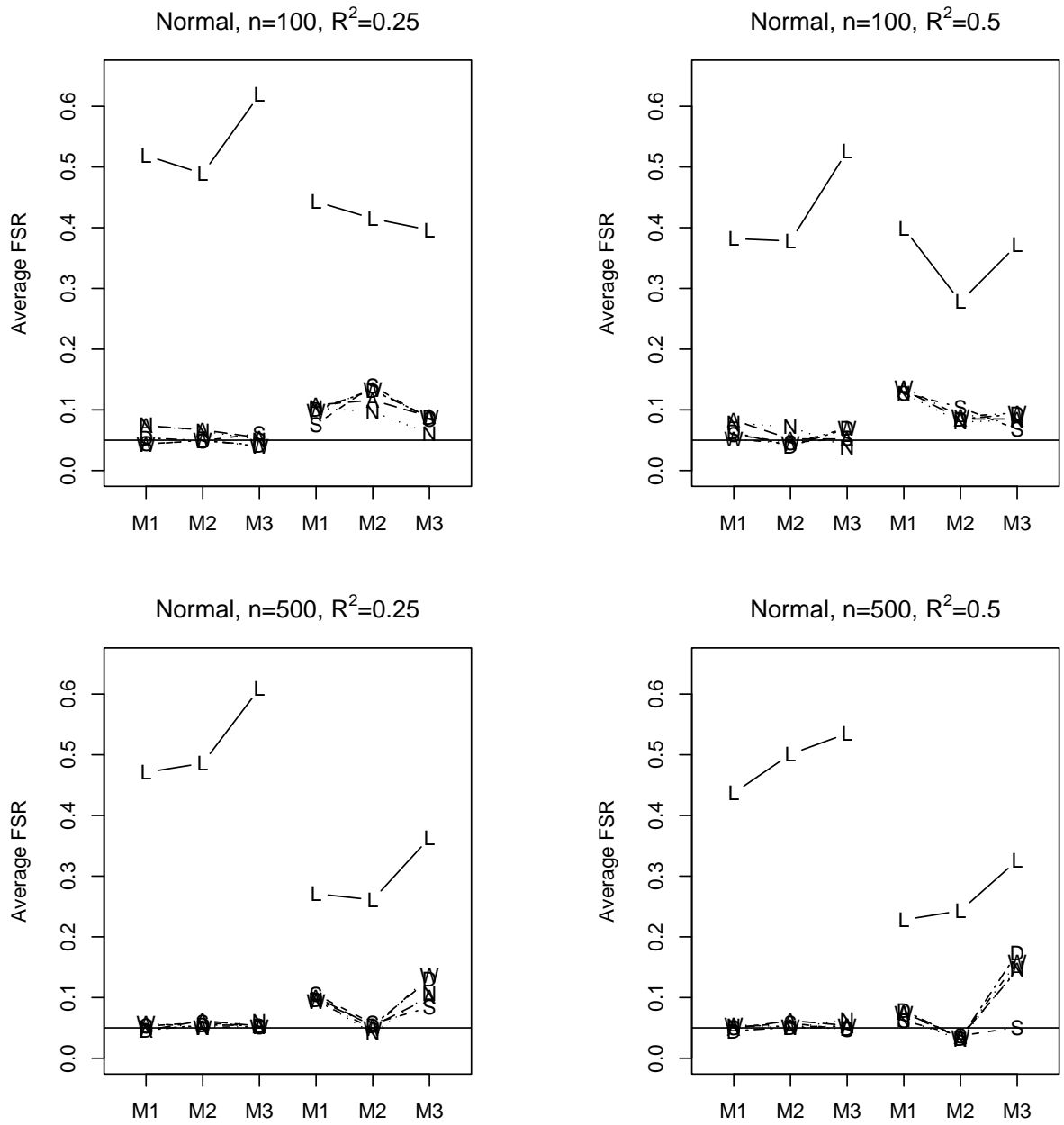


Figure 1: Average false selection rates (AFSR) for each method using normal predictors. First three points are uncorrelated and second three points are correlated predictors. LASSO (L), FFSR-NH (N), FFSR-WH (W), FFSR-SH (S), FFSR-NH<sub>adj</sub> (A), and FFSR-WH<sub>adj</sub> (D). The standard errors of all plotted points are bounded by 0.02.

The simulation showed that if the original  $p \times 1$  vector of predictors were uncorrelated, then  $\text{AFSR}_m$  and  $\text{AFSR}_q$  were approximately equal when using the adjustment methods. However, when the predictors were correlated, the AFSR contributions from the two groups were often unequal. For example, at  $n = 100$ ,  $R^2 = 0.25$ , and Model 2, FFSR-NH<sub>adj</sub> had  $\text{AFSR}_m = 0.104$  and  $\text{AFSR}_q = 0.012$ . For  $n = 500$ , those rates improved to 0.035 and 0.018, respectively.

## 5.2 Simulation for Response Optimization

For the response optimization study, two response surface designs were used to generate data. The first design was a 73-run, small composite, design with  $p = 10$  factors, where main effects are orthogonal but interactions are correlated with main effects and/or other interactions. The second design was a 100-run, orthogonal, central composite, design with  $p = 8$  factors, where all  $k_T = p_q = (2)(8) + 28 = 44$  variables are orthogonal.

For each design, responses were generated from two models: Models 1a and 1b used the small composite design, whereas Models 2a and 2b used the central composite design. The models are defined as follows.

$$1a : Y = 15 - 3X_1 + 1.5X_1^2 + 3X_1X_9 + X_2 - 2X_3 + 1.5X_4 + X_5 - 2X_6 + X_7 + X_8 - X_9 - 5X_9^2 + \epsilon$$

$$1b : Y = 15 - 5X_1 + 1.5X_1^2 + 3X_1X_9 + X_2 - 2X_3 + 1.5X_4 + X_5 - 2X_6 + X_7 + X_8 - 7X_9 - 5X_9^2 + \epsilon$$

$$\begin{aligned} 2a : Y &= 20 + 2X_1 - 4X_1^2 + 5X_1X_2 + 3X_1X_3 - 3X_2 - 3X_2^2 + 1.5X_2X_4 + 2X_3 + 4X_3^2 \\ &- 3X_4 - 2X_4^2 + 2X_4X_5 + 2.5X_5 + 2X_6 + 1.5X_7 + \epsilon \end{aligned}$$

$$\begin{aligned} 2b : Y &= 20 + 5X_1 - 3.5X_1^2 - X_1X_2 + 3X_1X_3 - 3X_2 - 3X_2^2 + 1.5X_2X_4 + 2X_3 + X_3^2 \\ &- 4X_4 - 2X_4^2 + 2X_4X_5 + 3.5X_5 + 2X_6 + 1.5X_7 + \epsilon. \end{aligned}$$

For each model,  $\epsilon \sim N(0, \sigma^2)$ , where  $\sigma$  was chosen to achieve theoretical  $R^2$  values .050, 0.75, or 0.90. As in the first study,  $N = 100$  independent data sets were generated from each model.

Because we are mimicking the situation where screening is conducted prior to the response surface design, we created models with most main effects present. In all the models, only one variable has no effect on the response ( $X_{10}$  in Models 1a and 1b,  $X_8$  in Models 2a and 2b). For

Model 1a, the variable  $X_9$  has a small main effect but a large interaction with  $X_1$  and a large quadratic effect. The purpose of this model is to illustrate the lack of power of the hierarchy-based approaches to select second-order effects when their parent main effects are small. In Model 1b, the main effects of  $X_1$  and  $X_9$  are larger. Therefore, we expect the hierarchy methods to perform better. For Model 2a, the variable  $X_1$  has a small main effect but a large interaction with  $X_3$  and a large quadratic effect. As for Model 1a, the hierarchy-based approaches are at a disadvantage for Model 2a because  $X_1$  must first enter before the large second-order effects have a chance to enter. In Model 2b, the effects of  $X_1$ ,  $X_4$ , and  $X_5$  are larger to give the hierarchy methods an advantage.

The goal of response surface modeling is usually to estimate the levels of a process that yield an optimal response. After fitting by Fast FSR, LASSO, or a standard approach where a full model was fit and terms removed if not significant at the  $\alpha = 0.05$  level, each fitted model was optimized to get a set of optimal factor levels. The optimization was carried out subject to the constraint that each  $X$  lies in  $(-2, 2)$ ; we call this constrained factor space the *region of interest*.

Tables 5 and 6 give the optimal factor levels for the four models as well as the maximum response. For most factors, the optimal level lies on the boundary of the region of interest. In Models 1a and 1b, only variable  $X_9$  has an optimal level in the interior the region. For Models 2a and 2b, variables  $X_1$ ,  $X_2$ , and  $X_4$  all have optimal levels in the interior of the region.

Table 5: Optimal Levels for 10-Factor Small Composite Design

Model	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$\mu(\mathbf{X}_{\text{opt}})$
1a	-2	2	-2	2	2	-2	2	2	-0.7	-	48.45
1b	-2	2	-2	2	2	-2	2	2	-1.3	-	58.45

Table 6: Optimal Levels for Central 8-Factor Composite Design

Model	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$\mu(\mathbf{X}_{\text{opt}})$
2a	1.64	1.03	2	0.64	2	2	2	-	57.35
2b	1.70	-0.86	2	-0.32	2	2	2	-	52.62

To compare the methods we need a measure of how well a method identifies the optimum levels. The true mean optimal response is  $\mu(\mathbf{X}_{\text{opt}})$ , and the true mean response using  $\widehat{\mathbf{X}}_{\text{opt}}$  is  $\mu(\widehat{\mathbf{X}}_{\text{opt}})$ .

For any factor not selected we set the optimal level at the center point, 0. We refer to  $\mu(\widehat{\mathbf{X}}_{\text{opt}})$  as the actual performance, whereas  $\mu(\mathbf{X}_{\text{opt}})$  is the optimal performance. The standardized difference,  $\{\mu(\widehat{\mathbf{X}}_{\text{opt}}) - \mu(\mathbf{X}_{\text{opt}})\}/\mu(\mathbf{X}_{\text{opt}})$ , is a measure of how close a method performs relative to the true optimal performance. When analyzing a real data set, the estimate of the optimal response is  $\widehat{\mu}(\widehat{\mathbf{X}}_{\text{opt}})$ , but that is not used in this simulation.

Figure 2 illustrates the mean performance for each method. Because Fast FSR with Sequential Adjustment for Weak Hierarchy (FFSR-WH<sub>adj</sub>) was always better than Fast FSR with Weak Hierarchy (FFSR-WH), and Fast FSR with Sequential Adjustment for No Hierarchy (FFSR-NH<sub>adj</sub>) was always better than Fast FSR with No Hierarchy (FFSR-NH), FFSR-WH and FFSR-NH were left off the figure. For Model 1a, Fast FSR with Strong Hierarchy (FFSR-SH) and FFSR-WH<sub>adj</sub> performed poorly. The reason is that  $X_9^2$  and  $X_1X_9$  are both large effects, but the main effect  $X_9$  is relatively small, thus making it hard for these second-order terms to enter. The LASSO and FFSR-NH<sub>adj</sub> performed the best in Model 1a, with the LASSO better for  $R^2 = 0.5$  and FFSR-NH<sub>adj</sub> better for  $R^2 = 0.9$ . For Model 1b, the methods performed fairly equally with FFSR-SH, FFSR-NH<sub>adj</sub>, and LASSO among the best. The LASSO does best with smaller  $R^2$ , and FFSR-NH<sub>adj</sub> is better for larger  $R^2$ . For Model 2a, FFSR-SH and FFSR-WH<sub>adj</sub> performed poorly. In this model,  $X_1^2$  is very important, but its main effect is relatively small. Therefore, FFSR-NH<sub>adj</sub> performed best. For Model 2b, FFSR-NH<sub>adj</sub> again performed the best overall.

BSW use bagging (Breiman, 1996) to improve predictions. The basic idea of bagging is to take a random sample with replacement of size  $n$  from the full data set and use this bootstrap sample to obtain  $\widehat{\beta}^*$ . After repeating the process  $B$  times, average the  $\widehat{\beta}^*$  to obtain  $\overline{\beta}^*$ . BSW note that  $\overline{\beta}^*$  typically has no zeros, so there is no variable selection in the averaged model. However, the model can be used for prediction or for determining optimal factor levels. Bagged versions of the Fast FSR methods were also used in the response surface simulations. Generally, bagging improved performance when using FFSR-SH, but the improvement was not as large for FFSR-NH<sub>adj</sub>. In general, when  $R^2 = 0.5$ , the bagged Fast FSR methods were superior to their regular versions in estimating the optimal factor levels. However, as  $R^2$  increased to 0.9, bagging yielded little or no improvement. The only exception was for the hierarchy-based approaches on Models 1a and 2a. In these models, bagging overcomes the problems of fitting large second-order terms with weak parent main effects.

A standard approach is to fit the full response surface and eliminate terms not significant at the

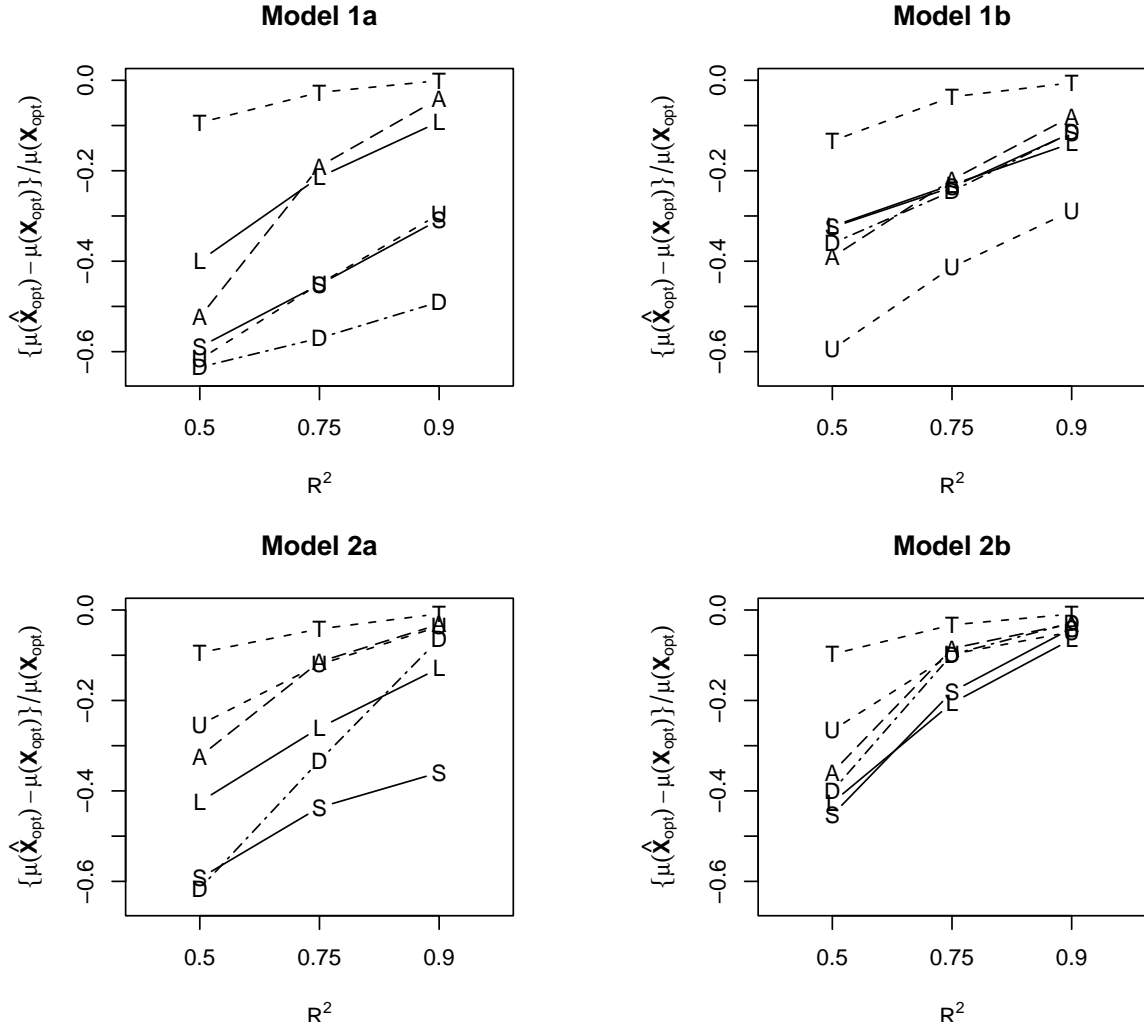


Figure 2: Scaled average difference in actual and optimal performance. Values close to zero are better. True Model (T), Standard Approach (U), LASSO (L), FFSR-SH (S), FFSR-NH<sub>adj</sub> (A), and FFSR-WH<sub>adj</sub> (D). The standard errors of all plotted points are bounded by 0.03.

$\alpha = 0.05$  level. This approach performed poorly for Models 1a and 1b, but performed very well for Models 2a and 2b. Possible reasons for the poor performance in Models 1a and 1b are the sparsity of the true models, the large number of factors, and the correlation between interactions in the design matrix. Even when the standard approach performed well, it still had large false selection rates. Therefore, we recommend FFSR-NH<sub>adj</sub>, especially in studies with a large number of factors; or a bagged version of FFSR-SH. From this study, it is clear that the power of a method to select informative quadratic and interaction terms is important when optimizing a response.

## 6 Example

*Dual Response Optimization in the Lipase Study.* Lipase is an enzyme used in industrial and food processes for its ability to break down lipids. Rathi et al. (2002) used response surface modeling to maximize both the production of lipase and its ability to break down fatty acids or specific activity. In order to produce lipase, the bacteria *Burkholderia cepacia* was cultivated with concentrations of glucose and palm oil added as nutrients. In addition to the nutrient factors, Rathi et al. (2002) were interested in the effect of incubation time, inoculum density, and agitation on the two response variables. Table 7 lists the variables in their study.

Table 7: Variables in Lipase Study

Variable	Name	Measurement Units
$X_1$	glucose	mg/mL
$X_2$	palm oil	% by volume (% v/v)
$X_3$	incubation time	hours
$X_4$	inoculum density	%
$X_5$	agitation	rev/min
$Y_1$	lipase	units/mL (U/mL)
$Y_2$	specific activity	units/mg (U/mg)

Rathi et al. (2002) used a 32-run face-centered, central composite design and fit a second-order linear model in all five factors excluding their interactions. Using their models, the estimated maximum lipase production is 31 U/mL and maximum activity is 110 units/mg.

Table 8: Model Summaries for Lipase Production

Method	Effects in Model	$R^2$	Adjusted $R^2$
Rathi et al.	$X_1, X_1^2, X_2, X_2^2, X_3, X_3^2, X_4, X_4^2, X_5, X_5^2$	0.74	0.62
FFSR-SH	No effects	0.00	0.00
FFSR-WH	No effects	0.00	0.00
FFSR-WH <sub>adj</sub>	No effects	0.00	0.00
FFSR-NH	$X_2^2$	0.26	0.24
FFSR-NH <sub>adj</sub>	$X_2^2, X_3, X_4^2$	0.53	0.48
Standard Approach	$X_2, X_2^2, X_3, X_4, X_4^2$	0.58	0.50

Table 9: Model Summaries for Specific Activity

Method	Effects in Model	$R^2$	Adjusted $R^2$
Rathi et al.	$X_1, X_1^2, X_2, X_2^2, X_3, X_3^2, X_4, X_4^2, X_5, X_5^2$	0.81	0.71
FFSR-SH	$X_3$	0.17	0.14
FFSR-WH	$X_3$	0.17	0.14
FFSR-WH <sub>adj</sub>	$X_3$	0.17	0.14
FFSR-NH	$X_2, X_2^2, X_3, X_4^2, X_5^2$	0.75	0.71
FFSR-NH <sub>adj</sub>	$X_2, X_2^2, X_3$	0.57	0.52
Standard Approach	$X_2, X_2^2, X_3, X_4, X_4^2, X_5, X_5^2$	0.78	0.71

We used Fast FSR methods with  $\gamma_0 = 0.05$  to analyze the data. The selected variables for each method and response are listed in Tables 8 and 9. Unlike the cutinase example, the Fast FSR methods did not choose the same effects in their final models. FFSR-SH and FFSR-WH appear to underfit the data. Because the main effects were not selected, these approaches were unable to fit the significant quadratic terms. Conversely, FFSR-NH<sub>adj</sub> fit larger, more reasonable models. For lipase production no effects were common to all models, although it is likely that palm oil, incubation time, and inoculum density all influence lipase production in some manner. For specific activity only incubation time is common to all the models, whereas glucose was the only factor not selected by any Fast FSR method.

The standard approach models in Tables 8 and 9 show which variables had Type III  $p$ -values

less than 0.05. Thus, of the full ten-variable model used by Rathi et al. (2002), only  $X_2^2$ ,  $X_3$ ,  $X_4^2$ , and  $X_5^2$  are statistically significant at 0.05 level for both responses. Additional simulations in Crews (2008) suggest that their ten-variable models are possibly too large.

Following Rathi et al. (2002), we maximized lipase production and specific activity subject to  $-1 \leq X_j \leq 1$ . Tables 10 and 11 give the coded optimal factor levels and estimated maximum for each method and response. Because the hierarchy-based approaches fit very small models, it is likely that they underestimate maximum lipase production and activity. The models fit using FFSR-NH<sub>adj</sub> are reasonable but differ somewhat for the two responses. The quadratic term for inoculum density,  $X_4^2$ , is not chosen in the model for lipase activity. Without this term the model estimates maximum activity to be approximately 73 U/mg, whereas including this term increases the estimate to approximately 90 U/mg. Further investigation shows that the next possible model in the forward sequence, which adds  $X_4^2$  and  $X_5^2$ , has  $\hat{\gamma}_F$  just over the 0.05 limit. Using this model for maximization leads to 100.63 U/mg. Based on these results, the optimal factor levels for glucose, palm oil and agitation are close to the center point. The optimal factor level for incubation time is the low level, and the optimal level for inoculum density is the high level. Using these optimal levels, the estimated maximum lipase production is approximately 23 U/mL, and the estimated maximum activity is approximately 90 – 100 U/mg. Therefore, it is likely that the maximum lipase production and specific activity observed in practice would be smaller than the estimates 31 U/mL and 110 units/mg. provided by Rathi et al.

Table 10: Optimal Levels for Maximum Lipase Production

Method	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$\hat{\mu}_1(\hat{\mathbf{X}}_{\text{opt}})$
Rathi et al.	0.01	0.09	-1	1	0.09	31.11
FFSR-SH	-	-	-	-	-	10.01
FFSR-WH	-	-	-	-	-	10.01
FFSR-WH <sub>adj</sub>	-	-	-	-	-	10.01
FFSR-NH	-	0	-	-	-	13.67
FFSR-NH <sub>adj</sub>	-	0	-1	1	-	22.98
Standard Approach	-	0.09	-1	1	-	26.59

Table 11: Optimal Levels for Maximum Specific Activity

Method	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$\hat{\mu}_2(\hat{\mathbf{X}}_{\text{opt}})$
Rathi et al.	0.02	0.16	-1	1	0.10	110.99
FFSR-SH	-	-	-1	-	-	59.94
FFSR-WH	-	-	-1	-	-	59.94
FFSR-WH <sub>adj</sub>	-	-	-1	-	-	59.94
FFSR-NH	-	0.16	-1	1	0	100.63
FFSR-NH <sub>adj</sub>	-	0.19	-1	-	-	73.61
Standard Approach	-	0.16	-1	-1	0.10	100.45

## 7 Conclusion

Fast FSR methods provide a simple approach for variable selection in applications where quadratic terms are of interest. Although the Strong Hierarchy restriction (FFSR-SH) has intuitive appeal and performed best in our first simulation study that had  $p = 20$  original variables, it can perform poorly in models where there are important second-order effects but weak parent main effects. In particular it did not perform very well in our second simulation study involving response optimization. Using no hierarchy restrictions (FFSR-NH) can prevent strong second-order effects from being missed. However, second-order terms often dominate the forward sequence, so adjusting the  $p$ -values with FFSR-NH<sub>adj</sub> is recommended. In general, bagging both FFSR-SH and FFSR-NH<sub>adj</sub> show improvement when estimating optimal factor levels.

### ACKNOWLEDGEMENTS

This work was supported by NSF grant DMS-0504283. SAS macros for computing Fast FSR variable selection are available at <http://www4.stat.ncsu.edu/~boos/var.select>.

## REFERENCES

Boos, D. D., Stefanski, L. A., Wu, Y. (2008), "Fast FSR Variable Selection with Applications to Clinical Trials." To appear in *Biometrics*.

- Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123-140.
- Chipman, H. A. (1996), "Bayesian Variable Selection with Related Predictors," *The Canadian Journal of Statistics*, 24, 17-36.
- Chipman, H. A. (1997), "A Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing," *Technometrics*, 39, 372-381.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2006), "BART: Bayesian Additive Regression Trees," preprint.
- Crews, H. B. (2008), "Fast FSR Methods for Second-Order Linear Regression Models," unpublished Ph.D. dissertation, North Carolina State University, Dept. of Statistics.
- Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, Chichester: John Wiley and Sons.
- Hamada, M., and Wu, C. F. J. (1992), "Analysis of Designed Experiments with Complex Aliasing," *Journal of Quality Technology*, 24, 130-137.
- Peixoto, J. L. (1990), "A Property of Well-Formulated Polynomial Regression Models," *The American Statistician*, 44, 26-30.
- Rathi, P., Goswami, V. K., Sahai, V., and Gupta, R. (2002), "Statistical Medium Optimization and Production of a Hyperthermostable Lipase from *Burkholderia cepacia* in a Bioreactor." *Journal of Applied Microbiology*, 35, 59-67.
- Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistic Society, Series B*, 58, 267-288.
- 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, M., Joseph, V. R., and Lin, Y. (2007), "An Efficient Variable Selection Approach for Analyzing Designed Experiments," *Technometrics*, 49, 430-439.
- Zhang, H. H., and Lu, W. (2007), "Adaptive-LASSO for Cox's Proportional Hazard Model," *Biometrika*, 94, 691-703.