An Efficient Variable Selection Approach for Analyzing Designed Experiments

Ming Yuan and V. Roshan Joseph

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Yi Lin

Department of Statistics
University of Wisconsin
Madison, WI 53706

The analysis of experiments in which numerous potential variables are examined is driven by the principles of effect sparsity, effect hierarchy, and effect heredity. We propose an efficient variable selection strategy to specifically address the unique challenges faced by such analysis. The proposed methods are natural extensions of the LARS general-purpose variable selection algorithm. They can be computed very rapidly and can find sparse models that better satisfy the goals of experiments. Simulations and real examples are used to illustrate the wide applicability of the proposed methods.

KEY WORDS: Effect heredity; Least angle regression; Variable selection.

1. INTRODUCTION

We consider the analysis of experiments in which numerous potential variables are examined. In most practical situations, however, only a relatively small number of observations are affordable. Because of their run size economy and flexibility, fractional factorial designs are widely used in such experiments. But the analysis of such designs is complicated by the aliasing of effects. The analysis is driven by the principles of effect sparsity, effect hierarchy, and effect heredity (Wu and Hamada 2000). The effect sparsity principle states that only a small number of effects are significant. The effect hierarchy principle states that lower-order effects are more important than higher-order effects. Using this principle, we can focus on lower-order effects, say, main effects and two-factor interactions, assuming that the higher-order interactions are negligible. The effect heredity principle, indicates that an interaction can be active only if one or both of its parent effects are also active. For example, a two-factor interaction can be active only if one or both of the corresponding main effects are active. These principles have proven to be effective tools for resolving the aliasing patterns.

The analysis of experiments can be formulated in the form of the general linear regression where we have n observations on a response Y and p explanatory variables $(X_1, X_2, \dots, X_p)$, and

$$Y = X\beta + \epsilon, \tag{1}$$

where $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I)$ and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$. Throughout this article, we center the response and each variable so that the observed mean is 0, and scale each variable so that the sample standard deviation is 1. Because each variable corresponds to an effect in the experiments (main effects, two-factor interactions, and so on), we use the two terms interchangeably in later discussions.

The principle of effect sparsity can be achieved by the variable selection, the goal of which is to search for the model that best describes the data-generating mechanism among the 2^p candidate models. However, as pointed out by Chipman, Hamada, and Wu (1997), the analysis of designed experiment poses several new challenges for variable selection. First, the number of explanatory variables greatly exceeds the number of runs; for example, in the 12-run Plackett–Burman design described in Table 1, 11 main effects and 55 two-factor interactions are considered. Second, due to the large number of potential variables, the number of candidate models usually is huge, which calls for computationally efficient methods. Third, the effects are always related due to the presence of interactions or polynomial terms of factors; consequently, the principle of effect heredity is required to achieve a reasonable model. For example, a general-purpose variable selection method may select two-factor interactions without the corresponding main effects. Such models are difficult to interpret in practice. This problem can be avoided by conforming with the effect heredity principle. Effect heredity is also closely related to the notion of marginality (Nelder 1977, 1994; McCullagh and Nelder 1989) which ensures that the response surface is invariant under scaling and translation of the factors of an experiment.

Classical variable selection methods, such as C_p, the Akaike information criterion, and Bayes information criterion, choose among possible models using penalized sum of squares criteria, with the penalty being an increasing function of the model dimension. But these methods are computationally infeasible, and their stepwise implementation is inappropriate for analyzing the designed experiments (Westfall, Young, and Lin 1998). Various other variable selection methods also have been introduced in recent years (e.g., George and McCulloch 1993; Foster and George 1994; Breiman 1995; Tibshirani 1996; George and Foster 2000; Fan and Li 2001; Shen and Ye 2002; Efron, Johnston, Hastie, and Tibshirani 2004; Yuan and Lin 2005). In particular, the stochastic search variable selection method developed by George and McCulloch (1993) has been adopted by Chipman et al. (1997) to analyze experiments with complex aliasing patterns. As noted by Chipman et al. (1997), their proposal remains computationally demanding. More recently, Li and Lin (2002) applied the variable selection procedure of Fan and Li (2001) to analyze supersaturated designs. Despite its nice theoretical properties, their approach does not impose the heredity principle.

TECHNOMETRICS, NOVEMBER 2007, VOL. 49, NO. 4
DOI 10.1198/004017007000000173

Table 1. A 12-run Plackett–Burman design for Example 1

Run	A	B	C	D	E	F	G	H	I	J	K
1	+	+	−	+	+	+	−	−	−	+	−
2	+	−	+	+	+	−	−	−	+	−	+
3	−	+	+	+	−	−	−	+	−	+	+
4	+	+	+	−	−	−	+	−	+	+	−
5	+	+	−	−	−	+	−	+	+	−	+
6	+	−	−	−	+	−	+	+	−	+	+
7	−	−	−	+	−	+	+	−	+	+	+
8	−	−	+	−	+	+	−	+	+	+	−
9	−	+	−	+	+	−	+	+	+	−	−
10	+	−	+	+	−	+	+	+	−	−	−
11	−	+	+	−	+	+	+	−	−	−	+
12	−	−	−	−	−	−	−	−	−	−	−

The lack of a fully satisfactory variable selection strategy for analyzing experiments motivates our work here. We consider the extension of an effective variable selection algorithm, LARS (least angle regression), proposed by Efron et al. (2004). The LARS algorithm is very fast and is closely connected with boosting and another popular variable selection method, the LASSO (Tibshirani 1996). Whereas the LARS enjoys great computational advantages and excellent predictive performance, it is devised for general-purpose variable selection and often produces models that are hard to interpret in practice. In this article we propose modified LARS algorithms so that the heredity principle can be taken into account in the variable selection. It is demonstrated that incorporating such constraints in variable selection often leads to a model better satisfying the goals of experiment.

The rest of the article is organized as follows. We review the LARS methodology in the next section. In Section 3 we present different modifications to the LARS algorithm so that the heredity principles can be taken into account. Although our main focus in this article is on main effects and their two-factor interactions, in Section 4 we also explain how the methods can be extended to the case of more complicated situations. We demonstrate the wide applicability of the proposed methods through three examples in Section 5, and conclude with a discussion in Section 6.

2. LARS

The LARS algorithm uses a variable selection strategy similar to forward selection. Starting with all coefficients equal to 0, the algorithm finds the variable that is most correlated with the response and proceeds in that direction. Instead of taking a full step toward the projection of Y on the variable, as would be done in forward selection, LARS takes the largest step possible in this direction only until some other variable has as much correlation with the current residual. Then this new variable is entered, and the process is continued. The great computational advantage of LARS comes from the fact that the LARS path is piecewise linear, and all we need to do is to locate the change points. More specifically, the LARS algorithm can be described as follows:

LARS algorithm.

1. Start from $\beta^{[0]} = 0$, $k = 1$, and $r^{[0]} = Y$.
2. Find a variable, X_j, that is most correlated with $r^{[0]}$ and set $\mathcal{B}_k = \{j\}$.
3. Compute the current direction, γ, which is a p-dimensional vector with $\gamma_{\mathcal{B}_k^c} = 0$ and
$$\gamma_{\mathcal{B}_k} = \left(X'_{\mathcal{B}_k} X_{\mathcal{B}_k}\right)^{-} X'_{\mathcal{B}_k} r^{[k-1]}. \tag{2}$$
4. For every $i \notin \mathcal{B}_k$, compute how far the algorithm will march in direction γ before X_i has the same amount of correlation with the residual as the variables in $\mathcal{B}_k$. This can be measured by the smallest $\alpha_i \in [0, 1]$ such that
$$\left|X'_i\left(r^{[k-1]} - \alpha_i X\gamma\right)\right| = \left|X'_{\mathcal{B}_1}\left(r^{[k-1]} - \alpha_i X\gamma\right)\right|. \tag{3}$$
5. If $\mathcal{B}_k \neq \{1, \ldots, p\}$, then let $\alpha = \min_{i \notin \mathcal{B}_k} \alpha_i \equiv \alpha_{i^*}$ and update $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{i^*\}$. Otherwise, set $\alpha = 1$.
6. Update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$, and $k = k + 1$. Go back to step 3 until $\alpha = 1$.

Here $\mathcal{B}_k$ keeps track of the variables that are included in the model at the kth stage, γ determines the direction in which the coefficient estimate will move along, and α measures how far the algorithm will march along that direction. Note that (3) is equivalent to

$$X'_i\left(r^{[k-1]} - \alpha_i X\gamma\right) = \pm X'_{\mathcal{B}_1}\left(r^{[k-1]} - \alpha_i X\gamma\right), \tag{4}$$

which can be easily solved for α_i. (See Efron et al. 2004 for more details.)

3. LARS UNDER HEREDITY PRINCIPLES

A LARS-type algorithm is driven by the measurement of "predictability." In its original form, "predictability" is measured by the correlation with the residual. At any point on the solution path of the LARS algorithm, the variables selected are the those that are the most correlated with the current residual. Define $\theta(r, X_i)$ as the angle between the two n-vectors, r and X_i. It is clear that the squared correlation between X_i and r can be written as $\cos^2(\theta(r, X_i)) = \|X'_i r\|^2/\|r\|^2$. This is also the proportion of the total variation in r that is explained by the regression on X_i, that is, the R^2 when r is regressed on X_i. In other words, a variable enters the LARS path if it has the highest "predictability" on its own. Now that the heredity principles are in place, some adjustment to the LARS algorithm is needed.

We consider two versions of the heredity principle (Chipman 1996). Under *strong heredity*, for a two-factor interaction to be active both its parent effects should be active, whereas under *weak heredity*, only one of its parent effects need to be active. We now propose modifications to the LARS algorithm so that the selected models will obey either the strong heredity or the weak heredity principles. This will lead to better models, provided that the true model, which is unknown to the experimenter, obeys the heredity principles. Exceptions are possible, but many empirical studies have confirmed the use of these principles. A Bayesian justification of the effect heredity has been provided by Joseph (2006).

To develop the idea, we consider only the main effects and two-factor interactions for the moment. The methods can also be applied to more general cases of models with an arbitrary number of terms, each of arbitrary order. We give such extensions in a later section.

3.1 Strong Heredity Principle

We begin with the strong heredity principle. In this case, the corresponding parent effects should be selected if an interaction is selected. To account for such dependence, when determining whether an interaction should be entered, it is natural to measure the average "predictability" of all variables that must be included. Adopting the idea of the original LARS algorithm, the predictability of a set of variables can be measured by the squared cosine of the angle between the residual and the linear space spanned by the set of variables. This idea is illustrated by Figure 1. To measure the predictability of $X_1 = (X_{11}, X_{12})$, which is bivariate, we look at the squared cosine of α_1, the angle between Y^* and the two-dimensional linear space spanned by the two components of X_1. This is a natural extension of the idea behind LARS. When the variable is one-dimensional, such as X_2 in the diagram, LARS looks at the squared cosine of α_2, which is the angle between two vectors, Y^* and X_2.

Once the measure of predictability for a set of variables is obtained, we have to adjust for the fact that different sets of variables have different numbers of degrees of freedom. It is clear that the more variables a set has, the better it can explain the residual for the given data. One way to adjust for this is to measure the predictability per degree of freedom, which can be defined as $\cos^2(\theta(r, X_{\mathcal{A}}))/n_{\mathcal{A}}$, where r is the current residual, $X_{\mathcal{A}}$ is the set of variables to be entertained, and $n_{\mathcal{A}}$ is the cardinality of $\mathcal{A}$. Recall that $\cos^2(\theta(r, X_{\mathcal{A}}))$ is the R^2 when r is regressed on $X_{\mathcal{A}}$, which can be computed by fitting one linear regression. A similar idea was also used by Yuan and Lin (2006) in a different context.

In so defining the measure of predictability, we implicitly assume that adding any variable, interaction or main effect, increases the model complexity in the same way. Although this can be motivated by the definition of degrees of freedom in the ANOVA analysis as elaborated by Yuan and Lin (2006), in practice, it might be desirable to ascribe different weights to different variables, that is, to distinguish between main effects and interactions. One possibility is to associate each variable with a different degree of freedom, then define the measure of predictability as $\cos^2(\theta(r, X_{\mathcal{A}}))/\sum_{j\in\mathcal{A}} n_j$, where n_j is the degree of freedom for the jth effect. For example, by the hierarchy principle, two-factor interactions are less important than main effects. To incorporate this principle, one may set greater degrees of freedom for a two-factor interaction than for main effects. Clearly, this extension reduces to the average predictability if $n_j = 1$ for all j. To fix ideas, in the rest of this article, we assume that $n_j = 1$.

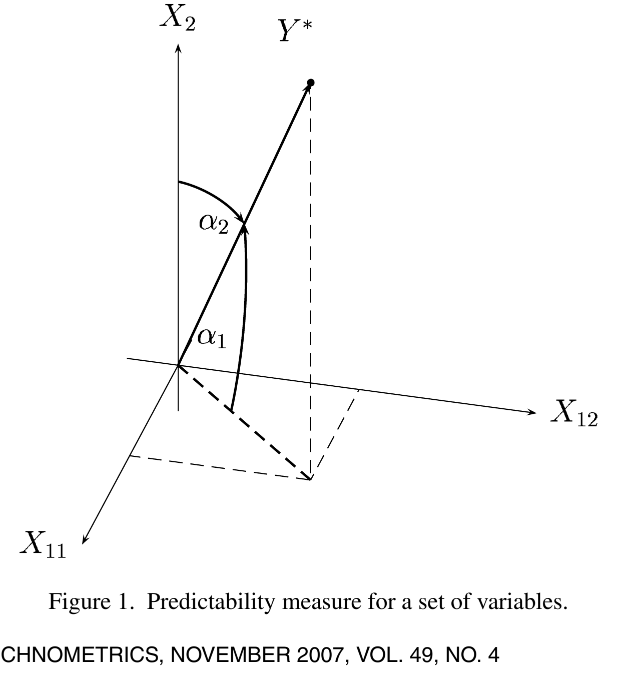

Figure 1. Predictability measure for a set of variables.

With such a notion of average "predictability," the proposed LARS-type algorithm proceeds in the following way:

Strong heredity algorithm.

0. For main effects, initialize the dependence set $\mathcal{D}$ as an empty set. For interactions, let the dependence set $\mathcal{D}$ be the set of corresponding parent effects.
1. Start from $\beta^{[0]} = 0$, $k = 1$, and $r^{[0]} = Y$.
2. Compute the "prediction score" for each candidate variable i,
$$s_i = \frac{\cos^2(\theta(r^{[0]}, X_{\{i\}\cup\mathcal{D}_i}))}{1 + n_{\mathcal{D}_i}}. \tag{5}$$
Denote $i^* = \arg\max_i s_i$. Define the current "most predictive variable" as $\mathcal{A}_1 = \{i^*\}$ and the "active set" as $\mathcal{B}_1 = \mathcal{A}_1 \cup \mathcal{D}_{i^*}$.
3. Compute the current direction γ, which is a p-dimensional vector with $\gamma_{\mathcal{B}_k^c} = 0$ and
$$\gamma_{\mathcal{B}_k} = \left(X'_{\mathcal{B}_k} X_{\mathcal{B}_k}\right)^- X'_{\mathcal{B}_k} r^{[k-1]}. \tag{6}$$
4. For every $i \notin \mathcal{B}_k$, update $\mathcal{D}_i = \mathcal{D}_i \cap \mathcal{B}_k^c$ and compute how far the algorithm will march in direction γ before X_i enters the most predictive set. This can be measured by the smallest $\alpha_i \in [0, 1]$ such that
$$\frac{\|X'_{\{i\}\cup\mathcal{D}_i}(r^{[k-1]} - \alpha_i X\gamma)\|^2}{1 + n_{\mathcal{D}_i}} \geq \frac{\|X'_{\mathcal{B}_1}(r^{[k-1]} - \alpha_i X\gamma)\|^2}{n_{\mathcal{B}_1}}. \tag{7}$$
5. If $\mathcal{B}_k \neq \{1, \dots, p\}$, then let $\alpha = \min_{i\notin\mathcal{B}_k} \alpha_i \equiv \alpha_{i^*}$ and update $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{i^*\}$ and $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{i^*\} \cup \mathcal{D}_{i^*}$. Otherwise, set $\alpha = 1$.
6. Update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$, and $k = k + 1$. Go back to step 3 until $\alpha = 1$.

As in the LARS algorithm, here we start with all coefficients being 0, then compare the candidate variables in terms of the average predictability (5). For main effects, the average predictability is defined as the magnitude of the correlation between the variable and the residual. For two-factor interactions, this is defined as the average predictability of all variables from $\{i\} \cup \mathcal{D}_i$, because all of them must enter the model as a group if the ith variable is selected. After identifying the most predictive variable, we form two different sets to keep track of the most predictive variables and the variables that enter the model (i.e., $\mathcal{A}_k$ and $\mathcal{B}_k$). In the case of LARS, the two sets coincide. But in our case, the two sets may differ, because some variables enter the model only because its child is highly predictive. The algorithm continues along the least squares estimate with only the variables from the active set, a direction that reduces the residual sum of squares the most. We march in this direction until another variable has at least the same amount of predictability as the variables from the current most predictive set.

By the definition of γ, (7) holds when $\alpha_i = 1$, because its right side equals 0 in this case. Therefore, α_i in step 4 is always well defined. Different from (3), finding α_i amounts to solving a quadratic equation, which also can be obtained in explicit form.

Another difference from the LARS algorithm is that the amount of progression measured by α is now defined through an inequality (7) rather than an equality. Such a modification is necessary because $\mathcal{D}_i$ may change in the process. The averaged predictability for the ith variable can increase as a result of the inclusion of an element of $\mathcal{D}_i$. For example, consider the case of two main effects A and B and a two-factor interaction AB. In the beginning, the predictability of AB is measured by $s_{AB,1} = \cos^2(\theta(Y, X_{\{A,B,AB\}}))/3$. Suppose that $s_{AB,1}$ is dominated by the predictive score of A, $s_A \equiv \cos^2(\theta(Y, X_{\{A\}}))$. Because A enters the model, the predictability of AB should now be measured by $s_{AB,2} = \cos^2(\theta(Y, X_{\{B,AB\}}))/2$, which might be even greater than s_A. In this case the solution to (7) is $\alpha = 0$, and $\{AB, B\}$ enter the model immediately after A enters the model.

Alternative approaches for incorporating the strong heredity principle in the LARS algorithm also have been introduced by Efron et al. (2004) and Turlach (2004). Efron et al. (2004) suggested a two-step procedure in which the main effects are considered first and their interactions are examined only in a subsequent step. As illustrated by Turlach (2004), such a procedure may exhibit less-than-optimal behavior in certain nontrivial situations. Turlach's proposal is similar in spirit to our proposal here. Using our notation, he suggested measuring the predictability of variables $\{i\} \cup \mathcal{D}_i$ by $|X_i' r|$. Unlike the averaged predictability that we used, this criterion does not account for the number of variables that enter simultaneously and also may lead to a suboptimal solution. Also note that Efron et al. (2004) and Turlach (2004) considered only strong heredity, not weak heredity.

3.2 Weak Heredity Principle

Unlike the strong heredity principle, here it is not predetermined which main effect must be included so that an interaction can be entered. Therefore, any element from $\mathcal{D}_i$ can enter the model together with the ith variable. We pick the one that yields the highest predictive score. More specifically, the predictive score for the ith variable is now defined as

$$\max_{j \in \mathcal{D}_i} \frac{\cos^2(\theta(r^{k-1}, X_{\{i,j\}}))}{2}. \tag{8}$$

Thus we have the following algorithm:

Weak heredity algorithm.

0. Initialize $\mathcal{D}_i = \phi$ if the ith variable is a main effect. Otherwise, let $\mathcal{D}_i$ be the set of the main effects corresponding to the ith variable.
1. Start from $\beta^{[0]} = 0$, $k = 1$, and $r^{[0]} = Y$.
2. If $\mathcal{D}_i = \phi$, then define the "predictive score" of a candidate variable as $s_i = \cos^2(\theta(r^{k-1}, X_i))$. If $\mathcal{D}_i \neq \phi$, then compute the "prediction scores" for each candidate variable i and each variable in $\mathcal{D}_i$,
$$s_{ij} = \frac{\cos^2(\theta(r^{k-1}, X_{\{i,j\}}))}{2}, \tag{9}$$
and define $s_i = \max_j s_{ij}$. Denote $i^* = \arg\max_i s_i$. Define the current "most predictive set" as $\mathcal{A}_1 = \{i^*\}$. If $\mathcal{D}_{i^*} = \phi$, then define the "active set" as $\mathcal{B}_1 = \mathcal{A}_1$. Otherwise, denote $j^* = \arg\max_j s_{i^*j}$ and define $\mathcal{B}_1 = \{i^*, j^*\}$.
3. Compute the current direction, γ, which is a p-dimensional vector, with $\gamma_{\mathcal{B}_k^c} = 0$ and
$$\gamma_{\mathcal{B}_k} = \left(X_{\mathcal{B}_k}' X_{\mathcal{B}_k}\right)^- X_{\mathcal{B}_k}' r^{[k-1]}. \tag{10}$$
4. For every $i \notin \mathcal{B}_k$, update $\mathcal{D}_i = \mathcal{D}_i \cap \mathcal{B}_k^c$. Compute how far the algorithm will march in direction γ before X_i enters the most predictive set. This can be measured by $\alpha_i \in [0, 1]$, defined as follows:

   a. If $\mathcal{D}_i = \phi$, then α_i is the smallest value such that
   $$\left\| X_i'\left(r^{[k-1]} - \alpha_i X\gamma\right)\right\|^2 \geq \frac{\|X_{\mathcal{B}_1}'(r^{[k-1]} - \alpha_i X\gamma)\|^2}{n_{\mathcal{B}_1}}. \tag{11}$$

   b. If $\mathcal{D}_i \neq \phi$, then, for each $j \in \mathcal{D}_i$, define α_{ij} as the smallest value in [0, 1] such that
   $$\frac{\|X_{\{i,j\}}'(r^{[k-1]} - \alpha_{ij} X\gamma)\|^2}{2} \geq \frac{\|X_{\mathcal{B}_1}'(r^{[k-1]} - \alpha_{ij} X\gamma)\|^2}{n_{\mathcal{B}_1}}. \tag{12}$$

   and $\alpha_i = \min_j \alpha_{ij}$.
5. Denote $i^* = \arg\min_i \alpha_i$ and update $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{i^*\}$. If $\mathcal{D}_{i^*} = \phi$, then set $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{i^*\}$. Otherwise, define $j^* = \arg\min_j \alpha_{i^*j}$ and update $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{i^*, j^*\}$.
6. Denote $\alpha = \min_{i \notin \mathcal{B}_k} \alpha_i$ and update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$, and $k = k + 1$. Go back to step 3 until $\alpha = 1$.

Not every variable in $\mathcal{D}_i$ necessarily enters the model together with the ith variable under the weak heredity principle. Only the parent effect that yields the high predictability score together with the ith variable enter the active set $\mathcal{B}_k$. The algorithm proceeds in the same fashion as that under the strong heredity principle.

3.3 Further Discussions

To incorporate the effect heredity principles, we took advantage of the nice geometric interpretation of the LARS. But this property is not shared by its cousin, the LASSO, which is given as a constrained least squares estimate. Despite the LASSO's close connection with the LARS, it is not immediately clear how the LASSO can accommodate the effect heredity principles. Because the LASSO shares a similar geometric interpretation with the LARS (Osborne, Presnell, and Turlach 2000; Efron et al. 2004), it is tempting to make adjustments to the solution path of the LASSO similar to those proposed here. But then it loses the interpretation as a constrained least squares estimate, and, as noted by Turlach (2004), such a modification is not trivial.

While incorporating the effect heredity principles, the proposed modifications also inherit the main advantages of the original LARS. In contrast to the "winner takes all" strategy adopted by the subset selection (i.e., regressors are either retained or dropped from the model), our methods yield a continuous solution path like the original LARS. It is known (Breiman

1996) that a discrete estimating procedure such as the subset selection can be extremely variable; that is, small changes in the data may lead to quite different models. Although a group of variables may enter the model simultaneously in our proposals due to heredity, they are entered in an incremental fashion as the original LARS and thus retains its stability.

Similar to the original LARS, the proposed algorithms are computationally thrifty. After the dependence sets, $\mathcal{D}$'s, are determined, the entire solution path is constructed in $O((p \wedge n)^3 + np(p \wedge n))$ computations, where $p \wedge n = \min(p, n)$. Note that this is also the cost of a least squares fit on all p variables when $p \leq n$. More specifically, let m denote the number of total steps. In the original LARS, $m = p \wedge n$. Because multiple variables can enter simultaneously, $m \leq p \wedge n$ in our proposals. At the kth step, there are at most $p - k$ candidates for the next active variable. Thus identification of the next active variable requires the computation of $O(p - k)$ inner products between the candidate variables and the residual. After the active variable is selected, we need to invert the Gram matrix of all of the selected variables to find the next direction. The same as for the original LARS, all m such calculations can be viewed as a Cheolesky factorization with all variables ordered appropriately (Efron et al. 2004) and can be done with a total of $O((p \wedge n)^3)$ computations. In practice, we observe that the proposed algorithms are comparable with or even faster than the original LARS in terms of computational speed. For example, the computational times for constructing the whole solution path in Example 1 of Section 5 were .19, .10, and .16 second for the original LARS, LARS with strong heredity, and LARS with weak heredity when the program was run on the same desktop computer.

4. BEYOND TWO–WAY INTERACTIONS

Generally, we can represent the heredity principles by sets $\{\mathcal{D}_i : i = 1, \ldots, p\}$, where $\mathcal{D}_i$ contains a set of variables. So that the ith variable can be considered for inclusion, all elements of $\mathcal{D}_i$ must be included under the strong heredity principle, and at least one element of $\mathcal{D}_i$ should be included under the weak heredity principle. It is worth pointing out that our definitions of strong and weak heredity principles are more general than their traditional versions. For example, Nelder (1998) mentioned a heredity principle that requires inclusion of a certain main effect so that an interaction can be considered. Such a *partial heredity* principle can be induced by the strong heredity principle with the choices of $\mathcal{D}_{AB} = \{A\}$ or $\mathcal{D}_{AB} = \{B\}$. In our previous discussion, we focused on dealing with two-factor interactions. More generally, both algorithms work for the case in which a variable does not depend on any other variables if it is in the dependent set of some variables, that is, $\mathcal{D}_j = \phi$ if $j \in \mathcal{D}_i$ for any i. If this is not the case (e.g., in the case when the polynomial factor interaction such as A^2B^2 is also entertained), then modifications to the foregoing algorithms are necessary.

It is helpful to think of the dependence structure described by the $\mathcal{D}$'s as a directed graph where all p variables are the nodes and an edge from i to j is present if and only if $j \in \mathcal{D}_i$. To handle the strong heredity principle, we first reevaluate the dependence set $\mathcal{D}'$s so that $\mathcal{D}_i$ contains all nodes that can be reached from the ith node. This is can be done efficiently using, for example, the breadth first algorithm (Cormen, Leiserson, Rivest, and Rivest 1990). After this step, the LARS algorithm with the strong heredity principle presented earlier can be applied.

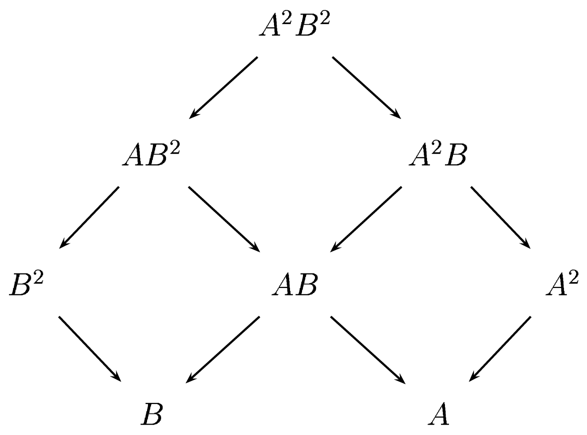

Figure 2. Dependence structure for two-way interaction between three-level factors.

The situation for the weak heredity principle is more complicated, because we need to first determine which variables are to be included together with the ith variable. We first determine which nodes are terminal nodes, that is, the nodes whose dependence set is empty. Then the candidate variable sets to be considered for the ith variable to be included can be described by all of the possible paths from it to any of the terminal nodes. This can be done efficiently using the depth first algorithm (Cormen et al. 1990). Let $\{P_1, \ldots, P_k\}$ denote the collection of such paths. We compare these paths again using the averaged predictive scores when all nodes on the path are included. The variables on the path with the highest averaged predictive score will enter the model, and these variables will be eliminated from the dependence sets of the remaining variables. The process then continues as the weak heredity principle algorithm presented in Section 3.2.

To elaborate, consider two three-level factors, A and B. The hierarchy among all variables can be described by the diagram in Figure 2 (Chipman 1996).

Under the strong heredity principle, the algorithm developed in Section 3.1 can be applied directly with

$$\begin{aligned}
\mathcal{D}_{B^2} &= \{B\}, \\
\mathcal{D}_{AB} &= \{A, B\}, \\
\mathcal{D}_{A^2} &= \{A\}, \\
\mathcal{D}_{AB^2} &= \{B^2, AB, A, B\}, \\
\mathcal{D}_{A^2B} &= \{AB, A^2, A, B\},
\end{aligned}$$

and

$$\mathcal{D}_{A^2B^2} = \{AB^2, A^2B, B^2, AB, A^2, A, B\}.$$

Under the weak heredity principle, the candidate paths for each interaction also can be obtained easily,

$$\begin{aligned}
B^2&: \quad \{(B, B^2)\}, \\
AB&: \quad \{(A, AB), (B, AB)\}, \\
A^2&: \quad \{(A, A^2)\}, \\
AB^2&: \quad \{(A, AB, AB^2), (B, AB, AB^2), (B, B^2, AB^2)\}, \\
A^2B&: \quad \{(A, AB, A^2B), (B, AB, A^2B), (A, A^2, A^2B)\},
\end{aligned}$$

and

$$A^2B^2\colon\ \big\{(A, AB, AB^2, A^2B^2), (B, AB, AB^2, A^2B^2), (B, B^2, AB^2, A^2B^2), (A, A^2, A^2B, A^2B^2)\big\},$$

where the variables within each pair of parentheses compose one candidate path. Now the weak heredity algorithm presented in Section 3.2 can be applied.

5. EXAMPLES

In this section we demonstrate the proposed variable selection strategy using three examples. The first example uses a 12-run, 2-level nonregular design; the second uses a 16-run, 2-level regular design; and the third uses an 18-run, nonregular, mixed-level design. These examples are selected to show the wide applicability of our method.

Example 1. This is a simulated example proposed by Hamada and Wu (1992). Eleven two-level factors and their second-order interactions are considered. The design is given in Table 1. The response is simulated according to the following linear model:

$$Y = A + 2AB + 2AC + \epsilon, \tag{13}$$

where $\epsilon \sim \mathcal{N}(0, .25^2)$.

There are a total of 66 candidate effects (11 main effects and 55 2-factor interactions). Figure 3 compares the solution paths obtained by the new methods and the LARS algorithm. The figure plots the traces of the estimated regression coefficients. Here the weak heredity version of the new method is able to pick up the correct effects in only two steps, whereas the LARS algorithm could not identify the main effect of A. In this example the strong heredity version did not work, which should be expected because the true model does not contain the main effects of B and C. In practice, we will not know which version of the heredity principle to use; therefore, we should run both of them. It will be easy to select the right one by looking at the solution paths. Ideally, we want to choose paths in which a small number of coefficients increase quickly at the early stages. In this example, comparing the solution paths generated by the strong and weak heredity versions of the algorithm, we can immediately understand that we should be using the weak heredity version. We also note that one of the ordinary forward-selection methods proposed by Hamada and Wu (1992) could not identify any of the important effects. This clearly shows the advantages of the proposed method.

Example 2. Consider a 2^{9-5} experiment reported by Raghavarao and Altan (2003). The design and data are given in Table 2.

The results of the analysis are plotted in Figure 4. The effects selected in the first five steps are given in Table 4. We see that

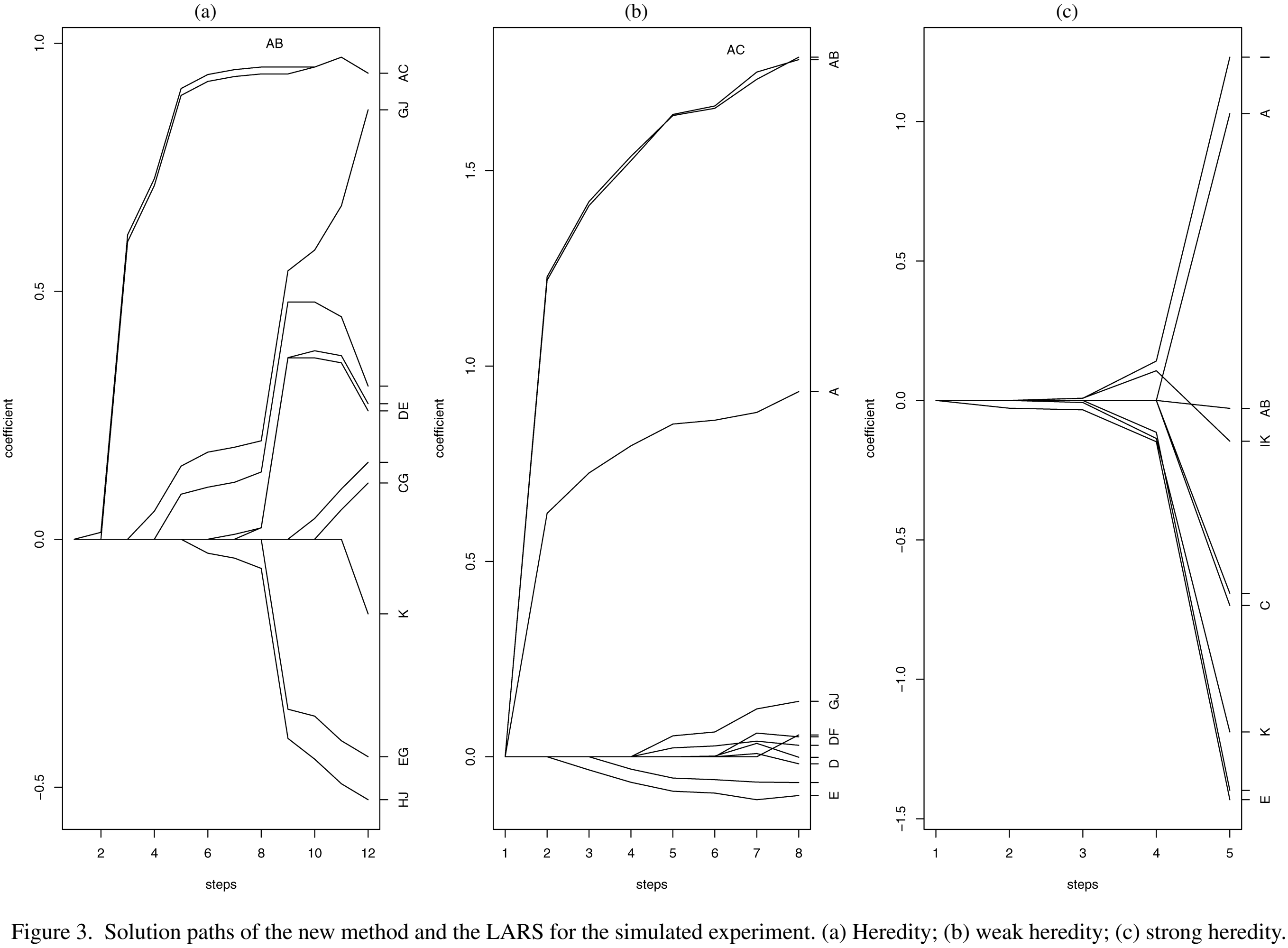

Figure 3. Solution paths of the new method and the LARS for the simulated experiment. (a) Heredity; (b) weak heredity; (c) strong heredity.

Table 2. The 2^{9-5} design and data for Example 3

Run	A	B	C	D	E	F	G	H	J	Y
1	−	−	−	−	−	−	−	−	−	136.475
2	+	+	−	+	+	−	−	−	−	147.775
3	+	−	−	+	−	−	+	+	−	142.425
4	+	−	+	+	+	−	+	+	+	141.800
5	+	+	+	+	−	−	−	−	+	136.675
6	−	+	−	−	+	−	+	+	−	150.725
7	−	+	+	−	−	−	+	+	+	142.800
8	−	−	+	−	+	−	−	−	+	135.825
9	+	+	+	−	−	+	−	+	−	143.476
10	−	+	−	+	+	+	+	−	+	145.150
11	+	−	+	−	+	+	+	−	−	142.600
12	−	−	−	+	−	+	−	+	+	139.375
13	+	+	−	−	+	+	−	+	+	139.650
14	+	−	−	−	−	+	+	−	+	144.775
15	−	−	+	+	+	+	−	+	−	148.275
16	−	+	+	+	−	+	+	−	−	141.075

Table 3. $OA(18, 2^1 3^7)$ and data from the blood glucose experiment

Run	A	G	B	C	D	E	F	H	Y
1	1	1	1	1	1	1	1	1	97.94
2	1	1	2	2	2	2	2	2	83.40
3	1	1	3	3	3	3	3	3	95.88
4	1	2	1	1	2	2	3	3	88.86
5	1	2	2	2	3	3	1	1	106.58
6	1	2	3	3	1	1	2	2	89.57
7	1	3	1	2	1	3	2	3	91.98
8	1	3	2	3	2	1	3	1	98.41
9	1	3	3	1	3	2	1	2	87.56
10	2	1	1	3	3	2	2	1	88.11
11	2	1	2	1	1	3	3	2	83.81
12	2	1	3	2	2	1	1	3	98.27
13	2	2	1	2	3	1	3	2	115.52
14	2	2	2	3	1	2	1	3	94.89
15	2	2	3	1	2	3	2	1	94.70
16	2	3	1	3	2	3	1	2	121.62
17	2	3	2	1	3	1	2	3	93.86
18	2	3	3	2	1	2	3	1	96.10

LARS identifies AH, J, E, G, and CH as significant. In a 2^{9-5} design, the effects are either orthogonal or completely aliased with others. Ignoring three- and higher-order interactions, we can obtain the following aliasing relationships for the foregoing five effects:

$$AH = EJ = DG = BG,$$

$$J = -CF,$$

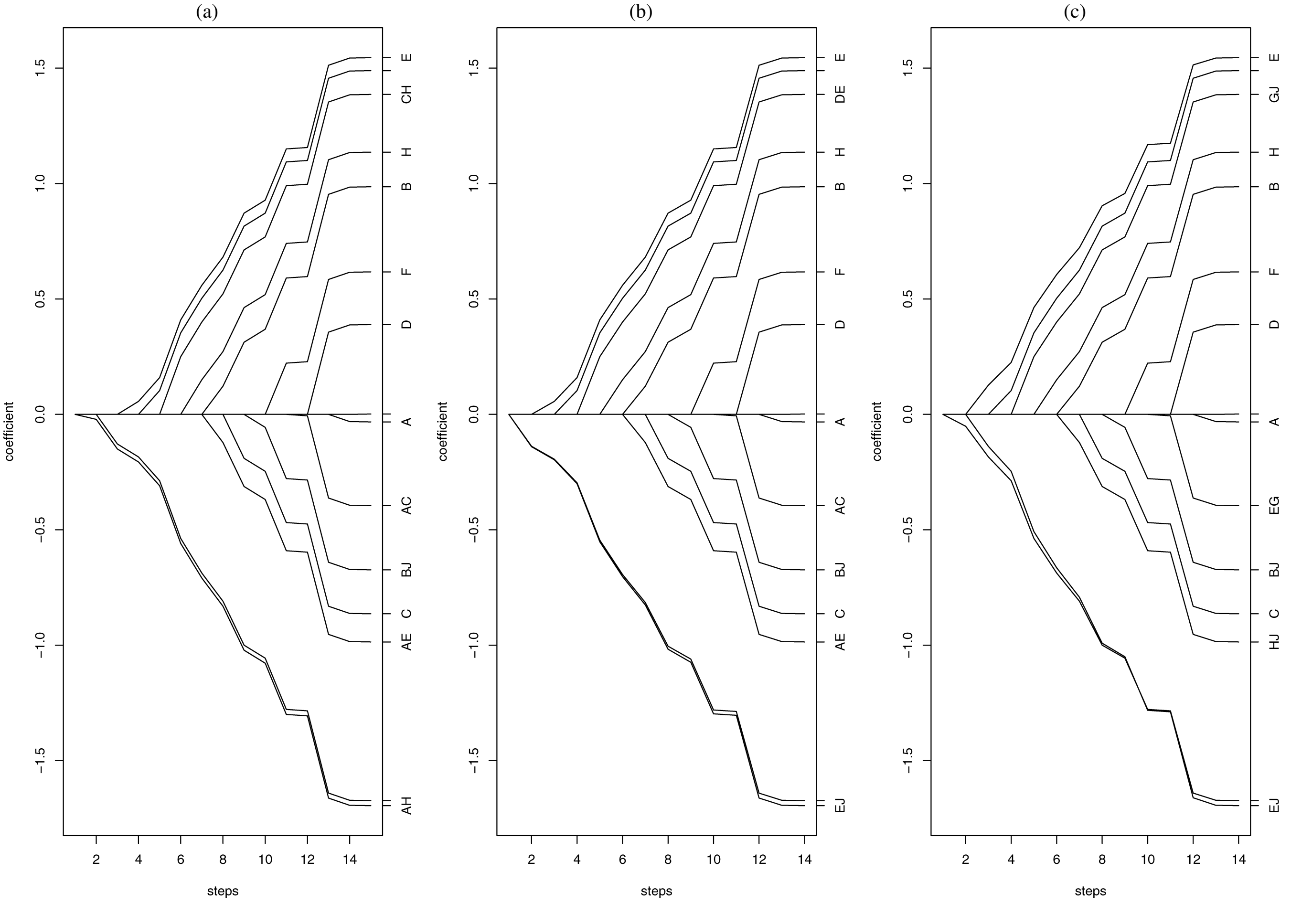

Figure 4. Solution paths for the 2^{9-5} factorial experiment. (a) No heredity; (b) weak heredity; (c) strong heredity.

Table 4. Effects selected at the first five steps

	Simulated			Factorial			Blood Glucose		
Step	None	Weak	Strong	None	Weak	Strong	None	Weak	Strong
1	AC	A, AC	E	AH	J, EJ	J	BH^2	B, BH, BH^2, B^2H^2	E, F, EF
2	AB	AB	I, K, IK	J	E	E, EJ	B^2H^2	E, EF	E^2
3	GJ	E	H	E	G	G	EF	E^2	F^2
4	DF	H	A, C, AC	G	DE	GJ	AH^2	GE	All of B, H
5	HJ	G, GJ	B, AB	CH	H	H	GE	BE^2	

$$E = -BC,$$

$$G = -FH = -AB,$$

and

$$CH = GJ = DE.$$

Thus the effect AH could actually be EJ, DG, or BG. Any one of these effects can produce the same fit; thus which one to choose is unclear. LARS selected AH because it appears before EJ, DG, and BG in the list of effects (they are listed in alphabetical order). Thus the foregoing selection is inconclusive. The same is true with the selection of J, E, G, and CH. In the literature, follow-up experiments are usually recommended for dealiasing the effects (see Meyer, Steinberg, and Box 1996). On the other hand, Wu and Hamada (2000) suggested that by applying effect hierarchy and effect heredity principles, some of the effects can be dealiased. Our methods incorporate both of these principles, and thus will entail less confusion due to aliasing.

As given in Table 4, the first five effects identified by strong heredity are J, E, EJ, G, and GJ. Note that this is the only set of effects from the five aliasing relationships that satisfy strong heredity. Thus our method has no ambiguity in selecting the effects. The final model from our method seems to be more meaningful and interpretable. Applying the weak heredity algorithm, we obtained the same effects except the last one. Instead of GJ, it identifies DE. But note that these two effects are completely aliased. Weak heredity cannot break it, because one of the parent effects from both of these interactions is significant. In such a situation, we recommend using strong heredity. Although DE can be significant under weak heredity, GJ is more likely to be significant because both of its parent effects, G and J, are significant. Interestingly, the heuristic analysis of Raghavarao and Altan (2003) also identified the same five effects, J, E, EJ, G, and GJ, as significant.

Example 3. The blood glucose experiment was studied by Hamada and Wu (1992), among many others. It has one two-level factor and seven three-level factors. The experimental design is the mixed-level orthogonal array, $OA(18, 2^1 3^7)$. The design and the response are given in Table 3.

Each three-level factor is divided into linear and quadratic effects using the orthogonal polynomial coding (Wu and Hamada 2000); thus there are 15 main effects and 96 2-factor interactions. We use this example to illustrate how complicated heredity principles as described in Figure 2 can be handled using the proposed methodology.

Figure 5 gives the solution paths of the LARS and the two proposed methods. The plot indicates that the weak heredity principle is more likely to be true and that the corresponding result is also in accordance with the previous analysis (Hamada and Wu 1992; Chipman et al. 1997), whereas LARS identifies a model that does not satisfy any of the heredity principles (see Table 4).

Because of the frequentist nature of our approach, the heredity rule that we considered is deterministic, and we search only models that satisfy heredity principles. In this sense, our approach is not as flexible as the Bayesian formulation of Chipman et al. (1997), which, through different prior specifications, can identify a model that does not satisfy heredity with strong support from the data. But our approach is faster than that of Chipman et al. (1997). Moreover, our approach is more user friendly in that it does not need sophisticated prior elicitation and avoids convergence issues of the Gibbs sampling.

6. DISCUSSION

Because of the large number of candidate variables, it is imperative to use an efficient variable selection algorithm for the analysis of experiments. The LARS algorithm is a good choice. But because the effects in experiments are related due to the presence of polynomial and interaction terms, the ordinary application of LARS may lead to models that are not interpretable. To overcome this problem, we have proposed a novel extension of the LARS algorithm that incorporates the effect heredity principles. Two versions of the algorithm, weak and strong heredity, have been presented. The proposed algorithms are computationally efficient and are able to select models that better satisfy the goals of the experiment.

We have demonstrated the advantages of the new algorithm by analyzing a wide range of experimental designs. In some cases the weak heredity version performed better, whereas in other cases the strong heredity version performed better. In practice, we do not know which version to use. Therefore, our recommendation is to apply both and select the best one based on the solution paths generated by them.

The analysis of the 2^{9-5} fractional factorial design reiterated the importance of using heredity principle in the analysis of experiments. The ordinary LARS algorithm produced a set of aliased effects that could not be distinguished; in contrast, our proposed approach could identify a unique model. Ambiguities are possible with the application of our algorithm, but the likelihood is much lower.

It is important to be able to select the final model after a solution path is constructed using the proposed method. This is commonly done by minimizing the unbiased risk estimate

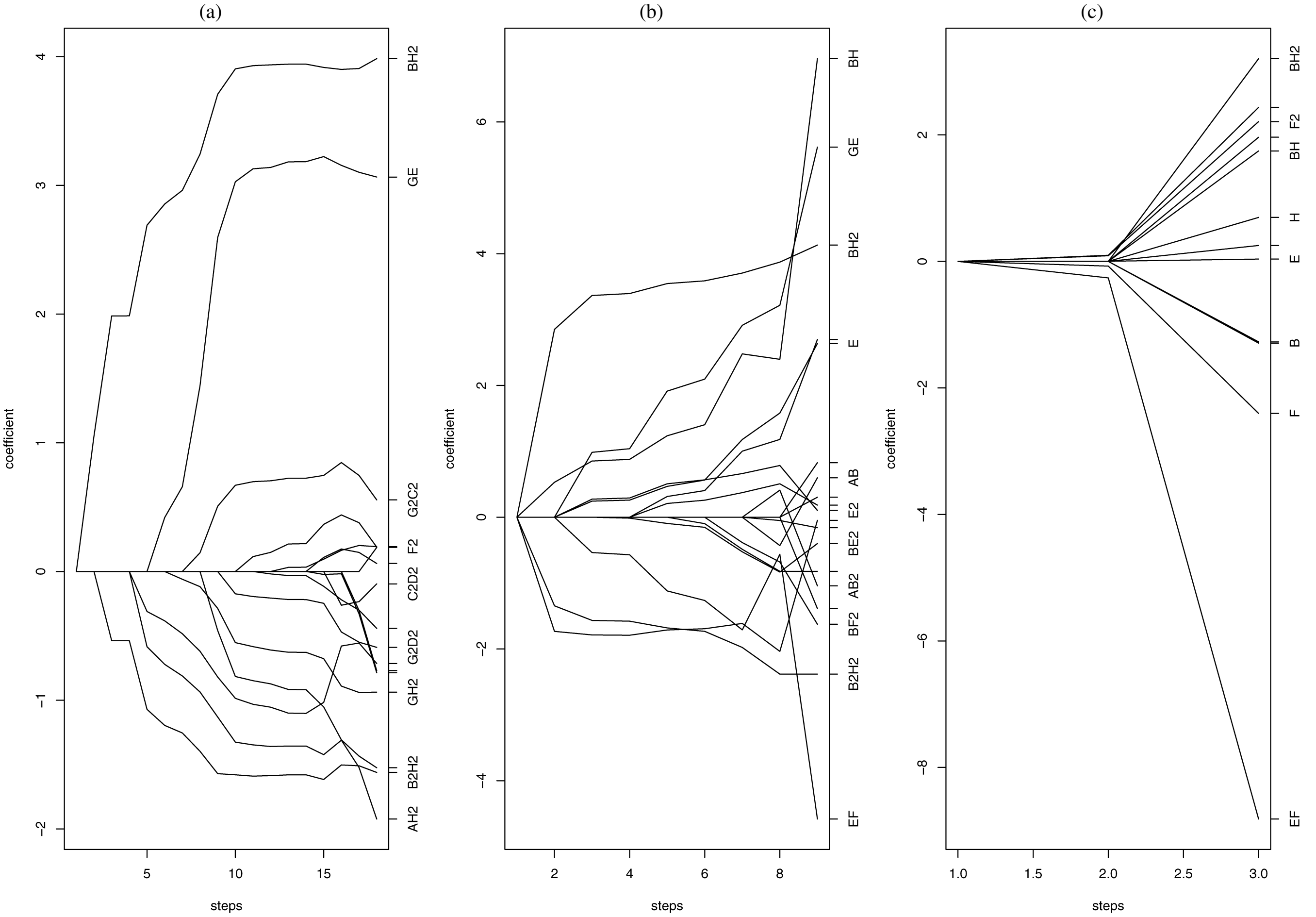

Figure 5. Solution paths for the blood glucose experiment. (a) No heredity; (b) weak heredity; (c) strong heredity.

or its proxies, such as the cross-validation score. In principle, these approaches can be applied to our methods as well. But the limited run size often makes these automatic selection methods questionable in practice. To illustrate the difficulty, again consider the simulation example for which we know the true model. The top panels of Figure 6 give the leave-out-one cross-validation scores together with their error bars (± one standard deviation). The horizontal axis represents the fraction of movement, defined as $\sum_j \int^t |\beta_j'(s)|\,ds / \sum_j \int^\infty |\beta_j'(s)|\,ds$. Clearly, all indicate that either a null mode or a model with 12 effects should be chosen, which certainly is not true. A practical solution is to look at the solution path, as we illustrated in the examples of the previous section. One may wish to choose paths in which a small number of coefficients increase quickly at the early stages, and the optimal model may be the point at which the increase slows significantly. According to our experience, this simple strategy works very well in practice. Other practical approaches also can be taken. For example, one may consider using the so-called "one standard error rule" (Breiman, Friedman, Stone, and Olshen 1984), where instead of choosing the model that minimizes the cross-validation score, one chooses the simplest model with a cross-validation score within one standard error from the smallest. Alternatively, one may use criteria that put more penalty on complicated models than the cross-validation. We leave a more thorough and rigorous investigation for future research.

Finally, we want to point out that the techniques developed here apply to the general linear regression variable selection problems. We focused on the analysis of designed experiments here only because effect heredity is most commonly applied in this context.

ACKNOWLEDGMENT

The authors thank the associate editor and three referees for their insightful comments, which greatly improved the manuscript. Yuan's research was supported in part by National Science Foundation (NSF) grant DMS-06-24841. Joseph's research was supported in part by NSF grant DMI-0448774. Lin's research was supported in part by NSF grant DMS-01-34987.

[Received December 2005. Revised February 2007.]

REFERENCES

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

——— (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350–2383.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984), *Classification and Regression Trees*, New York: Chapman & Hall/CRC.

Chipman, H. (1996), "Bayesian Variable Selection With Related Predictors," *Canadian Journal of Statistics*, 24, 17–36.

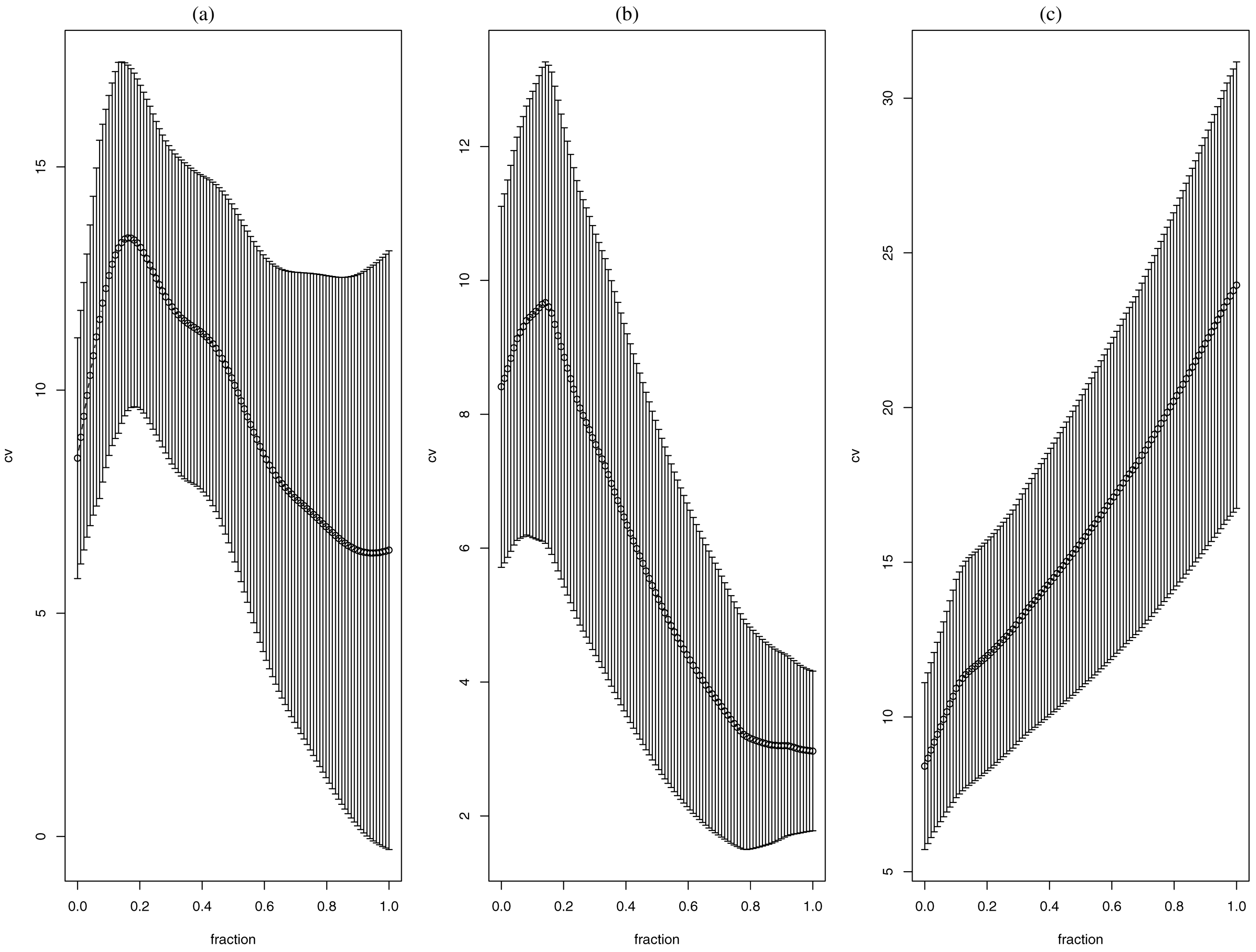

Figure 6. Failure of the traditional method in selecting the final model. (a) No heredity; (b) weak heredity; (c) strong heredity.

Chipman, H., Hamada, M., and Wu, C. F. J. (1997), "A Bayesian Variable Selection Approach for Analyzing Designed Experiments With Complex Aliasing," *Technometrics*, 39, 372–381.

Cormen, T., Leiserson, C., Rivest, R. L., and Rivest, R. (1990), *Introduction to Algorithms*, New York: McGraw-Hill.

Efron, B., Johnstone, I., Hastie, T., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499.

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

Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

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

Hamada, M., and Wu, C. F. J. (1992), "Analysis of Designed Experiments With Complex Aliasing," *Journal of Quality Technology*, 24, 130–137.

Joseph, V. R. (2006), "A Bayesian Approach to the Design and Analysis of Fractionated Experiments," *Technometrics*, 48, 219–229.

Li, R., and Lin, D. (2002), "Data Analysis in Supersaturated Designs," *Statistics and Probability Letters*, 59, 135–144.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

Meyer, R. D., Steinberg, D. M., and Box, G. (1996), "Follow-up Designs to Resolve Confounding in Multifactor Experiments" (with discussion), *Technometrics*, 38, 303–332.

Nelder, J. (1977), "A Reformulation of Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 140, 48–77.

——— (1994), "The Statistics of Linear Models," *Statistics and Computing*, 4, 221–234.

——— (1998), "The Selection of Terms in Response-Surface Models—How Strong Is the Weak Heredity Principle?" *The American Statistician*, 52, 315–318.

Osborne, M., Presnell, B., and Turlach, B. (2000), "On the LASSO and Its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337.

Raghavarao, D., and Altan, S. (2003), "A Heuristic Analysis of Highly Fractionated 2^n Factorial Experiments," *Metrika*, 58, 185–191.

Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221.

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

Turlach, B. (2004), Discussion of "Least Angle Regression," by B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani, *The Annals of Statistics*, 32, 481–490.

Westfall, P. H., Young, S., and Lin, D. (1998), "Forward Selection Error Control in Analysis of Supersaturated Designs," *Statistica Sinica*, 8, 101–117.

Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Yuan, M., and Lin, Y. (2005), "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models," *Journal of the American Statistical Association*, 100, 1215–1225.

——— (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Ser. B, 68, 49–67.