The Annals of Statistics
2004, Vol. 32, No. 2, 407–499

LEAST ANGLE REGRESSION

BY BRADLEY EFRON,[1] TREVOR HASTIE,[2] IAIN JOHNSTONE[3] AND ROBERT TIBSHIRANI[4]

Stanford University

The purpose of model selection algorithms such as *All Subsets*, *Forward Selection* and *Backward Elimination* is to choose a linear model on the basis of the same set of data to which the model will be applied. Typically we have available a large collection of possible covariates from which we hope to select a parsimonious set for the efficient prediction of a response variable. *Least Angle Regression* (LARS), a new model selection algorithm, is a useful and less greedy version of traditional forward selection methods. Three main properties are derived: (1) A simple modification of the LARS algorithm implements the Lasso, an attractive version of ordinary least squares that constrains the sum of the absolute regression coefficients; the LARS modification calculates all possible Lasso estimates for a given problem, using an order of magnitude less computer time than previous methods. (2) A different LARS modification efficiently implements Forward Stagewise linear regression, another promising new model selection method; this connection explains the similar numerical results previously observed for the Lasso and Stagewise, and helps us understand the properties of both methods, which are seen as constrained versions of the simpler LARS algorithm. (3) A simple approximation for the degrees of freedom of a LARS estimate is available, from which we derive a C_p estimate of prediction error; this allows a principled choice among the range of possible LARS estimates. LARS and its variants are computationally efficient: the paper describes a publicly available algorithm that requires only the same order of magnitude of computational effort as ordinary least squares applied to the full set of covariates.

1. Introduction. Automatic model-building algorithms are familiar, and sometimes notorious, in the linear model literature: Forward Selection, Backward Elimination, All Subsets regression and various combinations are used to automatically produce "good" linear models for predicting a response y on the basis of some measured covariates $x_1, x_2, \ldots, x_m$. Goodness is often defined in terms of prediction accuracy, but parsimony is another important criterion: simpler models are preferred for the sake of scientific insight into the $x - y$ relationship. Two promising recent model-building algorithms, the Lasso and Forward Stagewise lin-

Received March 2002; revised January 2003.
[1]Supported in part by NSF Grant DMS-00-72360 and NIH Grant 8R01-EB002784.
[2]Supported in part by NSF Grant DMS-02-04162 and NIH Grant R01-EB0011988-08.
[3]Supported in part by NSF Grant DMS-00-72661 and NIH Grant R01-EB001988-08.
[4]Supported in part by NSF Grant DMS-99-71405 and NIH Grant 2R01-CA72028.
AMS 2000 subject classification. 62J07.
Key words and phrases. Lasso, boosting, linear regression, coefficient paths, variable selection.

ear regression, will be discussed here, and motivated in terms of a computationally simpler method called Least Angle Regression.

Least Angle Regression (LARS) relates to the classic model-selection method known as Forward Selection, or "forward stepwise regression," described in Weisberg [(1980), Section 8.5]: given a collection of possible predictors, we select the one having largest absolute correlation with the response y, say x_{j_1}, and perform simple linear regression of y on x_{j_1}. This leaves a residual vector orthogonal to x_{j_1}, now considered to be the response. We project the other predictors orthogonally to x_{j_1} and repeat the selection process. After k steps this results in a set of predictors $x_{j_1}, x_{j_2}, \ldots, x_{j_k}$ that are then used in the usual way to construct a k-parameter linear model. Forward Selection is an aggressive fitting technique that can be overly greedy, perhaps eliminating at the second step useful predictors that happen to be correlated with x_{j_1}.

Forward Stagewise, as described below, is a much more cautious version of Forward Selection, which may take thousands of tiny steps as it moves toward a final model. It turns out, and this was the original motivation for the LARS algorithm, that a simple formula allows Forward Stagewise to be implemented using fairly large steps, though not as large as a classic Forward Selection, greatly reducing the computational burden. The geometry of the algorithm, described in Section 2, suggests the name "Least Angle Regression." It then happens that this same geometry applies to another, seemingly quite different, selection method called the Lasso [Tibshirani (1996)]. The LARS–Lasso–Stagewise connection is conceptually as well as computationally useful. The Lasso is described next, in terms of the main example used in this paper.

Table 1 shows a small part of the data for our main example.

Ten baseline variables, age, sex, body mass index, average blood pressure and six blood serum measurements, were obtained for each of $n = 442$ diabetes

TABLE 1

Diabetes study: 442 *diabetes patients were measured on* 10 *baseline variables*; *a prediction model was desired for the response variable*, *a measure of disease progression one year after baseline*

	AGE	**SEX**	**BMI**	**BP**	**Serum measurements**						**Response**
Patient	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$	$\mathbf{x_6}$	$\mathbf{x_7}$	$\mathbf{x_8}$	$\mathbf{x_9}$	$\mathbf{x_{10}}$	**y**
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. The statisticians were asked to construct a model that predicted response y from covariates $x_1, x_2, \ldots, x_{10}$. Two hopes were evident here, that the model would produce accurate baseline predictions of response for future patients and that the form of the model would suggest which covariates were important factors in disease progression.

The Lasso is a constrained version of ordinary least squares (OLS). Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be n-vectors representing the covariates, $m = 10$ and $n = 442$ in the diabetes study, and let $\mathbf{y}$ be the vector of responses for the n cases. By location and scale transformations we can always assume that the covariates have been standardized to have mean 0 and unit length, and that the response has mean 0,

$$\text{(1.1)} \qquad \sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1 \qquad \text{for } j = 1, 2, \ldots, m.$$

This is assumed to be the case in the theory which follows, except that numerical results are expressed in the original units of the diabetes example.

A candidate vector of regression coefficients $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m)'$ gives prediction vector $\widehat{\boldsymbol{\mu}}$,

$$\text{(1.2)} \qquad \widehat{\boldsymbol{\mu}} = \sum_{j=1}^{m} \mathbf{x}_j \widehat{\beta}_j = X\widehat{\boldsymbol{\beta}} \qquad [X_{n\times m} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)]$$

with total squared error

$$\text{(1.3)} \qquad S(\widehat{\boldsymbol{\beta}}) = \|\mathbf{y} - \widehat{\boldsymbol{\mu}}\|^2 = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2.$$

Let $T(\widehat{\boldsymbol{\beta}})$ be the absolute norm of $\widehat{\boldsymbol{\beta}}$,

$$\text{(1.4)} \qquad T(\widehat{\boldsymbol{\beta}}) = \sum_{j=1}^{m} |\widehat{\beta}_j|.$$

The Lasso chooses $\widehat{\boldsymbol{\beta}}$ by minimizing $S(\widehat{\boldsymbol{\beta}})$ subject to a bound t on $T(\widehat{\boldsymbol{\beta}})$,

$$\text{(1.5)} \qquad \textit{Lasso}\text{: minimize } S(\widehat{\boldsymbol{\beta}}) \quad \text{subject to } T(\widehat{\boldsymbol{\beta}}) \le t.$$

Quadratic programming techniques can be used to solve (1.5) though we will present an easier method here, closely related to the "homotopy method" of Osborne, Presnell and Turlach (2000a).

The left panel of Figure 1 shows all Lasso solutions $\widehat{\boldsymbol{\beta}}(t)$ for the diabetes study, as t increases from 0, where $\widehat{\boldsymbol{\beta}} = 0$, to $t = 3460.00$, where $\widehat{\boldsymbol{\beta}}$ equals the OLS regression vector, the constraint in (1.5) no longer binding. We see that the Lasso tends to shrink the OLS coefficients toward 0, more so for small values of t. Shrinkage often improves prediction accuracy, trading off decreased variance for increased bias as discussed in Hastie, Tibshirani and Friedman (2001).

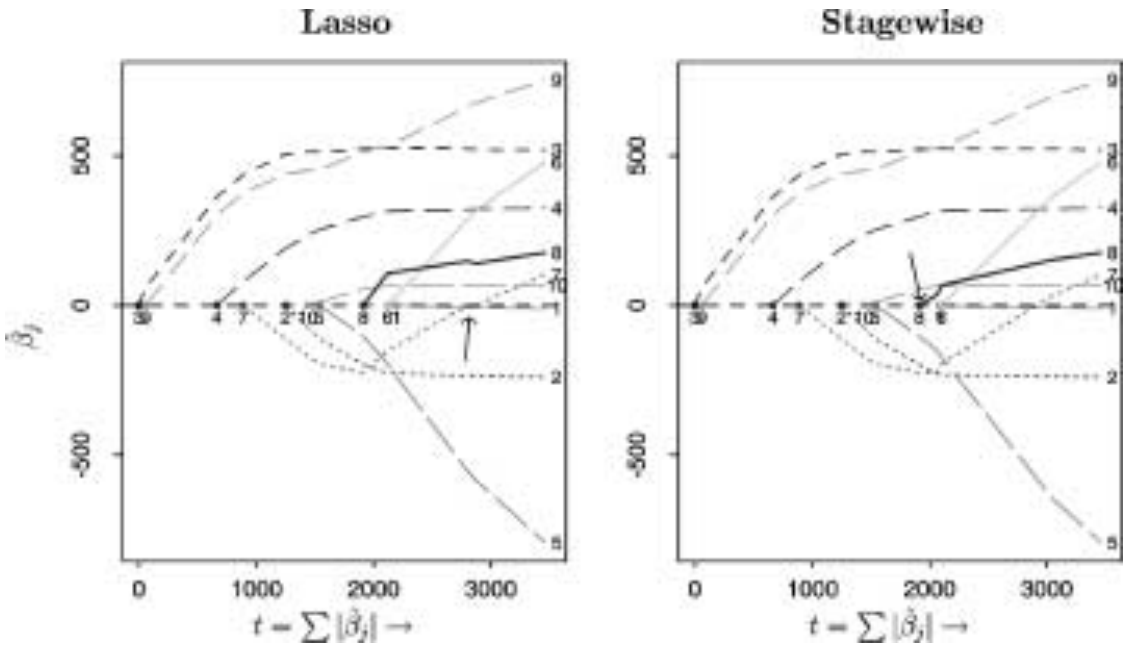

FIG. 1. *Estimates of regression coefficients* $\widehat{\beta}_j$, $j = 1, 2, \ldots, 10$, *for the diabetes study.* (Left panel) *Lasso estimates, as a function of* $t = \sum_j |\widehat{\beta}_j|$. *The covariates enter the regression equation sequentially as* t *increases, in order* $j = 3, 9, 4, 7, \ldots, 1$. (Right panel) *The same plot for Forward Stagewise Linear Regression. The two plots are nearly identical, but differ slightly for large* t *as shown in the track of covariate* 8.

The Lasso also has a parsimony property: for any given constraint value t, only a subset of the covariates have nonzero values of $\widehat{\beta}_j$. At $t = 1000$, for example, only variables 3, 9, 4 and 7 enter the Lasso regression model (1.2). If this model provides adequate predictions, a crucial question considered in Section 4, the statisticians could report these four variables as the important ones.

Forward Stagewise Linear Regression, henceforth called *Stagewise*, is an iterative technique that begins with $\widehat{\boldsymbol{\mu}} = 0$ and builds up the regression function in successive small steps. If $\widehat{\boldsymbol{\mu}}$ is the current Stagewise estimate, let $\mathbf{c}(\widehat{\boldsymbol{\mu}})$ be the vector of *current correlations*

$$\widehat{\mathbf{c}} = \mathbf{c}(\widehat{\boldsymbol{\mu}}) = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}}), \tag{1.6}$$

so that $\widehat{c}_j$ is proportional to the correlation between covariate x_j and the current residual vector. The next step of the Stagewise algorithm is taken in the direction of the greatest current correlation,

$$\widehat{j} = \arg\max |\widehat{c}_j| \quad \text{and} \quad \widehat{\boldsymbol{\mu}} \rightarrow \widehat{\boldsymbol{\mu}} + \varepsilon \cdot \operatorname{sign}(\widehat{c}_{\widehat{j}}) \cdot \mathbf{x}_{\widehat{j}}, \tag{1.7}$$

with ε some small constant. "Small" is important here: the "big" choice $\varepsilon = |\widehat{c}_{\widehat{j}}|$ leads to the classic Forward Selection technique, which can be overly greedy, impulsively eliminating covariates which are correlated with $x_{\widehat{j}}$. The Stagewise procedure is related to boosting and also to Friedman's MART algorithm

[Friedman (2001)]; see Section 8, as well as Hastie, Tibshirani and Friedman [(2001), Chapter 10 and Algorithm 10.4].

The right panel of Figure 1 shows the coefficient plot for Stagewise applied to the diabetes data. The estimates were built up in 6000 Stagewise steps [making ε in (1.7) small enough to conceal the "Etch-a-Sketch" staircase seen in Figure 2, Section 2]. The striking fact is the similarity between the Lasso and Stagewise estimates. Although their definitions look completely different, the results are nearly, *but not exactly*, identical.

The main point of this paper is that both Lasso and Stagewise are variants of a basic procedure called Least Angle Regression, abbreviated LARS (the "S" suggesting "Lasso" and "Stagewise"). Section 2 describes the LARS algorithm while Section 3 discusses modifications that turn LARS into Lasso or Stagewise, reducing the computational burden by at least an order of magnitude for either one. Sections 5 and 6 verify the connections stated in Section 3.

Least Angle Regression is interesting in its own right, its simple structure lending itself to inferential analysis. Section 4 analyzes the "degrees of freedom" of a LARS regression estimate. This leads to a C_p type statistic that suggests which estimate we should prefer among a collection of possibilities like those in Figure 1. A particularly simple C_p approximation, requiring no additional computation beyond that for the $\widehat{\boldsymbol{\beta}}$ vectors, is available for LARS.

Section 7 briefly discusses computational questions. An efficient *S* program for all three methods, LARS, Lasso and Stagewise, is available. Section 8 elaborates on the connections with boosting.

2. The LARS algorithm. Least Angle Regression is a stylized version of the Stagewise procedure that uses a simple mathematical formula to accelerate the computations. Only m steps are required for the full set of solutions, where m is the number of covariates: $m = 10$ in the diabetes example compared to the 6000 steps used in the right panel of Figure 1. This section describes the LARS algorithm. Modifications of LARS that produce Lasso and Stagewise solutions are discussed in Section 3, and verified in Sections 5 and 6. Section 4 uses the simple structure of LARS to help analyze its estimation properties.

The LARS procedure works roughly as follows. As with classic Forward Selection, we start with all coefficients equal to zero, and find the predictor most correlated with the response, say x_{j_1}. We take the largest step possible in the direction of this predictor until some other predictor, say x_{j_2}, has as much correlation with the current residual. At this point LARS parts company with Forward Selection. Instead of continuing along x_{j_1}, LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j_3} earns its way into the "most correlated" set. LARS then proceeds equiangularly between x_{j_1}, x_{j_2} and x_{j_3}, that is, along the "least angle direction," until a fourth variable enters, and so on.

The remainder of this section describes the algebra necessary to execute the equiangular strategy. As usual the algebraic details look more complicated than the simple underlying geometry, but they lead to the highly efficient computational algorithm described in Section 7.

LARS builds up estimates $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$, (1.2), in successive steps, each step adding one covariate to the model, so that after k steps just k of the $\widehat{\beta}_j$'s are nonzero. Figure 2 illustrates the algorithm in the situation with $m = 2$ covariates, $X = (\mathbf{x}_1, \mathbf{x}_2)$. In this case the current correlations (1.6) depend only on the projection $\bar{\mathbf{y}}_2$ of $\mathbf{y}$ into the linear space $\mathcal{L}(X)$ spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$\mathbf{c}(\widehat{\boldsymbol{\mu}}) = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}}) = X'(\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}). \tag{2.1}$$

The algorithm begins at $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}$ [remembering that the response has had its mean subtracted off, as in (1.1)]. Figure 2 has $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_0$ making a smaller angle with $\mathbf{x}_1$ than $\mathbf{x}_2$, that is, $c_1(\widehat{\boldsymbol{\mu}}_0) > c_2(\widehat{\boldsymbol{\mu}}_0)$. LARS then augments $\widehat{\boldsymbol{\mu}}_0$ in the direction of $\mathbf{x}_1$, to

$$\widehat{\boldsymbol{\mu}}_1 = \widehat{\boldsymbol{\mu}}_0 + \widehat{\gamma}_1 \mathbf{x}_1. \tag{2.2}$$

Stagewise would choose $\widehat{\gamma}_1$ equal to some small value ε, and then repeat the process many times. Classic Forward Selection would take $\widehat{\gamma}_1$ large enough to make $\widehat{\boldsymbol{\mu}}_1$ equal $\bar{\mathbf{y}}_1$, the projection of $\mathbf{y}$ into $\mathcal{L}(\mathbf{x}_1)$. LARS uses an intermediate value of $\widehat{\gamma}_1$, the value that makes $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}$, *equally* correlated with $\mathbf{x}_1$ and $\mathbf{x}_2$; that is, $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$, so $c_1(\widehat{\boldsymbol{\mu}}_1) = c_2(\widehat{\boldsymbol{\mu}}_1)$.

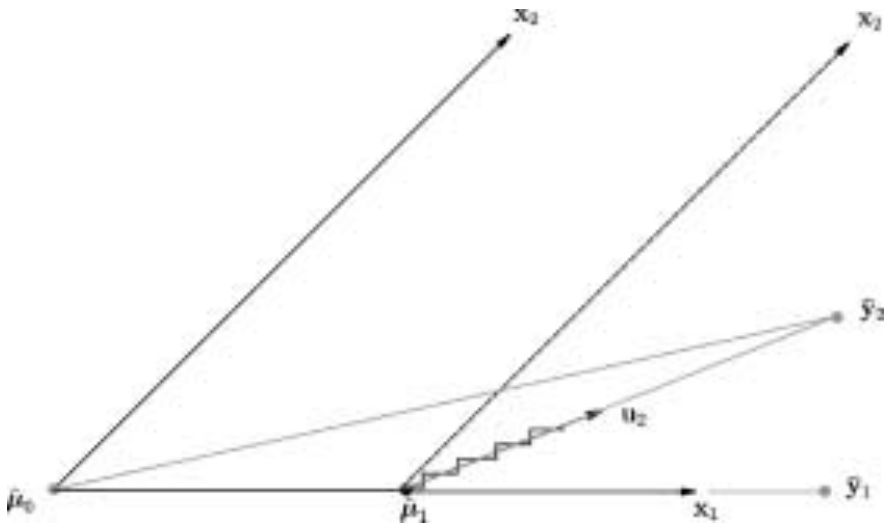

FIG. 2. *The LARS algorithm in the case of $m = 2$ covariates; $\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$. Beginning at $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}$, the residual vector $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_0$ has greater correlation with $\mathbf{x}_1$ than $\mathbf{x}_2$; the next LARS estimate is $\widehat{\boldsymbol{\mu}}_1 = \widehat{\boldsymbol{\mu}}_0 + \widehat{\gamma}_1 \mathbf{x}_1$, where $\widehat{\gamma}_1$ is chosen such that $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$; then $\widehat{\boldsymbol{\mu}}_2 = \widehat{\boldsymbol{\mu}}_1 + \widehat{\gamma}_2 \mathbf{u}_2$, where $\mathbf{u}_2$ is the unit bisector; $\widehat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$ in the case $m = 2$, but not for the case $m > 2$; see Figure 4. The staircase indicates a typical Stagewise path. Here LARS gives the Stagewise track as $\varepsilon \to 0$, but a modification is necessary to guarantee agreement in higher dimensions; see Section 3.2.*

Let $\mathbf{u}_2$ be the unit vector lying along the bisector. The next LARS estimate is

$$\widehat{\boldsymbol{\mu}}_2 = \widehat{\boldsymbol{\mu}}_1 + \widehat{\gamma}_2 \mathbf{u}_2, \tag{2.3}$$

with $\widehat{\gamma}_2$ chosen to make $\widehat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$ in the case $m = 2$. With $m > 2$ covariates, $\widehat{\gamma}_2$ would be smaller, leading to another change of direction, as illustrated in Figure 4. The "staircase" in Figure 2 indicates a typical Stagewise path. LARS is motivated by the fact that it is easy to calculate the step sizes $\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots$ theoretically, short-circuiting the small Stagewise steps.

Subsequent LARS steps, beyond two covariates, are taken along *equiangular vectors*, generalizing the bisector $\mathbf{u}_2$ in Figure 2. *We assume that the covariate vectors* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ *are linearly independent*. For $\mathcal{A}$ a subset of the indices $\{1, 2, \ldots, m\}$, define the matrix

$$X_{\mathcal{A}} = (\cdots s_j \mathbf{x}_j \cdots)_{j \in \mathcal{A}}, \tag{2.4}$$

where the signs s_j equal ± 1. Let

$$\mathcal{G}_{\mathcal{A}} = X'_{\mathcal{A}} X_{\mathcal{A}} \quad \text{and} \quad A_{\mathcal{A}} = (\mathbf{1}'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2}, \tag{2.5}$$

$\mathbf{1}_{\mathcal{A}}$ being a vector of 1's of length equaling $|\mathcal{A}|$, the size of $\mathcal{A}$. The

$$\textit{equiangular vector} \quad \mathbf{u}_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}} \qquad \text{where } w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}, \tag{2.6}$$

is the unit vector making equal angles, less than 90°, with the columns of $X_{\mathcal{A}}$,

$$X'_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} \quad \text{and} \quad \|\mathbf{u}_{\mathcal{A}}\|^2 = 1. \tag{2.7}$$

We can now fully describe the LARS algorithm. As with the Stagewise procedure we begin at $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and build up $\widehat{\boldsymbol{\mu}}$ by steps, larger steps in the LARS case. Suppose that $\widehat{\boldsymbol{\mu}}_{\mathcal{A}}$ is the current LARS estimate and that

$$\widehat{\mathbf{c}} = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{\mathcal{A}}) \tag{2.8}$$

is the vector of current correlations (1.6). The *active set* $\mathcal{A}$ is the set of indices corresponding to covariates with the greatest absolute current correlations,

$$\widehat{C} = \max_j \{|\widehat{c}_j|\} \quad \text{and} \quad \mathcal{A} = \{j : |\widehat{c}_j| = \widehat{C}\}. \tag{2.9}$$

Letting

$$s_j = \operatorname{sign}\{\widehat{c}_j\} \qquad \text{for } j \in \mathcal{A}, \tag{2.10}$$

we compute $X_{\mathcal{A}}$, $A_{\mathcal{A}}$ and $\mathbf{u}_{\mathcal{A}}$ as in (2.4)–(2.6), and also the inner product vector

$$\mathbf{a} \equiv X' \mathbf{u}_{\mathcal{A}}. \tag{2.11}$$

Then the next step of the LARS algorithm updates $\widehat{\boldsymbol{\mu}}_{\mathcal{A}}$, say to

$$\widehat{\boldsymbol{\mu}}_{\mathcal{A}_+} = \widehat{\boldsymbol{\mu}}_{\mathcal{A}} + \widehat{\gamma} \mathbf{u}_{\mathcal{A}}, \tag{2.12}$$

where

$$\widehat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\widehat{C} - \widehat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\widehat{C} + \widehat{c}_j}{A_{\mathcal{A}} + a_j} \right\}; \tag{2.13}$$

"$\min^+$" indicates that the minimum is taken over only positive components within each choice of j in (2.13).

Formulas (2.12) and (2.13) have the following interpretation: define

$$\boldsymbol{\mu}(\gamma) = \widehat{\boldsymbol{\mu}}_{\mathcal{A}} + \gamma \mathbf{u}_{\mathcal{A}}, \tag{2.14}$$

for $\gamma > 0$, so that the current correlation

$$c_j(\gamma) = \mathbf{x}_j'(\mathbf{y} - \boldsymbol{\mu}(\gamma)) = \widehat{c}_j - \gamma a_j. \tag{2.15}$$

For $j \in \mathcal{A}$, (2.7)–(2.9) yield

$$|c_j(\gamma)| = \widehat{C} - \gamma A_{\mathcal{A}}, \tag{2.16}$$

showing that all of the maximal absolute current correlations decline equally. For $j \in \mathcal{A}^c$, equating (2.15) with (2.16) shows that $c_j(\gamma)$ equals the maximal value at $\gamma = (\widehat{C} - \widehat{c}_j)/(A_{\mathcal{A}} - a_j)$. Likewise $-c_j(\gamma)$, the current correlation for the reversed covariate $-\mathbf{x}_j$, achieves maximality at $(\widehat{C} + \widehat{c}_j)/(A_{\mathcal{A}} + a_j)$. Therefore $\widehat{\gamma}$ *in* (2.13) *is the smallest positive value of* γ *such that some new index* $\widehat{j}$ *joins the active set*; $\widehat{j}$ is the minimizing index in (2.13), and the new active set $\mathcal{A}_+$ is $\mathcal{A} \cup \{\widehat{j}\}$; the new maximum absolute correlation is $\widehat{C}_+ = \widehat{C} - \widehat{\gamma} A_{\mathcal{A}}$.

Figure 3 concerns the LARS analysis of the diabetes data. The complete algorithm required only $m = 10$ steps of procedure (2.8)–(2.13), with the variables

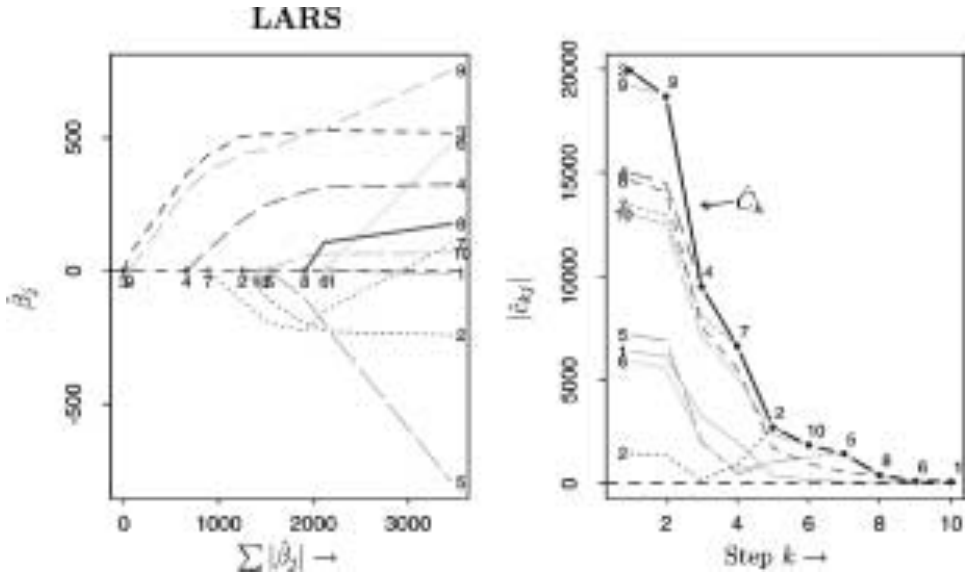

FIG. 3. *LARS analysis of the diabetes study*: (left) *estimates of regression coefficients* $\widehat{\beta}_j$, $j = 1, 2, \ldots, 10$; *plotted versus* $\sum |\widehat{\beta}_j|$; *plot is slightly different than either Lasso or Stagewise*, *Figure* 1; (right) *absolute current correlations as function of LARS step*; *variables enter active set* (2.9) *in order* $3, 9, 4, 7, \ldots, 1$; *heavy curve shows maximum current correlation* $\widehat{C}_k$ *declining with* k.

joining the active set $\mathcal{A}$ in the same order as for the Lasso: 3, 9, 4, 7, ..., 1. Tracks of the regression coefficients $\widehat{\beta}_j$ are nearly but not exactly the same as either the Lasso or Stagewise tracks of Figure 1.

The right panel shows the absolute current correlations

$$|\widehat{c}_{kj}| = |\mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1})| \tag{2.17}$$

for variables $j = 1, 2, \ldots, 10$, as a function of the LARS step k. The maximum correlation

$$\widehat{C}_k = \max\{|\widehat{c}_{kj}|\} = \widehat{C}_{k-1} - \widehat{\gamma}_{k-1} A_{k-1} \tag{2.18}$$

declines with k, as it must. At each step a new variable j joins the active set, henceforth having $|\widehat{c}_{kj}| = \widehat{C}_k$. The sign s_j of each $\mathbf{x}_j$ in (2.4) stays constant as the active set increases.

Section 4 makes use of the relationship between Least Angle Regression and Ordinary Least Squares illustrated in Figure 4. Suppose LARS has just completed step $k-1$, giving $\widehat{\boldsymbol{\mu}}_{k-1}$, and is embarking upon step k. The active set $\mathcal{A}_k$, (2.9), will have k members, giving X_k, $\mathcal{G}_k$, A_k and $\mathbf{u}_k$ as in (2.4)–(2.6) (here replacing subscript $\mathcal{A}$ with "k"). Let $\bar{\mathbf{y}}_k$ indicate the projection of $\mathbf{y}$ into $\mathcal{L}(X_k)$, which, since $\widehat{\boldsymbol{\mu}}_{k-1} \in \mathcal{L}(X_{k-1})$, is

$$\bar{\mathbf{y}}_k = \widehat{\boldsymbol{\mu}}_{k-1} + X_k \mathcal{G}_k^{-1} X_k'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1}) = \widehat{\boldsymbol{\mu}}_{k-1} + \frac{\widehat{C}_k}{A_k}\,\mathbf{u}_k, \tag{2.19}$$

the last equality following from (2.6) and the fact that the signed current correlations in $\mathcal{A}_k$ all equal $\widehat{C}_k$,

$$X_k'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1}) = \widehat{C}_k \mathbf{1}_{\mathcal{A}}. \tag{2.20}$$

Since $\mathbf{u}_k$ is a unit vector, (2.19) says that $\bar{\mathbf{y}}_k - \widehat{\boldsymbol{\mu}}_{k-1}$ has length

$$\bar{\gamma}_k \equiv \frac{\widehat{C}_k}{A_k}. \tag{2.21}$$

Comparison with (2.12) shows that the LARS estimate $\widehat{\boldsymbol{\mu}}_k$ lies on the line

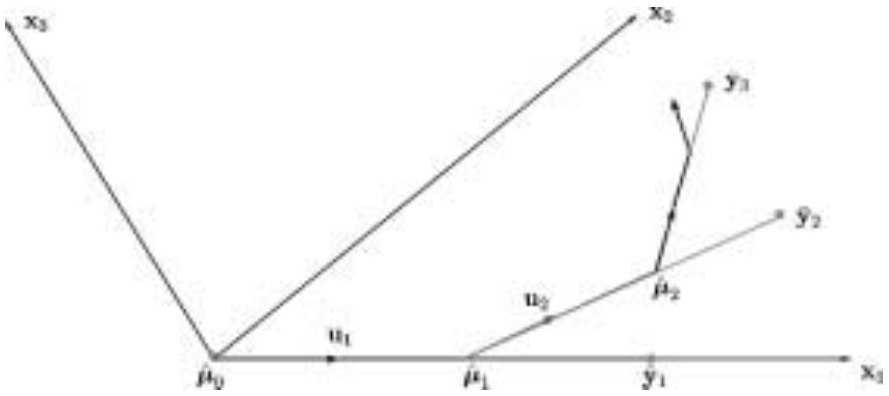

FIG. 4. *At each stage the LARS estimate $\widehat{\boldsymbol{\mu}}_k$ approaches, but does not reach, the corresponding OLS estimate $\bar{\mathbf{y}}_k$.*

from $\widehat{\boldsymbol{\mu}}_{k-1}$ to $\bar{\mathbf{y}}_k$,

$$\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_{k-1} = \frac{\widehat{\gamma}_k}{\bar{\gamma}_k}\,(\bar{\mathbf{y}}_k - \widehat{\boldsymbol{\mu}}_{k-1}). \tag{2.22}$$

It is easy to see that $\widehat{\gamma}_k$, (2.12), is always less than $\bar{\gamma}_k$, so that $\widehat{\boldsymbol{\mu}}_k$ lies closer than $\bar{\mathbf{y}}_k$ to $\widehat{\boldsymbol{\mu}}_{k-1}$. Figure 4 shows the successive LARS estimates $\widehat{\boldsymbol{\mu}}_k$ always approaching but never reaching the OLS estimates $\bar{\mathbf{y}}_k$.

The exception is at the last stage: since $\mathcal{A}_m$ contains all covariates, (2.13) is not defined. By convention the algorithm takes $\widehat{\gamma}_m = \bar{\gamma}_m = \widehat{C}_m / A_m$, making $\widehat{\boldsymbol{\mu}}_m = \bar{\mathbf{y}}_m$ and $\widehat{\boldsymbol{\beta}}_m$ equal the OLS estimate for the full set of m covariates.

The LARS algorithm is computationally thrifty. Organizing the calculations correctly, the computational cost for the entire m steps is of the same order as that required for the usual Least Squares solution for the full set of m covariates. Section 7 describes an efficient LARS program available from the authors. With the modifications described in the next section, this program also provides economical Lasso and Stagewise solutions.

3. Modified versions of Least Angle Regression. Figures 1 and 3 show Lasso, Stagewise and LARS yielding remarkably similar estimates for the diabetes data. The similarity is no coincidence. This section describes simple modifications of the LARS algorithm that produce Lasso or Stagewise estimates. Besides improved computational efficiency, these relationships elucidate the methods' rationale: all three algorithms can be viewed as moderately greedy forward stepwise procedures whose forward progress is determined by compromise among the currently most correlated covariates. LARS moves along the most obvious compromise direction, the equiangular vector (2.6), while Lasso and Stagewise put some restrictions on the equiangular strategy.

3.1. *The LARS–Lasso relationship*. The full set of Lasso solutions, as shown for the diabetes study in Figure 1, can be generated by a minor modification of the LARS algorithm (2.8)–(2.13). Our main result is described here and verified in Section 5. It closely parallels the homotopy method in the papers by Osborne, Presnell and Turlach (2000a, b), though the LARS approach is somewhat more direct.

Let $\widehat{\boldsymbol{\beta}}$ be a Lasso solution (1.5), with $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$. Then it is easy to show that the sign of any nonzero coordinate $\widehat{\beta}_j$ must agree with the sign s_j of the current correlation $\widehat{c}_j = \mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$,

$$\operatorname{sign}(\widehat{\beta}_j) = \operatorname{sign}(\widehat{c}_j) = s_j; \tag{3.1}$$

see Lemma 8 of Section 5. The LARS algorithm does not enforce restriction (3.1), but it can easily be modified to do so.

Suppose we have just completed a LARS step, giving a new active set $\mathcal{A}$ as in (2.9), and that the corresponding LARS estimate $\widehat{\boldsymbol{\mu}}_{\mathcal{A}}$ corresponds to a Lasso solution $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$. Let

$$w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}, \tag{3.2}$$

a vector of length the size of $\mathcal{A}$, and (somewhat abusing subscript notation) define $\widehat{\mathbf{d}}$ to be the m-vector equaling $s_j w_{\mathcal{A}j}$ for $j \in \mathcal{A}$ and zero elsewhere. Moving in the positive γ direction along the LARS line (2.14), we see that

$$\boldsymbol{\mu}(\gamma) = X\boldsymbol{\beta}(\gamma), \qquad \text{where } \beta_j(\gamma) = \widehat{\beta}_j + \gamma \widehat{d}_j \tag{3.3}$$

for $j \in \mathcal{A}$. Therefore $\beta_j(\gamma)$ will change sign at

$$\gamma_j = -\widehat{\beta}_j / \widehat{d}_j, \tag{3.4}$$

the first such change occurring at

$$\widetilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\}, \tag{3.5}$$

say for covariate $x_{\tilde{j}}$; $\widetilde{\gamma}$ equals infinity by definition if there is no $\gamma_j > 0$.

If $\widetilde{\gamma}$ is less than $\widehat{\gamma}$, (2.13), then $\beta_j(\gamma)$ cannot be a Lasso solution for $\gamma > \widetilde{\gamma}$ since the sign restriction (3.1) must be violated: $\beta_{\tilde{j}}(\gamma)$ has changed sign while $c_{\tilde{j}}(\gamma)$ has not. [The continuous function $c_{\tilde{j}}(\gamma)$ cannot change sign within a single LARS step since $|c_{\tilde{j}}(\gamma)| = \widehat{C} - \gamma A_{\mathcal{A}} > 0$, (2.16).]

LASSO MODIFICATION. If $\widetilde{\gamma} < \widehat{\gamma}$, stop the ongoing LARS step at $\gamma = \widetilde{\gamma}$ and remove $\widetilde{j}$ from the calculation of the next equiangular direction. That is,

$$\widehat{\boldsymbol{\mu}}_{\mathcal{A}_+} = \widehat{\boldsymbol{\mu}}_{\mathcal{A}} + \widetilde{\gamma}\mathbf{u}_{\mathcal{A}} \quad \text{and} \quad \mathcal{A}_+ = \mathcal{A} - \{\widetilde{j}\} \tag{3.6}$$

rather than (2.12).

THEOREM 1. *Under the Lasso modification, and assuming the "one at a time" condition discussed below, the LARS algorithm yields all Lasso solutions.*

The active sets $\mathcal{A}$ grow monotonically larger as the original LARS algorithm progresses, but the Lasso modification allows $\mathcal{A}$ to decrease. "One at a time" means that the increases and decreases never involve more than a single index j. This is the usual case for quantitative data and can always be realized by adding a little jitter to the y values. Section 5 discusses tied situations.

The Lasso diagram in Figure 1 was actually calculated using the modified LARS algorithm. Modification (3.6) came into play only once, at the arrowed point in the left panel. There $\mathcal{A}$ contained all 10 indices while $\mathcal{A}_+ = \mathcal{A} - \{7\}$. Variable 7 was restored to the active set one LARS step later, the next and last step then taking $\widehat{\boldsymbol{\beta}}$ all the way to the full OLS solution. The brief absence of variable 7 had an effect on the tracks of the others, noticeably $\widehat{\beta}_8$. The price of using Lasso instead of unmodified LARS comes in the form of added steps, 12 instead of 10 in this example. For the more complicated "quadratic model" of Section 4, the comparison was 103 Lasso steps versus 64 for LARS.

3.2. *The LARS–Stagewise relationship*. The staircase in Figure 2 indicates how the Stagewise algorithm might proceed forward from $\widehat{\boldsymbol{\mu}}_1$, a point of equal current correlations $\widehat{c}_1 = \widehat{c}_2$, (2.8). The first small step has (randomly) selected index $j = 1$, taking us to $\widehat{\boldsymbol{\mu}}_1 + \varepsilon \mathbf{x}_1$. Now variable 2 is more correlated,

$$\mathbf{x}_2'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_1 - \varepsilon \mathbf{x}_1) > \mathbf{x}_1'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_1 - \varepsilon \mathbf{x}_1), \tag{3.7}$$

forcing $j = 2$ to be the next Stagewise choice and so on.

We will consider an idealized Stagewise procedure in which the step size ε goes to zero. This collapses the staircase along the direction of the bisector $\mathbf{u}_2$ in Figure 2, making the Stagewise and LARS estimates agree. They always agree for $m = 2$ covariates, but another modification is necessary for LARS to produce Stagewise estimates in general. Section 6 verifies the main result described next.

Suppose that the Stagewise procedure has taken N steps of infinitesimal size ε from some previous estimate $\widehat{\boldsymbol{\mu}}$, with

$$N_j \equiv \#\{\text{steps with selected index } j\}, \qquad j = 1, 2, \ldots, m. \tag{3.8}$$

It is easy to show, as in Lemma 11 of Section 6, that $N_j = 0$ for j not in the active set $\mathcal{A}$ defined by the current correlations $\mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$, (2.9). Letting

$$P \equiv (N_1, N_2, \ldots, N_m)/N, \tag{3.9}$$

with $P_{\mathcal{A}}$ indicating the coordinates of P for $j \in \mathcal{A}$, the new estimate is

$$\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}} + N\varepsilon X_{\mathcal{A}} P_{\mathcal{A}} \qquad [(2.4)]. \tag{3.10}$$

(Notice that the Stagewise steps are taken along the directions $s_j \mathbf{x}_j$.)

The LARS algorithm (2.14) progresses along

$$\boldsymbol{\mu}_{\mathcal{A}} + \gamma X_{\mathcal{A}} w_{\mathcal{A}}, \qquad \text{where } w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} \qquad [(2.6)\text{–}(3.2)]. \tag{3.11}$$

Comparing (3.10) with (3.11) shows that LARS cannot agree with Stagewise if $w_{\mathcal{A}}$ has negative components, since $P_{\mathcal{A}}$ is nonnegative. To put it another way, the direction of Stagewise progress $X_{\mathcal{A}} P_{\mathcal{A}}$ must lie in the convex cone generated by the columns of $X_{\mathcal{A}}$,

$$\mathcal{C}_{\mathcal{A}} = \left\{ \mathbf{v} = \sum_{j \in \mathcal{A}} s_j \mathbf{x}_j P_j, \ P_j \geq 0 \right\}. \tag{3.12}$$

If $\mathbf{u}_{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}$ then there is no contradiction between (3.12) and (3.13). If not it seems natural to replace $\mathbf{u}_{\mathcal{A}}$ with its projection into $\mathcal{C}_{\mathcal{A}}$, that is, the nearest point in the convex cone.

STAGEWISE MODIFICATION. Proceed as in (2.8)–(2.13), except with $\mathbf{u}_{\mathcal{A}}$ replaced by $\mathbf{u}_{\hat{\mathcal{B}}}$, the unit vector lying along the projection of $\mathbf{u}_{\mathcal{A}}$ into $\mathcal{C}_{\mathcal{A}}$. (See Figure 9 in Section 6.)

THEOREM 2. *Under the Stagewise modification, the LARS algorithm yields all Stagewise solutions.*

The vector $\mathbf{u}_{\hat{\mathcal{B}}}$ in the Stagewise modification is the equiangular vector (2.6) for the subset $\widehat{\mathcal{B}} \subseteq \mathcal{A}$ corresponding to the face of $\mathcal{C}_{\mathcal{A}}$ into which the projection falls. Stagewise is a LARS type algorithm that allows the active set to decrease by one or more indices. This happened at the arrowed point in the right panel of Figure 1: there the set $\mathcal{A} = \{3, 9, 4, 7, 2, 10, 5, 8\}$ was decreased to $\widehat{\mathcal{B}} = \mathcal{A} - \{3, 7\}$. It took a total of 13 modified LARS steps to reach the full OLS solution $\bar{\boldsymbol{\beta}}_m = (X'X)^{-1}X'\mathbf{y}$. The three methods, LARS, Lasso and Stagewise, always reach OLS eventually, but LARS does so in only m steps while Lasso and, especially, Stagewise can take longer. For the $m = 64$ quadratic model of Section 4, Stagewise took 255 steps.

According to Theorem 2 the difference between successive Stagewise–modified LARS estimates is

$$\widehat{\boldsymbol{\mu}}_{\mathcal{A}_+} - \widehat{\boldsymbol{\mu}}_{\mathcal{A}} = \widehat{\gamma}\mathbf{u}_{\hat{\mathcal{B}}} = \widehat{\gamma}X_{\hat{\mathcal{B}}}w_{\hat{\mathcal{B}}}, \tag{3.13}$$

as in (3.13). Since $\mathbf{u}_{\hat{\mathcal{B}}}$ exists in the convex cone $\mathcal{C}_{\mathcal{A}}$, $w_{\hat{\mathcal{B}}}$ must have nonnegative components. This says that the difference of successive coefficient estimates for coordinate $j \in \widehat{\mathcal{B}}$ satisfies

$$\operatorname{sign}(\widehat{\beta}_{+j} - \widehat{\beta}_j) = s_j, \tag{3.14}$$

where $s_j = \operatorname{sign}\{\mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}})\}$.

We can now make a useful comparison of the three methods:

1. *Stagewise*—successive differences of $\widehat{\beta}_j$ agree in sign with the current correlation $\widehat{c}_j = \mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$;
2. *Lasso*—$\widehat{\beta}_j$ agrees in sign with $\widehat{c}_j$;
3. *LARS*—no sign restrictions (but see Lemma 4 of Section 5).

From this point of view, Lasso is intermediate between the LARS and Stagewise methods.

The successive difference property (3.14) makes the Stagewise $\widehat{\beta}_j$ estimates move monotonically away from 0. Reversals are possible only if $\widehat{c}_j$ changes sign while $\widehat{\beta}_j$ is "resting" between two periods of change. This happened to variable 7 in Figure 1 between the 8th and 10th Stagewise-modified LARS steps.

3.3. *Simulation study.* A small simulation study was carried out comparing the LARS, Lasso and Stagewise algorithms. The X matrix for the simulation was based on the diabetes example of Table 1, but now using a "Quadratic Model" having $m = 64$ predictors, including interactions and squares of the 10 original covariates:

(3.15) *Quadratic Model* 10 main effects, 45 interactions, 9 squares,

the last being the squares of each $\mathbf{x}_j$ except the dichotomous variable $\mathbf{x}_2$. The true mean vector $\boldsymbol{\mu}$ for the simulation was $\boldsymbol{\mu} = X\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ was obtained by running LARS for 10 steps on the original $(X, \mathbf{y})$ diabetes data (agreeing in this case with the 10-step Lasso or Stagewise analysis). Subtracting $\boldsymbol{\mu}$ from a centered version of the original $\mathbf{y}$ vector of Table 1 gave a vector $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\mu}$ of $n = 442$ residuals. The "true R^2" for this model, $\|\boldsymbol{\mu}\|^2/(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\varepsilon}\|^2)$, equaled 0.416.

100 simulated response vectors $\mathbf{y}^*$ were generated from the model

$$\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\varepsilon}^*, \tag{3.16}$$

with $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*)$ a random sample, with replacement, from the components of $\boldsymbol{\varepsilon}$. The LARS algorithm with $K = 40$ steps was run for each simulated data set $(X, \mathbf{y}^*)$, yielding a sequence of estimates $\widehat{\boldsymbol{\mu}}^{(k)*}$, $k = 1, 2, \ldots, 40$, and likewise using the Lasso and Stagewise algorithms.

Figure 5 compares the LARS, Lasso and Stagewise estimates. For a given estimate $\widehat{\boldsymbol{\mu}}$ define the *proportion explained* $\mathrm{pe}(\widehat{\boldsymbol{\mu}})$ to be

$$\mathrm{pe}(\widehat{\boldsymbol{\mu}}) = 1 - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/\|\boldsymbol{\mu}\|^2, \tag{3.17}$$

so $\mathrm{pe}(\mathbf{0}) = 0$ and $\mathrm{pe}(\boldsymbol{\mu}) = 1$. The solid curve graphs the average of $\mathrm{pe}(\widehat{\boldsymbol{\mu}}^{(k)*})$ over the 100 simulations, versus step number k for LARS, $k = 1, 2, \ldots, 40$. The corresponding curves are graphed for Lasso and Stagewise, except that the horizontal axis is now the average number of nonzero $\widehat{\beta}_j^*$ terms composing $\widehat{\boldsymbol{\mu}}^{(k)*}$. For example, $\widehat{\boldsymbol{\mu}}^{(40)*}$ averaged 33.23 nonzero terms with Stagewise, compared to 35.83 for Lasso and 40 for LARS.

Figure 5's most striking message is that the three algorithms performed almost identically, and rather well. The average proportion explained rises quickly,

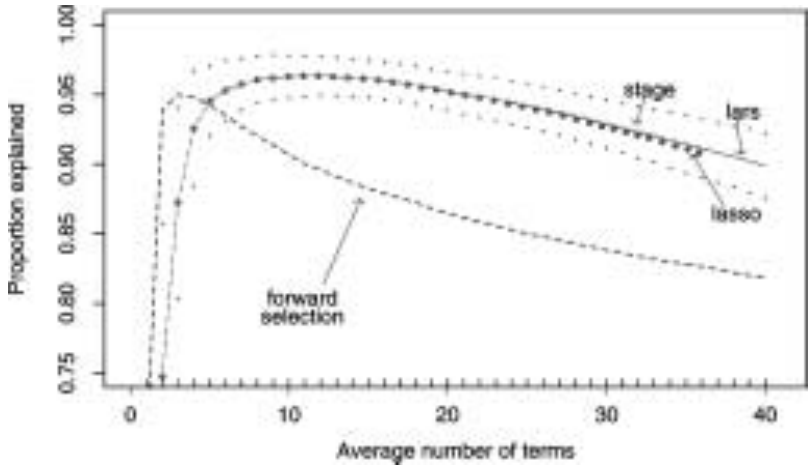

FIG. 5. *Simulation study comparing LARS, Lasso and Stagewise algorithms*; 100 *replications of model* (3.15)–(3.16). *Solid curve shows average proportion explained*, (3.17), *for LARS estimates as function of number of steps* $k = 1, 2, \ldots, 40$; *Lasso and Stagewise give nearly identical results*; *small dots indicate plus or minus one standard deviation over the* 100 *simulations*. *Classic Forward Selection* (*heavy dashed curve*) *rises and falls more abruptly.*

reaching a maximum of 0.963 at $k = 10$, and then declines slowly as k grows to 40. The light dots display the small standard deviation of $\text{pe}(\widehat{\boldsymbol{\mu}}^{(k)*})$ over the 100 simulations, roughly ± 0.02. Stopping at any point between $k = 5$ and 25 typically gave a $\widehat{\boldsymbol{\mu}}^{(k)*}$ with true predictive R^2 about 0.40, compared to the ideal value 0.416 for $\boldsymbol{\mu}$.

The dashed curve in Figure 5 tracks the average proportion explained by classic Forward Selection. It rises very quickly, to a maximum of 0.950 after $k = 3$ steps, and then falls back more abruptly than the LARS–Lasso–Stagewise curves. This behavior agrees with the characterization of Forward Selection as a dangerously greedy algorithm.

3.4. *Other LARS modifications*. Here are a few more examples of LARS type model-building algorithms.

POSITIVE LASSO. Constraint (1.5) can be strengthened to

$$\text{minimize} \quad S(\widehat{\boldsymbol{\beta}}) \qquad \text{subject to} \quad T(\widehat{\boldsymbol{\beta}}) \le t \text{ and all } \widehat{\beta}_j \ge 0. \tag{3.18}$$

This would be appropriate if the statisticians or scientists believed that the variables x_j *must* enter the prediction equation in their defined directions. Situation (3.18) is a more difficult quadratic programming problem than (1.5), but it can be solved by a further modification of the Lasso-modified LARS algorithm: change $|\widehat{c}_j|$ to $\widehat{c}_j$ at both places in (2.9), set $s_j = 1$ instead of (2.10) and change (2.13) to

$$\widehat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\widehat{C} - \widehat{c}_j}{A_{\mathcal{A}} - a_j} \right\}. \tag{3.19}$$

The positive Lasso usually does *not* converge to the full OLS solution $\bar{\beta}_m$, even for very large choices of t.

The changes above amount to considering the $\mathbf{x}_j$ as generating half-lines rather than full one-dimensional spaces. A positive Stagewise version can be developed in the same way, and has the property that the $\widehat{\beta}_j$ tracks are always monotone.

LARS–OLS hybrid. After k steps the LARS algorithm has identified a set $\mathcal{A}_k$ of covariates, for example, $\mathcal{A}_4 = \{3, 9, 4, 7\}$ in the diabetes study. Instead of $\widehat{\boldsymbol{\beta}}_k$ we might prefer $\bar{\boldsymbol{\beta}}_k$, the OLS coefficients based on the linear model with covariates in $\mathcal{A}_k$—using LARS to find the model but not to estimate the coefficients. Besides looking more familiar, this will always increase the usual empirical R^2 measure of fit (though not necessarily the true fitting accuracy),

$$R^2(\bar{\boldsymbol{\beta}}_k) - R^2(\widehat{\boldsymbol{\beta}}_k) = \frac{1 - \rho_k^2}{\rho_k(2 - \rho_k)} [R^2(\widehat{\boldsymbol{\beta}}_k) - R^2(\widehat{\boldsymbol{\beta}}_{k-1})], \tag{3.20}$$

where $\rho_k = \widehat{\gamma}_k / \bar{\gamma}_k$ as in (2.22).

The increases in R^2 were small in the diabetes example, on the order of 0.01 for $k \geq 4$ compared with $R^2 \doteq 0.50$, which is expected from (3.20) since we would usually continue LARS until $R^2(\widehat{\boldsymbol{\beta}}_k) - R^2(\widehat{\boldsymbol{\beta}}_{k-1})$ was small. For the same reason $\bar{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\beta}}_k$ are likely to lie near each other as they did in the diabetes example.

Main effects first. It is straightforward to restrict the order in which variables are allowed to enter the LARS algorithm. For example, having obtained $\mathcal{A}_4 = \{3, 9, 4, 7\}$ for the diabetes study, we might *then* wish to check for interactions. To do this we begin LARS again, replacing $\mathbf{y}$ with $\mathbf{y} - \widehat{\boldsymbol{\mu}}_4$ and $\mathbf{x}$ with the $n \times 6$ matrix whose columns represent the interactions $\mathbf{x}_{3:9}, \mathbf{x}_{3:4}, \ldots, \mathbf{x}_{4:7}$.

Backward Lasso. The Lasso–modified LARS algorithm can be run backward, starting from the full OLS solution $\bar{\boldsymbol{\beta}}_m$. Assuming that all the coordinates of $\bar{\boldsymbol{\beta}}_m$ are nonzero, their signs must agree with the signs s_j that the current correlations had during the final LARS step. This allows us to calculate the last equiangular direction $\mathbf{u}_{\mathcal{A}}$, (2.4)–(2.6). Moving backward from $\widehat{\boldsymbol{\mu}}_m = X\bar{\boldsymbol{\beta}}_m$ along the line $\boldsymbol{\mu}(\gamma) = \widehat{\boldsymbol{\mu}}_m - \gamma \mathbf{u}_{\mathcal{A}}$, we eliminate from the active set the index of the first $\widehat{\beta}_j$ that becomes zero. Continuing backward, we keep track of all coefficients $\widehat{\beta}_j$ and current correlations $\widehat{c}_j$, following essentially the same rules for changing $\mathcal{A}$ as in Section 3.1. As in (2.3), (3.5) the calculation of $\widetilde{\gamma}$ and $\widehat{\gamma}$ is easy.

The crucial property of the Lasso that makes backward navigation possible is (3.1), which permits calculation of the correct equiangular direction $\mathbf{u}_{\mathcal{A}}$ at each step. In this sense Lasso can be just as well thought of as a backward-moving algorithm. This is not the case for LARS or Stagewise, both of which are inherently forward-moving algorithms.

4. Degrees of freedom and C_p estimates. Figures 1 and 3 show all possible Lasso, Stagewise or LARS estimates of the vector $\boldsymbol{\beta}$ for the diabetes data. The scientists want just a single $\widehat{\boldsymbol{\beta}}$ of course, so we need some rule for selecting among the possibilities. This section concerns a C_p-type selection criterion, especially as it applies to the choice of LARS estimate.

Let $\widehat{\boldsymbol{\mu}} = g(\mathbf{y})$ represent a formula for estimating $\boldsymbol{\mu}$ from the data vector $\mathbf{y}$. Here, as usual in regression situations, we are considering the covariate vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ fixed at their observed values. We assume that given the $\mathbf{x}$'s, $\mathbf{y}$ is generated according to an homoskedastic model

$$\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \tag{4.1}$$

meaning that the components y_i are uncorrelated, with mean μ_i and variance σ^2. Taking expectations in the identity

$$(\widehat{\mu}_i - \mu_i)^2 = (y_i - \widehat{\mu}_i)^2 - (y_i - \mu_i)^2 + 2(\widehat{\mu}_i - \mu_i)(y_i - \mu_i), \tag{4.2}$$

and summing over i, yields

$$E\left\{\frac{\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2}{\sigma^2}\right\}=E\left\{\frac{\|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2}{\sigma^2}-n\right\}+2\sum_{i=1}^{n}\frac{\operatorname{cov}(\hat{\mu}_i,\,y_i)}{\sigma^2}. \tag{4.3}$$

The last term of (4.3) leads to a convenient definition of the *degrees of freedom* for an estimator $\hat{\boldsymbol{\mu}}=g(\mathbf{y})$,

$$df_{\mu,\sigma^2}=\sum_{i=1}^{n}\operatorname{cov}(\hat{\mu}_i,\,y_i)/\sigma^2, \tag{4.4}$$

and a C_p-type risk estimation formula,

$$C_p(\hat{\boldsymbol{\mu}})=\frac{\|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2}{\sigma^2}-n+2df_{\mu,\sigma^2}. \tag{4.5}$$

If σ^2 and df_{μ,σ^2} are known, $C_p(\hat{\boldsymbol{\mu}})$ is an unbiased estimator of the true risk $E\{\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2/\sigma^2\}$. For linear estimators $\hat{\boldsymbol{\mu}}=M\mathbf{y}$, model (4.1) makes $df_{\mu,\sigma^2}=\operatorname{trace}(M)$, equaling the usual definition of degrees of freedom for OLS, and coinciding with the proposal of Mallows (1973). Section 6 of Efron and Tibshirani (1997) and Section 7 of Efron (1986) discuss formulas (4.4) and (4.5) and their role in C_p, Akaike information criterion (AIC) and Stein's unbiased risk estimated (SURE) estimation theory, a more recent reference being Ye (1998).

Practical use of C_p formula (4.5) requires preliminary estimates of $\boldsymbol{\mu}$, σ^2 and df_{μ,σ^2}. In the numerical results below, the usual OLS estimates $\bar{\boldsymbol{\mu}}$ and $\bar{\sigma}^2$ from the full OLS model were used to calculate bootstrap estimates of df_{μ,σ^2}; bootstrap samples $\mathbf{y}^*$ and replications $\hat{\boldsymbol{\mu}}^*$ were then generated according to

$$\mathbf{y}^*\sim N(\bar{\boldsymbol{\mu}},\bar{\sigma}^2)\quad\text{and}\quad\hat{\boldsymbol{\mu}}^*=g(\mathbf{y}^*). \tag{4.6}$$

Independently repeating (4.6) say B times gives straightforward estimates for the covariances in (4.4),

$$\widehat{\operatorname{cov}}_i=\frac{\sum_{b=1}^{B}\hat{\boldsymbol{\mu}}_i^*(b)[\,\mathbf{y}_i^*(b)-\mathbf{y}_i^*(\cdot)]}{B-1},\qquad\text{where }\mathbf{y}^*(\cdot)=\frac{\sum_{b=1}^{B}\mathbf{y}^*(b)}{B}, \tag{4.7}$$

and then

$$\widehat{df}=\sum_{i=1}^{n}\widehat{\operatorname{cov}}_i/\bar{\sigma}^2. \tag{4.8}$$

Normality is not crucial in (4.6). Nearly the same results were obtained using $\mathbf{y}^*=\bar{\boldsymbol{\mu}}^*+\mathbf{e}^*$, where the components of $\mathbf{e}^*$ were resampled from $\mathbf{e}=\mathbf{y}-\bar{\boldsymbol{\mu}}$.

The left panel of Figure 6 shows $\widehat{df}_k$ for the diabetes data LARS estimates $\hat{\boldsymbol{\mu}}_k$, $k=1,2,\ldots,m=10$. It portrays a startlingly simple situation that we will call the "simple approximation,"

$$df(\hat{\boldsymbol{\mu}}_k)\doteq k. \tag{4.9}$$

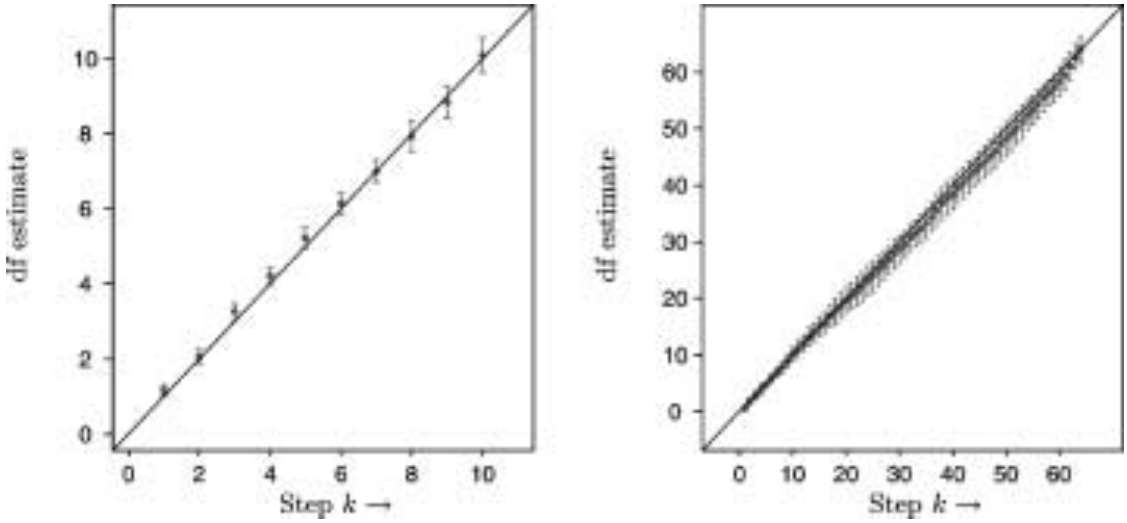

FIG. 6. *Degrees of freedom for LARS estimates* $\hat{\boldsymbol{\mu}}_k$: (left) *diabetes study*, *Table* 1, $k = 1, 2, \ldots, m = 10$; (right) *quadratic model* (3.15) *for the diabetes data*, $m = 64$. *Solid line is simple approximation* $df_k = k$. *Dashed lines are approximate* 95% *confidence intervals for the bootstrap estimates. Each panel based on* $B = 500$ *bootstrap replications.*

The right panel also applies to the diabetes data, but this time with the quadratic model (3.15), having $m = 64$ predictors. We see that the simple approximation (4.9) is again accurate within the limits of the bootstrap computation (4.8), where $B = 500$ replications were divided into 10 groups of 50 each in order to calculate Student-t confidence intervals.

If (4.9) can be believed, and we will offer some evidence in its behalf, we can estimate the risk of a k-step LARS estimator $\hat{\boldsymbol{\mu}}_k$ by

$$C_p(\hat{\boldsymbol{\mu}}_k) \doteq \|\mathbf{y} - \hat{\boldsymbol{\mu}}_k\|^2/\bar{\sigma}^2 - n + 2k. \tag{4.10}$$

The formula, which is the same as the C_p estimate of risk for an OLS estimator based on a subset of k preselected predictor vectors, has the great advantage of not requiring any further calculations beyond those for the original LARS estimates. The formula applies only to LARS, and not to Lasso or Stagewise.

Figure 7 displays $C_p(\hat{\boldsymbol{\mu}}_k)$ as a function of k for the two situations of Figure 6. Minimum C_p was achieved at steps $k = 7$ and $k = 16$, respectively. Both of the minimum C_p models looked sensible, their first several selections of "important" covariates agreeing with an earlier model based on a detailed inspection of the data assisted by medical expertise.

The simple approximation becomes a theorem in two cases.

THEOREM 3. *If the covariate vectors* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ *are mutually orthogonal, then the* k*-step LARS estimate* $\hat{\boldsymbol{\mu}}_k$ *has* $df(\hat{\boldsymbol{\mu}}_k) = k$.

To state the second more general setting we introduce the following condition.

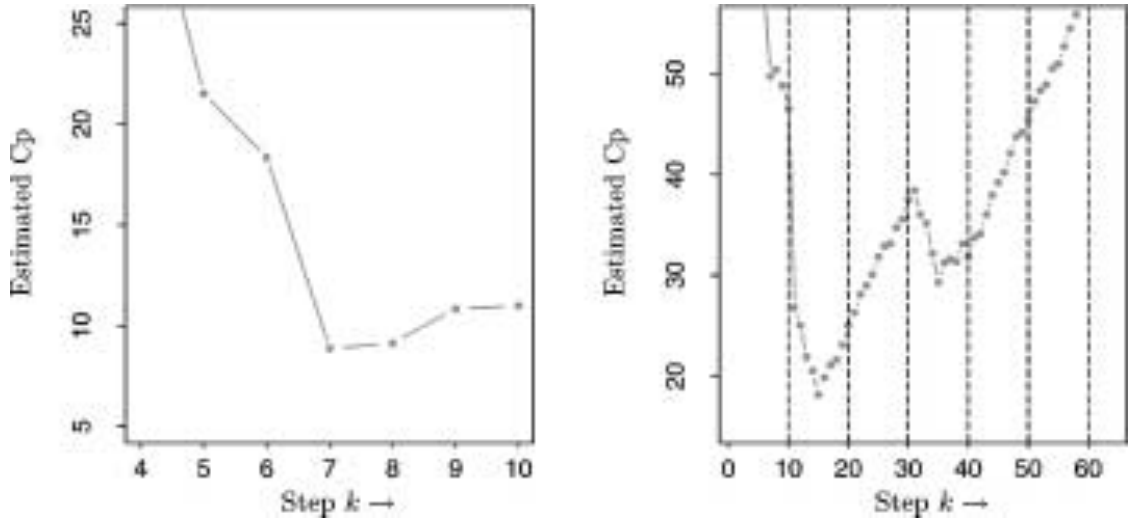

FIG. 7. *C_p estimates of risk* (4.10) *for the two situations of Figure* 6: (left) $m = 10$ *model has smallest* C_p *at* $k = 7$; (right) $m = 64$ *model has smallest* C_p *at* $k = 16$.

POSITIVE CONE CONDITION. For all possible subsets $X_{\mathcal{A}}$ of the full design matrix X,

$$G_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}} > 0, \tag{4.11}$$

where the inequality is taken element-wise.

The positive cone condition holds if X is orthogonal. It is strictly more general than orthogonality, but counterexamples (such as the diabetes data) show that not all design matrices X satisfy it.

It is also easy to show that LARS, Lasso and Stagewise all coincide under the positive cone condition, so the degrees-of-freedom formula applies to them too in this case.

THEOREM 4. *Under the positive cone condition,* $df(\hat{\boldsymbol{\mu}}_k) = k$.

The proof, which appears later in this section, is an application of Stein's unbiased risk estimate (SURE) [Stein (1981)]. Suppose that $g : \mathbb{R}^n \to \mathbb{R}^n$ is almost differentiable (see Remark A.1 in the Appendix) and set $\nabla \cdot g = \sum_{i=1}^n \partial g_i/\partial x_i$. If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, then Stein's formula states that

$$\sum_{i=1}^{n} \operatorname{cov}(g_i, y_i)/\sigma^2 = E[\nabla \cdot g(\mathbf{y})]. \tag{4.12}$$

The left-hand side is $df(g)$ for the general estimator $g(\mathbf{y})$. Focusing specifically on LARS, it will turn out that $\nabla \cdot \hat{\boldsymbol{\mu}}_k(\mathbf{y}) = k$ in *all* situations with probability 1, but that the continuity assumptions underlying (4.12) and SURE can fail in certain nonorthogonal cases where the positive cone condition does not hold.

A range of simulations suggested that the simple approximation is quite accurate even when the $\mathbf{x}_j$'s are highly correlated and that it requires concerted effort at pathology to make $df(\widehat{\boldsymbol{\mu}}_k)$ much different than k.

Stein's formula assumes normality, $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. A cruder "delta method" rationale for the simple approximation requires only homoskedasticity, (4.1). The geometry of Figure 4 implies

$$\widehat{\boldsymbol{\mu}}_k = \bar{\mathbf{y}}_k - \cot_k \cdot \|\bar{\mathbf{y}}_{k+1} - \bar{\mathbf{y}}_k\|, \tag{4.13}$$

where $\cot_k$ is the cotangent of the angle between $\mathbf{u}_k$ and $\mathbf{u}_{k+1}$,

$$\cot_k = \frac{\mathbf{u}_k'\mathbf{u}_{k+1}}{[1 - (\mathbf{u}_k'\mathbf{u}_{k+1})^2]^{1/2}}. \tag{4.14}$$

Let $\mathbf{v}_k$ be the unit vector orthogonal to $\mathcal{L}(X_b)$, the linear space spanned by the first k covariates selected by LARS, and pointing into $\mathcal{L}(X_{k+1})$ along the direction of $\bar{\mathbf{y}}_{k+1} - \bar{\mathbf{y}}_k$. For $\mathbf{y}^*$ near $\mathbf{y}$ we can reexpress (4.13) as a locally linear transformation,

$$\widehat{\boldsymbol{\mu}}_k^* = \widehat{\boldsymbol{\mu}}_k + M_k(\mathbf{y}^* - \mathbf{y}) \qquad \text{with } M_k = P_k - \cot_k \cdot \mathbf{u}_k\mathbf{v}_k', \tag{4.15}$$

P_k being the usual projection matrix from $\mathbb{R}^n$ into $\mathcal{L}(X_k)$; (4.15) holds within a neighborhood of $\mathbf{y}$ such that the LARS choices $\mathcal{L}(X_k)$ and $\mathbf{v}_k$ remain the same.

The matrix M_k has $\operatorname{trace}(M_k) = k$. Since the trace equals the degrees of freedom for linear estimators, the simple approximation (4.9) is seen to be a delta method approximation to the bootstrap estimates (4.6) and (4.7).

It is clear that (4.9) $df(\widehat{\boldsymbol{\mu}}_k) \doteq k$ cannot hold for the Lasso, since the degrees of freedom is m for the full model but the total number of steps taken can exceed m. However, we have found empirically that an intuitively plausible result holds: the degrees of freedom is well approximated by the number of nonzero predictors in the model. Specifically, starting at step 0, let $\ell(k)$ be the index of the last model in the Lasso sequence containing k predictors. Then $df(\widehat{\boldsymbol{\mu}}_{\ell(k)}) \doteq k$. We do not yet have any mathematical support for this claim.

4.1. *Orthogonal designs.* In the orthogonal case, we assume that $\mathbf{x}_j = \mathbf{e}_j$ for $j = 1, \ldots, m$. The LARS algorithm then has a particularly simple form, reducing to soft thresholding at the order statistics of the data.

To be specific, define the soft thresholding operation on a scalar y_1 at threshold t by

$$\eta(y_1; t) = \begin{cases} y_1 - t, & \text{if } y_1 > t, \\ 0, & \text{if } |y_1| \le t, \\ y_1 + t, & \text{if } y_1 < -t. \end{cases}$$

The order statistics of the absolute values of the data are denoted by

$$|y|_{(1)} \ge |y|_{(2)} \ge \cdots \ge |y|_{(n)} \ge |y|_{(n+1)} := 0. \tag{4.16}$$

We note that $y_{m+1}, \ldots, y_n$ do not enter into the estimation procedure, and so we may as well assume that $m = n$.

LEMMA 1. *For an orthogonal design with* $\mathbf{x}_j = \mathbf{e}_j$, $j = 1, \ldots, n$, *the kth LARS estimate* $(0 \le k \le n)$ *is given by*

$$\hat{\mu}_{k,i}(\mathbf{y}) = \begin{cases} y_i - |y|_{(k+1)}, & \text{if } y_i > |y|_{(k+1)}, \\ 0, & \text{if } |y_i| \le |y|_{(k+1)}, \\ y_i + |y|_{(k+1)}, & \text{if } y_i < -|y|_{(k+1)}, \end{cases} \tag{4.17}$$

$$= \eta\bigl(y_i; |y|_{(k+1)}\bigr). \tag{4.18}$$

PROOF. The proof is by induction, stepping through the LARS sequence. First note that the LARS parameters take a simple form in the orthogonal setting:

$$G_{\mathcal{A}} = I_{\mathcal{A}}, \qquad A_{\mathcal{A}} = |\mathcal{A}|^{-1/2}, \qquad \mathbf{u}_{\mathcal{A}} = |\mathcal{A}|^{-1/2}\mathbf{1}_{\mathcal{A}}, \qquad a_{k,j} = 0, \qquad j \notin \mathcal{A}_k.$$

We assume for the moment that there are no ties in the order statistics (4.16), so that the variables enter one at a time. Let $j(l)$ be the index corresponding to the lth order statistic, $|y|_{(l)} = s_l y_{j(l)}$: we will see that $\mathcal{A}_k = \{j(1), \ldots, j(k)\}$.

We have $\mathbf{x}'_j \mathbf{y} = y_j$, and so at the first step LARS picks variable $j(1)$ and sets $\hat{C}_1 = |y|_{(1)}$. It is easily seen that

$$\hat{\gamma}_1 = \min_{j \ne j(1)} \bigl\{|y|_{(1)} - |y_j|\bigr\} = |y|_{(1)} - |y|_{(2)}$$

and so

$$\hat{\boldsymbol{\mu}}_1 = \bigl[|y|_{(1)} - |y|_{(2)}\bigr]\mathbf{e}_{j(1)},$$

which is precisely (4.17) for $k = 1$.

Suppose now that step $k-1$ has been completed, so that $\mathcal{A}_k = \{j(1), \ldots, j(k)\}$ and (4.17) holds for $\hat{\boldsymbol{\mu}}_{k-1}$. The current correlations $\hat{C}_k = |y|_{(k)}$ and $\hat{c}_{k,j} = y_j$ for $j \notin \mathcal{A}_k$. Since $A_k - a_{k,j} = k^{-1/2}$, we have

$$\hat{\gamma}_k = \min_{j \notin \mathcal{A}_k} k^{1/2}\bigl\{|y|_{(k)} - |y_j|\bigr\}$$

and

$$\hat{\gamma}_k \mathbf{u}_k = \bigl[|y|_{(k)} - |y|_{(k+1)}\bigr]\mathbf{1}\{j \in \mathcal{A}_k\}.$$

Adding this term to $\hat{\boldsymbol{\mu}}_{k-1}$ yields (4.17) for step k.

The argument clearly extends to the case in which there are ties in the order statistics (4.16): if $|y|_{(k+1)} = \cdots = |y|_{(k+r)}$, then $\mathcal{A}_k(\mathbf{y})$ expands by r variables at step $k+1$ and $\hat{\boldsymbol{\mu}}_{k+\nu}(\mathbf{y})$, $\nu = 1, \ldots, r$, are all determined at the same time and are equal to $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y})$. □

PROOF OF THEOREM 4 (Orthogonal case). The argument is particularly simple in this setting, and so worth giving separately. First we note from (4.17) that $\hat{\boldsymbol{\mu}}_k$ is continuous and Lipschitz(1) and so certainly almost differentiable.

Hence (4.12) shows that we simply have to calculate $\nabla \cdot \hat{\boldsymbol{\mu}}_k$. Inspection of (4.17) shows that

$$\begin{aligned}\nabla \cdot \hat{\boldsymbol{\mu}}_k &= \sum_i \frac{\partial \hat{\mu}_{k,i}}{\partial y_i}(\mathbf{y}) \\ &= \sum_i I\{|y_i| > |y|_{(k+1)}\} = k\end{aligned}$$

almost surely, that is, except for ties. This completes the proof. □

4.2. *The divergence formula.* While for the most general design matrices X, it can happen that $\hat{\boldsymbol{\mu}}_k$ fails to be almost differentiable, we will see that the divergence formula

$$\nabla \cdot \hat{\boldsymbol{\mu}}_k(\mathbf{y}) = k \tag{4.19}$$

does hold almost everywhere. Indeed, certain authors [e.g., Meyer and Woodroofe (2000)] have argued that the divergence $\nabla \cdot \widehat{\boldsymbol{\mu}}$ of an estimator provides itself a useful measure of the effective dimension of a model.

Turning to LARS, we shall say that $\hat{\boldsymbol{\mu}}(\mathbf{y})$ is locally linear at a data point y_0 if there is some small open neighborhood of y_0 on which $\hat{\boldsymbol{\mu}}(\mathbf{y}) = M\mathbf{y}$ is exactly linear. Of course, the matrix $M = M(y_0)$ can depend on y_0—in the case of LARS, it will be seen to be constant on the interior of polygonal regions, with jumps across the boundaries. We say that a set G has full measure if its complement has Lebesgue measure zero.

LEMMA 2. *There is an open set G_k of full measure such that, at all $\mathbf{y} \in G_k$, $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ is locally linear and $\nabla \cdot \hat{\boldsymbol{\mu}}_k(\mathbf{y}) = k$.*

PROOF. We give here only the part of the proof that relates to actual calculation of the divergence in (4.19). The arguments establishing continuity and local linearity are delayed to the Appendix.

So, let us fix a point $\mathbf{y}$ in the interior of G_k. From Lemma 13 in the Appendix, this means that near $\mathbf{y}$ the active set $\mathcal{A}_k(\mathbf{y})$ is locally constant, that a single variable enters at the next step, this variable being the same near $\mathbf{y}$. In addition, $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ is locally linear, and hence in particular differentiable. Since $G_k \subset G_l$ for $l < k$, the same story applies at all previous steps and we have

$$\hat{\boldsymbol{\mu}}_k(\mathbf{y}) = \sum_{l=1}^{k} \gamma_l(\mathbf{y})\mathbf{u}_l. \tag{4.20}$$

Differentiating the jth component of vector $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ yields

$$\frac{\partial \hat{\mu}_{k,j}}{\partial y_i}(\mathbf{y}) = \sum_{l=1}^{k} \frac{\partial \gamma_l(\mathbf{y})}{\partial y_i} u_{l,j}.$$

In particular, for the divergence

$$\nabla \cdot \hat{\boldsymbol{\mu}}_k(\mathbf{y}) = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_{k,i}}{\partial y_i} = \sum_{l=1}^{k} \langle \nabla \gamma_l, \mathbf{u}_l \rangle, \tag{4.21}$$

the brackets indicating inner product.

The active set is $\mathcal{A}_k = \{1, 2, \ldots, k\}$ and $\mathbf{x}_{k+1}$ is the variable to enter next. For $k \geq 2$, write $\boldsymbol{\delta}_k = \mathbf{x}_l - \mathbf{x}_k$ for any choice $l < k$—as remarked in the Conventions in the Appendix, the choice of l is immaterial (e.g., $l = 1$ for definiteness). Let $b_{k+1} = \langle \boldsymbol{\delta}_{k+1}, \mathbf{u}_k \rangle$, which is nonzero, as argued in the proof of Lemma 13. As shown in (A.4) in the Appendix, (2.13) can be rewritten

$$\gamma_k(\mathbf{y}) = b_{k+1}^{-1} \langle \boldsymbol{\delta}_{k+1}, \mathbf{y} - \hat{\boldsymbol{\mu}}_{k-1} \rangle. \tag{4.22}$$

For $k \geq 2$, define the linear space of vectors equiangular with the active set

$$\mathcal{L}_k = \mathcal{L}_k(\mathbf{y}) = \big\{ \mathbf{u} : \langle \mathbf{x}_1, \mathbf{u} \rangle = \cdots = \langle \mathbf{x}_k, \mathbf{u} \rangle \ \text{ for } \mathbf{x}_l \text{ with } l \in \mathcal{A}_k(\mathbf{y}) \big\}.$$

[We may drop the dependence on $\mathbf{y}$ since $\mathcal{A}_k(\mathbf{y})$ is locally fixed.] Clearly $\dim \mathcal{L}_k = n - k + 1$ and

$$\mathbf{u}_k \in \mathcal{L}_k, \qquad \mathcal{L}_{k+1} \subset \mathcal{L}_k. \tag{4.23}$$

We shall now verify that, for each $k \geq 1$,

$$\langle \nabla \gamma_k, \mathbf{u}_k \rangle = 1 \quad \text{and} \quad \langle \nabla \gamma_k, \mathbf{u} \rangle = 0 \qquad \text{for } \mathbf{u} \in \mathcal{L}_{k+1}. \tag{4.24}$$

Formula (4.21) shows that this suffices to prove Lemma 2.

First, for $k = 1$ we have $\gamma_1(\mathbf{y}) = b_2^{-1} \langle \boldsymbol{\delta}_2, \mathbf{y} \rangle$ and $\langle \nabla \gamma_1, \mathbf{u} \rangle = b_2^{-1} \langle \boldsymbol{\delta}_2, \mathbf{u} \rangle$, and that

$$\langle \boldsymbol{\delta}_2, \mathbf{u} \rangle = \langle \mathbf{x}_1 - \mathbf{x}_2, \mathbf{u} \rangle = \begin{cases} b_2, & \text{if } \mathbf{u} = \mathbf{u}_1, \\ 0, & \text{if } \mathbf{u} \in \mathcal{L}_2. \end{cases}$$

Now, for general k, combine (4.22) and (4.20):

$$b_{k+1} \gamma_k(\mathbf{y}) = \langle \boldsymbol{\delta}_{k+1}, \mathbf{y} \rangle - \sum_{l=1}^{k-1} \langle \boldsymbol{\delta}_{k+1}, \mathbf{u}_l \rangle \gamma_l(\mathbf{y}),$$

and hence

$$b_{k+1} \langle \nabla \gamma_k, \mathbf{u} \rangle = \langle \boldsymbol{\delta}_{k+1}, \mathbf{u} \rangle - \sum_{l=1}^{k-1} \langle \boldsymbol{\delta}_{k+1}, \mathbf{u}_l \rangle \langle \nabla \gamma_l, \mathbf{u} \rangle.$$

From the definitions of b_{k+1} and $\mathcal{L}_{k+1}$ we have

$$\langle \boldsymbol{\delta}_{k+1}, \mathbf{u} \rangle = \langle \mathbf{x}_l - \mathbf{x}_{k+1} \rangle = \begin{cases} b_{k+1}, & \text{if } \mathbf{u} = \mathbf{u}_k, \\ 0, & \text{if } \mathbf{u} \in \mathcal{L}_{k+1}. \end{cases}$$

Hence the truth of (4.24) for step k follows from its truth at step $k - 1$ because of the containment properties (4.23). □

4.3. *Proof of Theorem* 4. To complete the proof of Theorem 4, we state the following regularity result, proved in the Appendix.

LEMMA 3. *Under the positive cone condition*, $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ *is continuous and almost differentiable*.

This guarantees that Stein's formula (4.12) is valid for $\hat{\boldsymbol{\mu}}_k$ under the positive cone condition, so the divergence formula of Lemma 2 then immediately yields Theorem 4.

5. LARS and Lasso properties. The LARS and Lasso algorithms are described more carefully in this section, with an eye toward fully understanding their relationship. Theorem 1 of Section 3 will be verified. The latter material overlaps results in Osborne, Presnell and Turlach (2000a), particularly in their Section 4. Our point of view here allows the Lasso to be described as a quite simple modification of LARS, itself a variation of traditional Forward Selection methodology, and in this sense should be more accessible to statistical audiences. In any case we will stick to the language of regression and correlation rather than convex optimization, though some of the techniques are familiar from the optimization literature.

The results will be developed in a series of lemmas, eventually lending to a proof of Theorem 1 and its generalizations. The first three lemmas refer to attributes of the LARS procedure that are not specific to its Lasso modification.

Using notation as in (2.17)–(2.20), suppose LARS has completed step $k-1$, giving estimate $\widehat{\boldsymbol{\mu}}_{k-1}$ and active set $\mathcal{A}_k$ for step k, with covariate $\mathbf{x}_k$ the newest addition to the active set.

LEMMA 4. *If* $\mathbf{x}_k$ *is the only addition to the active set at the end of step* $k-1$, *then the coefficient vector* $w_k = A_k \mathcal{G}_k^{-1}\mathbf{1}_k$ *for the equiangular vector* $\mathbf{u}_k = X_k w_k$, (2.6), *has its* k*th component* w_{kk} *agreeing in sign with the current correlation* $c_{kk} = \mathbf{x}_k'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1})$. *Moreover, the regression vector* $\widehat{\boldsymbol{\beta}}_k$ *for* $\widehat{\boldsymbol{\mu}}_k = X\widehat{\boldsymbol{\beta}}_k$ *has its* k*th component* $\widehat{\beta}_{kk}$ *agreeing in sign with* c_{kk}.

Lemma 4 says that new variables *enter* the LARS active set in the "correct" direction, a weakened version of the Lasso requirement (3.1). This will turn out to be a crucial connection for the LARS–Lasso relationship.

PROOF OF LEMMA 4. The case $k = 1$ is apparent. Note that since

$$X_k'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1}) = \widehat{C}_k \mathbf{1}_k,$$

(2.20), from (2.6) we have

$$w_k = A_k \widehat{C}_k^{-1}[(X_k'X_k)^{-1}X_k'(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{k-1})] := A_k \widehat{C}_k^{-1} w_k^*. \tag{5.1}$$

The term in square braces is the least squares coefficient vector in the regression of the current residual on X_k, and the term preceding it is positive.

Note also that

$$X_k'(\mathbf{y} - \bar{\mathbf{y}}_{k-1}) = (\mathbf{0}, \delta)' \qquad \text{with } \delta > 0, \tag{5.2}$$

since $X_{k-1}'(\mathbf{y} - \bar{\mathbf{y}}_{k-1}) = \mathbf{0}$ by definition (this $\mathbf{0}$ has $k-1$ elements), and $c_k(\gamma) = \mathbf{x}_k'(\mathbf{y} - \gamma \mathbf{u}_{k-1})$ decreases more slowly in γ than $c_j(\gamma)$ for $j \in \mathcal{A}_{k-1}$:

$$c_k(\gamma) \begin{cases} < c_j(\gamma), & \text{for } \gamma < \widehat{\gamma}_{k-1}, \\ = c_j(\gamma) = \widehat{C}_k, & \text{for } \gamma = \widehat{\gamma}_{k-1}, \\ > c_j(\gamma), & \text{for } \widehat{\gamma}_{k-1} < \gamma < \bar{\gamma}_{k-1}. \end{cases} \tag{5.3}$$

Thus

$$\widehat{w}_k^* = (X_k'X_k)^{-1}X_k'(\mathbf{y} - \bar{\mathbf{y}}_{k-1} + \bar{\mathbf{y}}_{k-1} - \widehat{\boldsymbol{\mu}}_{k-1}) \tag{5.4}$$

$$= (X_k'X_k)^{-1}\begin{pmatrix} \mathbf{0} \\ \delta \end{pmatrix} + (X_k'X_k)^{-1}X_k'[(\bar{\gamma}_{k-1} - \widehat{\gamma}_{k-1})\mathbf{u}_{k-1}]. \tag{5.5}$$

The kth element of $\widehat{w}_k^*$ is positive, because it is in the first term in (5.5) [$(X_k'X_k)$ is positive definite], and in the second term it is 0 since $\mathbf{u}_{k-1} \in \mathcal{L}(X_{k-1})$.

This proves the first statement in Lemma 4. The second follows from

$$\widehat{\beta}_{kk} = \widehat{\beta}_{k-1,k} + \widehat{\gamma}_k w_{kk}, \tag{5.6}$$

and $\widehat{\beta}_{k-1,k} = 0$, $\mathbf{x}_k$ not being active before step k. □

Our second lemma interprets the quantity $A_{\mathcal{A}} = (\mathbf{1}'\mathcal{G}_{\mathcal{A}}^{-1}\mathbf{1})^{-1/2}$, (2.4) and (2.5). Let $\mathcal{S}_{\mathcal{A}}$ indicate the extended simplex generated by the columns of $X_{\mathcal{A}}$,

$$\mathcal{S}_{\mathcal{A}} = \left\{ \mathbf{v} = \sum_{j \in \mathcal{A}} s_j \mathbf{x}_j P_j : \sum_{j \in \mathcal{A}} P_j = 1 \right\}, \tag{5.7}$$

"extended" meaning that the coefficients P_j are allowed to be negative.

LEMMA 5. *The point in $\mathcal{S}_{\mathcal{A}}$ nearest the origin is*

$$\mathbf{v}_{\mathcal{A}} = A_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}}X_{\mathcal{A}}w_{\mathcal{A}} \qquad \textit{where } w_{\mathcal{A}} = A_{\mathcal{A}}\mathcal{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}, \tag{5.8}$$

with length $\|\mathbf{v}_{\mathcal{A}}\| = A_{\mathcal{A}}$. If $\mathcal{A} \subseteq \mathcal{B}$, then $A_{\mathcal{A}} \geq A_{\mathcal{B}}$, the largest possible value being $A_{\mathcal{A}} = 1$ for $\mathcal{A}$ a singleton.

PROOF. For any $\mathbf{v} \in \mathcal{S}_{\mathcal{A}}$, the squared distance to the origin is $\|X_{\mathcal{A}}P\|^2 = P'\mathcal{G}_{\mathcal{A}}P$. Introducing a Lagrange multiplier to enforce the summation constraint, we differentiate

$$P'\mathcal{G}_{\mathcal{A}}P - \lambda(\mathbf{1}_{\mathcal{A}}'P - 1), \tag{5.9}$$

and find that the minimizing $P_{\mathcal{A}} = \lambda \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$. Summing, we get $\lambda \mathbf{1}'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = 1$, and hence

$$P_{\mathcal{A}} = A_{\mathcal{A}}^2 \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}} w_{\mathcal{A}}. \tag{5.10}$$

Hence $\mathbf{v}_{\mathcal{A}} = X_{\mathcal{A}} P_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$ and

$$\|\mathbf{v}_{\mathcal{A}}\|^2 = P'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} P_{\mathcal{A}} = A_{\mathcal{A}}^4 \mathbf{1}'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}}^2, \tag{5.11}$$

verifying (5.8). If $\mathcal{A} \subseteq \mathcal{B}$, then $\mathcal{S}_{\mathcal{A}} \subseteq \mathcal{S}_{\mathcal{B}}$, so the nearest distance $A_{\mathcal{B}}$ must be equal to or less than the nearest distance $A_{\mathcal{A}}$. $A_{\mathcal{A}}$ obviously equals 1 if and only if $\mathcal{A}$ has only one member. □

The LARS algorithm and its various modifications proceed in piecewise linear steps. For m-vectors $\widehat{\boldsymbol{\beta}}$ and $\mathbf{d}$, let

$$\boldsymbol{\beta}(\gamma) = \widehat{\boldsymbol{\beta}} + \gamma \mathbf{d} \quad \text{and} \quad S(\gamma) = \|\mathbf{y} - X\boldsymbol{\beta}(\gamma)\|^2. \tag{5.12}$$

LEMMA 6. *Letting* $\widehat{\mathbf{c}} = X'(\mathbf{y} - X\widehat{\boldsymbol{\beta}})$ *be the current correlation vector at* $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$,

$$S(\gamma) - S(0) = -2\widehat{\mathbf{c}}'\mathbf{d}\gamma + \mathbf{d}'X'X\mathbf{d}\gamma^2. \tag{5.13}$$

PROOF. $\mathcal{S}(\gamma)$ is a quadratic function of γ, with first two derivatives at $\gamma = 0$,

$$\dot{S}(0) = -2\widehat{\mathbf{c}}'\mathbf{d} \quad \text{and} \quad \ddot{S}(0) = 2\mathbf{d}'X'X\mathbf{d}. \tag{5.14}$$

□

The remainder of this section concerns the LARS–Lasso relationship. Now $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(t)$ will indicate a Lasso solution (1.5), and likewise $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(t) = X\widehat{\boldsymbol{\beta}}(t)$. Because $S(\widehat{\boldsymbol{\beta}})$ and $T(\widehat{\boldsymbol{\beta}})$ are both convex functions of $\widehat{\boldsymbol{\beta}}$, with S strictly convex, standard results show that $\widehat{\boldsymbol{\beta}}(t)$ and $\widehat{\boldsymbol{\mu}}(t)$ are unique and continuous functions of t.

For a given value of t let

$$\mathcal{A} = \{j : \widehat{\beta}_j(t) \neq 0\}. \tag{5.15}$$

We will show later that $\mathcal{A}$ is also the active set that determines the equiangular direction $\mathbf{u}_{\mathcal{A}}$, (2.6), for the LARS–Lasso computations.

We wish to characterize the track of the Lasso solutions $\widehat{\boldsymbol{\beta}}(t)$ or equivalently of $\widehat{\boldsymbol{\mu}}(t)$ as t increases from 0 to its maximum effective value. Let $\mathcal{T}$ be an open interval of the t axis, with infimum t_0, within which the set $\mathcal{A}$ of nonzero Lasso coefficients $\widehat{\beta}_j(t)$ remains constant.

LEMMA 7. *The Lasso estimates* $\widehat{\boldsymbol{\mu}}(t)$ *satisfy*

$$\widehat{\boldsymbol{\mu}}(t) = \widehat{\boldsymbol{\mu}}(t_0) + A_{\mathcal{A}}(t - t_0)\mathbf{u}_{\mathcal{A}} \tag{5.16}$$

for $t \in \mathcal{T}$, *where* $\mathbf{u}_{\mathcal{A}}$ *is the equiangular vector* $X_{\mathcal{A}} w_{\mathcal{A}}$, $w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$, (2.7).

PROOF. The lemma says that, for t in $\mathcal{T}$, $\widehat{\boldsymbol{\mu}}(t)$ moves linearly along the equiangular vector $\mathbf{u}_{\mathcal{A}}$ determined by $\mathcal{A}$. We can also state this in terms of the nonzero regression coefficients $\widehat{\beta}_{\mathcal{A}}(t)$,

$$\widehat{\beta}_{\mathcal{A}}(t) = \widehat{\beta}_{\mathcal{A}}(t_0) + S_{\mathcal{A}} A_{\mathcal{A}} (t - t_0) w_{\mathcal{A}}, \tag{5.17}$$

where $S_{\mathcal{A}}$ is the diagonal matrix with diagonal elements s_j, $j \in \mathcal{A}$. [$S_{\mathcal{A}}$ is needed in (5.17) because definitions (2.4), (2.10) require $\widehat{\boldsymbol{\mu}}(t) = X\widehat{\boldsymbol{\beta}}(t) = X_{\mathcal{A}} S_{\mathcal{A}} \widehat{\beta}_{\mathcal{A}}(t)$.]

Since $\widehat{\boldsymbol{\beta}}(t)$ satisfies (1.5) and has nonzero set $\mathcal{A}$, it also minimizes

$$S(\widehat{\beta}_{\mathcal{A}}) = \|\mathbf{y} - X_{\mathcal{A}} S_{\mathcal{A}} \widehat{\beta}_{\mathcal{A}}\|^2 \tag{5.18}$$

subject to

$$\sum_{\mathcal{A}} s_j \widehat{\beta}_j = t \quad \text{and} \quad \operatorname{sign}(\widehat{\beta}_j) = s_j \qquad \text{for } j \in \mathcal{A}. \tag{5.19}$$

[The inequality in (1.5) can be replaced by $T(\widehat{\boldsymbol{\beta}}) = t$ as long as t is less than $\sum |\bar{\beta}_j|$ for the full m-variable OLS solution $\bar{\boldsymbol{\beta}}_m$.] Moreover, the fact that the minimizing point $\widehat{\beta}_{\mathcal{A}}(t)$ occurs strictly *inside* the simplex (5.19), combined with the strict convexity of $S(\widehat{\beta}_{\mathcal{A}})$, implies we can drop the second condition in (5.19) so that $\widehat{\beta}_{\mathcal{A}}(t)$ solves

$$\text{minimize} \quad \{S(\widehat{\beta}_{\mathcal{A}})\} \qquad \text{subject to} \quad \sum_{\mathcal{A}} s_j \widehat{\beta}_j = t. \tag{5.20}$$

Introducing a Lagrange multiplier, (5.20) becomes

$$\text{minimize} \quad \tfrac{1}{2}\|\mathbf{y} - X_{\mathcal{A}} S_{\mathcal{A}} \widehat{\beta}_{\mathcal{A}}\|^2 + \lambda \sum_{\mathcal{A}} s_j \widehat{\beta}_j. \tag{5.21}$$

Differentiating we get

$$-S_{\mathcal{A}} X'_{\mathcal{A}} (\mathbf{y} - X_{\mathcal{A}} S_{\mathcal{A}} \widehat{\beta}_{\mathcal{A}}) + \lambda S_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} = 0. \tag{5.22}$$

Consider two values t_1 and t_2 in $\mathcal{T}$ with $t_0 < t_1 < t_2$. Corresponding to each of these are values for the Lagrange multiplier λ such that $\lambda_1 > \lambda_2$, and solutions $\widehat{\beta}_{\mathcal{A}}(t_1)$ and $\widehat{\beta}_{\mathcal{A}}(t_2)$. Inserting these into (5.22), differencing and premultiplying by $S_{\mathcal{A}}$ we get

$$X'_{\mathcal{A}} X_{\mathcal{A}} S_{\mathcal{A}} \big(\widehat{\beta}_{\mathcal{A}}(t_2) - \widehat{\beta}_{\mathcal{A}}(t_1)\big) = (\lambda_1 - \lambda_2) \mathbf{1}_{\mathcal{A}}. \tag{5.23}$$

Hence

$$\widehat{\beta}_{\mathcal{A}}(t_2) - \widehat{\beta}_{\mathcal{A}}(t_1) = (\lambda_1 - \lambda_2) S_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}. \tag{5.24}$$

However, $s'_{\mathcal{A}}[(\widehat{\beta}_{\mathcal{A}}(t_2) - \widehat{\beta}_{\mathcal{A}}(t_1)] = t_2 - t_1$ according to the Lasso definition, so

$$t_2 - t_1 = (\lambda_1 - \lambda_2) s'_{\mathcal{A}} S_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = (\lambda_1 - \lambda_2) \mathbf{1}'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = (\lambda_1 - \lambda_2) A_{\mathcal{A}}^{-2} \tag{5.25}$$

and

$$\widehat{\beta}_{\mathcal{A}}(t_2) - \widehat{\beta}_{\mathcal{A}}(t_1) = S_{\mathcal{A}} A_{\mathcal{A}}^2 (t_2 - t_1) \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = S_{\mathcal{A}} A_{\mathcal{A}} (t - t_1) w_{\mathcal{A}}. \tag{5.26}$$

Letting $t_2 = t$ and $t_1 \to t_0$ gives (5.17) by the continuity of $\widehat{\boldsymbol{\beta}}(t)$, and finally (5.16). Note that (5.16) implies that the maximum absolute correlation $\widehat{C}(t)$ equals $\widehat{C}(t_0) - A_{\mathcal{A}}^2(t - t_0)$, so that $\widehat{C}(t)$ is a piecewise linear decreasing function of the Lasso parameter t. $\square$

The Lasso solution $\widehat{\boldsymbol{\beta}}(t)$ occurs on the surface of the diamond-shaped convex polytope

$$\mathcal{D}(t) = \Big\{\boldsymbol{\beta} : \sum |\beta_j| \le t\Big\}, \tag{5.27}$$

$\mathcal{D}(t)$ increasing with t. Lemma 7 says that, for $t \in \mathcal{T}$, $\widehat{\boldsymbol{\beta}}(t)$ moves linearly along edge $\mathcal{A}$ of the polytope, the edge having $\beta_j = 0$ for $j \notin \mathcal{A}$. Moreover the regression estimates $\widehat{\boldsymbol{\mu}}(t)$ move in the LARS equiangular direction $\mathbf{u}_{\mathcal{A}}$, (2.6). It remains to show that "$\mathcal{A}$" changes according to the rules of Theorem 1, which is the purpose of the next three lemmas.

LEMMA 8. *A Lasso solution $\widehat{\boldsymbol{\beta}}$ has*

$$\widehat{c}_j = \widehat{C} \cdot \operatorname{sign}(\widehat{\beta}_j) \qquad \text{for } j \in \mathcal{A}, \tag{5.28}$$

where $\widehat{c}_j$ equals the current correlation $\mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}}) = \mathbf{x}_j'(\mathbf{y} - X\widehat{\boldsymbol{\beta}})$. In particular, this implies that

$$\operatorname{sign}(\widehat{\beta}_j) = \operatorname{sign}(\widehat{c}_j) \qquad \text{for } j \in \mathcal{A}. \tag{5.29}$$

PROOF. This follows immediately from (5.22) by noting that the jth element of the left-hand side is $\widehat{c}_j$, and the right-hand side is $\lambda \cdot \operatorname{sign}(\widehat{\beta}_j)$ for $j \in \mathcal{A}$. Likewise $\lambda = |\widehat{c}_j| = \widehat{C}$. $\square$

LEMMA 9. *Within an interval $\mathcal{T}$ of constant nonzero set $\mathcal{A}$, and also at $t_0 = \inf(\mathcal{T})$, the Lasso current correlations $c_j(t) = \mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}}(t))$ satisfy*

$$|c_j(t)| = \widehat{C}(t) \equiv \max\{|c_\ell(t)|\} \qquad \text{for } j \in \mathcal{A}$$

and

$$|c_j(t)| \le \widehat{C}(t) \qquad \text{for } j \notin \mathcal{A}. \tag{5.30}$$

PROOF. Equation (5.28) says that the $|c_j(t)|$ have identical values, say $\widehat{C}_t$, for $j \in \mathcal{A}$. It remains to show that $\widehat{C}_t$ has the extremum properties indicated in (5.30). For an m-vector $\mathbf{d}$ we define $\boldsymbol{\beta}(\gamma) = \widehat{\boldsymbol{\beta}}(t) + \gamma\mathbf{d}$ and $S(\gamma)$ as in (5.12), likewise $T(\gamma) = \sum |\beta_j(\gamma)|$, and

$$R_t(d) = -\dot{S}(0)/\dot{T}(0). \tag{5.31}$$

Again assuming $\widehat{\beta}_j > 0$ for $j \in \mathcal{A}$, by redefinition of $\mathbf{x}_j$ if necessary, (5.14) and (5.28) yield

$$R_t(\mathbf{d}) = 2\left[\widehat{C}_t \sum_{\mathcal{A}} d_j + \sum_{\mathcal{A}^c} c_j(t) d_j\right] \Big/ \left[\sum_{\mathcal{A}} d_j + \sum_{\mathcal{A}^c} |d_j|\right]. \tag{5.32}$$

If $d_j = 0$ for $j \notin \mathcal{A}$, and $\sum d_j \neq 0$,

$$R_t(\mathbf{d}) = 2\widehat{C}_t, \tag{5.33}$$

while if $\mathbf{d}$ has only component j nonzero we can make

$$R_t(\mathbf{d}) = 2|c_j(t)|. \tag{5.34}$$

According to Lemma 7 the Lasso solutions for $t \in \mathcal{T}$ use $d_{\mathcal{A}}$ proportional to $w_{\mathcal{A}}$ with $d_j = 0$ for $j \notin \mathcal{A}$, so

$$R_t \equiv R_t(w_{\mathcal{A}}) \tag{5.35}$$

is the downward slope of the curve $(T, S(T))$ at $T = t$, and by the definition of the Lasso must maximize $R_t(\mathbf{d})$. This shows that $\widehat{C}_t = \widehat{C}(t)$, and verifies (5.30), which also holds at $t_0 = \inf(\mathcal{T})$ by the continuity of the current correlations. □

We note that Lemmas 7–9 follow relatively easily from the Karush–Kuhn–Tucker conditions for optimality for the quadratic programming Lasso problem [Osborne, Presnell and Turlach (2000a)]; we have chosen a more geometrical argument here to demonstrate the nature of the Lasso path.

Figure 8 shows the (T, S) curve corresponding to the Lasso estimates in Figure 1. The arrow indicates the tangent to the curve at $t = 1000$, which has

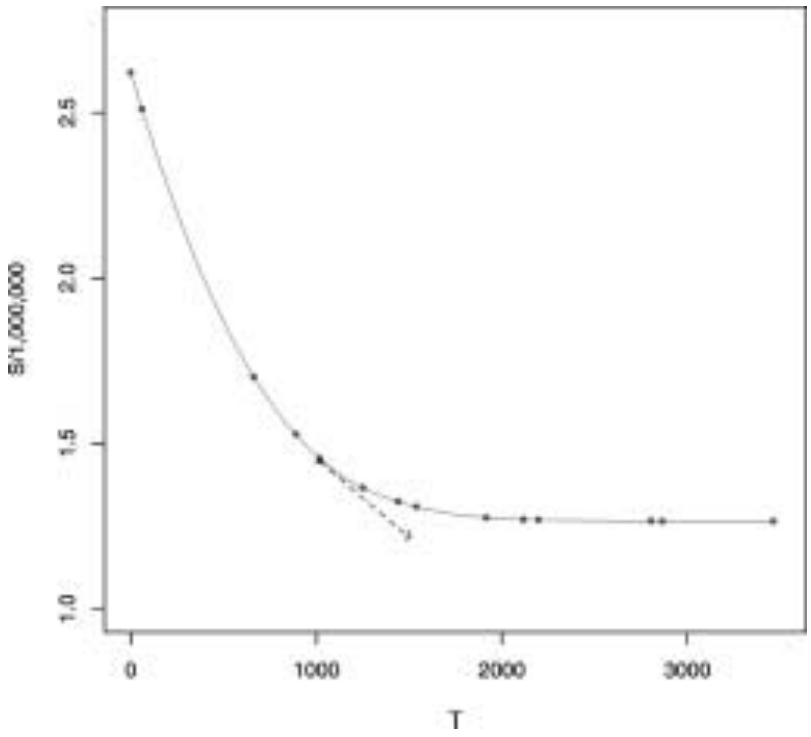

FIG. 8. *Plot of S versus T for Lasso applied to diabetes data*; *points indicate the* 12 *modified LARS steps of Figure* 1; *triangle is* (T, S) *boundary point at* $t = 1000$; *dashed arrow is tangent at* $t = 1000$, *negative slope* R_t, (5.31). *The* (T, S) *curve is a decreasing*, *convex*, *quadratic spline*.

downward slope R_{1000}. The argument above relies on the fact that $R_t(\mathbf{d})$ cannot be greater than R_t, or else there would be (T, S) values lying below the optimal curve. Using Lemmas 3 and 4 it can be shown that the (T, S) curve is always convex, as in Figure 8, being a quadratic spline with $\dot{S}(T) = -2\widehat{C}(T)$ and $\ddot{S}(T) = 2A_{\mathcal{A}}^2$.

We now consider in detail the choice of active set at a breakpoint of the piecewise linear Lasso path. Let $t = t_0$ indicate such a point, $t_0 = \inf(\mathcal{T})$ as in Lemma 9, with Lasso regression vector $\widehat{\boldsymbol{\beta}}$, prediction estimate $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$, current correlations $\widehat{\mathbf{c}} = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$, $s_j = \operatorname{sign}(\widehat{c}_j)$ and maximum absolute correlation $\widehat{C}$. Define

$$\mathcal{A}_1 = \{j : \widehat{\beta}_j \neq 0\}, \qquad \mathcal{A}_0 = \{j : \widehat{\beta}_j = 0 \text{ and } |\widehat{c}_j| = \widehat{C}\}, \tag{5.36}$$

$\mathcal{A}_{10} = \mathcal{A}_1 \cup \mathcal{A}_0$ and $\mathcal{A}_2 = \mathcal{A}_{10}^c$, and take $\boldsymbol{\beta}(\gamma) = \widehat{\boldsymbol{\beta}} + \gamma\mathbf{d}$ for some m-vector $\mathbf{d}$; also $S(\gamma) = \|\mathbf{y} - X\boldsymbol{\beta}(\gamma)\|^2$ and $T(\gamma) = \sum|\beta_j(\gamma)|$.

LEMMA 10. *The negative slope* (5.31) *at* t_0 *is bounded by* $2\widehat{C}$,

$$R(\mathbf{d}) = -\dot{S}(0)/\dot{T}(0) \leq 2\widehat{C}, \tag{5.37}$$

with equality only if $d_j = 0$ *for* $j \in \mathcal{A}_2$. *If so, the differences* $\Delta S = S(\gamma) - S(0)$ *and* $\Delta T = T(\gamma) - T(0)$ *satisfy*

$$\Delta S = -2\widehat{C}\Delta T + L(\mathbf{d})^2 \cdot (\Delta T)^2, \tag{5.38}$$

where

$$L(\mathbf{d}) = \|X\mathbf{d}/d_+\|. \tag{5.39}$$

PROOF. We can assume $\widehat{c}_j \geq 0$ for all j, by redefinition if necessary, so $\widehat{\beta}_j \geq 0$ according to Lemma 8. Proceeding as in (5.32),

$$R(\mathbf{d}) = 2\widehat{C}\left[\sum_{\mathcal{A}_{10}} d_j + \sum_{\mathcal{A}_2}(\widehat{c}_j/\widehat{C})d_j\right] \Big/ \left[\sum_{\mathcal{A}_1} d_j + \sum_{\mathcal{A}_0 \cup \mathcal{A}_2} |d_j|\right]. \tag{5.40}$$

We need $d_j \geq 0$ for $j \in \mathcal{A}_0 \cup \mathcal{A}_2$ in order to maximize (5.40), in which case

$$R(\mathbf{d}) = 2\widehat{C}\left[\sum_{\mathcal{A}_{10}} d_j + \sum_{\mathcal{A}_2}(\widehat{c}_j/\widehat{C})d_j\right] \Big/ \left[\sum_{\mathcal{A}_{10}} d_j + \sum_{\mathcal{A}_2} d_j\right]. \tag{5.41}$$

This is $< 2\widehat{C}$ unless $d_j = 0$ for $j \in \mathcal{A}_2$, verifying (5.37), and also implying

$$T(\gamma) = T(0) + \gamma \sum_{\mathcal{A}_{10}} d_j. \tag{5.42}$$

The first term on the right-hand side of (5.13) is then $-2\widehat{C}(\Delta T)$, while the second term equals $(\mathbf{d}/d_+)'X'X(\mathbf{d}/d_+)(\Delta T)^2 = L(\mathbf{d})^2$. □

Lemma 10 has an important consequence. Suppose that $\mathcal{A}$ is the current active set for the Lasso, as in (5.17), and that $\mathcal{A} \subseteq \mathcal{A}_{10}$. Then Lemma 5 says that $L(\mathbf{d})$ is $\geq A_{\mathcal{A}}$, and (5.38) gives

$$\Delta S \geq -2\widehat{C} \cdot \Delta T + A_{\mathcal{A}}^2 \cdot (\Delta T)^2, \tag{5.43}$$

with equality if $\mathbf{d}$ is chosen to give the equiangular vector $\mathbf{u}_{\mathcal{A}}$, $d_{\mathcal{A}} = S_{\mathcal{A}} w_{\mathcal{A}}$, $d_{\mathcal{A}^c} = 0$. The Lasso operates to minimize $S(T)$ so we want ΔS to be as negative as possible. Lemma 10 says that if the support of $\mathbf{d}$ is not confined to $\mathcal{A}_{10}$, then $\dot{S}(0)$ exceeds the optimum value $-2\widehat{C}$; if it is confined, then $\dot{S}(0) = -2\widehat{C}$ but $\ddot{S}(0)$ exceeds the minimum value $2A_{\mathcal{A}}$ unless $d_{\mathcal{A}}$ is proportional to $S_{\mathcal{A}} w_{\mathcal{A}}$ as in (5.17).

Suppose that $\widehat{\boldsymbol{\beta}}$, a Lasso solution, exactly equals a $\widehat{\boldsymbol{\beta}}$ obtained from the Lasso-modified LARS algorithm, henceforth called LARS–Lasso, as at $t = 1000$ in Figures 1 and 3. We know from Lemma 7 that subsequent Lasso estimates will follow a linear track determined by some subset $\mathcal{A}$, $\boldsymbol{\mu}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma \mathbf{u}_{\mathcal{A}}$, and so will the LARS–Lasso estimates, but to verify Theorem 1 we need to show that "$\mathcal{A}$" is the same set in both cases.

Lemmas 4–7 put four constraints on the Lasso choice of $\mathcal{A}$. Define $\mathcal{A}_1$, $\mathcal{A}_0$ and $\mathcal{A}_{10}$ as at (5.36).

CONSTRAINT 1. $\mathcal{A}_1 \subseteq \mathcal{A}$. This follows from Lemma 7 since for sufficiently small γ the subsequent Lasso coefficients (5.17),

$$\widehat{\beta}_{\mathcal{A}}(\gamma) = \widehat{\beta}_{\mathcal{A}} + \gamma S_{\mathcal{A}} w_{\mathcal{A}}, \tag{5.44}$$

will have $\widehat{\beta}_j(\gamma) \neq 0$, $j \in \mathcal{A}_1$.

CONSTRAINT 2. $\mathcal{A} \subseteq \mathcal{A}_{10}$. Lemma 10, (5.37) shows that the Lasso choice $\widehat{\mathbf{d}}$ in $\boldsymbol{\beta}(\gamma) = \widehat{\boldsymbol{\beta}} + \gamma \widehat{\mathbf{d}}$ must have its nonzero support in $\mathcal{A}_{10}$, or equivalently that $\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma \mathbf{u}_{\mathcal{A}}$ must have $\mathbf{u}_{\mathcal{A}} \in \mathcal{L}(X_{\mathcal{A}_{10}})$. (It is possible that $\mathbf{u}_{\mathcal{A}}$ happens to equal $\mathbf{u}_{\mathcal{B}}$ for some $\mathcal{B} \supset \mathcal{A}_{10}$, but that does not affect the argument below.)

CONSTRAINT 3. $w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$ cannot have $\operatorname{sign}(w_j) \neq \operatorname{sign}(\widehat{c}_j)$ for any coordinate $j \in \mathcal{A}_0$. If it does, then $\operatorname{sign}(\widehat{\beta}_j(\gamma)) \neq \operatorname{sign}(\widehat{c}_j(\gamma))$ for sufficiently small γ, violating Lemma 8.

CONSTRAINT 4. Subject to Constraints 1–3, $\mathcal{A}$ must minimize $A_{\mathcal{A}}$. This follows from Lemma 10 as in (5.43), and the requirement that the Lasso curve $S(T)$ declines at the fastest possible rate.

Theorem 1 follows by induction: beginning at $\widehat{\boldsymbol{\beta}}_0 = 0$, we follow the LARS–Lasso algorithm and show that at every succeeding step it must continue to agree with the Lasso definition (1.5). First of all, suppose that $\widehat{\boldsymbol{\beta}}$, our hypothesized Lasso and LARS–Lasso solution, has occurred strictly *within* a LARS–Lasso step. Then

$\mathcal{A}_0$ is empty so that Constraints 1 and 2 imply that $\mathcal{A}$ cannot change its current value: the equivalence between Lasso and LARS–Lasso must continue at least to the end of the step.

The one-at-a-time assumption of Theorem 1 says that at a LARS–Lasso breakpoint, $\mathcal{A}_0$ has exactly one member, say j_0, so $\mathcal{A}$ must equal $\mathcal{A}_1$ or $\mathcal{A}_{10}$. There are two cases: if j_0 has just been *added* to the set $\{|\widehat{c}_j| = \widehat{C}\}$, then Lemma 4 says that $\operatorname{sign}(w_{j_0}) = \operatorname{sign}(\widehat{c}_{j_0})$, so that Constraint 3 is not violated; the other three constraints and Lemma 5 imply that the Lasso choice $\mathcal{A} = \mathcal{A}_{10}$ agrees with the LARS–Lasso algorithm. The other case has j_0 *deleted* from the active set as in (3.6). Now the choice $\mathcal{A} = \mathcal{A}_{10}$ is ruled out by Constraint 3: it would keep $w_{\mathcal{A}}$ the same as in the previous LARS–Lasso step, and we know that that was stopped in (3.6) to prevent a sign contradiction at coordinate j_0. In other words, $\mathcal{A} = \mathcal{A}_1$, in accordance with the Lasso modification of LARS. This completes the proof of Theorem 1.

A LARS–Lasso algorithm is available even if the one-at-a-time condition does not hold, but at the expense of additional computation. Suppose, for example, *two* new members j_1 and j_2 are added to the set $\{|\widehat{c}_j| = \widehat{C}\}$, so $\mathcal{A}_0 = \{j_1, j_2\}$. It is possible but not certain that $\mathcal{A}_{10}$ does not violate Constraint 3, in which case $\mathcal{A} = \mathcal{A}_{10}$. However, if it does violate Constraint 3, then both possibilities $\mathcal{A} = \mathcal{A}_1 \cup \{j_1\}$ and $\mathcal{A} = \mathcal{A}_1 \cup \{j_2\}$ must be examined to see which one gives the smaller value of $A_{\mathcal{A}}$. Since one-at-a-time computations, perhaps with some added $\mathbf{y}$ jitter, apply to all practical situations, the LARS algorithm described in Section 7 is not equipped to handle many-at-a-time problems.

6. Stagewise properties. The main goal of this section is to verify Theorem 2. Doing so also gives us a chance to make a more detailed comparison of the LARS and Stagewise procedures. Assume that $\widehat{\boldsymbol{\beta}}$ is a Stagewise estimate of the regression coefficients, for example, as indicated at $\sum |\widehat{\beta}_j| = 2000$ in the right panel of Figure 1, with prediction vector $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$, current correlations $\widehat{\mathbf{c}} = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$, $\widehat{C} = \max\{|\widehat{c}_j|\}$ and maximal set $\mathcal{A} = \{j : |\widehat{c}_j| = \widehat{C}\}$. We must show that successive Stagewise estimates of $\boldsymbol{\beta}$ develop according to the modified LARS algorithm of Theorem 2, henceforth called LARS–Stagewise. For convenience we can assume, by redefinition of $\mathbf{x}_j$ as $-\mathbf{x_j}$, if necessary, that the signs $s_j = \operatorname{sign}(\widehat{c}_j)$ are all nonnegative.

As in (3.8)–(3.10) we suppose that the Stagewise procedure (1.7) has taken N additional ε-steps forward from $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$, giving new prediction vector $\widehat{\boldsymbol{\mu}}(N)$.

LEMMA 11. *For sufficiently small ε, only $j \in \mathcal{A}$ can have $P_j = N_j/N > 0$.*

PROOF. Letting $N\varepsilon \equiv \gamma$, $\|\widehat{\boldsymbol{\mu}}(N) - \widehat{\boldsymbol{\mu}}\| \le \gamma$ so that $\widehat{\mathbf{c}}(N) = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}}(N))$ satisfies

$$|\widehat{c}_j(N) - \widehat{c}_j| = \left|\mathbf{x}_j'(\widehat{\boldsymbol{\mu}}(N) - \widehat{\boldsymbol{\mu}})\right| \le \|\mathbf{x}_j\| \cdot \|\widehat{\boldsymbol{\mu}}(N) - \widehat{\boldsymbol{\mu}}\| \le \gamma. \tag{6.1}$$

For $\gamma < \frac{1}{2}[\widehat{C} - \max_{\mathcal{A}^c}\{\widehat{c}_j\}]$, j in $\mathcal{A}^c$ cannot have maximal current correlation and can never be involved in the N steps. $\square$

Lemma 11 says that we can write the developing Stagewise prediction vector as

$$\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma \mathbf{v}, \qquad \text{where } \mathbf{v} = X_{\mathcal{A}} P_{\mathcal{A}}, \tag{6.2}$$

$P_{\mathcal{A}}$ a vector of length $|\mathcal{A}|$, with components N_j/N for $j \in \mathcal{A}$. The nature of the Stagewise procedure puts three constraints on $\mathbf{v}$, the most obvious of which is the following.

CONSTRAINT I. The vector $\mathbf{v} \in \mathcal{S}_{\mathcal{A}}^+$, the nonnegative simplex

$$\mathcal{S}_{\mathcal{A}}^+ = \left\{ \mathbf{v} : \mathbf{v} = \sum_{j \in \mathcal{A}} \mathbf{x}_j P_j,\ P_j \geq 0,\ \sum_{j \in \mathcal{A}} P_j = 1 \right\}. \tag{6.3}$$

Equivalently, $\gamma \mathbf{v} \in \mathcal{C}_{\mathcal{A}}$, the convex cone (3.12).

The Stagewise procedure, unlike LARS, is not required to use all of the maximal set $\mathcal{A}$ as the active set, and can instead restrict the nonzero coordinates P_j to a subset $\mathcal{B} \subseteq \mathcal{A}$. Then $\mathbf{v} \in \mathcal{L}(X_{\mathcal{B}})$, the linear space spanned by the columns of $X_{\mathcal{B}}$, but not all such vectors $\mathbf{v}$ are allowable Stagewise forward directions.

CONSTRAINT II. The vector $\mathbf{v}$ must be proportional to the equiangular vector $\mathbf{u}_{\mathcal{B}}$, (2.6), that is, $\mathbf{v} = \mathbf{v}_{\mathcal{B}}$, (5.8),

$$\mathbf{v}_{\mathcal{B}} = A_{\mathcal{B}}^2 X_{\mathcal{B}} \mathcal{G}_{\mathcal{B}}^{-1} \mathbf{1}_{\mathcal{B}} = A_{\mathcal{B}} \mathbf{u}_{\mathcal{B}}. \tag{6.4}$$

Constraint II amounts to requiring that the current correlations in $\mathcal{B}$ decline at an equal rate: since

$$\widehat{c}_j(\gamma) = \mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}} - \gamma \mathbf{v}) = \widehat{c}_j - \gamma \mathbf{x}_j' \mathbf{v}, \tag{6.5}$$

we need $X_{\mathcal{B}}' \mathbf{v} = \lambda \mathbf{1}_{\mathcal{B}}$ for some $\lambda > 0$, implying $\mathbf{v} = \lambda \mathcal{G}_{\mathcal{B}}^{-1} \mathbf{1}_{\mathcal{B}}$; choosing $\lambda = A_{\mathcal{B}}^2$ satisfies Constraint II. Violating Constraint II makes the current correlations $\widehat{c}_j(\gamma)$ unequal so that the Stagewise algorithm as defined at (1.7) could not proceed in direction $\mathbf{v}$.

Equation (6.4) gives $X_{\mathcal{B}}' \mathbf{v}_{\mathcal{B}} = A_{\mathcal{B}}^2 \mathbf{1}_{\mathcal{B}}$, or

$$\mathbf{x}_j' \mathbf{v}_{\mathcal{B}} = A_{\mathcal{B}}^2 \qquad \text{for } j \in \mathcal{B}. \tag{6.6}$$

CONSTRAINT III. The vector $\mathbf{v} = \mathbf{v}_{\mathcal{B}}$ must satisfy

$$\mathbf{x}_j' \mathbf{v}_{\mathcal{B}} \geq A_{\mathcal{B}}^2 \qquad \text{for } j \in \mathcal{A} - \mathcal{B}. \tag{6.7}$$

Constraint III follows from (6.5). It says that the current correlations for members of $\mathcal{A} = \{j : |\widehat{c}_j| = \widehat{C}\}$ *not* in $\mathcal{B}$ must decline at least as quickly as those in $\mathcal{B}$. If this were not true, then $\mathbf{v}_{\mathcal{B}}$ would not be an allowable direction for Stagewise development since variables in $\mathcal{A} - \mathcal{B}$ would immediately reenter (1.7).

To obtain strict inequality in (6.7), let $\mathcal{B}_0 \subset \mathcal{A} - \mathcal{B}$ be the set of indices for which $\mathbf{x}_j'\mathbf{v}_{\mathcal{B}} = A_{\mathcal{B}}^2$. It is easy to show that $\mathbf{v}_{\mathcal{B}\cup\mathcal{B}_o} = \mathbf{v}_{\mathcal{B}}$. In other words, if we take $\mathcal{B}$ to be the *largest* set having a given $\mathbf{v}_{\mathcal{B}}$ proportional to its equiangular vector, then $\mathbf{x}_j'\mathbf{v}_{\mathcal{B}} > A_{\mathcal{B}}^2$ for $j \in \mathcal{A} - \mathcal{B}$.

Writing $\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma\mathbf{v}$ as in (6.2) presupposes that the Stagewise solutions follow a piecewise linear track. However, the presupposition can be reduced to one of piecewise differentiability by taking γ infinitesimally small. We can always express the family of Stagewise solutions as $\widehat{\boldsymbol{\beta}}(z)$, where the real-valued parameter Z plays the role of T for the Lasso, increasing from 0 to some maximum value as $\widehat{\boldsymbol{\beta}}(z)$ goes from $\mathbf{0}$ to the full OLS estimate. [The choice $Z = T$ used in Figure 1 may not necessarily yield a one-to-one mapping; $Z = S(\mathbf{0}) - S(\widehat{\boldsymbol{\beta}})$, the reduction in residual squared error, always does.] We suppose that the Stagewise estimate $\widehat{\boldsymbol{\beta}}(z)$ is everywhere right differentiable with respect to z. Then the right derivative

$$\widehat{\mathbf{v}} = d\widehat{\boldsymbol{\beta}}(z)/dz \tag{6.8}$$

must obey the three constraints.

The definition of the idealized Stagewise procedure in Section 3.2, in which $\varepsilon \to 0$ in rule (1.7), is somewhat vague but the three constraints apply to any reasonable interpretation. It turns out that the LARS–Stagewise algorithm satisfies the constraints and is unique in doing so. This is the meaning of Theorem 2. [Of course the LARS–Stagewise algorithm is also supported by direct numerical comparisons with (1.7), as in Figure 1's right panel.]

If $\mathbf{u}_{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}$, then $\mathbf{v} = \mathbf{v}_{\mathcal{A}}$ obviously satisfies the three constraints. The interesting situation for Theorem 2 is $\mathbf{u}_{\mathcal{A}} \notin \mathcal{C}_{\mathcal{A}}$, which we now assume to be the case. Any subset $\mathcal{B} \subset \mathcal{A}$ determines a face of the convex cone of dimension $|\mathcal{B}|$, the face having $P_j > 0$ in (3.12) for $j \in \mathcal{B}$ and $P_j = 0$ for $j \in \mathcal{A} - \mathcal{B}$. The orthogonal projection of $\mathbf{u}_{\mathcal{A}}$ into the linear subspace $\mathcal{L}(X_{\mathcal{B}})$, say $\mathrm{Proj}_{\mathcal{B}}(\mathbf{u}_{\mathcal{A}})$, is proportional to $\mathcal{B}$'s equiangular vector $\mathbf{u}_{\mathcal{B}}$: using (2.7),

$$\mathrm{Proj}_{\mathcal{B}}(\mathbf{u}_{\mathcal{A}}) = X_{\mathcal{B}}\mathcal{G}_{\mathcal{B}}^{-1}X_{\mathcal{B}}'\mathbf{u}_{\mathcal{A}} = X_{\mathcal{B}}\mathcal{G}_{\mathcal{B}}^{-1}A_{\mathcal{A}}\mathbf{1}_{\mathcal{B}} = (A_{\mathcal{A}}/A_{\mathcal{B}})\cdot\mathbf{u}_{\mathcal{B}}, \tag{6.9}$$

or equivalently

$$\mathrm{Proj}_{\mathcal{B}}(\mathbf{v}_{\mathcal{A}}) = (A_{\mathcal{A}}/A_{\mathcal{B}})^2\mathbf{v}_{\mathcal{B}}. \tag{6.10}$$

The nearest point to $\mathbf{u}_{\mathcal{A}}$ in $\mathcal{C}_{\mathcal{A}}$, say $\widehat{\mathbf{u}}_{\mathcal{A}}$, is of the form $\Sigma_{\mathcal{A}}\mathbf{x}_j\widehat{P}_j$ with $\widehat{P}_j \geq 0$. Therefore $\widehat{\mathbf{u}}_{\mathcal{A}}$ exists strictly within face $\widehat{\mathcal{B}}$, where $\widehat{\mathcal{B}} = \{j : \widehat{P}_j > 0\}$, and must equal $\mathrm{Proj}_{\widehat{\mathcal{B}}}(\mathbf{u}_{\mathcal{A}})$. According to (6.9), $\widehat{\mathbf{u}}_{\mathcal{A}}$ is proportional to $\widehat{\mathcal{B}}$'s equiangular vector $\mathbf{u}_{\widehat{\mathcal{B}}}$, and also to $\mathbf{v}_{\widehat{\mathcal{B}}} = A_{\mathcal{B}}\mathbf{u}_{\mathcal{B}}$. In other words $\mathbf{v}_{\widehat{\mathcal{B}}}$ satisfies Constraint II, and it obviously also satisfies Constraint I. Figure 9 schematically illustrates the geometry.

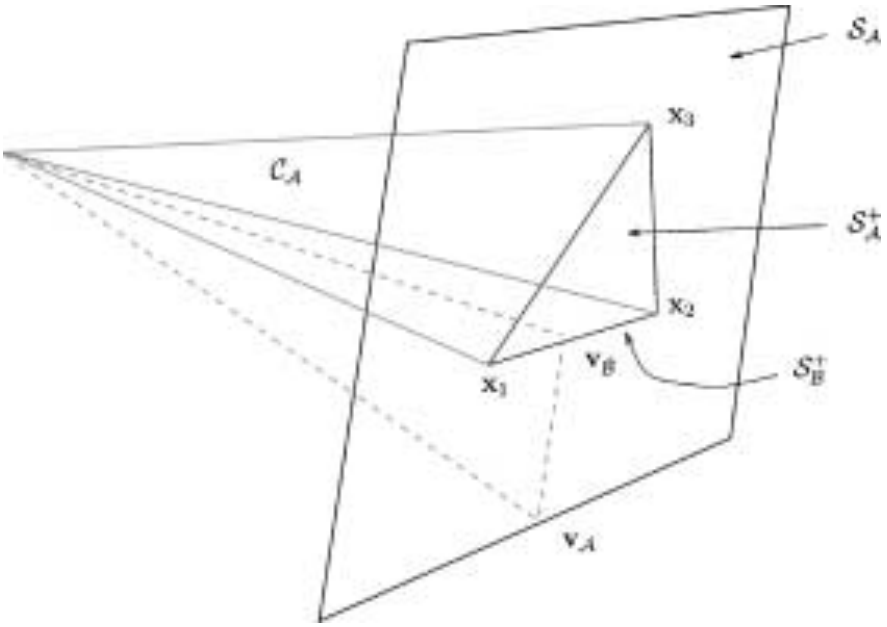

FIG. 9. *The geometry of the LARS–Stagewise modification.*

LEMMA 12. *The vector* $\mathbf{v}_{\hat{\mathcal{B}}}$ *satisfies Constraints* I–III, *and conversely if* $\mathbf{v}$ *satisfies the three constraints, then* $\mathbf{v} = \mathbf{v}_{\hat{\mathcal{B}}}$.

PROOF. Let $\mathrm{Cos} \equiv A_{\mathcal{A}}/A_{\mathcal{B}}$ and $\mathrm{Sin} = [1 - \mathrm{Cos}^2]^{1/2}$, the latter being greater than zero by Lemma 5. For any face $\mathcal{B} \subset \mathcal{A}$, (6.9) implies

$$\mathbf{u}_{\mathcal{A}} = \mathrm{Cos} \cdot \mathbf{u}_{\mathcal{B}} + \mathrm{Sin} \cdot \mathbf{z}_{\mathcal{B}}, \tag{6.11}$$

where $\mathbf{z}_{\mathcal{B}}$ is a unit vector orthogonal to $\mathcal{L}(X_{\mathcal{B}})$, pointing away from $\mathcal{C}_{\mathcal{A}}$. By an n-dimensional coordinate rotation we can make $\mathcal{L}(X_{\mathcal{B}}) = \mathcal{L}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_J)$, $J = |\mathcal{B}|$, the space of n-vectors with last $n - J$ coordinates zero, and also

$$\mathbf{u}_{\mathcal{B}} = (1, \mathbf{0}, 0, \mathbf{0}), \qquad \mathbf{u}_{\mathcal{A}} = (\mathrm{Cos}, \mathbf{0}, \mathrm{Sin}, \mathbf{0}), \tag{6.12}$$

the first $\mathbf{0}$ having length $J - 1$, the second $\mathbf{0}$ length $n - J - 1$. Then we can write

$$\mathbf{x}_j = \big(A_{\mathcal{B}}, \mathbf{x}_{j_2}, 0, \mathbf{0}\big) \qquad \text{for } j \in \mathcal{B}, \tag{6.13}$$

the first coordinate $A_{\mathcal{B}}$ being required since $\mathbf{x}_j' \mathbf{u}_{\mathcal{B}} = A_{\mathcal{B}}$, (2.7). Notice that $\mathbf{x}_j' \mathbf{u}_{\mathcal{A}} = \mathrm{Cos} \cdot A_{\mathcal{B}} = A_{\mathcal{A}}$, as also required by (2.7).

For $\ell \in \mathcal{A} - \mathcal{B}$ denote $\mathbf{x}_\ell$ as

$$\mathbf{x}_\ell = \big(x_{\ell_1}, \mathbf{x}_{\ell_2}, x_{\ell_3}, \mathbf{x}_{\ell_4}\big), \tag{6.14}$$

so (2.7) yields

$$A_{\mathcal{A}} = \mathbf{x}_\ell' \mathbf{u}_{\mathcal{A}} = \mathrm{Cos} \cdot x_{\ell_1} + \mathrm{Sin} \cdot x_{\ell_3}. \tag{6.15}$$

Now assume $\mathcal{B} = \widehat{\mathcal{B}}$. In this case a separating hyperplane $\mathcal{H}$ orthogonal to $\mathbf{z}_{\hat{\mathcal{B}}}$ in (6.11) passes between the convex cone $\mathcal{C}_{\mathcal{A}}$ and $\mathbf{u}_{\mathcal{A}}$, through $\widehat{\mathbf{u}}_{\mathcal{A}} = \mathrm{Cos} \cdot \mathbf{u}_{\hat{\mathcal{B}}}$, implying $x_{\ell_3} \le 0$ [i.e., $\mathbf{x}_\ell$ and $\mathbf{u}_{\mathcal{A}}$ are on opposite sides of $\mathcal{H}$, x_{ℓ_3} being negative since the corresponding coordinate of $\mathbf{u}_{\mathcal{A}}$, "Sin" in (6.12), is positive]. Equation (6.15) gives $\mathrm{Cos} \cdot x_{\ell_1} \ge A_{\mathcal{A}} = \mathrm{Cos} \cdot A_{\hat{\mathcal{B}}}$ or

$$\mathbf{x}_\ell' \mathbf{v}_{\hat{\mathcal{B}}} = x_\ell' (A_{\hat{\mathcal{B}}} \mathbf{u}_{\hat{\mathcal{B}}}) = A_{\hat{\mathcal{B}}} x_{\ell_1} \ge A_{\hat{\mathcal{B}}}^2, \tag{6.16}$$

verifying that Constraint III is satisfied.

Conversely suppose that $\mathbf{v}$ satisfies Constraints I–III so that $\mathbf{v} \in \mathcal{S}_{\mathcal{A}}^+$ and $\mathbf{v} = \mathbf{v}_{\mathcal{B}}$ for the nonzero coefficient set $\mathcal{B}$: $\mathbf{v}_{\mathcal{B}} = \Sigma_{\mathcal{B}} \mathbf{x}_j P_j$, $P_j > 0$. Let $\mathcal{H}$ be the hyperplane passing through $\mathrm{Cos} \cdot \mathbf{u}_{\mathcal{B}}$ orthogonally to $\mathbf{z}_{\mathcal{B}}$, (6.9), (6.11). If $\mathbf{v}_{\mathcal{B}} \ne \mathbf{v}_{\hat{\mathcal{B}}}$, then at least one of the vectors $\mathbf{x}_\ell$, $\ell \in \mathcal{A} - \mathcal{B}$, must lie on the same side of $\mathcal{H}$ as $\mathbf{u}_{\mathcal{A}}$, so that $x_{\ell_3} > 0$ (or else $\mathcal{H}$ would be a separating hyperplane between $\mathbf{u}_{\mathcal{A}}$ and $\mathcal{C}_{\mathcal{A}}$, and $\mathbf{v}_{\mathcal{B}}$ would be proportional to $\widehat{\mathbf{u}}_{\mathcal{A}}$, the nearest point to $\mathbf{u}_{\mathcal{A}}$ in $\mathcal{C}_{\mathcal{A}}$, implying $\mathbf{v}_{\mathcal{B}} = \mathbf{v}_{\hat{\mathcal{B}}}$). Now (6.15) gives $\mathrm{Cos} \cdot x_{\ell_1} < A_{\mathcal{A}} = \mathrm{Cos} \cdot A_{\mathcal{B}}$, or

$$\mathbf{x}_\ell' \mathbf{v}_{\mathcal{B}} = \mathbf{x}_\ell' (A_{\mathcal{B}} \mathbf{u}_{\mathcal{B}}) = A_{\mathcal{B}} x_{\ell_1} < A_{\mathcal{B}}^2. \tag{6.17}$$

This violates Constraint III, showing that $\mathbf{v}$ must equal $\mathbf{v}_{\hat{\mathcal{B}}}$. □

Notice that the direction of advance $\widehat{\mathbf{v}} = \mathbf{v}_{\hat{\mathcal{B}}}$ of the idealized Stagewise procedure is a function only of the current maximal set $\widehat{\mathcal{A}} = \{j : |\widehat{c}_j| = \widehat{C}\}$, say $\widehat{\mathbf{v}} = \phi(\widehat{\mathcal{A}})$. In the language of (6.7),

$$\frac{d\widehat{\boldsymbol{\beta}}(z)}{dz} = \phi(\widehat{\mathcal{A}}). \tag{6.18}$$

The LARS–Stagewise algorithm of Theorem 2 produces an evolving family of estimates $\widehat{\boldsymbol{\beta}}$ that everywhere satisfies (6.18). This is true at every LARS–Stagewise breakpoint by the definition of the Stagewise modification. It is also true between breakpoints. Let $\widehat{\mathcal{A}}$ be the maximal set at the breakpoint, giving $\widehat{\mathbf{v}} = \mathbf{v}_{\hat{\mathcal{B}}} = \phi(\widehat{\mathcal{A}})$. In the succeeding LARS–Stagewise interval $\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma \mathbf{v}_{\hat{\mathcal{B}}}$, the maximal set is immediately reduced to $\widehat{\mathcal{B}}$, according to properties (6.6), (6.7) of $\mathbf{v}_{\hat{\mathcal{B}}}$, at which it stays during the entire interval. However, $\phi(\widehat{\mathcal{B}}) = \phi(\widehat{\mathcal{A}}) = \mathbf{v}_{\hat{\mathcal{B}}}$ since $\mathbf{v}_{\hat{\mathcal{B}}} \in \mathcal{C}_{\hat{\mathcal{B}}}$, so the LARS–Stagewise procedure, which continues in the direction $\widehat{\mathbf{v}}$ until a new member is added to the active set, continues to obey the idealized Stagewise equation (6.18).

All of this shows that the LARS–Stagewise algorithm produces a legitimate version of the idealized Stagewise track. The converse of Lemma 12 says that there are no other versions, verifying Theorem 2.

The Stagewise procedure has its potential generality as an advantage over LARS and Lasso: it is easy to define forward Stagewise methods for a wide variety of nonlinear fitting problems, as in Hastie, Tibshirani and Friedman [(2001), Chapter 10, which begins with a Stagewise analysis of "boosting"]. Comparisons

with LARS and Lasso within the linear model framework, as at the end of Section 3.2, help us better understand Stagewise methodology. This section's results permit further comparisons.

Consider proceeding forward from $\widehat{\boldsymbol{\mu}}$ along unit vector $\mathbf{u}$, $\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma\mathbf{u}$, two interesting choices being the LARS direction $\mathbf{u}_{\hat{\mathcal{A}}}$ and the Stagewise direction $\widehat{\boldsymbol{\mu}}_{\hat{\mathcal{B}}}$. For $\mathbf{u} \in \mathcal{L}(X_{\hat{\mathcal{A}}})$, the rate of change of $S(\gamma) = \|\mathbf{y} - \widehat{\boldsymbol{\mu}}(\gamma)\|^2$ is

$$-\frac{\partial S(\gamma)}{\partial \gamma}\bigg|_0 = 2\widehat{C} \cdot \frac{\mathbf{u}'_{\mathcal{A}} \cdot \mathbf{u}}{A_{\hat{\mathcal{A}}}}, \tag{6.19}$$

(6.19) following quickly from (5.14). This shows that the LARS direction $\mathbf{u}_{\hat{\mathcal{A}}}$ maximizes the instantaneous decrease in S. The ratio

$$\frac{\partial S_{\text{Stage}}(\gamma)}{\partial \gamma}\bigg|_0 \Big/ \frac{\partial S_{\text{LARS}}(\gamma)}{\partial \gamma}\bigg|_0 = \frac{A_{\hat{\mathcal{A}}}}{A_{\hat{\mathcal{B}}}}, \tag{6.20}$$

equaling the quantity "Cos" in (6.15).

The comparison goes the other way for the maximum absolute correlation $\widehat{C}(\gamma)$. Proceeding as in (2.15),

$$-\frac{\partial \widehat{C}(\gamma)}{\partial \gamma}\bigg|_0 = \min_{\hat{\mathcal{A}}}\{|x'_j\mathbf{u}|\}. \tag{6.21}$$

The argument for Lemma 12, using Constraints II and III, shows that $\mathbf{u}_{\hat{\mathcal{B}}}$ maximizes (6.21) at $A_{\hat{\mathcal{B}}}$, and that

$$\frac{\partial \widehat{C}_{\text{LARS}}(\gamma)}{\partial \gamma}\bigg|_0 \Big/ \frac{\partial \widehat{C}_{\text{Stage}}(\gamma)}{\partial \gamma}\bigg|_0 = \frac{A_{\hat{\mathcal{A}}}}{A_{\hat{\mathcal{B}}}}. \tag{6.22}$$

The original motivation for the Stagewise procedure was to minimize residual squared error within a framework of parsimonious forward search. However, (6.20) shows that Stagewise is less greedy than LARS in this regard, it being more accurate to describe Stagewise as striving to minimize the maximum absolute residual correlation.

7. Computations. The entire sequence of steps in the LARS algorithm with $m < n$ variables requires $O(m^3 + nm^2)$ computations—the cost of a least squares fit on m variables.

In detail, at the kth of m steps, we compute $m - k$ inner products c_{jk} of the nonactive $\mathbf{x}_j$ with the current residuals to identify the next active variable, and then invert the $k \times k$ matrix $\mathcal{G}_k = X'_k X_k$ to find the next LARS direction. We do this by updating the Cholesky factorization R_{k-1} of $\mathcal{G}_{k-1}$ found at the previous step [Golub and Van Loan (1983)]. At the final step m, we have computed the Cholesky $R = R_m$ for the full cross-product matrix, which is the dominant calculation for a least squares fit. Hence the LARS sequence can be seen as a Cholesky factorization with a guided ordering of the variables.

The computations can be reduced further by recognizing that the inner products above can be updated at each iteration using the cross-product matrix $X'X$ and the current directions. For $m \gg n$, this strategy is counterproductive and is not used.

For the *lasso* modification, the computations are similar, except that occasionally one has to drop a variable, and hence *downdate* R_k [costing at most $O(m^2)$ operations per downdate]. For the *stagewise* modification of LARS, we need to check at each iteration that the components of w are all positive. If not, one or more variables are dropped [using the *inner loop* of the NNLS algorithm described in Lawson and Hanson (1974)], again requiring downdating of R_k. With many correlated variables, the stagewise version can take many more steps than LARS because of frequent dropping and adding of variables, increasing the computations by a factor up to 5 or more in extreme cases.

The LARS algorithm (in any of the three states above) works gracefully for the case where there are many more variables than observations: $m \gg n$. In this case LARS terminates at the saturated least squares fit after $n - 1$ variables have entered the active set [at a cost of $O(n^3)$ operations]. (This number is $n - 1$ rather than n, because the columns of X have been mean centered, and hence it has row-rank $n - 1$.) We make a few more remarks about the $m \gg n$ case in the *lasso* state:

1. The LARS algorithm continues to provide Lasso solutions along the way, and the final solution highlights the fact that a Lasso fit can have no more than $n - 1$ (mean centered) variables with nonzero coefficients.
2. Although the model involves no more than $n - 1$ variables at any time, the number of *different* variables ever to have entered the model during the entire sequence can be—and typically is—greater than $n - 1$.
3. The model sequence, particularly near the saturated end, tends to be quite variable with respect to small changes in $\mathbf{y}$.
4. The estimation of σ^2 may have to depend on an auxiliary method such as nearest neighbors (since the final model is saturated). We have not investigated the accuracy of the simple approximation formula (4.12) for the case $m > n$.

Documented S-PLUS implementations of LARS and associated functions are available from www-stat.stanford.edu/~hastie/Papers/; the diabetes data also appears there.

8. Boosting procedures. One motivation for studying the Forward Stagewise algorithm is its usefulness in adaptive fitting for data mining. In particular, Forward Stagewise ideas are used in "boosting," an important class of fitting methods for data mining introduced by Freund and Schapire (1997). These methods are one of the hottest topics in the area of machine learning, and one of the most effective prediction methods in current use. Boosting can use any adaptive fitting procedure as its "base learner" (model fitter): trees are a popular choice, as implemented in CART [Breiman, Friedman, Olshen and Stone (1984)].

Friedman, Hastie and Tibshirani (2000) and Friedman (2001) studied boosting and proposed a number of procedures, the most relevant to this discussion being *least squares boosting*. This procedure works by successive fitting of regression trees to the current residuals. Specifically we start with the residual $\mathbf{r} = \mathbf{y}$ and the fit $\hat{\mathbf{y}} = 0$. We fit a tree in $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ to the response $\mathbf{y}$ giving a fitted tree $\mathbf{t}_1$ (an n-vector of fitted values). Then we update $\hat{\mathbf{y}}$ to $\hat{\mathbf{y}} + \varepsilon \cdot \mathbf{t}_1$, $\mathbf{r}$ to $\mathbf{y} - \hat{\mathbf{y}}$ and continue for many iterations. Here ε is a small positive constant. Empirical studies show that small values of ε work better than $\varepsilon = 1$: in fact, for prediction accuracy "the smaller the better." The only drawback in taking very small values of ε is computational slowness.

A major research question has been why boosting works so well, and specifically why is ε-shrinkage so important? To understand boosted trees in the present context, we think of our predictors not as our original variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, but instead as the set of all trees $\mathbf{t}_k$ that could be fitted to our data. There is a strong similarity between least squares boosting and Forward Stagewise regression as defined earlier. Fitting a tree to the current residual is a numerical way of finding the "predictor" most correlated with the residual. Note, however, that the greedy algorithms used in CART do not search among all possible trees, but only a subset of them. In addition the set of all trees, including a parametrization for the predicted values in the terminal nodes, is infinite. Nevertheless one can define idealized versions of least-squares boosting that look much like Forward Stagewise regression.

Hastie, Tibshirani and Friedman (2001) noted the the striking similarity between Forward Stagewise regression and the Lasso, and conjectured that this may help explain the success of the Forward Stagewise process used in least squares boosting. That is, in some sense least squares boosting may be carrying out a Lasso fit on the infinite set of tree predictors. Note that direct computation of the Lasso via the LARS procedure would not be feasible in this setting because the number of trees is infinite and one could not compute the optimal step length. However, Forward Stagewise regression is feasible because it only need find the the most correlated predictor among the infinite set, where it approximates by numerical search.

In this paper we have established the connection between the Lasso and Forward Stagewise regression. We are now thinking about how these results can help to understand and improve boosting procedures. One such idea is a modified form of Forward Stagewise: we find the best tree as usual, but rather than taking a small step in only that tree, we take a small least squares step in all trees currently in our model. One can show that for small step sizes this procedure approximates LARS; its advantage is that it can be carried out on an infinite set of predictors such as trees.

APPENDIX

A.1. Local linearity and Lemma 2.

CONVENTIONS. We write $\mathbf{x}_l$ with subscript l for members of the active set $\mathcal{A}_k$. Thus $\mathbf{x}_l$ denotes the lth variable to enter, being an abuse of notation for $s_l\mathbf{x}_{j(l)} = \operatorname{sgn}(\hat{c}_{j(l)})\mathbf{x}_{j(l)}$. Expressions $\mathbf{x}_l'(\mathbf{y} - \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y})) = \hat{C}_k(\mathbf{y})$ and $\mathbf{x}_l'\mathbf{u}_k = A_k$ clearly do not depend on which $\mathbf{x}_l \in \mathcal{A}_k$ we choose.

By writing $j \notin \mathcal{A}_k$, we intend that both $\mathbf{x}_j$ and $-\mathbf{x}_j$ are candidates for inclusion at the next step. One could think of negative indices $-j$ corresponding to "new" variables $\mathbf{x}_{-j} = -\mathbf{x}_j$.

The active set $\mathcal{A}_k(\mathbf{y})$ depends on the data $\mathbf{y}$. When $\mathcal{A}_k(\mathbf{y})$ is the same for all $\mathbf{y}$ in a neighborhood of $\mathbf{y}_0$, we say that $\mathcal{A}_k(\mathbf{y})$ is locally fixed [at $\mathcal{A}_k = \mathcal{A}_k(\mathbf{y}_0)$].

A function $g(\mathbf{y})$ is locally Lipschitz at $\mathbf{y}$ if for all sufficiently small vectors $\Delta\mathbf{y}$,

$$\|\Delta g\| = \|g(\mathbf{y} + \Delta\mathbf{y}) - g(\mathbf{y})\| \le L\|\Delta\mathbf{y}\|. \tag{A.1}$$

If the constant L applies for all $\mathbf{y}$, we say that g is uniformly locally Lipschitz (L), and the word "locally" may be dropped.

LEMMA 13. *For each k, $0 \le k \le m$, there is an open set G_k of full measure on which $\mathcal{A}_k(\mathbf{y})$ and $\mathcal{A}_{k+1}(\mathbf{y})$ are locally fixed and differ by 1, and $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ is locally linear. The sets G_k are decreasing as k increases.*

PROOF. The argument is by induction. The induction hypothesis states that for each $\mathbf{y}_0 \in G_{k-1}$ there is a small ball $B(\mathbf{y}_0)$ on which (a) the active sets $\mathcal{A}_{k-1}(\mathbf{y})$ and $\mathcal{A}_k(\mathbf{y})$ are fixed and equal to $\mathcal{A}_{k-1}$ and $\mathcal{A}_k$, respectively, (b) $|\mathcal{A}_k \setminus \mathcal{A}_{k-1}| = 1$ so that the same single variable enters locally at stage $k-1$ and (c) $\hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}) = M\mathbf{y}$ is linear. We construct a set G_k with the same property.

Fix a point $\mathbf{y}_0$ and the corresponding ball $B(\mathbf{y}_0) \subset G_{k-1}$, on which $\mathbf{y} - \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}) = \mathbf{y} - M\mathbf{y} = R\mathbf{y}$, say. For indices $j_1, j_2 \notin \mathcal{A}$, let $N(j_1, j_2)$ be the set of $\mathbf{y}$ for which there exists a γ such that

$$w'(R\mathbf{y} - \gamma\mathbf{u}_k) = \mathbf{x}_{j_1}'(R\mathbf{y} - \gamma\mathbf{u}_k) = \mathbf{x}_{j_2}'(R\mathbf{y} - \gamma\mathbf{u}_k). \tag{A.2}$$

Setting $\boldsymbol{\delta}_1 = \mathbf{x}_l - \mathbf{x}_{j_1}$, the first equality may be written $\boldsymbol{\delta}_1' R\mathbf{y} = \gamma\boldsymbol{\delta}_1'\mathbf{u}_k$ and so when $\boldsymbol{\delta}_1'\mathbf{u}_k \ne 0$ determines

$$\gamma = \boldsymbol{\delta}_1' R\mathbf{y}/\boldsymbol{\delta}_1'\mathbf{u}_k =: \boldsymbol{\eta}_1'\mathbf{y}.$$

[If $\boldsymbol{\delta}_1'\mathbf{u}_k = 0$, there are no qualifying $\mathbf{y}$, and $N(j_1, j_2)$ is empty.] Now using the second equality and setting $\boldsymbol{\delta}_2 = \mathbf{x}_l - \mathbf{x}_{j_2}$, we see that $N(j_1, j_2)$ is contained in the set of $\mathbf{y}$ for which

$$\boldsymbol{\delta}_2' R\mathbf{y} = \boldsymbol{\eta}_1'\mathbf{y}\ \boldsymbol{\delta}_2'\mathbf{u}_k.$$

In other words, setting $\eta_2 = R'\delta_2 - (\delta_2'\mathbf{u}_k)\eta_1$, we have

$$N(j_1, j_2) \subset \{\mathbf{y} : \eta_2'\mathbf{y} = 0\}.$$

If we define

$$N(\mathbf{y}_0) = \bigcup \{N(j_1, j_2) : j_1, j_2 \notin \mathcal{A}, j_1 \neq j_2\},$$

it is evident that $N(\mathbf{y}_0)$ is a finite union of hyperplanes and hence closed. For $\mathbf{y} \in B(\mathbf{y}_0) \setminus N(\mathbf{y}_0)$, a unique new variable joins the active set at step k. Near each such $\mathbf{y}$ the "joining" variable is locally the same and $\gamma_k(\mathbf{y})\mathbf{u}_k$ is locally linear.

We then define $G_k \subset G_{k-1}$ as the union of such sets $B(\mathbf{y}) \setminus N(\mathbf{y})$ over $\mathbf{y} \in G_{k-1}$. Thus G_k is open and, on G_k, $\mathcal{A}_{k+1}(\mathbf{y})$ is locally constant and $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ is locally linear. Thus properties (a)–(c) hold for G_k.

The same argument works for the initial case $k = 0$: since $\hat{\boldsymbol{\mu}}_0 = 0$, there is no circularity.

Finally, since the intersection of G_k with any compact set is covered by a finite number of $B(y_i) \setminus N(y_i)$, it is clear that G_k has full measure. □

LEMMA 14. *Suppose that, for* $\mathbf{y}$ *near* $\mathbf{y}_0$, $\hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y})$ *is continuous* (*resp. linear*) *and that* $\mathcal{A}_k(\mathbf{y}) = \mathcal{A}_k$. *Suppose also that, at* $\mathbf{y}_0$, $\mathcal{A}_{k+1}(\mathbf{y}_0) = \mathcal{A} \cup \{k+1\}$.

Then for $\mathbf{y}$ *near* $\mathbf{y}_0$, $\mathcal{A}_{k+1}(\mathbf{y}) = \mathcal{A}_k \cup \{k+1\}$ *and* $\hat{\gamma}_k(\mathbf{y})$ *and hence* $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ *are continuous* (*resp. linear*) *and uniformly Lipschitz.*

PROOF. Consider first the situation at $\mathbf{y}_0$, with $\widehat{C}_k$ and $\widehat{c}_{kj}$ defined in (2.18) and (2.17), respectively. Since $k+1 \notin \mathcal{A}_k$, we have $|\hat{C}_k(\mathbf{y}_0)| > \hat{c}_{k,k+1}(\mathbf{y}_0)$, and $\hat{\gamma}_k(\mathbf{y}_0) > 0$ satisfies

$$\text{(A.3)} \qquad \hat{C}_k(\mathbf{y}_0) - \hat{\gamma}_k(\mathbf{y}_0)A_k \left\{ \begin{matrix} = \\ > \end{matrix} \right\} \hat{c}_{k,j}(\mathbf{y}_0) - \hat{\gamma}_k(\mathbf{y}_0)a_{k,j} \qquad \text{as } \left\{ \begin{matrix} j = k+1 \\ j > k+1 \end{matrix} \right. .$$

In particular, it must be that $A_k \neq a_{k,k+1}$, and hence

$$\hat{\gamma}_k(\mathbf{y}_0) = \frac{\hat{C}_k(\mathbf{y}_0) - \hat{c}_{k,k+1}(\mathbf{y}_0)}{A_k - a_{k,k+1}} > 0.$$

Call an index j admissible if $j \notin \mathcal{A}_k$ and $a_{k,j} \neq A_k$. For $\mathbf{y}$ near $\mathbf{y}_0$, this property is independent of $\mathbf{y}$. For admissible j, define

$$R_{k,j}(\mathbf{y}) = \frac{\hat{C}_k(\mathbf{y}) - \hat{c}_{k,j}(\mathbf{y})}{A_k - a_{k,j}},$$

which is continuous (resp. linear) near $\mathbf{y}_0$ from the assumption on $\hat{\boldsymbol{\mu}}_{k-1}$. By definition,

$$\hat{\gamma}_k(\mathbf{y}) = \min_{j \in \mathcal{P}_k(\mathbf{y})} R_{k,j}(\mathbf{y}),$$

where

$$\mathcal{P}_k(\mathbf{y}) = \{j \text{ admissible and } R_{k,j}(\mathbf{y}) > 0\}.$$

For admissible j, $R_{k,j}(\mathbf{y}_0) \neq 0$, and near $\mathbf{y}_0$ the functions $\mathbf{y} \to R_{k,j}(\mathbf{y})$ are continuous and of fixed sign. Thus, near $\mathbf{y}_0$ the set $\mathcal{P}_k(\mathbf{y})$ stays fixed at $\mathcal{P}_k(\mathbf{y}_0)$ and (A.3) implies that

$$R_{k,k+1}(\mathbf{y}) < R_{k,j}(\mathbf{y}), \qquad j > k+1,\ j \in \mathcal{P}_k(\mathbf{y}).$$

Consequently, for $\mathbf{y}$ near $\mathbf{y}_0$, only variable $k+1$ joins the active set, and so $\mathcal{A}_{k+1}(\mathbf{y}) = \mathcal{A}_k \cup \{k+1\}$, and

$$\hat{\gamma}_k(\mathbf{y}) = R_{k,k+1}(\mathbf{y}) = \frac{(\mathbf{x}_l - \mathbf{x}_{k+1})'(\mathbf{y} - \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}))}{(\mathbf{x}_l - \mathbf{x}_{k+1})'\mathbf{u}_k}. \tag{A.4}$$

This representation shows that both $\hat{\gamma}_k(\mathbf{y})$ and hence $\hat{\boldsymbol{\mu}}_k(\mathbf{y}) = \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}) + \hat{\gamma}_k(\mathbf{y})\mathbf{u}_k$ are continuous (resp. linear) near $\mathbf{y}_0$.

To show that $\hat{\gamma}_k$ is locally Lipschitz at $\mathbf{y}$, we set $\boldsymbol{\delta} = \mathbf{w} - \mathbf{x}_{k+1}$ and write, using notation from (A.1),

$$\Delta\hat{\gamma}_k = \frac{\boldsymbol{\delta}'(\Delta\mathbf{y} - \Delta\hat{\boldsymbol{\mu}}_{k-1})}{\boldsymbol{\delta}'\mathbf{u}_k}.$$

As $\mathbf{y}$ varies, there is a finite list of vectors $(\mathbf{x}_l, \mathbf{x}_{k+1}, \mathbf{u}_k)$ that can occur in the denominator term $\boldsymbol{\delta}'\mathbf{u}_k$, and since all such terms are positive [as observed below (A.3)], they have a uniform positive lower bound, $a_{\min}$ say. Since $\|\boldsymbol{\delta}\| \leq 2$ and $\hat{\boldsymbol{\mu}}_{k-1}$ is Lipschitz (L_{k-1}) by assumption, we conclude that

$$\frac{|\Delta\hat{\gamma}_k|}{\|\Delta\mathbf{y}\|} \leq 2a_{\min}^{-1}(1 + L_{k-1}) =: L_k. \qquad \square$$

A.2. Consequences of the positive cone condition.

LEMMA 15. *Suppose that* $|\mathcal{A}_+| = |\mathcal{A}| + 1$ *and that* $X_{\mathcal{A}+} = [X_{\mathcal{A}}\ \ \mathbf{x}_+]$ (*where* $\mathbf{x}_+ = s_j\mathbf{x}_j$ *for some* $j \notin \mathcal{A}$). *Let* $P_{\mathcal{A}} = X_{\mathcal{A}}G_{\mathcal{A}}^{-1}X'_{\mathcal{A}}$ *denote projection on* $\operatorname{span}(X_{\mathcal{A}})$, *so that* $a = \mathbf{x}'_+P_{\mathcal{A}}\mathbf{x}_+ < 1$. *The* $+$*-component of* $G_{\mathcal{A}+}^{-1}\mathbf{1}_{\mathcal{A}+}$ *is*

$$(G_{\mathcal{A}+}^{-1}\mathbf{1}_{\mathcal{A}+})_+ = (1-a)^{-1}\left(1 - \frac{\mathbf{x}'_+\mathbf{u}_{\mathcal{A}}}{A_{\mathcal{A}}}\right). \tag{A.5}$$

Consequently, under the positive cone condition (4.11),

$$\mathbf{x}'_+\mathbf{u}_{\mathcal{A}} < A_{\mathcal{A}}. \tag{A.6}$$

PROOF. Write $G_{\mathcal{A}+}$ as a partitioned matrix

$$G_{\mathcal{A}+} = \begin{pmatrix} X'X & X'\mathbf{x}_+ \\ \mathbf{x}'_+X & \mathbf{x}'_+\mathbf{x}_+ \end{pmatrix} = \begin{pmatrix} A & B \\ B' & D \end{pmatrix}.$$

Applying the formula for the inverse of a partitioned matrix [e.g., Rao (1973), page 33],

$$(G_{\mathcal{A}+}^{-1}\mathbf{1}_{\mathcal{A}+})_+ = -E^{-1}F'\mathbf{1} + E^{-1},$$

where

$$E = D - B'A^{-1}B = 1 - \mathbf{x}_+'P_{\mathcal{A}}\mathbf{x}_+,$$
$$F = A^{-1}B = G_{\mathcal{A}}^{-1}X'\mathbf{x}_+,$$

from which (A.5) follows. The positive cone condition implies that $G_{\mathcal{A}+}^{-1}\mathbf{1}_{\mathcal{A}+} > 0$, and so (A.6) is immediate. □

A.3. Global continuity and Lemma 3. We shall call $\mathbf{y}_0$ a multiple point at step k if two or more variables enter at the same time. Lemma 14 shows that such points form a set of measure zero, but they can and do cause discontinuities in $\hat{\boldsymbol{\mu}}_{k+1}$ at $\mathbf{y}_0$ in general. We will see, however, that the positive cone condition prevents such discontinuities.

We confine our discussion to double points, hoping that these arguments will be sufficient to establish the same pattern of behavior at points of multiplicity 3 or higher. In addition, by renumbering, we shall suppose that indices $k+1$ and $k+2$ are those that are added at double point $\mathbf{y}_0$. Similarly, for convenience only, we assume that $\mathcal{A}_k(\mathbf{y})$ is constant near $\mathbf{y}_0$. Our task then is to show that, for $\mathbf{y}$ near a double point $\mathbf{y}_0$, both $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ and $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y})$ are continuous and uniformly locally Lipschitz.

LEMMA 16. *Suppose that $\mathcal{A}_k(\mathbf{y}) = \mathcal{A}_k$ is constant near $\mathbf{y}_0$ and that $\mathcal{A}_{k+}(\mathbf{y}_0) = \mathcal{A}_k \cup \{k+1, k+2\}$. Then for $\mathbf{y}$ near $\mathbf{y}_0$, $\mathcal{A}_{k+}(\mathbf{y}) \setminus \mathcal{A}_k$ can only be one of three possibilities, namely $\{k+1\}$, $\{k+2\}$ or $\{k+1, k+2\}$. In all cases $\hat{\boldsymbol{\mu}}_k(\mathbf{y}) = \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}) + \hat{\gamma}_k(\mathbf{y})\mathbf{u}_k$ as usual, and both $\gamma_k(\mathbf{y})$ and $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ are continuous and locally Lipschitz.*

PROOF. We use notation and tools from the proof of Lemma 14. Since $\mathbf{y}_0$ is a double point and the positivity set $\mathcal{P}_k(\mathbf{y}) = \mathcal{P}_k$ near $\mathbf{y}_0$, we have

$$0 < R_{k,k+1}(\mathbf{y}_0) = R_{k,k+2}(\mathbf{y}_0) < R_{k,j}(\mathbf{y}_0) \qquad \text{for } j \in \mathcal{P}_k \setminus \{k+1, k+2\}.$$

Continuity of $R_{k,j}$ implies that near $\mathbf{y}_0$ we still have

$$0 < R_{k,k+1}(\mathbf{y}), R_{k,k+2}(\mathbf{y}) < \min\{R_{k,j}(\mathbf{y});\ j \in \mathcal{P}_k \setminus \{k+1, k+2\}\}.$$

Hence $\mathcal{A}_{k+} \setminus \mathcal{A}_k$ must equal $\{k+1\}$ or $\{k+2\}$ or $\{k+1, k+2\}$ according as $R_{k,k+1}(\mathbf{y})$ is less than, greater than or equal to $R_{k,k+2}(\mathbf{y})$. The continuity of

$$\hat{\gamma}_k(\mathbf{y}) = \min\{R_{k,k+1}(\mathbf{y}), R_{k,k+2}(\mathbf{y})\}$$

is immediate, and the local Lipschitz property follows from the arguments of Lemma 14. □

LEMMA 17. *Assume the conditions of Lemma* 16 *and in addition that the positive cone condition* (4.11) *holds. Then* $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y})$ *is continuous and locally Lipschitz near* $\mathbf{y}_0$.

PROOF. Since $\mathbf{y}_0$ is a double point, property (A.3) holds, but now with equality when $j = k+1$ or $k+2$ and strict inequality otherwise. In other words, there exists $\delta_0 > 0$ for which

$$\hat{C}_{k+1}(\mathbf{y}_0) - \hat{c}_{k+1,j}(\mathbf{y}_0) \begin{cases} = 0, & \text{if } j = k+2, \\ \geq \delta_0, & \text{if } j > k+2. \end{cases}$$

Consider a neighborhood $B(\mathbf{y}_0)$ of $\mathbf{y}_0$ and let $N(\mathbf{y}_0)$ be the set of double points in $B(\mathbf{y}_0)$, that is, those for which $\mathcal{A}_{k+1}(\mathbf{y}) \setminus \mathcal{A}_k = \{k+1, k+2\}$. We establish the convention that at such double points $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y}) = \hat{\boldsymbol{\mu}}_k(\mathbf{y})$; at other points $\mathbf{y}$ in $B(\mathbf{y}_0)$, $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y})$ is defined by $\hat{\boldsymbol{\mu}}_k(\mathbf{y}) + \hat{\gamma}_{k+1}(\mathbf{y})\mathbf{u}_{k+1}$ as usual.

Now consider those $\mathbf{y}$ near $\mathbf{y}_0$ for which $\mathcal{A}_{k+1}(\mathbf{y}) \setminus \mathcal{A}_k = \{k+1\}$, and so, from the previous lemma, $\mathcal{A}_{k+2}(\mathbf{y}) \setminus \mathcal{A}_{k+1} = \{k+2\}$. For such $\mathbf{y}$, continuity and the local Lipschitz property for $\hat{\boldsymbol{\mu}}_k$ imply that

$$\hat{C}_{k+1}(\mathbf{y}) - \hat{c}_{k+1,j}(\mathbf{y}) \begin{cases} = O(\|\mathbf{y} - \mathbf{y}_0\|), & \text{if } \ j = k+2, \\ > \delta_0/2, & \text{if } \ j > k+2. \end{cases}$$

It is at this point that we use the positive cone condition (via Lemma 15) to guarantee that $A_{k+1} > a_{k+1,k+2}$. Also, since $\mathcal{A}_{k+1}(\mathbf{y}) \setminus \mathcal{A}_k = \{k+1\}$, we have

$$\hat{C}_{k+1}(\mathbf{y}) > \hat{c}_{k+1,k+2}(\mathbf{y}).$$

These two facts together show that $k+2 \in \mathcal{P}_{k+1}(\mathbf{y})$ and hence that

$$\hat{\gamma}_{k+1}(\mathbf{y}) = \frac{\hat{C}_{k+1}(\mathbf{y}) - \hat{c}_{k+1,k+2}(\mathbf{y})}{A_{k+1} - a_{k+1,k+2}} = O(\|\mathbf{y} - \mathbf{y}_0\|)$$

is continuous and locally Lipschitz. In particular, as $\mathbf{y}$ approaches $N(\mathbf{y}_0)$, we have $\hat{\gamma}_{k+1}(\mathbf{y}) \to 0$. □

REMARK A.1. We say that a function $g : \mathbb{R}^n \to \mathbb{R}$ is *almost differentiable* if it is absolutely continuous on almost all line segments parallel to the coordinate axes, and its partial derivatives (which consequently exist a.e.) are locally integrable. This definition of almost differentiability appears superficially to be weaker than that given by Stein, but it is in fact precisely the property used in his proof. Furthermore, this definition is equivalent to the standard definition of weak differentiability used in analysis.

PROOF OF LEMMA 3. We have shown explicitly that $\hat{\boldsymbol{\mu}}_k(\mathbf{y})$ is continuous and uniformly locally Lipschitz near single and double points. Similar arguments

extend the property to points of multiplicity 3 and higher, and so all points $\mathbf{y}$ are covered. Finally, absolute continuity of $\mathbf{y} \to \hat{\boldsymbol{\mu}}_k(\mathbf{y})$ on line segments is a simple consequence of the uniform Lipschitz property, and so $\hat{\boldsymbol{\mu}}_k$ is almost differentiable. □

Acknowledgments. The authors thank Jerome Friedman, Bogdan Popescu, Saharon Rosset and Ji Zhu for helpful discussions.

REFERENCES

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560.

FREUND, Y. and SCHAPIRE, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

FRIEDMAN, J. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407.

GOLUB, G. and VAN LOAN, C. (1983). *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, MD.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference and Prediction*. Springer, New York.

LAWSON, C. and HANSON, R. (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104.

OSBORNE, M., PRESNELL, B. and TURLACH, B. (2000a). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403.

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. (2000b). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.* **58** 267–288.

WEISBERG, S. (1980). *Applied Linear Regression*. Wiley, New York.

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: brad@stat.stanford.edu

DISCUSSION

BY HEMANT ISHWARAN

Cleveland Clinic Foundation

Being able to reliably, and automatically, select variables in linear regression models is a notoriously difficult problem. This research attacks this question head on, introducing not only a computationally efficient algorithm and method, LARS (and its derivatives), but at the same time introducing comprehensive theory explaining the intricate details of the procedure as well as theory to guide its practical implementation. This is a fascinating paper and I commend the authors for this important work.

Automatic variable selection, the main theme of this paper, has many goals. So before embarking upon a discussion of the paper it is important to first sit down and clearly identify what the objectives are. The authors make it clear in their introduction that, while often the goal in variable selection is to select a "good" linear model, where goodness is measured in terms of prediction accuracy performance, it is also important at the same time to choose models which lean toward the parsimonious side. So here the goals are pretty clear: we want good prediction error performance but also simpler models. These are certainly reasonable objectives and quite justifiable in many scientific settings. At the same, however, one should recognize the difficulty of the task, as the two goals, low prediction error and smaller models, can be diametrically opposed. By this I mean that certainly from an oracle point of view it is true that minimizing prediction error will identify the true model, and thus, by going after prediction error (in a perfect world), we will also get smaller models by default. However, in practice, what happens is that small gains in prediction error often translate into larger models and less dimension reduction. So as procedures get better at reducing prediction error, they can also get worse at picking out variables accurately.

Unfortunately, I have some misgivings that LARS might be falling into this trap. Mostly my concern is fueled by the fact that Mallows' C_p is the criterion used for determining the optimal LARS model. The use of C_p often leads to overfitting, and this coupled with the fact that LARS is a forward optimization procedure, which is often found to be greedy, raises some potential flags. This, by the way, does not necessarily mean that LARS per se is overfitting, but rather that I think C_p may be an inappropriate model selection criterion for LARS. It is this point that will be the focus of my discussion. I will offer some evidence that C_p can sometimes be used effectively if *model uncertainty* is accounted for, thus pointing to ways for its more appropriate use within LARS. Mostly I will make my arguments by way of high-dimensional simulations. My focus on high dimensions is motivated in part by the increasing interest in such problems, but also because it is in such problems that performance breakdowns become magnified and are more easily identified.

Note that throughout my discussion I will talk only about LARS, but, given the connections outlined in the paper, the results should also naturally apply to the Lasso and Stagewise derivatives.

1. Is C_p the correct stopping rule for LARS? The C_p criterion was introduced by Mallows (1973) to be used with the OLS as an unbiased estimator for the model error. However, it is important to keep in mind that it was not intended to be used *when the model is selected by the data* as this can lead to selection bias and in some cases poor subset selection [Breiman (1992)]. Thus, choosing the model with lowest C_p value is only a heuristic technique with sometimes bad performance. Indeed, ultimately, this leads to an inconsistent procedure for the OLS [Shao (1993)]. Therefore, while I think it is reasonable to assume that the C_p formula (4.10) is correct [i.e., that it is reasonable to expect that $df(\widehat{\boldsymbol{\mu}}_k) \approx k$ under a wide variety of settings], there is really no reason to expect that minimizing the C_p value will lead to an optimal procedure for LARS.

In fact, using C_p in a Forward Stagewise procedure of any kind seems to me to be a risky thing to do given that C_p often overfits and that Stagewise procedures are typically greedy. Figure 5 of the paper is introduced (partly) to dispel these types of concerns about LARS being greedy. The message there is that $\text{pe}(\widehat{\boldsymbol{\mu}})$, a performance measurement related to prediction error, declines slowly from its maximum value for LARS compared to the quick drop seen with standard forward stepwise regression. Thus, LARS acts differently than well-known greedy algorithms and so we should not be worried. However, I see the message quite differently. If the maximum proportion explained for LARS is roughly the same over a large range of steps, and hence models of different dimension, then this implies that there is not much to distinguish between higher- and lower-dimensional models. Combine this with the use of C_p which could provide poor estimates for the prediction error due to selection bias and there is real concern for estimating models that are too large.

To study this issue, let me start by reanalyzing the diabetes data (which was the basis for generating Figure 5). In this analysis I will compare LARS to a Bayesian method developed in Ishwaran and Rao (2000), referred to as SVS (short for Stochastic Variable Selection). The SVS procedure is a hybrid of the spike-and-slab model approach pioneered by Mitchell and Beauchamp (1988) and later developed in George and McCulloch (1993). Details for SVS can be found in Ishwaran and Rao (2000, 2003). My reason for using SVS as a comparison procedure is that, like LARS, its coefficient estimates are derived via shrinkage. However, unlike LARS, these estimates are based on model averaging *in combination* with shrinkage. The use of model averaging is a way of accounting for model uncertainty, and my argument will be that models selected via C_p based on SVS coefficients will be more stable than those found using LARS thanks to the extra benefit of model averaging.

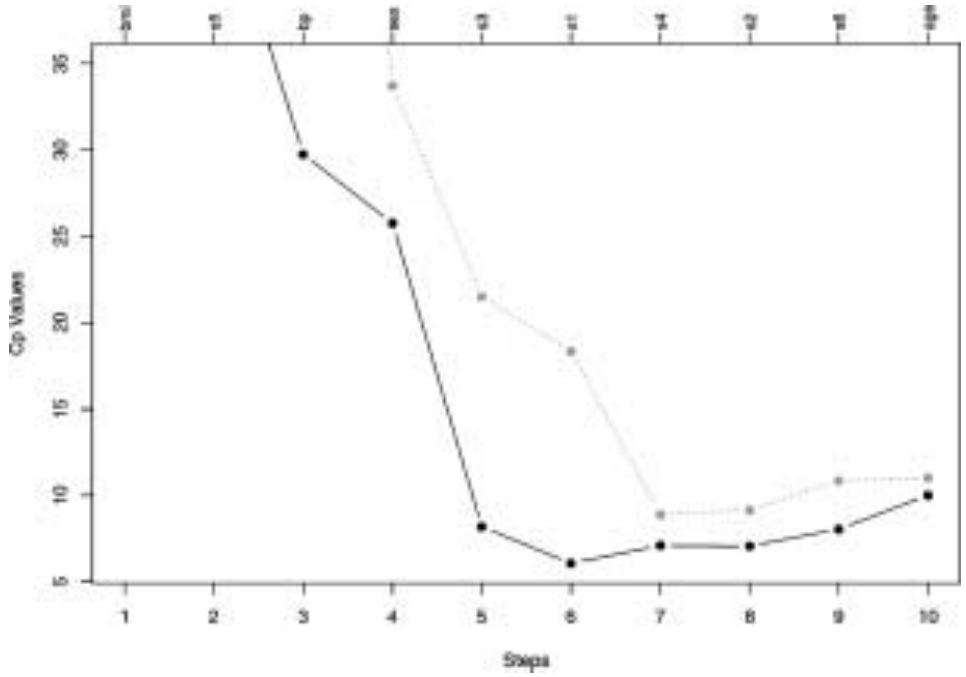

FIG. 1. *C_p values from main effects model for diabetes data*: *thick line is values from SVS*; *thin dashed line is from LARS. Covariates listed at the top of the graph are ordered by importance as measured by their absolute posterior mean.*

Figures 1 and 2 present the C_p values for the main effects model and the quadratic model from both procedures (the analysis for LARS was based on S-PLUS code kindly provided by Trevor Hastie). The C_p values for SVS were computed by (a) finding the posterior mean values for coefficients, (b) ranking covariates by the size of their absolute posterior mean coefficient values (with the top rank going to the largest absolute mean) and (c) computing the C_p value $C_p(\widetilde{\boldsymbol{\mu}}_k) = \|\mathbf{y} - \widetilde{\boldsymbol{\mu}}_k\|/\overline{\sigma}^2 - n + 2k$, where $\widetilde{\boldsymbol{\mu}}_k$ is the OLS estimate based on the k top ranked covariates. All covariates were standardized. This technique of using C_p with SVS was discussed in Ishwaran and Rao (2000).

We immediately see some differences in the figures. In Figure 1, the final model selected by SVS had $k = 6$ variables, while LARS had $k = 7$ variables. More interesting, though, are the discrepancies for the quadratic model seen in Figure 2. Here the optimal SVS model had $k = 8$ variables in contrast to the much higher $k = 15$ variables found by LARS. The top eight variables from SVS (some of these can be read off the top of the plot) are bmi, ltg, map, hdl, sex, age.sex, bmi.map and glu.2. The last three variables are interaction effects and a squared main effects term. The top eight variables from LARS are bmi, ltg, map, hdl, bmi.map, age.sex, glu.2 and bmi.2. Although there is a reasonable overlap in variables, there is still enough of a discrepancy to be concerned. The different model sizes are also cause for concern. Another worrisome aspect for LARS seen in Figure 2 is that its C_p values remain bounded away from zero. This should be compared to the C_p values for SVS, which attain a near-zero mininum value, as we would hope for.

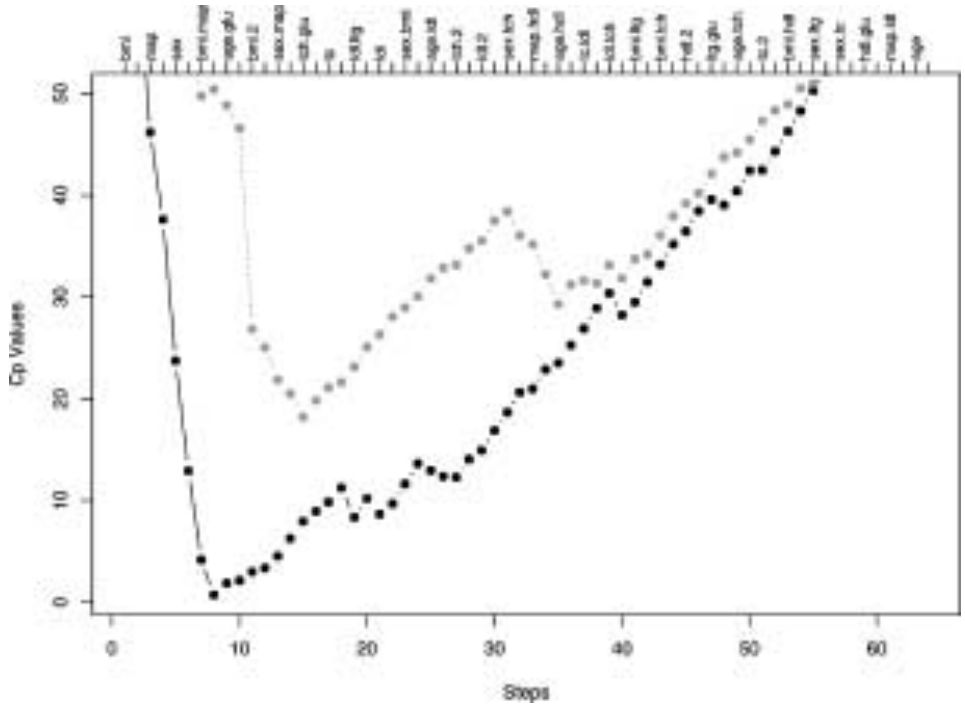

FIG. 2. *C_p values from quadratic model*: *best model from SVS is $k = 8$ (thick line) compared with $k = 15$ from LARS (thin dashed line). Note how the minimum value for SVS is nearly zero.*

2. High-dimensional simulations. Of course, since we do not know the true answer in the diabetes example, we cannot definitively assess if the LARS models are too large. Instead, it will be helpful to look at some simulations for a more systematic study. The simulations I used were designed following the recipe given in Breiman (1992). Data was simulated in all cases by using i.i.d. N(0, 1) variables for ε_i. Covariates x_i, for $i = 1, \ldots, n$, were generated independently from a multivariate normal distribution with zero mean and with covariance satisfying $E(x_{i,j}x_{i,k}) = \rho^{|j-k|}$. I considered two settings for ρ: (i) $\rho = 0$ (uncorrelated); (ii) $\rho = 0.90$ (correlated). In all simulations, $n = 800$ and $m = 400$. Nonzero β_j coefficients were in 15 clusters of 7 adjacent variables centered at every 25th variable. For example, for the variables clustered around the 25th variable, the coefficient values were given by $\beta_{25+j} = |h - j|^{1.25}$ for $|j| < h$, where $h = 4$. The other 14 clusters were defined similarly. All other coefficients were set to zero. This gave a total of 105 nonzero values and 295 zero values. Coefficient values were adjusted by multiplying by a common constant to make the theoretical R^2 value equal to 0.75 [see Breiman (1992) for a discussion of this point]. Please note that, while the various parameters chosen for the simulations might appear specific, I also experimented with other simulations (not reported) by considering different configurations for the dimension m, sample size n, correlation ρ and the number of nonzero coefficients. What I found was consistent with the results presented here.

For each ρ correlation setting, simulations were repeated 100 times independently. Results are recorded in Table 1. There I have recorded what I call TotalMiss,

TABLE 1
Breiman simulation: $m = 400, n = 800$ *and* 105 *nonzero* β_j

	$\rho = 0$ (uncorrelated X)					$\rho = 0.9$ (correlated X)				
	$\widehat{m}$	pe($\widehat{\mu}$)	TotalMiss	FDR	FNR	$\widehat{m}$	pe($\widehat{\mu}$)	TotalMiss	FDR	FNR
LARS	210.69	0.907	126.63	0.547	0.055	99.51	0.962	75.77	0.347	0.135
svsCp	126.66	0.887	61.14	0.323	0.072	58.86	0.952	66.38	0.153	0.164
svsBMA	400.00	0.918	295.00	0.737	0.000	400.00	0.966	295.00	0.737	0.000
Step	135.53	0.876	70.35	0.367	0.075	129.24	0.884	137.10	0.552	0.208

FDR and FNR. TotalMiss is the total number of misclassified variables, that is, the total number of falsely identified nonzero β_j coefficients and falsely identified zero coefficients; FDR and FNR are the false discovery and false nondiscovery rates defined as the false positive and false negative rates for those coefficients identified as nonzero and zero, respectively. The TotalMiss, FDR and FNR values reported are the averaged values from the 100 simulations. Also recorded in the table is $\widehat{m}$, the average number of variables selected by a procedure, as well as the performance value pe($\widehat{\mu}$) [cf. (3.17)], again averaged over the 100 simulations.

Table 1 records the results from various procedures. The entry "svsCp" refers to the C_p-based SVS method used earlier; "Step" is standard forward stepwise regression using the C_p criterion; "svsBMA" is the Bayesian model averaged estimator from SVS. My only reason for including svsBMA is to gauge the prediction error performance of the other procedures. Its variable selection performance is not of interest. Pure Bayesian model averaging leads to improved prediction, but because it does no dimension reduction at all it cannot be considered as a serious candidate for selecting variables.

The overall conclusions from Table 1 are summarized as follows:

1. The total number of misclassified coefficients and FDR values is high in the uncorrelated case for LARS and high in the correlated case for stepwise regression. Their estimated models are just too large. In comparison, svsCp does well in both cases. Overall it does the best in terms of selecting variables by maintaining low FDR and TotalMiss values. It also maintains good performance values.
2. LARS's performance values are good, second only to svsBMA. However, low prediction error does not necessarily imply good variable selection.

3. LARS C_p values in orthogonal models. Figure 3 shows the C_p values for LARS from the two sets of simulations. It is immediately apparent that the C_p curve in the uncorrelated case is too flat, leading to models which are too large. These simulations were designed to reflect an orthogonal design setting (at least asymptotically), so what is it about the orthogonal case that is adversely affecting LARS?

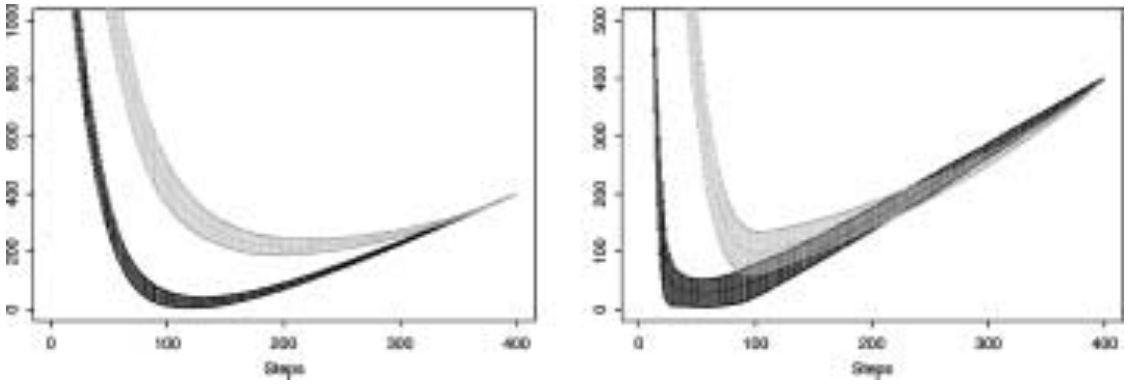

FIG. 3. *C_p values from simulations where $\rho = 0$ (left) and $\rho = 0.9$ (right): bottom curves are from SVS; top curves are from LARS. The lines seen on each curve are the mean C_p values based on 100 simulations. Note how the minimum value for SVS is near zero in both cases. Also superimposed on each curve are error bars representing mean values plus or minus one standard deviation.*

We can use Lemma 1 of the paper to gain some insight into this. For this argument I will assume that m is fixed (the lemma is stated for $m = n$ but applies in general) and I will need to assume that $X_{n\times m}$ is a random orthogonal matrix, chosen so that its rows are exchangeable. To produce such an X, choose m values $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}$ without replacement from $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_j$ is defined as in Section 4.1, and set $X = [\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}]$. It is easy to see that this ensures row-exchangeability. Hence, $\mu_1, \ldots, \mu_n$ are exchangeable and, therefore, $Y_i = \mu_i + \varepsilon_i$ are exchangeable since ε_i are i.i.d. I will assume, as in (4.1), that ε_i are independent $\mathrm{N}(0, \sigma^2)$ variables.

For simplicity take $\sigma^2 = \overline{\sigma}^2 = 1$. Let V_j, for $j = 0, \ldots, m-1$, denote the $(j+1)$st largest value from the set of values $\{|Y_{i_1}|, \ldots, |Y_{i_m}|\}$. Let k_0 denote the true dimension, that is, the number of nonzero coordinates of the true $\boldsymbol{\beta}$, and suppose that k is some dimension larger than k_0 such that $1 \le k_0 < k \le m \le n$. Notice that $V_k \le V_{k_0}$, and thus, by Lemma 1 and (4.10),

$$\begin{aligned} C_p(\widehat{\boldsymbol{\mu}}_k) - C_p(\widehat{\boldsymbol{\mu}}_{k_0}) &= (V_k^2 - V_{k_0}^2) \sum_{j=1}^{m} \mathbb{1}\{|Y_{i_j}| > V_{k_0}\} + V_k^2 \sum_{j=1}^{m} \mathbb{1}\{V_k < |Y_{i_j}| \le V_{k_0}\} \\ &\quad - \sum_{j=1}^{m} Y_{i_j}^2 \mathbb{1}\{V_k < |Y_{i_j}| \le V_{k_0}\} + 2(k - k_0) \\ &\le -\Delta_k B_k + 2(k - k_0), \end{aligned}$$

where $\Delta_k = V_{k_0}^2 - V_k^2 \ge 0$ and $B_k = \sum_{j=1}^{m} \mathbb{1}\{|Y_{i_j}| > V_{k_0}\}$. Observe that by exchangeability B_k is a Binomial$(m, k_0/m)$ random variable. It is a little messy to work out the distribution for Δ_k explicitly. However, it is not hard to see that Δ_k can be reasonably large with high probability. Now if $k_0 > k - k_0$ and k_0 is large, then B_k, which has a mean of k_0, will become the dominant term in $\Delta_k B_k$

and $\Delta_k B_k$ will become larger than $2(k - k_0)$ with high probability. This suggests, at least in this setting, that C_p will overfit if the dimension of the problem is high. In this case there will be too much improvement in the residual sums of squares when moving from k_0 to k because of the nonvanishing difference between the squared order statistics $V_{k_0}^2$ and V_k^2.

4. Summary. The use of C_p seems to encourage large models in LARS, especially in high-dimensional orthogonal problems, and can adversely affect variable selection performance. It can also be unreliable when used with stepwise regression. The use of C_p with SVS, however, seems better motivated due to the benefits of model averaging, which mitigates the selection bias effect. This suggests that C_p can be used effectively if model uncertainty is accounted for. This might be one remedy. Another remedy would be simply to use a different model selection criteria when using LARS.

REFERENCES

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87** 738–754.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

ISHWARAN, H. and RAO, J. S. (2000). Bayesian nonparametric MCMC for large variable selection problems. Unpublished manuscript.

ISHWARAN, H. and RAO, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.* **98** 438–455.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83** 1023–1036.

SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.

DEPARTMENT OF BIOSTATISTICS/WB4
CLEVELAND CLINIC FOUNDATION
9500 EUCLID AVENUE
CLEVELAND, OHIO 44195
USA
E-MAIL: ishwaran@bio.ri.ccf.org

DISCUSSION

BY KEITH KNIGHT

University of Toronto

First, I congratulate the authors for a truly stimulating paper. The paper resolves a number of important questions but, at the same time, raises many others. I would

like to focus my comments to two specific points.

1. The similarity of Stagewise and LARS fitting to the Lasso suggests that the estimates produced by Stagewise and LARS fitting may minimize an objective function that is similar to the appropriate Lasso objective function. It is not at all (at least to me) obvious how this might work though. I note, though, that the construction of such an objective function may be easier than it seems. For example, in the case of bagging [Breiman (1996)] or subagging [Bühlmann and Yu (2002)], an "implied" objective function can be constructed. Suppose that $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ are estimates (e.g., computed from subsamples or bootstrap samples) that minimize, respectively, objective functions $Z_1, \ldots, Z_m$ and define

$$\widehat{\theta} = g(\widehat{\theta}_1, \ldots, \widehat{\theta}_m);$$

then $\widehat{\theta}$ minimizes the objective function

$$Z(t) = \inf\{Z_1(t_1) + \cdots + Z_m(t_m) : g(t_1, \ldots, t_m) = t\}.$$

(Thanks to Gib Bassett for pointing this out to me.) A similar construction for stagewise fitting (or LARS in general) could facilitate the analysis of the statistical properties of the estimators obtained via these algorithms.

2. When I first started experimenting with the Lasso, I was impressed by its robustness to small changes in its tuning parameter relative to more classical stepwise subset selection methods such as Forward Selection and Backward Elimination. (This is well illustrated by Figure 5; at its best, Forward Selection is comparable to LARS, Stagewise and the Lasso but the performance of Forward Selection is highly dependent on the model size.) Upon reflection, I realized that there was a simple explanation for this robustness. Specifically, the strict convexity in $\boldsymbol{\beta}$ for each t in the Lasso objective function (1.5) together with the continuity (in the appropriate sense) in t of these objective functions implies that the Lasso solutions $\widehat{\boldsymbol{\beta}}(t)$ are continuous in t; this continuity breaks down for nonconvex objective functions. Of course, the same can be said of other penalized least squares estimates whose penalty is convex. What seems to make the Lasso special is (i) its ability to produce exact 0 estimates and (ii) the "fact" that its bias seems to be more controllable than it is for other methods (e.g., ridge regression, which naturally overshrinks large effects) in the sense that for a fixed tuning parameter the bias is bounded by a constant that depends on the design but not the true parameter values. At the same time, though, it is perhaps unfair to compare stepwise methods to the Lasso, LARS or Stagewise fitting since the space of models considered by the latter methods seems to be "nicer" than it is for the former and (perhaps more important) since the underlying motivation for using Forward Selection is typically not prediction. For example, bagged Forward Selection might perform as well as the other methods in many situations.

REFERENCES

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.
BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE ST.
TORONTO, ONTARIO M5S 3G3
CANADA
E-MAIL: keith@utstat.toronto.edu

DISCUSSION

BY JEAN-MICHEL LOUBES AND PASCAL MASSART

Université Paris-Sud

The issue of model selection has drawn the attention of both applied and theoretical statisticians for a long time. Indeed, there has been an enormous range of contribution in model selection proposals, including work by Akaike (1973), Mallows (1973), Foster and George (1994), Birgé and Massart (2001a) and Abramovich, Benjamini, Donoho and Johnstone (2000). Over the last decade, modern computer-driven methods have been developed such as All Subsets, Forward Selection, Forward Stagewise or Lasso. Such methods are useful in the setting of the standard linear model, where we observe noisy data and wish to predict the response variable using only a few covariates, since they provide automatically linear models that fit the data. The procedure described in this paper is, on the one hand, numerically very efficient and, on the other hand, very general, since, with slight modifications, it enables us to recover the estimates given by the Lasso and Stagewise.

1. Estimation procedure. The "LARS" method is based on a recursive procedure selecting, at each step, the covariates having largest absolute correlation with the response y. In the case of an orthogonal design, the estimates can then be viewed as an l^1-penalized estimator. Consider the linear regression model where we observe y with some random noise ε, with orthogonal design assumptions:

$$y = X\beta + \varepsilon.$$

Using the soft-thresholding form of the estimator, we can write it, equivalently, as the minimum of an ordinary least squares and an l^1 penalty over the coefficients of the regression. As a matter of fact, at step $k = 1, \ldots, m$, the estimators $\hat{\beta}^k = X^{-1}\hat{\mu}^k$ are given by

$$\hat{\mu}^k = \arg\min_{\mu \in \mathbb{R}^n} \left(\|Y - \mu\|_n^2 + 2\lambda_n^2(k)\|\mu\|_1 \right).$$

There is a trade-off between the two terms, balanced by the smoothing decreasing sequence $\lambda_n^2(k)$. The more stress is laid on the penalty, the more parsimonious the representation will be. The choice of the l^1 penalty enables us to keep the largest coefficients, while the smallest ones shrink toward zero in a soft-thresholding scheme. This point of view is investigated in the Lasso algorithm as well as in studying the false discovery rate (FDR).

So, choosing these weights in an optimal way determines the form of the estimator as well as its asymptotic behavior. In the case of the algorithmic procedure, the suggested level is the $(k+1)$-order statistic:

$$\lambda_n^2(k) = |y|_{(k+1)}.$$

As a result, it seems possible to study the asymptotic behavior of the LARS estimates under some conditions on the coefficients of β. For instance, if there exists a roughness parameter $\rho \in [0, 2]$, such that $\sum_{j=1}^m |\beta_j|^\rho \leq M$, metric entropy theory results lead to an upper bound for the mean square error $\|\hat{\beta} - \beta\|^2$. Here we refer to the results obtained in Loubes and van de Geer (2002). Consistency should be followed by the asymptotic distribution, as is done for the Lasso in Knight and Fu (2000).

The interest for such an investigation is double: first, it gives some insight into the properties of such estimators. Second, it suggests an approach for choosing the threshold λ_n^2 which can justify the empirical cross-validation method, developed later in the paper. Moreover, the asymptotic distributions of the estimators are needed for inference.

Other choices of penalty and loss functions can also be considered. First, for $\gamma \in (0, 1]$, consider

$$J_\gamma(\beta) = \sum_{j=1}^m (X\beta)_j^\gamma.$$

If $\gamma < 1$, the penalty is not convex anymore, but there exist algorithms to solve the minimization problem. Constraints on the l^γ norm of the coefficients are equivalent to lacunarity assumptions and may make estimation of sparse signals easier, which is often the case for high-dimensional data for instance.

Moreover, replacing the quadratic loss function with an l^1 loss gives rise to a robust estimator, the penalized absolute deviation of the form

$$\tilde{\mu}^k = \arg\min_{\mu \in \mathbb{R}^n} \big(\|Y - \mu\|_{n,1} + 2\lambda_n^2(k) \|\mu\|_1 \big).$$

Hence, it is possible to get rid of the problem of variance estimation for the model with these estimates whose asymptotic behavior can be derived from Loubes and van de Geer (2002), in the regression framework.

Finally, a penalty over both the number of coefficients and the smoothness of the coefficients can be used to study, from a theoretical point of view, the asymptotics

of the estimate. Such a penalty is analogous to complexity penalties studied in van de Geer (2001):

$$\mu^\star = \arg \min_{\mu \in \mathbb{R}^n, k \in [1,m]} \big(\|Y - \mu\|_n^2 + 2\lambda_n^2(k) \|\mu\|_1 + \mathrm{pen}(k) \big).$$

2. Mallows' C_p. We now discuss the crucial issue of selecting the number k of influential variables. To make this discussion clear, let us first assume the variance σ^2 of the regression errors to be known. Interestingly the penalized criterion which is proposed by the authors is exactly equivalent to Mallows' C_p when the design is orthogonal (this is indeed the meaning of their Theorem 3). More precisely, using the same notation as in the paper, let us focus on the following situation which illustrates what happens in the orthogonal case where LARS is equivalent to the Lasso. One observes some random vector y in $\mathbb{R}^n$, with expectation μ and covariance matrix $\sigma^2 I_n$. The variable selection problem that we want to solve here is to determine which components of y are influential. According to Lemma 1, given k, the kth LARS estimate $\widehat{\mu}_k$ of μ can be explicitly computed as a soft-thresholding estimator. Indeed, considering the order statistics of the absolute values of the data denoted by

$$|y|_{(1)} \geq |y|_{(2)} \geq \cdots \geq |y|_{(n)}$$

and defining the soft threshold function $\eta(\cdot, t)$ with level $t \geq 0$ as

$$\eta(x,t) = x \mathbb{1}_{|x|>t} \left(1 - \frac{t}{|x|} \right),$$

one has

$$\widehat{\mu}_{k,i} = \eta\big(y_i, |y|_{(k+1)}\big).$$

To select k, the authors propose to minimize the C_p criterion

$$C_p(\widehat{\mu}_k) = \|y - \widehat{\mu}_k\|^2 - n\sigma^2 + 2k\sigma^2. \tag{1}$$

Our purpose is to analyze this proposal with the help of the results on penalized model selection criteria proved in Birgé and Massart (2001a, b). In these papers some oracle type inequalities are proved for selection procedures among some arbitrary collection of projection estimators on linear models when the regression errors are Gaussian. In particular one can apply them to the variable subset selection problem above, assuming the random vector y to be Gaussian. If one decides to penalize in the same way the subsets of variables with the same cardinality, then the penalized criteria studied in Birgé and Massart (2001a, b) take the form

$$C'_p(\widetilde{\mu}_k) = \|y - \widetilde{\mu}_k\|^2 - n\sigma^2 + \mathrm{pen}(k), \tag{2}$$

where pen(k) is some penalty to be chosen and $\widetilde{\mu}_k$ denotes the hard threshold estimator with components

$$\widetilde{\mu}_{k,i} = \eta'\big(y_i, |y|_{(k+1)}\big),$$

where

$$\eta'(x,t) = x\mathbb{1}_{|x|>t}.$$

The essence of the results proved by Birgé and Massart (2001a, b) in this case is the following. Their analysis covers penalties of the form

$$\text{pen}(k) = 2k\sigma^2 C\left(\log\left(\frac{n}{k}\right) + C'\right)$$

[note that the FDR penalty proposed in Abramovich, Benjamini, Donoho and Johnstone (2000) corresponds to the case $C = 1$]. It is proved in Birgé and Massart (2001a) that if the penalty pen(k) is heavy enough (i.e., $C > 1$ and C' is an adequate absolute constant), then the model selection procedure works in the sense that, up to a constant, the selected estimator $\widetilde{\mu}_{\widetilde{k}}$ performs as well as the best estimator among the collection $\{\widetilde{\mu}_k, 1 \le k \le n\}$ in terms of quadratic risk. On the contrary, it is proved in Birgé and Massart (2001b) that if $C < 1$, then at least asymptotically, whatever C', the model selection does not work, in the sense that, even if $\mu = 0$, the procedure will systematically choose large values of k, leading to a suboptimal order for the quadratic risk of the selected estimator $\widetilde{\mu}_{\widetilde{k}}$. So, to summarize, some $2k\sigma^2\log(n/k)$ term should be present in the penalty, in order to make the model selection criterion (2) work. In particular, the choice $\text{pen}(k) = 2k\sigma^2$ is not appropriate, which means that Mallows' C_p does not work in this context. At first glance, these results seem to indicate that some problems could occur with the use of the Mallows' C_p criterion (1). Fortunately, however, this is not at all the case because a very interesting phenomenon occurs, due to the soft-thresholding effect. As a matter of fact, if we compare the residual sums of squares of the soft threshold estimator $\widehat{\mu}_k$ and the hard threshold estimator $\widetilde{\mu}_k$, we easily get

$$\|y - \widehat{\mu}_k\|^2 - \|y - \widetilde{\mu}_k\|^2 = \sum_{i=1}^{n} |y|^2_{(k+1)}\mathbb{1}_{|y_i|>|y|_{(k+1)}} = k|y|^2_{(k+1)}$$

so that the "soft" C_p criterion (1) can be interpreted as a "hard" criterion (2) with random penalty

$$\text{pen}(k) = k|y|^2_{(k+1)} + 2k\sigma^2. \tag{3}$$

Of course this kind of penalty escapes *stricto sensu* to the analysis of Birgé and Massart (2001a, b) as described above since the penalty is not deterministic. However, it is quite easy to realize that, in this penalty, $|y|^2_{(k+1)}$ plays the role of the apparently "missing" logarithmic factor $2\sigma^2\log(n/k)$. Indeed, let us consider the

pure noise situation where $\mu = 0$ to keep the calculations as simple as possible. Then, if we consider the order statistics of a sample $U_1, \ldots, U_n$ of the uniform distribution on [0, 1]

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)},$$

taking care of the fact that these statistics are taken according to the usual increasing order while the order statistics on the data are taken according to the reverse order, $\sigma^{-2}|y|^2_{(k+1)}$ has the same distribution as

$$Q^{-1}(U_{(k+1)}),$$

where Q denotes the tail function of the chi-square distribution with 1 degree of freedom. Now using the double approximations $Q^{-1}(u) \sim 2\log(|u|)$ as u goes to 0 and $U_{(k+1)} \approx (k+1)/n$ (which at least means that, given k, $nU_{(k+1)}$ tends to $k+1$ almost surely as n goes to infinity but can also be expressed with much more precise probability bounds) we derive that $|y|^2_{(k+1)} \approx 2\sigma^2 \log(n/(k+1))$. The conclusion is that it is possible to interpret the "soft" C_p criterion (1) as a randomly penalized "hard" criterion (2). The random part of the penalty $k|y|^2_{(k+1)}$ cleverly plays the role of the unavoidable logarithmic term $2\sigma^2 k \log(n/k)$, allowing the hope that the usual $2k\sigma^2$ term will be heavy enough to make the selection procedure work as we believe it does. A very interesting feature of the penalty (3) is that its random part depends neither on the scale parameter σ^2 nor on the tail of the errors. This means that one could think to adapt the data-driven strategy proposed in Birgé and Massart (2001b) to choose the penalty without knowing the scale parameter to this context, even if the errors are not Gaussian. This would lead to the following heuristics. For large values of k, one can expect the quantity $-\|y - \widehat{\mu}_k\|^2$ to behave as an affine function of k with slope $\alpha(n)\sigma^2$. If one is able to compute $\alpha(n)$, either theoretically or numerically (our guess is that it varies slowly with n and that it is close to 1.5), then one can just estimate the slope (for instance by making a regression of $-\|y - \widehat{\mu}_k\|^2$ with respect to k for large enough values of k) and plug the resulting estimate of σ^2 into (1). Of course, some more efforts would be required to complete this analysis and provide rigorous oracle inequalities in the spirit of those given in Birgé and Massart (2001a, b) or Abramovich, Benjamini, Donoho and Johnstone (2000) and also some simulations to check whether our proposal to estimate σ^2 is valid or not.

Our purpose here was just to mention some possible explorations starting from the present paper that we have found very stimulating. It seems to us that it solves practical questions of crucial interest and raises very interesting theoretical questions: consistency of LARS estimator; efficiency of Mallows' C_p in this context; use of random penalties in model selection for more general frameworks.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000–19, Dept. Statistics, Stanford Univ.

AKAIKE, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60** 255–265.

BIRGÉ, L. and MASSART, P. (2001a). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268.

BIRGÉ, L. and MASSART, P. (2001b). A generalized C_p criterion for Gaussian model selection. Technical Report 647, Univ. Paris 6 & 7.

FOSTER, D. and GEORGE, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.

KNIGHT, K. and FU, B. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378.

LOUBES, J.-M. and VAN DE GEER, S. (2002). Adaptive estimation with soft thresholding penalties. *Statist. Neerlandica* **56** 453–478.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Math. Methods Statist.* **10** 355–374.

CNRS AND LABORATOIRE DE MATHÉMATIQUES
UMR 8628
EQUIPE DE PROBABILITÉS, STATISTIQUE
ET MODÉLISATION
UNIVERSITÉ PARIS-SUD, BÂT. 425
91405 ORDAY CEDEX
FRANCE
E-MAIL: Jean-Michel.Loubes@math.u-psud.fr

LABORATOIRE DE MATHÉMATIQUES
UMR 8628
EQUIPE DE PROBABILITÉS, STATISTIQUE
ET MODÉLISATION
UNIVERSITÉ PARIS-SUD, BÂT. 425
91405 ORDAY CEDEX
FRANCE
E-MAIL: Pascal.Massart@math.u-psud.fr

DISCUSSION

BY DAVID MADIGAN AND GREG RIDGEWAY

Rutgers University and Avaya Labs Research, and RAND

Algorithms for simultaneous shrinkage and selection in regression and classification provide attractive solutions to knotty old statistical challenges. Nevertheless, as far as we can tell, Tibshirani's Lasso algorithm has had little impact on statistical practice. Two particular reasons for this may be the relative inefficiency of the original Lasso algorithm and the relative complexity of more recent Lasso algorithms [e.g., Osborne, Presnell and Turlach (2000)]. Efron, Hastie, Johnstone and Tibshirani have provided an efficient, simple algorithm for the Lasso as well as algorithms for stagewise regression and the new least angle regression. As such this paper is an important contribution to statistical computing.

1. Predictive performance. The authors say little about predictive performance issues. In our work, however, the relative out-of-sample predictive performance of LARS, Lasso and Forward Stagewise (and variants thereof) takes

TABLE 1
Stagewise, LARS and Lasso mean square error predictive performance, comparing cross-validation with C_p

	Diabetes			Boston			Servo	
	CV	**C_p**		**CV**	**C_p**		**CV**	**C_p**
Stagewise	3083	3082	Stagewise	25.7	25.8	Stagewise	1.33	1.32
LARS	3080	3083	LARS	25.5	25.4	LARS	1.33	1.30
Lasso	3083	3082	Lasso	25.8	25.7	Lasso	1.34	1.31

center stage. Interesting connections exist between boosting and stagewise algorithms so predictive comparisons with boosting are also of interest.

The authors present a simple C_p statistic for LARS. In practice, a cross-validation (CV) type approach for selecting the degree of shrinkage, while computationally more expensive, may lead to better predictions. We considered this using the LARS software. Here we report results for the authors' diabetes data, the Boston housing data and the Servo data from the UCI Machine Learning Repository. Specifically, we held out 10% of the data and chose the shrinkage level using either C_p or nine-fold CV using 90% of the data. Then we estimated mean square error (MSE) on the 10% hold-out sample. Table 1 shows the results for main-effects models.

Table 1 exhibits two particular characteristics. First, as expected, Stagewise, LARS and Lasso perform similarly. Second, C_p performs as well as cross-validation; if this holds up more generally, larger-scale applications will want to use C_p to select the degree of shrinkage.

Table 2 presents a reanalysis of the same three datasets but now considering

TABLE 2
Predictive performance of competing methods: *LM is a main-effects linear model with least squares fitting*; *LARS is least angle regression with main effects and CV shrinkage selection*; *LARS two-way C_p is least angle regression with main effects and all two-way interactions, shrinkage selection via C_p*; *GBM additive and GBM two-way use least squares boosting, the former using main effects only, the latter using main effects and all two-way interactions*; *MSE is mean square error on a* 10% *holdout sample*; *MAD is mean absolute deviation*

	Diabetes		Boston		Servo	
	MSE	**MAD**	**MSE**	**MAD**	**MSE**	**MAD**
LM	3000	44.2	23.8	3.40	1.28	0.91
LARS	3087	45.4	24.7	3.53	1.33	0.95
LARS two-way C_p	3090	45.1	14.2	2.58	0.93	0.60
GBM additive	3198	46.7	16.5	2.75	0.90	0.65
GBM two-way	3185	46.8	14.1	2.52	0.80	0.60

five different models: least squares; LARS using cross-validation to select the coefficients; LARS using C_p to select the coefficients and allowing for two-way interactions; least squares boosting fitting only main effects; least squares boosting allowing for two-way interactions. Again we used the authors' LARS software and, for the boosting results, the gbm package in R [Ridgeway (2003)]. We evaluated all the models using the same cross-validation group assignments.

A plain linear model provides the best out-of-sample predictive performance for the diabetes dataset. By contrast, the Boston housing and Servo data exhibit more complex structure and models incorporating higher-order structure do a better job.

While no general conclusions can emerge from such a limited analysis, LARS seems to be competitive with these particular alternatives. We note, however, that for the Boston housing and Servo datasets Breiman (2001) reports substantially better predictive performance using random forests.

2. Extensions to generalized linear models. The minimal computational complexity of LARS derives largely from the squared error loss function. Applying LARS-type strategies to models with nonlinear loss functions will require some form of approximation. Here we consider LARS-type algorithms for logistic regression.

Consider the logistic log-likelihood for a regression function $f(\mathbf{x})$ which will be linear in $\mathbf{x}$:

$$\ell(f) = \sum_{i=1}^{N} y_i f(\mathbf{x}_i) - \log\bigl(1 + \exp(f(\mathbf{x}_i))\bigr). \tag{1}$$

We can initialize $f(\mathbf{x}) = \log(\bar{y}/(1-\bar{y}))$. For some α we wish to find the covariate x_j that offers the greatest improvement in the logistic log-likelihood, $\ell(f(\mathbf{x}) + \mathbf{x}_j^t \alpha)$. To find this $\mathbf{x}_j$ we can compute the directional derivative for each j and choose the maximum,

$$j^* = \arg\max_j \left| \frac{d}{d\alpha} \ell\bigl(f(\mathbf{x}) + \mathbf{x}_j^t \alpha\bigr) \right|_{\alpha=0} \tag{2}$$

$$= \arg\max_j \left| \mathbf{x}_j^t \left(\mathbf{y} - \frac{1}{1 + \exp(-f(\mathbf{x}))} \right) \right|. \tag{3}$$

Note that as with LARS this is the covariate that is most highly correlated with the residuals. The selected covariate is the first member of the active set, A. For α small enough (3) implies

$$(s_{j^*}\mathbf{x}_{j^*} - s_j \mathbf{x}_j)^t \left(y - \frac{1}{1 + \exp(-f(\mathbf{x}) - \mathbf{x}_{j^*}^t \alpha)} \right) \geq 0 \tag{4}$$

for all $j \in A^C$, where s_j indicates the sign of the correlation as in the LARS development. Choosing α to have the largest magnitude while maintaining the constraint in (4) involves a nonlinear optimization. However, linearizing (4) yields

a fairly simple approximate solution. If x_2 is the variable with the second largest correlation with the residual, then

$$\hat{\alpha} = \frac{(s_{j^*}\mathbf{x}_{j^*} - s_2\mathbf{x}_2)^t(y - p(\mathbf{x}))}{(s_{j^*}\mathbf{x}_{j^*} - s_2\mathbf{x}_2)^t(p(\mathbf{x})(1 - p(\mathbf{x}))\mathbf{x}_{j^*})}. \tag{5}$$

The algorithm may need to iterate (5) to obtain the exact optimal $\hat{\alpha}$. Similar logic yields an algorithm for the full solution.

We simulated $N = 1000$ observations with 10 independent normal covariates $\mathbf{x}_i \sim N_{10}(\mathbf{0}, \mathbf{I})$ with outcomes $Y_i \sim \text{Bern}(1/(1 + \exp(-\mathbf{x}_i^t\beta)))$, where $\beta \sim N_{10}(0, 1)$. Figure 1 shows a comparison of the coefficient estimates using Forward Stagewise and the Least Angle method of estimating coefficients, the final estimates arriving at the MLE. While the paper presents LARS for squared error problems, the Least Angle approach seems applicable to a wider family of models. However, an out-of-sample evaluation of predictive performance is essential to assess the utility of such a modeling strategy.

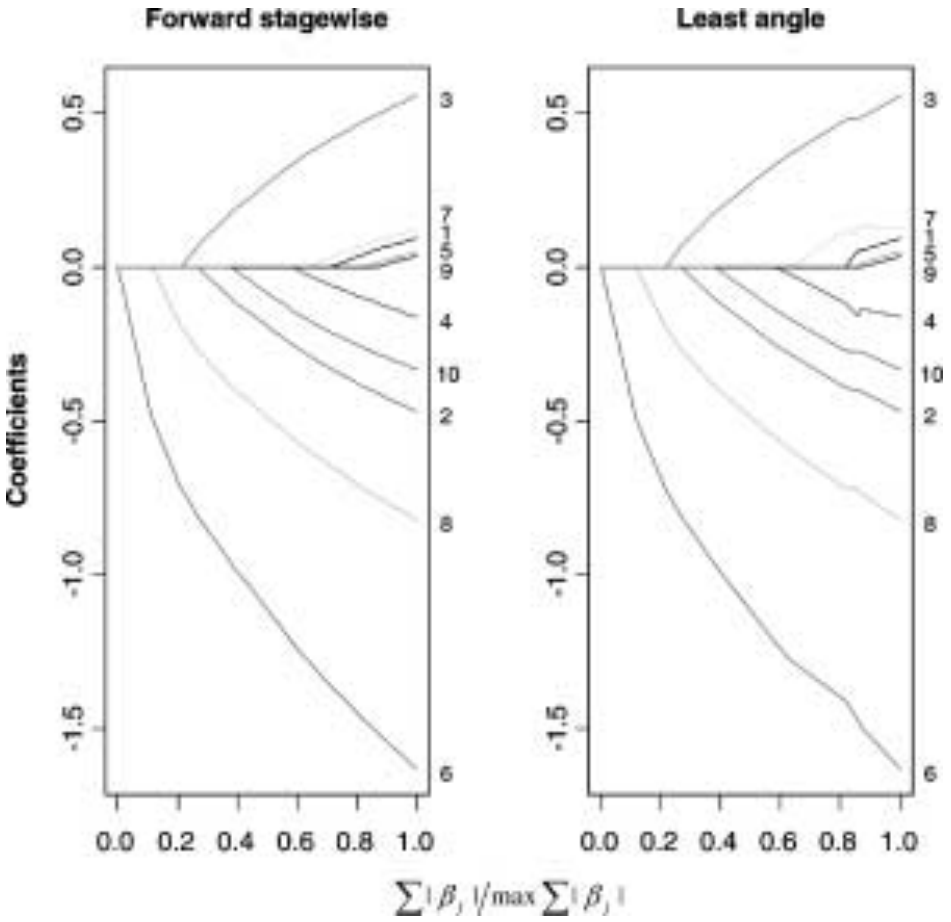

FIG. 1. *Comparison of coefficient estimation for Forward Stagewise and Least Angle Logistic Regression.*

Specifically for the Lasso, one alternative strategy for logistic regression is to use a quadratic approximation for the log-likelihood. Consider the Bayesian version of Lasso with hyperparameter γ (i.e., the penalized rather than constrained version of Lasso):

$$\begin{aligned}
&\log f(\boldsymbol{\beta}|y_1,\ldots,y_n)\\
&\quad\propto \sum_{i=1}^{n}\log\bigl(y_i\Lambda(\mathbf{x}_i\boldsymbol{\beta})+(1-y_i)\bigl(1-\Lambda(\mathbf{x}_i\boldsymbol{\beta})\bigr)\bigr)+d\log\biggl(\frac{\gamma^{1/2}}{2}\biggr)-\gamma^{1/2}\sum_{i=1}^{d}|\beta_i|\\
&\quad\approx \biggl(\sum_{i=1}^{n}a_i(\mathbf{x}_i\boldsymbol{\beta})^2+b_i(\mathbf{x}_i\boldsymbol{\beta})+c_i\biggr)+d\log\biggl(\frac{\gamma^{1/2}}{2}\biggr)-\gamma^{1/2}\sum_{i=1}^{d}|\beta_i|,
\end{aligned}$$

where Λ denotes the logistic link, d is the dimension of $\boldsymbol{\beta}$ and a_i, b_i and c_i are Taylor coefficients. Fu's elegant coordinatewise "Shooting algorithm" [Fu (1998)], can optimize this target starting from either the least squares solution or from zero. In our experience the shooting algorithm converges rapidly.

REFERENCES

BREIMAN, L. (2001). Random forests. Available at ftp://ftp.stat.berkeley.edu/pub/users/breiman/randomforest2001.pdf.

FU, W. J. (1998). Penalized regressions: The Bridge versus the Lasso. *J. Comput. Graph. Statist.* **7** 397–416.

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403.

RIDGEWAY, G. (2003). GBM 0.7-2 package manual. Available at http://cran.r-project.org/doc/packages/gbm.pdf.

AVAYA LABS RESEARCH
AND
DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08855
USA
E-MAIL: madigan@stat.rutgers.edu

RAND STATISTICS GROUP
SANTA MONICA, CALIFORNIA 90407-2138
USA
E-MAIL: gregr@rand.org

DISCUSSION

BY SAHARON ROSSET AND JI ZHU

IBM T. J. Watson Research Center and Stanford University

1. Introduction. We congratulate the authors on their excellent work. The paper combines elegant theory and useful practical results in an intriguing manner. The LAR–Lasso–boosting relationship opens the door for new insights on existing

methods' underlying statistical mechanisms and for the development of new and promising methodology. Two issues in particular have captured our attention, as their implications go beyond the squared error loss case presented in this paper, into wider statistical domains: robust fitting, classification, machine learning and more. We concentrate our discussion on these two results and their extensions.

2. Piecewise linear regularized solution paths. The first issue is the piecewise linear solution paths to regularized optimization problems. As the discussion paper shows, the path of optimal solutions to the "Lasso" regularized optimization problem

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{1}$$

is piecewise linear as a function of λ; that is, there exist $\infty > \lambda_0 > \lambda_1 > \cdots > \lambda_m = 0$ such that $\forall\, \lambda \geq 0$, with $\lambda_k \geq \lambda \geq \lambda_{k+1}$, we have

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_k) - (\lambda - \lambda_k)\gamma_k.$$

In the discussion paper's terms, γ_k is the "LAR" direction for the kth step of the LAR–Lasso algorithm.

This property allows the LAR–Lasso algorithm to generate the whole path of Lasso solutions, $\hat{\beta}(\lambda)$, for "practically" the cost of one least squares calculation on the data (this is exactly the case for LAR but not for LAR–Lasso, which may be significantly more computationally intensive on some data sets). The important practical consequence is that it is not necessary to select the regularization parameter λ in advance, and it is now computationally feasible to optimize it based on cross-validation (or approximate C_p, as presented in the discussion paper).

The question we ask is: what makes the solution paths piecewise linear? Is it the use of squared error loss? Or the Lasso penalty? The answer is that both play an important role. However, the family of (loss, penalty) pairs which facilitates piecewise linear solution paths turns out to contain many other interesting and useful optimization problems.

We now briefly review our results, presented in detail in Rosset and Zhu (2004). Consider the general regularized optimization problem

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_i L(y_i, \mathbf{x}_i^t\beta) + \lambda J(\beta), \tag{2}$$

where we only assume that the loss L and the penalty J are both convex functions of β for any sample $\{\mathbf{x}_i^t, y_i\}_{i=1}^n$. For our discussion, the data sample is assumed fixed, and so we will use the notation $L(\beta)$, where the dependence on the data is implicit.

Notice that piecewise linearity is equivalent to requiring that

$$\frac{\partial\hat{\beta}(\lambda)}{\partial\lambda} \in \mathcal{R}^p$$

is piecewise constant as a function of λ. If L, J are twice differentiable functions of β, then it is easy to derive that

$$\frac{\partial \hat{\beta}(\lambda)}{\partial \lambda} = -\big(\nabla^2 L(\hat{\beta}(\lambda)) + \lambda \nabla^2 J(\hat{\beta}(\lambda))\big)^{-1} \nabla J\big(\hat{\beta}(\lambda)\big). \tag{3}$$

With a little more work we extend this result to "almost twice differentiable" loss and penalty functions (i.e., twice differentiable everywhere except at a finite number of points), which leads us to the following *sufficient conditions for piecewise linear* $\hat{\beta}(\lambda)$:

1. $\nabla^2 L(\hat{\beta}(\lambda))$ is piecewise constant as a function of λ. This condition is met if L is a piecewise-quadratic function of β. This class includes the squared error loss of the Lasso, but also absolute loss and combinations of the two, such as Huber's loss.
2. $\nabla J(\hat{\beta}(\lambda))$ is piecewise constant as a function of λ. This condition is met if J is a piecewise-linear function of β. This class includes the l_1 penalty of the Lasso, but also the l_∞ norm penalty.

2.1. *Examples*. Our first example is the "Huberized" Lasso; that is, we use the loss

$$L(y, \mathbf{x}\beta) = \begin{cases} (y - \mathbf{x}^t\beta)^2, & \text{if } |y - \mathbf{x}^t\beta| \le \delta, \\ \delta^2 + 2\delta(|y - \mathbf{x}\beta| - \delta), & \text{otherwise,} \end{cases} \tag{4}$$

with the Lasso penalty. This loss is more robust than squared error loss against outliers and long-tailed residual distributions, while still allowing us to calculate the whole path of regularized solutions efficiently.

To illustrate the importance of robustness in addition to regularization, consider the following simple simulated example: take $n = 100$ observations and $p = 80$ predictors, where all x_{ij} are i.i.d. $N(0, 1)$ and the true model is

$$y_i = 10 \cdot x_{i1} + \varepsilon_i, \tag{5}$$

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} 0.9 \cdot N(0, 1) + 0.1 \cdot N(0, 100). \tag{6}$$

So the normality of residuals, implicitly assumed by using squared error loss, is violated.

Figure 1 shows the optimal coefficient paths $\hat{\beta}(\lambda)$ for the Lasso (right) and "Huberized" Lasso, using $\delta = 1$ (left). We observe that the Lasso fails in identifying the correct model $E(Y|x) = 10x_1$ while the robust loss identifies it almost exactly, *if we choose the appropriate regularization parameter*.

As a second example, consider a classification scenario where the loss we use depends on the margin $y\mathbf{x}^t\beta$ rather than on the residual. In particular, consider the 1-*norm* support vector machine regularized optimization problem, popular in the

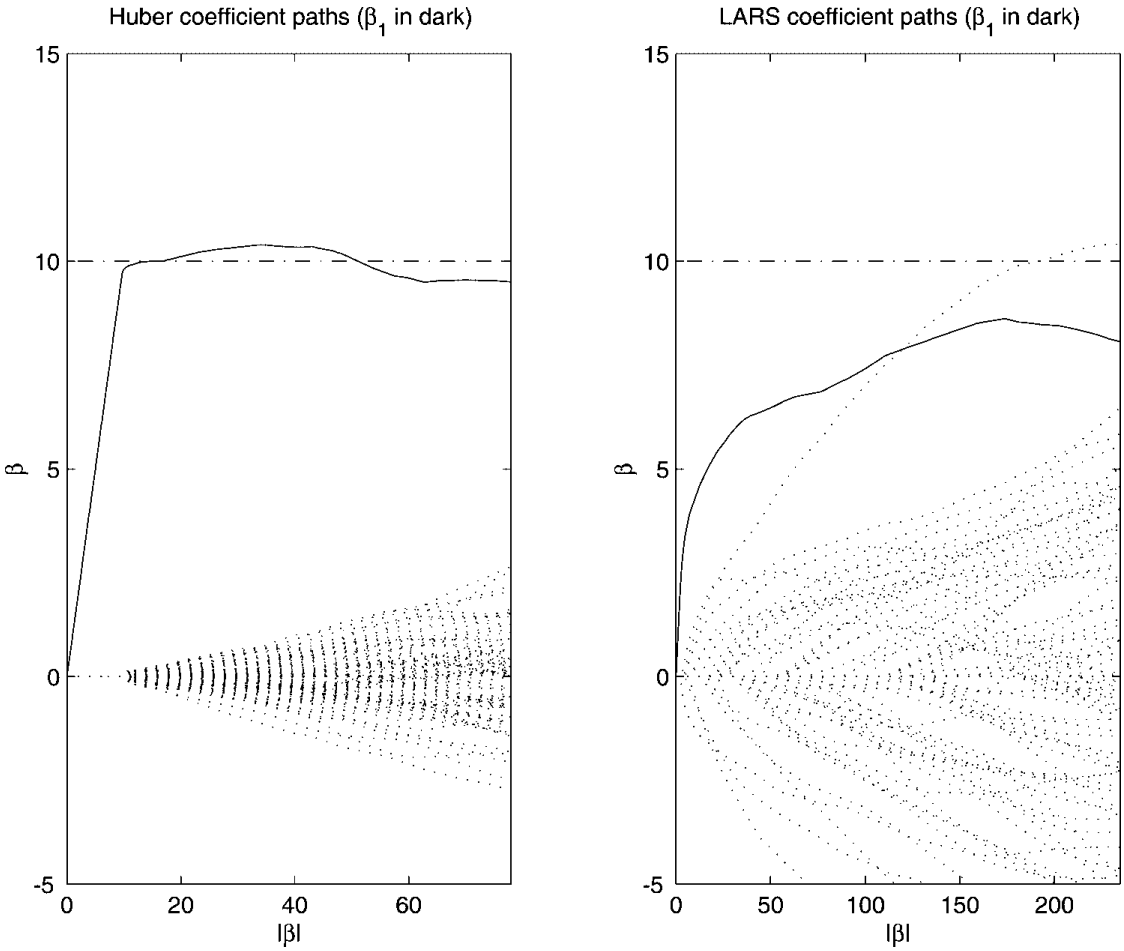

FIG. 1. *Coefficient paths for Huberized Lasso* (left) *and Lasso* (right) *for data example*: $\hat{\beta}_1(\lambda)$ *is the full line*; *the true model is* $E(Y|x) = 10x_1$.

machine learning community. It consists of minimizing the "hinge loss" with a Lasso penalty:

$$L(y, \mathbf{x}^t\beta) = \begin{cases} (1 - y\mathbf{x}^t\beta), & \text{if } y\mathbf{x}^t\beta \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

This problem obeys our conditions for piecewise linearity, and so we can generate all regularized solutions for this fitting problem efficiently. This is particularly advantageous in high-dimensional machine learning problems, where regularization is critical, and it is usually not clear in advance what a good regularization parameter would be. A detailed discussion of the computational and methodological aspects of this example appears in Zhu, Rosset, Hastie, and Tibshirani (2004).

3. Relationship between "boosting" algorithms and l_1-regularized fitting. The discussion paper establishes the close relationship between ε-stagewise linear regression and the Lasso. Figure 1 in that paper illustrates the near-equivalence in

the solution paths generated by the two methods, and Theorem 2 formally states a related result. It should be noted, however, that their theorem falls short of proving the global relation between the methods, which the examples suggest.

In Rosset, Zhu and Hastie (2003) we demonstrate that this relation between the path of l_1-regularized optimal solutions [which we have denoted above by $\hat{\beta}(\lambda)$] and the path of "generalized" ε-stagewise (AKA boosting) solutions extends beyond squared error loss and in fact applies to any convex differentiable loss.

More concretely, consider the following generic gradient-based "ε-boosting" algorithm [we follow Friedman (2001) and Mason, Baxter, Bartlett and Frean (2000) in this view of boosting], which iteratively builds the solution path $\beta^{(t)}$:

ALGORITHM 1 (Generic gradient-based boosting algorithm).

1. Set $\beta^{(0)} = 0$.
2. For $t = 1 : T$,
   (a) Let $j_t = \arg\max_j |\frac{\partial \sum_i L(y_i, \mathbf{x}_i^t \beta^{(t-1)})}{\partial \beta_j^{(t-1)}}|$.
   (b) Set $\beta_{j_t}^{(t)} = \beta_{j_t}^{(t-1)} - \varepsilon \operatorname{sign}(\frac{\partial \sum_i L(y_i, \mathbf{x}_i^t \beta^{(t-1)})}{\partial \beta_{j_t}^{(t-1)}})$ and $\beta_k^{(t)} = \beta_k^{(t-1)}, k \neq j_t$.

This is a coordinate descent algorithm, which reduces to forward stagewise, as defined in the discussion paper, if we take the loss to be squared error loss: $L(y_i, \mathbf{x}_i^t \beta^{(t-1)}) = (y_i - \mathbf{x}_i^t \beta^{(t-1)})^2$. If we take the loss to be the exponential loss,

$$L\big(y_i, \mathbf{x}_i^t \beta^{(t-1)}\big) = \exp\big(-y_i \mathbf{x}_i^t \beta^{(t-1)}\big),$$

we get a variant of AdaBoost [Freund and Schapire (1997)]—the original and most famous boosting algorithm.

Figure 2 illustrates the equivalence between Algorithm 1 and the optimal solution path for a simple logistic regression example, using five variables from the "spam" dataset. We can see that there is a perfect equivalence between the regularized solution path (left) and the "boosting" solution path (right).

In Rosset, Zhu and Hastie (2003) we formally state this equivalence, with the required conditions, as a conjecture. We also generalize the weaker result, proven by the discussion paper for the case of squared error loss, to any convex differentiable loss.

This result is interesting in the boosting context because it facilitates a view of boosting as approximate and implicit regularized optimization. The situations in which boosting is employed in practice are ones where explicitly solving regularized optimization problems is not practical (usually very high-dimensional predictor spaces). The approximate regularized optimization view which emerges from our results allows us to better understand boosting and its great empirical success [Breiman (1999)]. It also allows us to derive approximate convergence results for boosting.

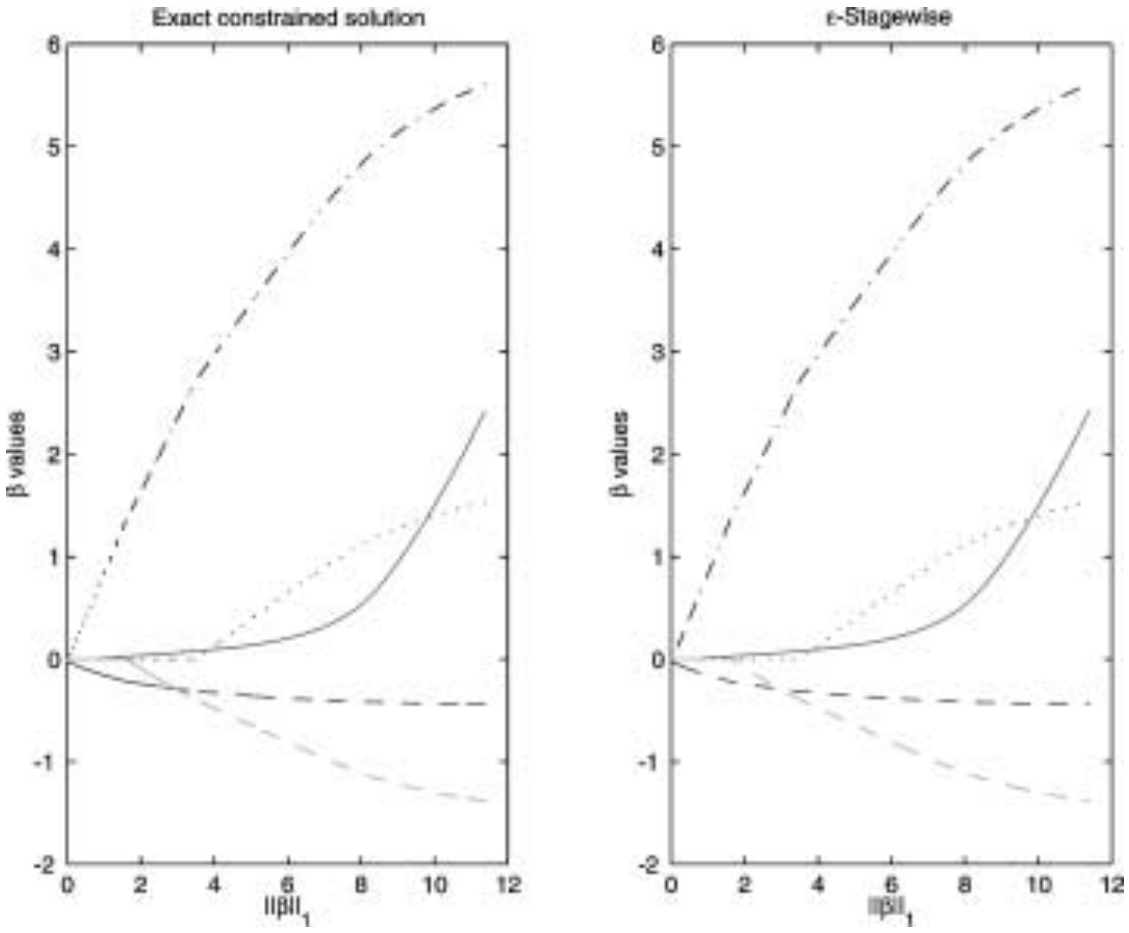

FIG. 2. *Exact coefficient paths* (left) *for* l_1*-constrained logistic regression and boosting coefficient paths* (right) *with binomial log-likelihood loss on five variables from the* "*spam*" *dataset. The boosting path was generated using* $\varepsilon = 0.003$ *and* 7000 *iterations.*

4. Conclusion. The computational and theoretical results of the discussion paper shed new light on variable selection and regularization methods for linear regression. However, we believe that variants of these results are useful and applicable beyond that domain. We hope that the two extensions that we have presented convey this message successfully.

Acknowledgment. We thank Giles Hooker for useful comments.

REFERENCES

BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Computation* **11** 1493–1517.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232.

MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (2000). Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems* **12** 512–518. MIT Press, Cambridge, MA.

ROSSET, S. and ZHU, J. (2004). Piecewise linear regularized solution paths. *Advances in Neural Information Processing Systems* **16**. To appear.

ROSSET, S., ZHU, J. and HASTIE, T. (2003). Boosting as a regularized path to a maximum margin classifier. Technical report, Dept. Statistics, Stanford Univ.

ZHU, J., ROSSET, S., HASTIE, T. and TIBSHIRANI, R. (2004). 1-norm support vector machines. *Neural Information Processing Systems* **16**. To appear.

IBM T. J. WATSON RESEARCH CENTER
P.O. BOX 218
YORKTOWN HEIGHTS, NEW YORK 10598
USA
E-MAIL: srosset@us.ibm.com

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
550 EAST UNIVERSITY
ANN ARBOR, MICHIGAN 48109-1092
USA
E-MAIL: jizhu@umich.edu

DISCUSSION

BY ROBERT A. STINE

University of Pennsylvania

I have enjoyed reading the work of each of these authors over the years, so it is a real pleasure to have this opportunity to contribute to the discussion of this collaboration. The geometry of LARS furnishes an elegant bridge between the Lasso and Stagewise regression, methods that I would not have suspected to be so related. Toward my own interests, LARS offers a rather different way to construct a regression model by gradually blending predictors rather than using a predictor all at once. I feel that the problem of "automatic feature generation" (proposing predictors to consider in a model) is a current challenge in building regression models that can compete with those from computer science, and LARS suggests a new approach to this task. In the examples of Efron, Hastie, Johnstone and Tibshirani (EHJT) (particularly that summarized in their Figure 5), LARS produces models with smaller predictive error than the old workhorse, stepwise regression. Furthermore, as an added bonus, the code supplied by the authors runs faster for me than the `step` routine for stepwise regression supplied with *R*, the generic version of S-PLUS that I use.

My discussion focuses on the use of C_p to choose the number of predictors. The bootstrap simulations in EHJT show that LARS reaches higher levels of "proportion explained" than stepwise regression. Furthermore, the goodness-of-fit obtained by LARS remains high over a wide range of models, in sharp contrast to the narrow peak produced by stepwise selection. Because the cost of overfitting with LARS appears less severe than with stepwise, LARS would seem to have a clear advantage in this respect. Even if we do overfit, the fit of LARS degrades

only slightly. The issue becomes learning how much LARS overfits, particularly in situations with many potential predictors (m as large as or larger than n).

To investigate the model-selection aspects of LARS further, I compared LARS to stepwise regression using a "reversed" five-fold cross-validation. The cross-validation is reversed in the sense that I estimate the models on one fold (88 observations) and then predict the other four. This division with more set aside for validation than used in estimation offers a better comparison of models. For example, Shao (1993) shows that one needs to let the proportion used for validation grow large in order to get cross validation to find the right model. This reversed design also adds a further benefit of making the variable selection harder. The quadratic model fitted to the diabetes data in EHJT selects from $m = 64$ predictors using a sample of $n = 442$ cases, or about 7 cases per predictor. Reversed cross-validation is closer to a typical data-mining problem. With only one fold of 88 observations to train the model, observation noise obscures subtle predictors. Also, only a few degrees of freedom remain to estimate the error variance $\overline{\sigma}^2$ that appears in C_p [equation (4.5)]. Because I also wanted to see what happens when $m > n$, I repeated the cross-validation with 5 additional possible predictors added to the 10 in the diabetes data. These 5 spurious predictors were simulated i.i.d. Gaussian noise; one can think of them as extraneous predictors that one might encounter when working with an energetic, creative colleague who suggests many ideas to explore. With these 15 base predictors, the search must consider $m = 15 + \binom{15}{2} + 14 = 134$ possible predictors.

Here are a few details of the cross-validation. To obtain the stepwise results, I ran forward stepwise using the hard threshold $2 \log m$, which is also known as the risk inflation criterion (RIC) [Donoho and Johnstone (1994) and Foster and George (1994)]. One begins with the most significant predictor. If the squared t-statistic for this predictor, say $t^2_{(1)}$, is less than the threshold $2 \log m$, then the search stops, leaving us with the "null" model that consists of just an intercept. If instead $t^2_{(1)} \geq 2 \log m$, the associated predictor, say $X_{(1)}$, joins the model and the search moves on to the next predictor. The second predictor $X_{(2)}$ joins the model if $t^2_{(2)} \geq 2 \log m$; otherwise the search stops with the one-predictor model. The search continues until reaching a predictor whose t-statistic fails this test, $t^2_{(q+1)} < 2 \log m$, leaving a model with q predictors. To obtain the results for LARS, I picked the order of the fit by minimizing C_p. Unlike the forward, sequential stepwise search, LARS globally searches a collection of models up to some large size, seeking the model which has the smallest C_p. I set the maximum model size to 50 (for $m = 64$) or 64 (for $m = 134$). In either case, the model is chosen from the collection of linear and quadratic effects in the 10 or 15 basic predictors. Neither search enforces the principle of marginality; an interaction can join the model without adding main effects.

I repeated the five-fold cross validation 20 times, each time randomly partitioning the 442 cases into 5 folds. This produces 100 stepwise and LARS fits. For each

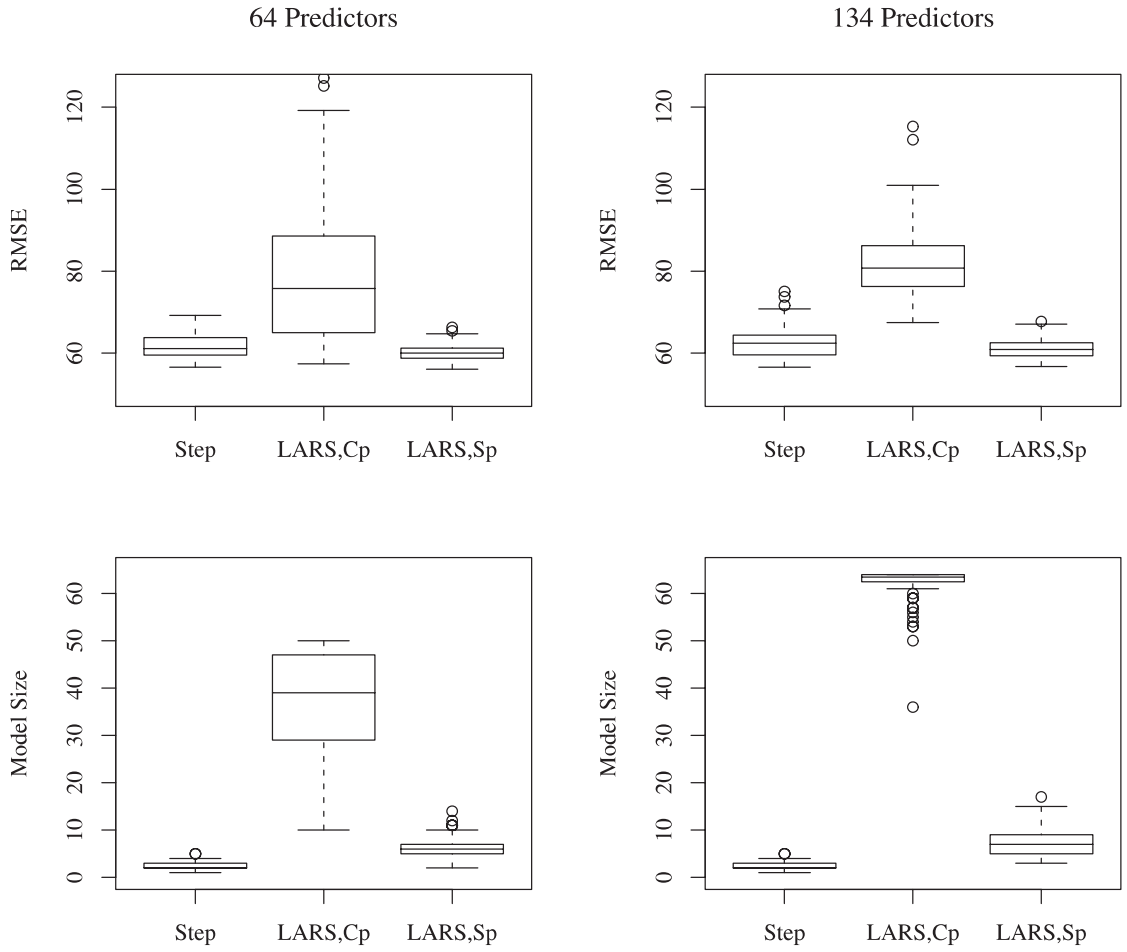

FIG. 1. *Five-fold cross-validation of the prediction error and size of stepwise regression and LARS when fitting models to a collection of* 64 (left) *or* 134 *predictors* (right). *LARS fits chosen by C_p overfit and have larger RMSE than stepwise*; *with C_p replaced by the alternative criterion S_p defined in* (3), *the LARS fits become more parsimonious with smaller RMSE. The random splitting into estimation and validation samples was repeated* 20 *times*, *for a total of* 100 *stepwise and LARS fits*.

of these, I computed the square root of the out-of-sample mean square error (MSE) when the model fit on one fold was used to predict the held-back 354 [$= 4(88) + 2$] observations. I also saved the size q for each fit.

Figure 1 summarizes the results of the cross-validation. The comparison boxplots on the left compare the square root of the MSE (top) and selected model order (bottom) of stepwise to LARS when picking from $m = 64$ candidate predictors; those on the right summarize what happens with $m = 134$. When choosing from among 64 predictors, the median size of a LARS model identified by C_p is 39. The median stepwise model has but 2 predictors. (I will describe the S_p criterion further below.) The effects of overfitting on the prediction error of LARS are clear: LARS has higher RMSE than stepwise. The median RMSE for stepwise is near 62. For LARS, the median RMSE is larger, about 78. Although the predictive accuracy of LARS declines more slowly than that of stepwise when it

overfits (imagine the fit of stepwise with 39 predictors), LARS overfits by enough in this case that it predicts worse than the far more parsimonious stepwise models. With more predictors ($m = 134$), the boxplots on the right of Figure 1 show that C_p tends to pick the largest possible model—here a model with 64 predictors.

Why does LARS overfit? As usual with variable selection in regression, it is simpler to try to understand when thinking about the utopian orthogonal regression with known σ^2. Assume, as in Lemma 1 of EHJT, that the predictors X_j are the columns of an identity matrix, $X_j = e_j = (0, \ldots, 0, 1_j, 0, \ldots, 0)$. Assume also that we know $\sigma^2 = 1$ and use it in place of the troublesome $\overline{\sigma}^2$ in C_p, so that for this discussion

$$C_p = \mathrm{RSS}(p) - n + 2p. \tag{1}$$

To define $\mathrm{RSS}(p)$ in this context, denote the ordered values of the response as

$$Y^2_{(1)} > Y^2_{(2)} > \cdots > Y^2_{(n)}.$$

The soft thresholding summarized in Lemma 1 of EHJT implies that the residual sum-of-squares of LARS with q predictors is

$$\mathrm{RSS}(q) = (q+1)Y^2_{(q+1)} + \sum_{j=q+2}^{n} Y^2_{(j)}.$$

Consequently, the drop in C_p when going from the model with q to the model with $q+1$ predictors is

$$C_q - C_{q+1} = (q+1)\, d_q - 2,$$

with

$$d_q = Y^2_{(q+1)} - Y^2_{(q+2)};$$

C_p adds X_{q+1} to the model if $C_q - C_{q+1} > 0$.

This use of C_p works well for the orthogonal "null" model, but overfits when the model includes much signal. Figure 2 shows a graph of the mean and standard deviation of $\mathrm{RSS}(q) - \mathrm{RSS}(0) + 2q$ for an orthogonal model with $n = m = 100$ and $Y_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. I subtracted $\mathrm{RSS}(0)$ rather than n to reduce the variation in the simulation. Figure 3 gives a rather different impression. The simulation is identical except that the data have some signal. Now, $EY_i = 3$ for $i = 1, \ldots, 5$. The remaining 95 observations are $N(0, 1)$. The "true" model has only 5 nonzero components, but the minimal expected C_p falls near 20.

This stylized example suggests an explanation for the overfitting—as well as motivates a way to avoid some of it. Consider the change in RSS for a null model when adding the sixth predictor. For this step, $\mathrm{RSS}(5) - \mathrm{RSS}(6) = 6(Y^2_{(6)} - Y^2_{(7)})$. Even though we multiply the difference between the squares by 6, adjacent order statistics become closer when removed from the extremes, and C_p tends to increase

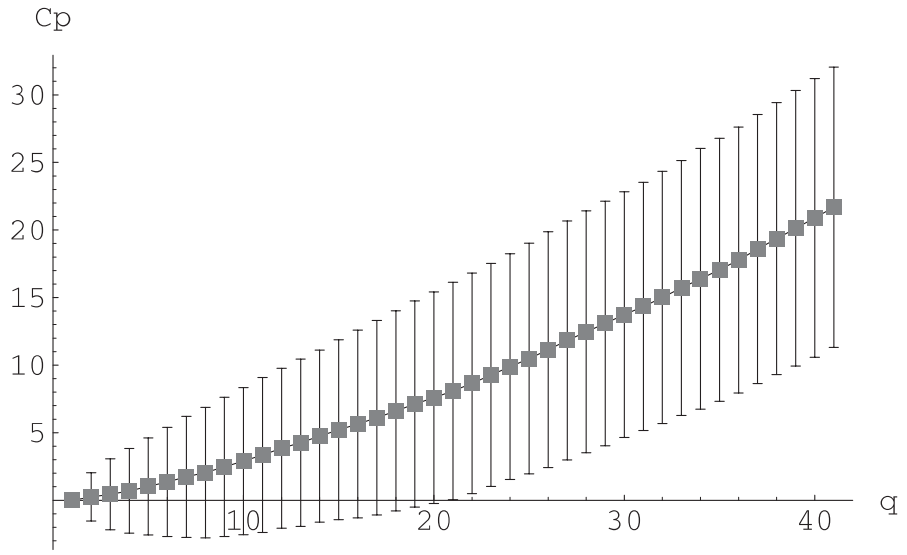

FIG. 2. *A simulation of C_p for LARS applied to orthogonal, normal data with no signal correctly identifies the null model. These results are from a simulation with* 1000 *replications, each consisting of a sample of* 100 *i.i.d. standard normal observations. Error bars indicate* ± 1 *standard deviation.*

as shown in Figure 2. The situation changes when signal is present. First, the five observations with mean 3 are likely to be the first five ordered observations. So, their spacing is likely to be larger because their order is determined by a sample of five normals; C_p adds these. When reaching the noise, the difference $Y^2_{(6)} - Y^2_{(7)}$ now behaves like the difference between the *first two* squared order statistics in an

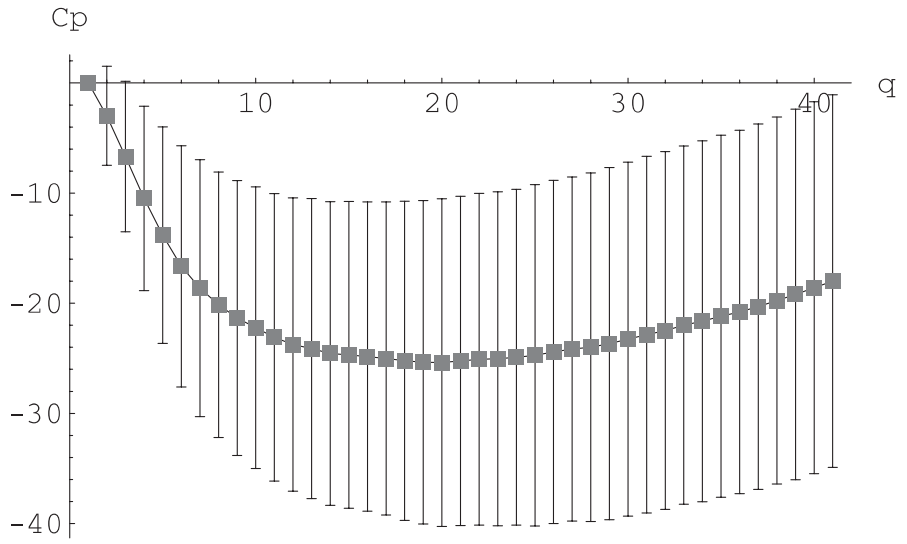

FIG. 3. *A simulation of C_p for LARS applied to orthogonal, normal data with signal present overfits. Results are from a simulation with* 1000 *replications, each consisting of* 5 *observations with mean* 3 *combined with a sample of* 95 *i.i.d. standard normal observations. Error bars indicate* ± 1 *standard deviation.*

i.i.d. sample of 95 standard normals. Consequently, this comparison involves the gap between the most extreme order statistics rather than those from within the sample, and as a result C_p drops to indicate a larger model.

This explanation of the overfitting suggests a simple alternative to C_p that leads to smaller LARS models. The idea is to compare the decreasing residual sum of squares $\mathrm{RSS}(q)$ to what is expected under a model that has fitted some signal *and* some noise. Since overfitting seems to have relatively benign effects on LARS, one does not want to take the hard-thresholding approach; my colleague Dean Foster suggested that the criterion might do better by assuming that some of the predictors already in the model are really noise. The criterion S_p suggested here adopts this notion. The form of S_p relies on approximations for normal order statistics commonly used in variable selection, particularly adaptive methods [Benjamini and Hochberg (1995) and Foster and Stine (1996)]. These approximate the size of the jth normal order statistic in a sample of n with $\sqrt{2\log(n/j)}$. To motivate the form of the S_p criterion, I return to the orthogonal situation and consider what happens when deciding whether to increase the size of the model from q to $q+1$ predictors. If I know that k of the already included q predictors represent signal and the rest of the predictors are noise, then $d_q = Y^2_{(q+1)} - Y^2_{(q+2)}$ is about

$$2\log\frac{m-k}{q+1-k} - 2\log\frac{m-k}{q+2-k}. \tag{2}$$

Since I do not know k, I will just set $k = q/2$ (i.e., assume that half of those already in the model are noise) and approximate d_q as

$$\delta(q) = 2\log\frac{q/2+2}{q/2+1}.$$

[Define $\delta(0) = 2\log 2$ and $\delta(1) = 2\log 3/2$.] This approximation suggests choosing the model that minimizes

$$S_q = \mathrm{RSS}(q) + \hat{\sigma}^2 \sum_{j=1}^{q} j\delta(j), \tag{3}$$

where $\hat{\sigma}^2$ represents an "honest" estimate of σ^2 that avoids selection bias. The S_p criterion, like C_p, penalizes the residual sum-of-squares, but uses a different penalty.

The results for LARS with this criterion define the third set of boxplots in Figure 1. To avoid selection bias in the estimate of σ^2, I used a two-step procedure. First, fit a forward stepwise regression using hard thresholding. Second, use the estimated error variance from this stepwise fit as $\hat{\sigma}^2$ in S_p and proceed with LARS. Because hard thresholding avoids overfitting in the stepwise regression, the resulting estimator $\hat{\sigma}^2$ is probably conservative—but this is just what is needed when modeling data with an excess of possible predictors. If the variance estimate from the largest LARS model is used instead, the S_p criterion also overfits (though

not so much as C_p). Returning to Figure 1, the combination of LARS with S_p obtains the smallest typical MSE with both $m = 64$ and 134 predictors. In either case, LARS includes more predictors than the parsimonious stepwise fits obtained by hard thresholding.

These results lead to more questions. What are the risk properties of the LARS predictor chosen by C_p or S_p? How is it that the number of possible predictors m does not appear in either criterion? This definition of S_p simply supposes half of the included predictors are noise; why half? What is a better way to set k in (2)? Finally, that the combination of LARS with either C_p or S_p has less MSE than stepwise when predicting diabetes is hardly convincing that such a pairing would do well in other applications. Statistics would be well served by having a repository of test problems comparable to those held at UC Irvine for judging machine learning algorithms [Blake and Merz (1998)].

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

BLAKE, C. and MERZ, C. (1998). UCI repository of machine learning databases. Technical report, School Information and Computer Science, Univ. California, Irvine. Available at www.ics.uci.edu/~mlearn/MLRepository.html.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.

FOSTER, D. P. and STINE, R. A. (1996). Variable selection via information theory. Technical Report Discussion Paper 1180, Center for Mathematical Studies in Economics and Management Science, Northwestern Univ.

SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: stine@wharton.upenn.edu

DISCUSSION

BY BERWIN A. TURLACH

University of Western Australia

I would like to begin by congratulating the authors (referred to below as EHJT) for their interesting paper in which they propose a new variable selection method

(LARS) for building linear models and show how their new method relates to other methods that have been proposed recently. I found the paper to be very stimulating and found the additional insight that it provides about the Lasso technique to be of particular interest.

My comments center around the question of how we can select linear models that conform with the marginality principle [Nelder (1977, 1994) and McCullagh and Nelder (1989)]; that is, the response surface is invariant under scaling and translation of the explanatory variables in the model. Recently one of my interests was to explore whether the Lasso technique or the nonnegative garrote [Breiman (1995)] could be modified such that it incorporates the marginality principle. However, it does not seem to be a trivial matter to change the criteria that these techniques minimize in such a way that the marginality principle is incorporated in a satisfactory manner.

On the other hand, it seems to be straightforward to modify the LARS technique to incorporate this principle. In their paper, EHJT address this issue somewhat in passing when they suggest toward the end of Section 3 that one first fit main effects only and interactions in a second step to control the order in which variables are allowed to enter the model. However, such a two-step procedure may have a somewhat less than optimal behavior as the following, admittedly artificial, example shows.

Assume we have a vector of explanatory variables $X = (X_1, X_2, \ldots, X_{10})$ where the components are independent of each other and $X_i, i = 1, \ldots, 10$, follows a uniform distribution on $[0, 1]$. Take as model

$$Y = (X_1 - 0.5)^2 + X_2 + X_3 + X_4 + X_5 + \varepsilon, \tag{1}$$

where ε has mean zero, has standard deviation 0.05 and is independent of X.

It is not difficult to verify that in this model X_1 and Y are uncorrelated. Moreover, since the X_i's are independent, X_1 is also uncorrelated with any residual vector coming from a linear model formed only by explanatory variables selected from $\{X_2, \ldots, X_{10}\}$.

Thus, if one fits a main effects only model, one would expect that the LARS algorithm has problems identifying that X_1 should be part of the model. That this is indeed the case is shown in Figure 1. The picture presents the result of the LARS analysis for a typical data set generated from model (1); the sample size was $n = 500$. Note that, unlike Figure 3 in EHJT, Figure 1 (and similar figures below) has been produced using the standardized explanatory variables and no back-transformation to the original scale was done.

For this realization, the variables are selected in the sequence X_2, X_5, X_4, X_3, X_6, X_{10}, X_7, X_8, X_9 and, finally, X_1. Thus, as expected, the LARS algorithm has problems identifying X_1 as part of the model. To further verify this, 1000 different data sets, each of size $n = 500$, were simulated from model (1) and a LARS analysis performed on each of them. For each of the 10 explanatory variables the

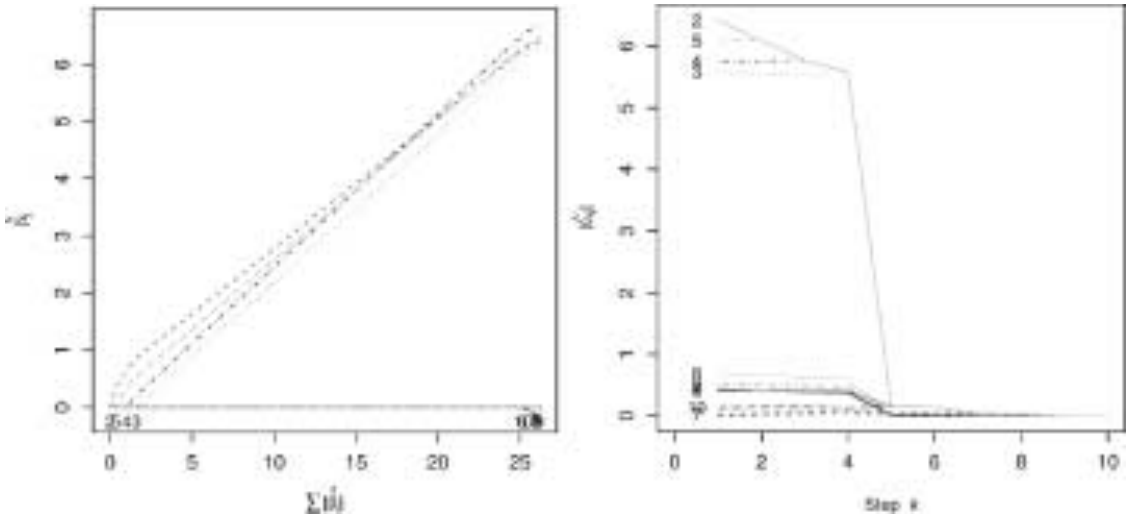

FIG. 1. *LARS analysis of simulated data with main terms only*: (left) *estimates of regression coefficients* $\hat{\beta}_j$, $j = 1, \ldots, 10$, *plotted versus* $\sum |\hat{\beta}_j|$; (right) *absolute current correlations as functions of LARS step*.

step at which it was selected to enter the model was recorded. Figure 2 shows for each of the variables the (percentage) histogram of these data.

It is clear that the LARS algorithm has no problems identifying that $X_2, \ldots, X_5$ should be in the model. These variables are all selected in the first four steps and, not surprisingly given the model, with more or less equal probability at any of these

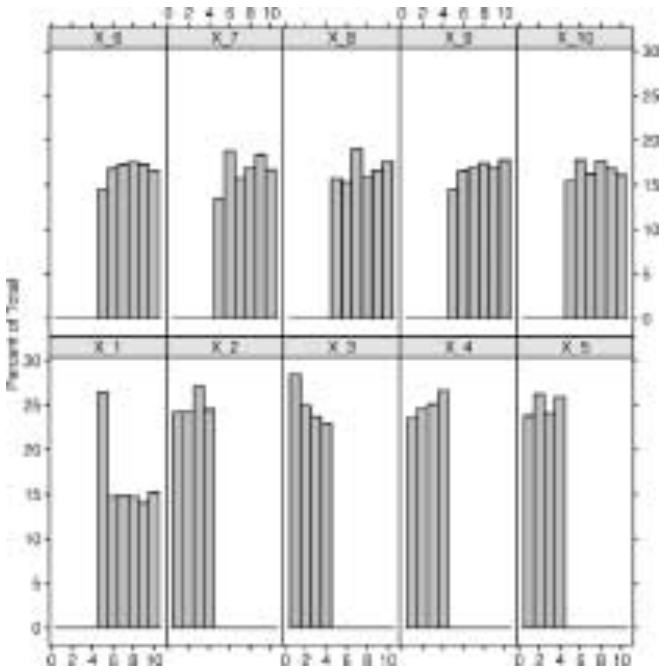

FIG. 2. *Percentage histogram of step at which each variable is selected based on* 1000 *replications*: *results shown for LARS analysis using main terms only*.

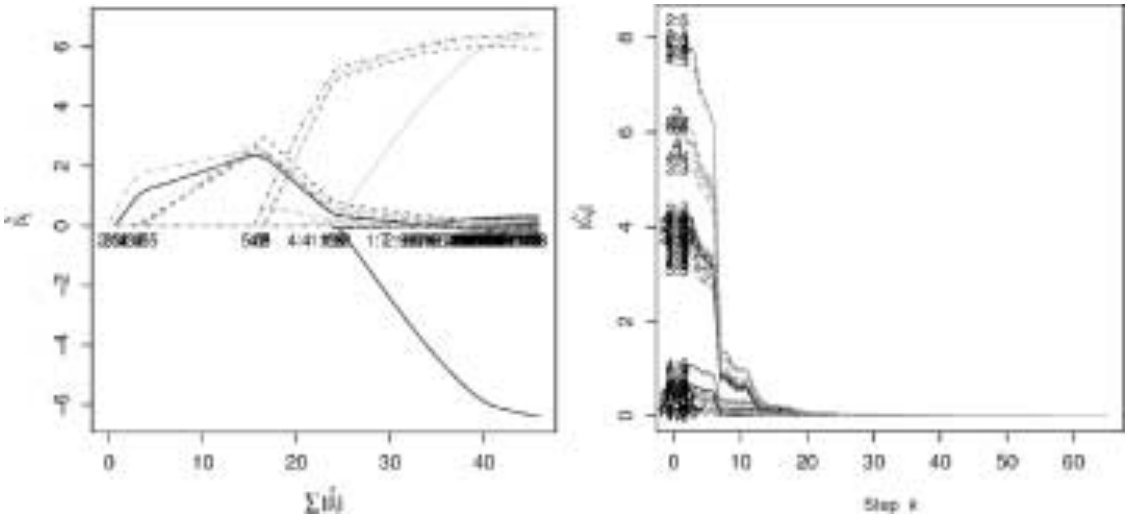

FIG. 3. *LARS analysis of simulated data with main terms and interaction terms*: (left) *estimates of regression coefficients* $\hat{\beta}_j$, $j = 1, \ldots, 65$, *plotted versus* $\sum |\hat{\beta}_j|$; (right) *absolute current correlations as functions of LARS step*.

steps. X_1 has a chance of approximately 25% of being selected as the fifth variable, otherwise it may enter the model at step 6, 7, 8, 9 or 10 (each with probability roughly 15%). Finally, each of the variables X_6 to X_{10} seems to be selected with equal probability anywhere between step 5 and step 10.

This example shows that a main effects first LARS analysis followed by a check for interaction terms would not work in such cases. In most cases the LARS analysis would miss X_1 as fifth variable and even in the cases where it was selected at step 5 it would probably be deemed to be unimportant and excluded from further analysis.

How does LARS perform if one uses from the beginning all 10 main effects and all 55 interaction terms? Figure 3 shows the LARS analysis for the same data used to produce Figure 1 but this time the design matrix was augmented to contain all main effects and all interactions. The order in which the variables enter the model is $X_{2:5} = X_2 \times X_5$, $X_{2:4}$, $X_{3:4}$, $X_{2:3}$, $X_{3:5}$, $X_{4:5}$, $X_{5:5} = X_5^2$, X_4, X_3, X_2, X_5, $X_{4:4}$, $X_{1:1}$, $X_{1:6}$, $X_{1:9}$, X_1, ... In this example the last of the six terms that are actually in model (1) was selected by the LARS algorithm in step 16.

Using the same 1000 samples of size $n = 500$ as above and performing a LARS analysis on them using a design matrix with all main and interaction terms shows a surprising result. Again, for each replication the step at which a variable was selected into the model by LARS was recorded and Figure 4 shows for each variable histograms for these data. To avoid cluttering, the histograms in Figure 4 were truncated to [1, 20]; the complete histograms are shown on the left in Figure 7.

The most striking feature of these histograms is that the six interaction terms $X_{i:j}$, $i, j \in \{2, 3, 4, 5\}$, $i < j$, were always selected first. In no replication was any

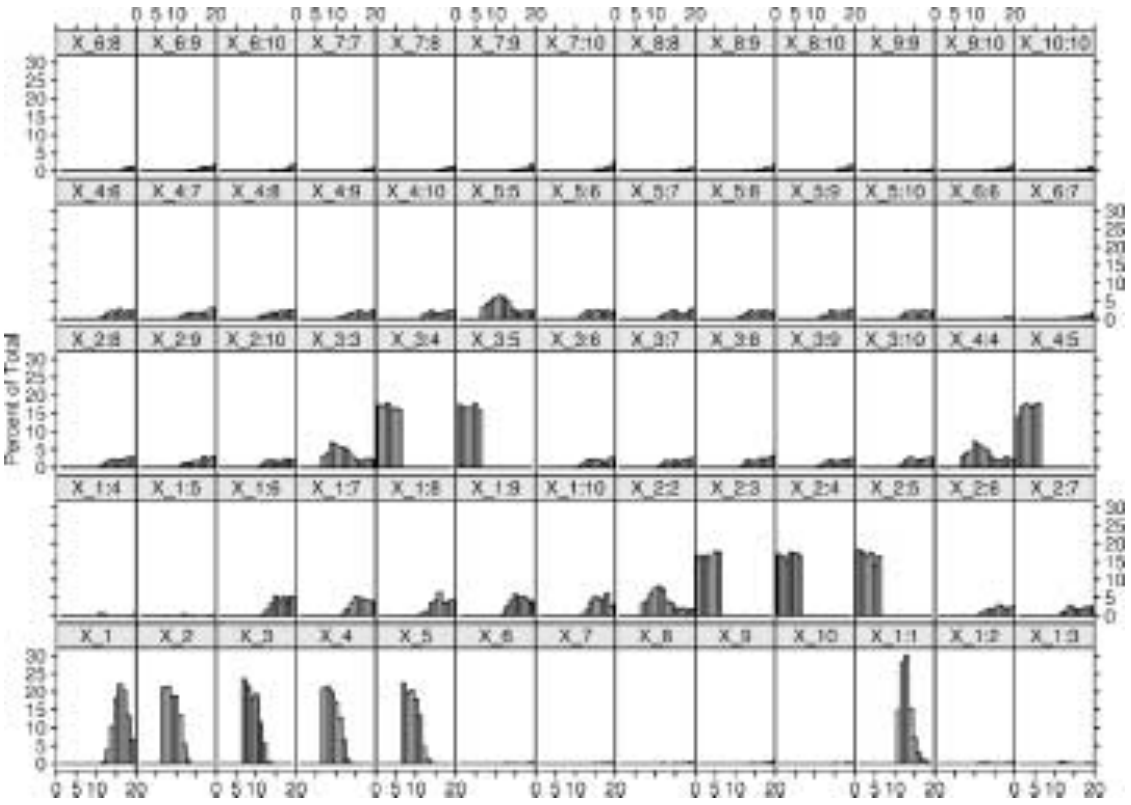

FIG. 4. *Percentage histogram of step at which each variable is selected based on* 1000 *replications: results shown for variables selected in the first* 20 *steps of a LARS analysis using main and interaction terms.*

of these terms selected after step 6 and no other variable was ever selected in the first six steps. That one of these terms is selected as the first term is not surprising as these variables have the highest correlation with the response variable Y. It can be shown that the covariance of these interaction terms with Y is by a factor $\sqrt{12/7} \approx 1.3$ larger than the covariance between X_i and Y for $i = 2, \ldots, 5$. But that these six interaction terms dominate the early variable selection steps in such a manner came as a bit as a surprise.

After selecting these six interaction terms, the LARS algorithm then seems to select mostly X_2, X_3, X_4 and X_5, followed soon by $X_{1:1}$ and X_1. However, especially the latter one seems to be selected rather late and other terms may be selected earlier. Other remarkable features in Figure 4 are the peaks in histograms of $X_{i:i}$ for $i = 2, 3, 4, 5$; each of these terms seems to have a fair chance of being selected before the corresponding main term and before $X_{1:1}$ and X_1.

One of the problems seems to be the large number of interaction terms that the LARS algorithm selects without putting the corresponding main terms into the model too. This behavior violates the marginality principle. Also, for this model, one would expect that ensuring that for each higher-order term the corresponding lower-order terms are in the model too would alleviate the problem that the six interaction terms $X_{i:j}, i, j \in \{2, 3, 4, 5\}, i < j$, are always selected first.

I give an alternative description of the LARS algorithm first before I show how it can be modified to incorporate the marginality principle. This description is based on the discussion in EHJT and shown in Algorithm 1.

ALGORITHM 1 (An alternative description of the LARS algorithm).

1. Set $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and $k = 0$.
2. **repeat**
3. Calculate $\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\boldsymbol{\mu}}_k)$ and set $\hat{C} = \max_j\{|\hat{c}_j|\}$.
4. Let $\mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}$.
5. Set $X_{\mathcal{A}} = (\cdots \mathbf{x}_j \cdots)_{j\in\mathcal{A}}$ for calculating $\bar{\mathbf{y}}_{k+1} = (X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}X'_{\mathcal{A}}\mathbf{y}$ and $\mathbf{a} = X'_{\mathcal{A}}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k)$.
6. Set

$$\hat{\boldsymbol{\mu}}_{k+1} = \hat{\boldsymbol{\mu}}_k + \hat{\gamma}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k),$$

where, if $\mathcal{A}^c \neq \varnothing$,

$$\hat{\gamma} = \min_{j\in\mathcal{A}^c}{}^{+}\left\{\frac{\hat{C} - \hat{c}_j}{\hat{C} - a_j}, \frac{\hat{C} + \hat{c}_j}{\hat{C} + a_j}\right\},$$

otherwise set $\hat{\gamma} = 1$.
7. $k \leftarrow k + 1$.
8. **until** $\mathcal{A}^c = \varnothing$.

We start with an estimated response $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and then iterate until all variables have been selected. In each iteration, we first determine (up to a constant factor) the correlation between all variables and the current residual vector. All variables whose absolute correlation with the residual vector equals the maximal achievable absolute correlation are chosen to be in the model and we calculate the least squares regression response, say $\bar{\mathbf{y}}_{k+1}$, using these variables. Then we move from our current estimated response $\hat{\boldsymbol{\mu}}_k$ toward $\bar{\mathbf{y}}_{k+1}$ until a new variable would enter the model.

Using this description of the LARS algorithm, it seems obvious how to modify the algorithm such that it respects the marginality principle. Assume that for each column i of the design matrix we set $d_{ij} = 1$ if column j should be in the model whenever column i is in the model and zero otherwise; here $j \neq i$ takes values in $\{1, \ldots, m\}$, where m is the number of columns of the design matrix. For example, abusing this notation slightly, for model (1) we might set $d_{1:1,1} = 1$ and all other $d_{1:1,j} = 0$; or $d_{1:2,1} = 1$, $d_{1:2,2} = 1$ and all other $d_{1:2,j}$ equal to zero.

Having defined such a dependency structure between the columns of the design matrix, the obvious modification of the LARS algorithm is that when adding, say, column i to the selected columns one also adds all those columns for which $d_{ij} = 1$. This modification is described in Algorithm 2.

ALGORITHM 2 (The modified LARS algorithm).

1. Set $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and $k = 0$.
2. **repeat**
3. Calculate $\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\boldsymbol{\mu}}_k)$ and set $\hat{C} = \max_j\{|\hat{c}_j|\}$.
4. Let $\mathcal{A}_0 = \{j : |\hat{c}_j| = \hat{C}\}$, $\mathcal{A}_1 = \{j : d_{ij} \neq 0, i \in \mathcal{A}_0\}$ and $\mathcal{A} = \mathcal{A}_0 \cup \mathcal{A}_1$.
5. Set $X_{\mathcal{A}} = (\cdots \mathbf{x}_j \cdots)_{j\in\mathcal{A}}$ for calculating $\bar{\mathbf{y}}_{k+1} = (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} X'_{\mathcal{A}} \mathbf{y}$ and $\mathbf{a} = X'_{\mathcal{A}}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k)$.
6. Set

$$\hat{\boldsymbol{\mu}}_{k+1} = \hat{\boldsymbol{\mu}}_k + \hat{\gamma}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k),$$

where, if $\mathcal{A}^c \neq \varnothing$,

$$\hat{\gamma} = \min_{j\in\mathcal{A}^c}{}^{+}\left\{\frac{\hat{C} - \hat{c}_j}{\hat{C} - a_j}, \frac{\hat{C} + \hat{c}_j}{\hat{C} + a_j}\right\},$$

otherwise set $\hat{\gamma} = 1$.

7. $k \leftarrow k + 1$.
8. **until** $\mathcal{A}^c = \varnothing$.

Note that compared with the original Algorithm 1 only the fourth line changes. Furthermore, for all $i \in \mathcal{A}$ it is obvious that for $0 \leq \gamma \leq 1$ we have

$$|\hat{c}_i(\gamma)| = (1 - \gamma)|\hat{c}_i|, \tag{2}$$

where $\hat{\mathbf{c}}(\gamma) = X'(\mathbf{y} - \hat{\boldsymbol{\mu}}(\gamma))$ and $\hat{\boldsymbol{\mu}}(\gamma) = \hat{\boldsymbol{\mu}}_k + \gamma(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k)$.

Note that, by definition, the value of $|\hat{c}_j|$ is the same for all $j \in \mathcal{A}_0$. Hence, the functions (2) for those variables are identical, namely $(1-\gamma)\hat{C}$, and for all $j \in \mathcal{A}_1$ the corresponding functions $|\hat{c}_j(\gamma)|$ will intersect $(1-\gamma)\hat{C}$ at $\gamma = 1$. This explains why in line 6 we only have to check for the first intersection between $(1 - \gamma)\hat{C}$ and $|\hat{c}_j(\gamma)|$ for $j \in \mathcal{A}^c$.

It also follows from (2) that, for all $j \in \mathcal{A}_0$, we have

$$\mathbf{x}'_j(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k) = \operatorname{sign}(\hat{c}_j)\hat{C}.$$

Thus, for those variables that are in $\mathcal{A}_0$ we move in line 6 of the modified algorithm in a direction that has a similar geometric interpretation as the direction along which we move in the original LARS algorithm. Namely that for each $j \in \mathcal{A}_0$ the angle between the direction in which we move and $\operatorname{sign}(\hat{c}_j)\mathbf{x}_j$ is the same and this angle is less than 90°.

Figure 5 shows the result of the modified LARS analysis for the same data used above. Putting variables that enter the model simultaneously into brackets, the order in which the variables enter the model is $(X_{2:5}, X_2, X_5)$, $(X_{3:4}, X_3, X_4)$, $X_{2:5}$, $X_{2:3}$, $(X_{1:1}, X_1)$, ... That is, the modified LARS algorithm selects in this case in five steps a model with 10 terms, 6 of which are the terms that are indeed in model (1).

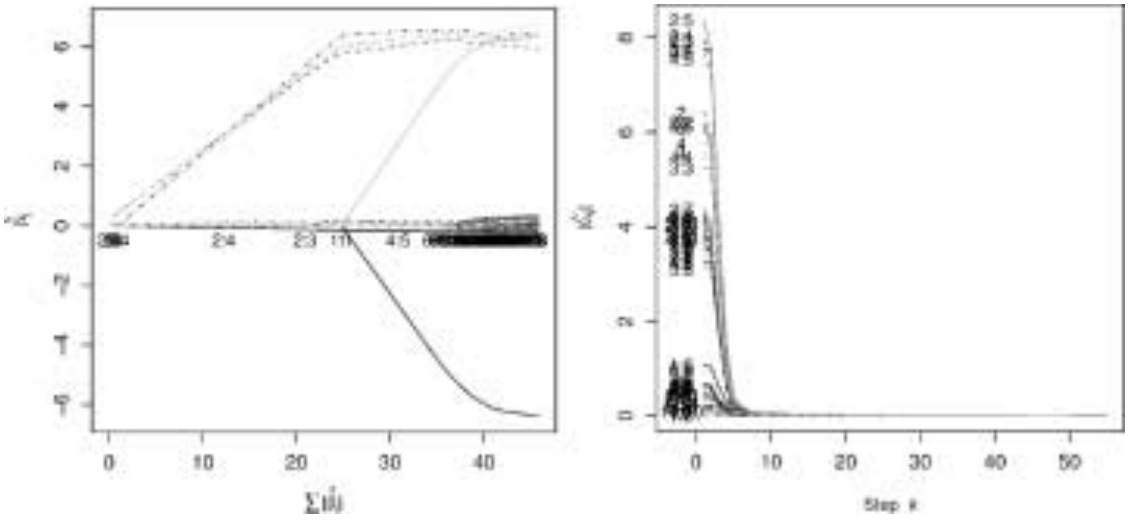

FIG. 5. *Modified LARS analysis of simulated data with main terms and interaction terms*: (left) *estimates of regression coefficients* $\hat{\beta}_j$, $j = 1, \ldots, 65$, *plotted versus* $\sum |\hat{\beta}_j|$; (right) *absolute current correlations as functions of* $k = \#\mathcal{A}^c$.

Using the same 1000 samples of size $n = 500$ as above and performing a modified LARS analysis on them using a design matrix with all main and interaction terms also shows markedly improved results. Again, for each replication the step at which a variable was selected into the model was recorded and Figure 6 shows for each variable histograms for these data. To avoid cluttering, the histograms in Figure 6 were truncated to $[1, 20]$; the complete histograms are shown on the right in Figure 7.

From Figure 6 it can be seen that now the variables X_2, X_3, X_4 and X_5 are all selected within the first three iterations of the modified LARS algorithm. Also $X_{1:1}$ and X_1 are picked up consistently and early. Compared with Figure 4 there are marked differences in the distribution of when the variable is selected for the interaction terms $X_{i:j}$, $i, j \in \{2, 3, 4, 5\}$, $i \leq j$, and the main terms X_i, $i = 6, \ldots, 10$. The latter can be explained by the fact that the algorithm now enforces the marginality principle. Thus, it seems that this modification does improve the performance of the LARS algorithm for model (1). Hopefully it would do so also for other models.

In conclusion, I offer two further remarks and a question. First, note that the modified LARS algorithm may also be used to incorporate factor variables with more than two levels. In such a situation, I would suggest that indicator variables for *all* levels are included in the initial design matrix; but this would be done mainly to easily calculate all the correlations. The dependencies d_{ij} would be set up such that if one indicator variable is selected, then all enter the model. However, to avoid redundancies one would only put all but one of these columns into the matrix $X_{\mathcal{A}}$. This would also avoid that $X_{\mathcal{A}}$ would eventually become singular if more than one explanatory variable is a factor variable.

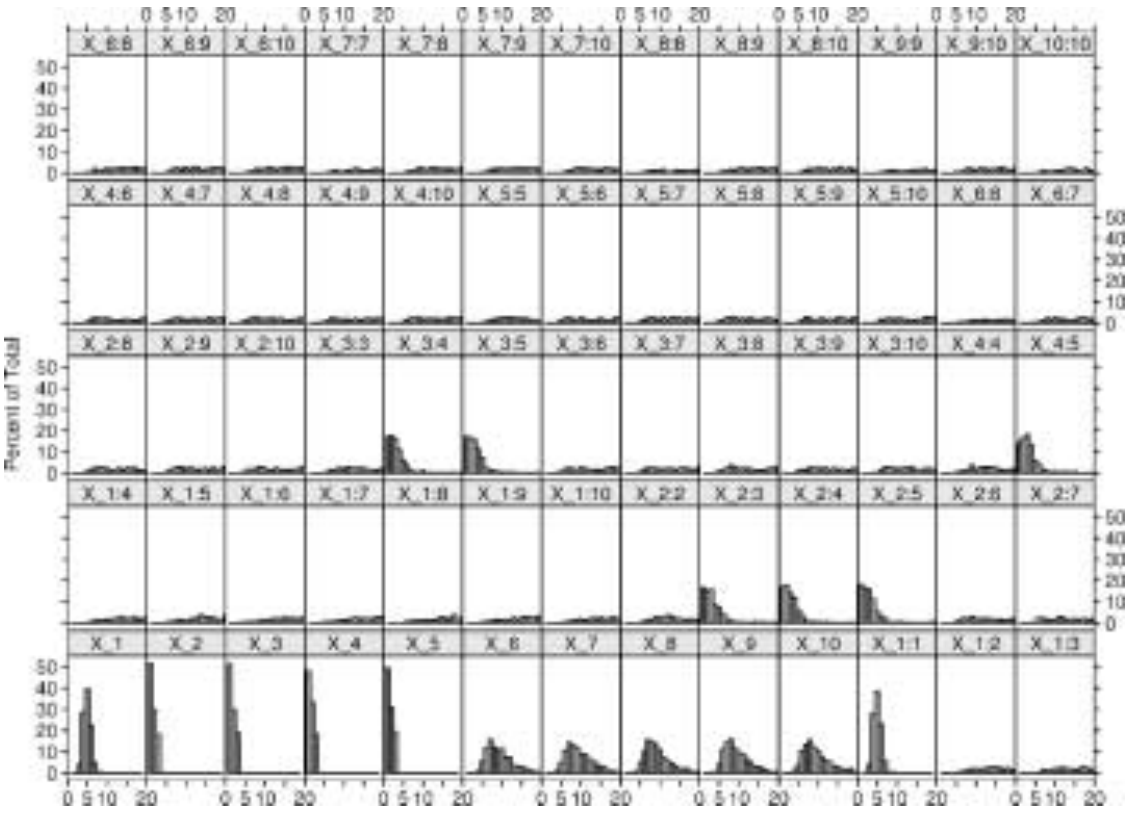

FIG. 6. *Percentage histogram of step at which each variable is selected based on* 1000 *replications*: *results shown for variables selected in the first* 20 *steps of a modified LARS analysis using main and interaction terms*.

Second, given the insight between the LARS algorithm and the Lasso algorithm described by EHJT, namely the sign constraint (3.1), it now seems also possible to modify the Lasso algorithm to incorporate the marginality principle by incorporating the sign constraint into Algorithm 2. However, whenever a variable

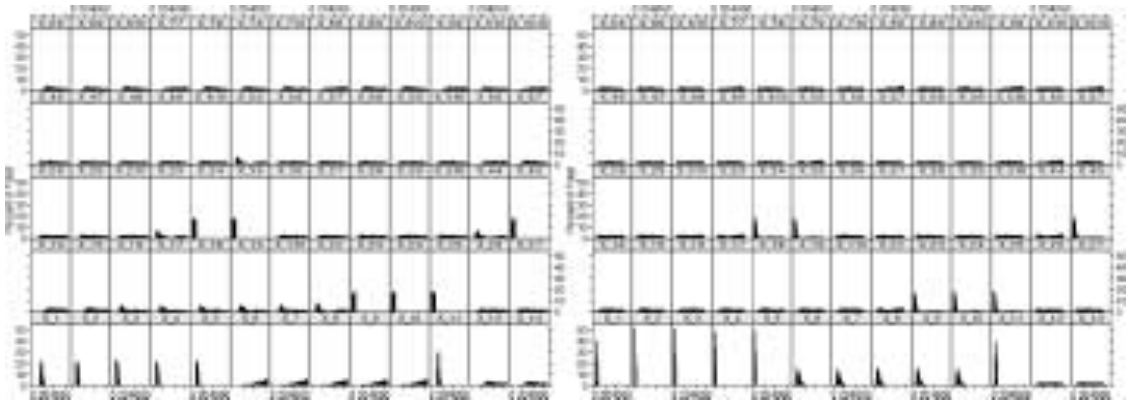

FIG. 7. *Percentage histogram of step at which each variable is selected based on* 1000 *replications*: (left) *LARS analysis*; (right) *modified LARS analysis*.

would be dropped from the set $\mathcal{A}_0$ due to violating the sign constraint there might also be variables dropped from $\mathcal{A}_1$. For the latter variables these might introduce discontinuities in the traces of the corresponding parameter estimates, a feature that does not seem to be desirable. Perhaps a better modification of the Lasso algorithm that incorporates the marginality principle can still be found?

Finally, the behavior of the LARS algorithm for model (1) when all main terms and interaction terms are used surprised me a bit. This behavior seems to raise a fundamental question, namely, although we try to build a linear model and, as we teach our students, correlation "measures the direction and strength of the linear relationship between two quantitative variables" [Moore and McCabe (1999)], one has to wonder whether selecting variables using correlation as a criterion is a sound principle? Or should we modify the algorithms to use another criterion?

REFERENCES

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MOORE, D. S. and MCCABE, G. P. (1999). *Introduction to the Practice of Statistics*, 3rd ed. Freeman, New York.

NELDER, J. A. (1977). A reformulation of linear models (with discussion). *J. Roy. Statist. Soc. Ser. A* **140** 48–76.

NELDER, J. A. (1994). The statistics of linear models: Back to basics. *Statist. Comput.* **4** 221–234.

SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF WESTERN AUSTRALIA
35 STIRLING HIGHWAY
CRAWLEY WA 6009
AUSTRALIA
E-MAIL: berwin@maths.uwa.edu.au

DISCUSSION

BY SANFORD WEISBERG[1]

University of Minnesota

Most of this article concerns the uses of LARS and the two related methods in the age-old, "somewhat notorious," problem of "[a]utomatic model-building algorithms..." for linear regression. In the following, I will confine my comments to this notorious problem and to the use of LARS and its relatives to solve it.

[1]Supported by NSF Grant DMS-01-03983.

1. The implicit assumption. Suppose the response is y, and we collect the m predictors into a vector x, the realized data into an $n \times m$ matrix X and the response is the n-vector Y. If P is the projection onto the column space of $(1, X)$, then LARS, like ordinary least squares (OLS), assumes that, for the purposes of model building, Y can be replaced by $\hat{Y} = PY$ without loss of information. In large samples, this is equivalent to the assumption that the conditional distributions $F(y|x)$ can be written as

$$F(y|x) = F(y|x'\beta) \tag{1}$$

for some unknown vector β. Efron, Hastie, Johnstone and Tibshirani use this assumption in the definition of the LARS algorithm and in estimating residual variance by $\hat{\sigma}^2 = \|(I - P)Y\|^2/(n - m - 1)$. For LARS to be reasonable, we need to have some assurance that this particular assumption holds or that it is relatively benign. If this assumption is not benign, then LARS like OLS is unlikely to produce useful results.

A more general alternative to (1) is

$$F(y|x) = F(y|x'B), \tag{2}$$

where B is an $m \times d$ rank d matrix. The smallest value of d for which (2) holds is called the structural dimension of the regression problem [Cook (1998)]. An obvious precursor to fitting linear regression is deciding on the structural dimension, not proceeding as if $d = 1$. For the diabetes data used in the article, the R package `dr` [Weisberg (2002)] can be used to estimate d using any of several methods, including sliced inverse regression [Li (1991)]. For these data, fitting these methods suggests that (1) is appropriate.

Expanding x to include functionally related terms is another way to provide a large enough model that (1) holds. Efron, Hastie, Johnstone and Tibshirani illustrate this in the diabetes example in which they expand the 10 predictors to 65 including all quadratics and interactions. This alternative does not include (2) as a special case, as it includes a few models of various dimensions, and this seems to be much more complex than (2).

Another consequence of assumption (1) is the reliance of LARS, and of OLS, on correlations. The correlation measures the degree of linear association between two variables particularly for normally distributed or at least elliptically contoured variables. This requires not only linearity in the conditional distributions of y given subsets of the predictors, but also linearity in the conditional distributions of $a'x$ given $b'x$ for all a and b [see, e.g., Cook and Weisberg (1999a)]. When the variables are not linearly related, bizarre results can follow; see Cook and Weisberg (1999b) for examples. Any method that replaces Y by PY cannot be sensitive to nonlinearity in the conditional distributions.

Methods based on PY alone may be strongly influenced by outliers and high leverage cases. As a simple example of this, consider the formula for C_p given by

Efron, Hastie, Johnstone and Tibshirani:

$$C_p(\hat{\mu}) = \frac{\|Y - \hat{\mu}\|^2}{\sigma^2} - n + 2\sum_{i=1}^{n} \frac{\operatorname{cov}(\hat{\mu}_i, y_i)}{\sigma^2}. \tag{3}$$

Estimating σ^2 by $\hat{\sigma}^2 = \|(I - P)Y\|^2/(n - m - 1)$, and adapting Weisberg (1981), (3) can be rewritten as a sum of n terms, the ith term given by

$$C_{pi}(\hat{\mu}) = \frac{(\hat{y}_i - \hat{\mu}_i)^2}{\hat{\sigma}^2} + \frac{\operatorname{cov}(\hat{\mu}_i, y_i)}{\hat{\sigma}^2} - \left(\frac{h_i - \operatorname{cov}(\hat{\mu}_i, y_i)}{\hat{\sigma}^2}\right),$$

where $\hat{y}_i$ is the ith element of PY and h_i is the ith leverage, a diagonal element of P. From the simulation reported in the article, a reasonable approximation to the covariance term is $\hat{\sigma}^2 u_i$, where u_i is the ith diagonal of the projection matrix on the columns of $(1, X)$ with nonzero coefficients at the current step of the algorithm. We then get

$$C_{pi}(\hat{\mu}) = \frac{(\hat{y}_i - \hat{\mu}_i)^2}{\hat{\sigma}^2} + u_i - (h_i - u_i),$$

which is the same as the formula given in Weisberg (1981) for OLS except that $\hat{\mu}_i$ is computed from LARS rather than from a projection. The point here is that the value of $C_{pi}(\hat{\mu})$ depends on the agreement between $\hat{\mu}_i$ and $\hat{y}_i$, on the leverage in the subset model and on the difference in the leverage between the full and subset models. Neither of these latter two terms has much to do with the problem of interest, which is the study of the conditional distribution of y given x, but they are determined by the predictors only.

2. Selecting variables. Suppose that we can write $x = (x_a, x_u)$ for some decomposition of x into two pieces, in which x_a represents the "active" predictors and x_u the unimportant or inactive predictors. The variable selection problem is to find the smallest possible x_a so that

$$F(y|x) = F(y|x_a) \tag{4}$$

thereby identifying the active predictors. Standard subset selection methods attack this problem by first assuming that (1) holds, and then fitting models with different choices for x_a, possibly all possible choices or a particular subset of them, and then using some sort of inferential method or criterion to decide if (4) holds, or more precisely if

$$F(y|x) = F(y|\gamma' x_a)$$

holds for some γ. Efron, Hastie, Johnstone and Tibshirani criticize the standard methods as being too greedy: once we put a variable, say, $x^* \in x_a$, then any predictor that is highly correlated with x^* will never be included. LARS, on the other hand, permits highly correlated predictors to be used.

LARS or any other methods based on correlations cannot be much better at finding x_a than are the standard methods. As a simple example of what can go wrong, I modified the diabetes data in the article by adding nine new predictors, created by multiplying each of the original predictors excluding the sex indicator by 2.2, and then rounding to the nearest integer. These rounded predictors are clearly less relevant than are the original predictors, since they are the original predictors with noise added by the rounding. We would hope that none of these would be among the active predictors.

Using the S-PLUS functions kindly provided by Efron, Hastie, Johnstone and Tibshirani, the LARS procedure applied to the original data selects a seven-predictor model, including, in order, BMI, S5, BP, S3, SEX, S6 and S1. LARS applied to the data augmented with the nine inferior predictors selects an eight-predictor model, including, in order, BMI, S5, rBP, rS3, BP, SEX, S6 and S1, where the prefix "r" indicates a rounded variable rather than the variable itself. LARS not only selects two of the inferior rounded variables, but it selects both BP and its rounded version rBP, effectively claiming that the rounding is informative with respect to the response.

Inclusion and exclusion of elements in x_a depends on the marginal distribution of x as much as on the conditional distribution of $y|x$. For example, suppose that the diabetes data were a random sample from a population. The variables S3 and S4 have a large sample correlation, and LARS selects one of them, S3, as an active variable. Suppose a therapy were available that could modify S4 without changing the value of S3, so in the future S3 and S4 would be nearly uncorrelated. Although this would arguably not change the distribution of $y|x$, it would certainly change the marginal distribution of x, and this could easily change the set of active predictors selected by LARS or any other method that starts with correlations.

A characteristic that LARS shares with the usual methodology for subset selection is that the results are invariant under rescaling of any individual predictor, but not invariant under reparameterization of functionally related predictors. In the article, the authors create more predictors by first rescaling predictors to have zero mean and common standard deviation, and then adding all possible cross-products and quadratics to the existing predictors. For this expanded definition of the predictors, LARS selects a 15 variable model, including 6 main-effects, 6 two-factor interactions and 3 quadratics. If we add quadratics and interactions first and then rescale, LARS picks an 8 variable model with 2 main-effects, 6 two-factor interactions, and only 3 variables in common with the model selected by scaling first. If we define the quadratics and interactions to be orthogonal to the main-effects, we again get a different result. The lack of invariance with regard to definition of functionally related predictors can be partly solved by considering the functionally related variables simultaneously rather than sequentially. This seems to be self-defeating, at least for the purpose of subset selection.

3. Summary. Long-standing problems often gain notoriety because solution of them is of wide interest and at the same time illusive. Automatic model building in linear regression is one such problem. My main point is that neither LARS nor, as near as I can tell, any other *automatic* method has any hope of solving this problem because automatic procedures by their very nature do not consider the context of the problem at hand. I cannot see any solution to this problem that is divorced from context. Most of the ideas in this discussion are not new, but I think they bear repeating when trying to understand LARS methodology in the context of linear regression. Similar comments can be found in Efron (2001) and elsewhere.

REFERENCES

COOK, R. D. (1998). *Regression Graphics*. Wiley, New York.

COOK, R. D. and WEISBERG, S. (1999a). *Applied Regression Including Computing and Graphics*. Wiley, New York.

COOK, R. D. and WEISBERG, S. (1999b). Graphs in statistical analysis: Is the medium the message? *Amer. Statist.* **53** 29–37.

EFRON, B. (2001). Discussion of "Statistical modeling: The two cultures," by L. Breiman. *Statist. Sci.* **16** 218–219.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

WEISBERG, S. (1981). A statistic for allocating C_p to individual cases. *Technometrics* **23** 27–31.

WEISBERG, S. (2002). Dimension reduction regression in R. *J. Statistical Software* **7**. (On-line journal available at www.jstatsoft.org. The software is available from cran.r-project.org.)

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
1994 BUFORD AVENUE
ST. PAUL, MINNESOTA 55108
USA
E-MAIL: sandy@stat.umn.edu

REJOINDER

BY BRADLEY EFRON, TREVOR HASTIE, IAIN JOHNSTONE AND ROBERT TIBSHIRANI

The original goal of this project was to explain the striking similarities between models produced by the Lasso and Forward Stagewise algorithms, as exemplified by Figure 1. LARS, the Least Angle Regression algorithm, provided the explanation and proved attractive in its own right, its simple structure permitting theoretical insight into all three methods. In what follows "LAR" will refer to the basic, unmodified form of Least Angle Regression developed in Section 2, while "LARS" is the more general version giving LAR, Lasso, Forward

Stagewise and other variants as in Section 3.4. Here is a summary of the principal properties developed in the paper:

1. LAR builds a regression model in piecewise linear forward steps, accruing explanatory variables one at a time; each step is taken along the equiangular direction between the current set of explanators. The step size is less greedy than classical forward stepwise regression, smoothly blending in new variables rather than adding them discontinuously.
2. Simple modifications of the LAR procedure produce all Lasso and Forward Stagewise solutions, allowing their efficient computation and showing that these methods also follow piecewise linear equiangular paths. The Forward Stagewise connection suggests that LARS-type methods may also be useful in more general "boosting" applications.
3. The LARS algorithm is computationally efficient; calculating the full set of LARS models requires the same order of computation as ordinary least squares.
4. A k-step LAR fit uses approximately k degrees of freedom, in the sense of added prediction error (4.5). This approximation is exact in the case of orthogonal predictors and is generally quite accurate. It permits C_p-type stopping rules that do not require auxiliary bootstrap or cross-validation computations.
5. For orthogonal designs, LARS models amount to a succession of soft thresholding estimates, (4.17).

All of this is rather technical in nature, showing how one might efficiently carry out a program of automatic model-building ("machine learning"). Such programs seem increasingly necessary in a scientific world awash in huge data sets having hundreds or even thousands of available explanatory variables.

What this paper, strikingly, does not do is justify any of the three algorithms as providing *good* estimators in some decision-theoretic sense. A few hints appear, as in the simulation study of Section 3.3, but mainly we are relying on recent literature to say that LARS methods are at least reasonable algorithms and that it is worthwhile understanding their properties. Model selection, the great underdeveloped region of classical statistics, deserves careful theoretical examination but that does not happen here. We are not as pessimistic as Sandy Weisberg about the potential of automatic model selection, but agree that it requires critical examination as well as (over) enthusiastic algorithm building.

The LARS algorithm in any of its forms produces a one-dimensional path of prediction vectors going from the origin to the full least-squares solution. (Figures 1 and 3 display the paths for the diabetes data.) In the LAR case we can label the predictors $\widehat{\boldsymbol{\mu}}(k)$, where k is identified with both the number of steps and the degrees of freedom. What the figures do not show is when to stop the model-building process and report $\widehat{\boldsymbol{\mu}}$ back to the investigator. The examples in our paper rather casually used stopping rules based on minimization of the C_p error prediction formula.

Robert Stine and Hemant Ishwaran raise some reasonable doubts about C_p minimization as an effective stopping rule. For any one value of k, C_p is an unbiased estimator of prediction error, so in a crude sense C_p minimization is trying to be an unbiased estimator of the optimal stopping point k_{opt}. As such it is bound to overestimate k_{opt} in a large percentage of the cases, perhaps near 100% if k_{opt} is near zero.

We can try to improve C_p by increasing the *df* multiplier "2" in (4.5). Suppose we change 2 to some value *mult*. In standard normal-theory model building situations, for instance choosing between linear, quadratic, cubic, ... regression models, the *mult* rule will prefer model $k+1$ to model k if the relevant t-statistic exceeds $\sqrt{mult}$ in absolute value (here we are assuming σ^2 known); $mult = 2$ amounts to using a rejection rule with $\alpha = 16\%$. Stine's interesting S_p method chooses *mult* closer to 4, $\alpha = 5\%$.

This works fine for Stine's examples, where k_{opt} is indeed close to zero. We tried it on the simulation example of Section 3.3. Increasing *mult* from 2 to 4 decreased the average selected step size from 31 to 15.5, but with a small increase in actual squared estimation error. Perhaps this can be taken as support for Ishwaran's point that since LARS estimates have a broad plateau of good behavior, one can often get by with much smaller models than suggested by C_p minimization. Of course no one example is conclusive in an area as multifaceted as model selection, and perhaps no 50 examples either. A more powerful theory of model selection is sorely needed, but until it comes along we will have to make do with simulations, examples and bits and pieces of theory of the type presented here.

Bayesian analysis of prediction problems tends to favor *much* bigger choices of *mult*. In particular the Bayesian information criterion (BIC) uses *mult* = log(sample size). This choice has favorable consistency properties, selecting the correct model with probability 1 as the sample size goes to infinity. However, it can easily select too-small models in nonasymptotic situations.

Jean-Michel Loubes and Pascal Massart provide two interpretations using penalized estimation criteria in the orthogonal regression setting. The first uses the link between soft thresholding and ℓ_1 penalties to motivate entropy methods for asymptotic analysis. The second is a striking perspective on the use of C_p with LARS. Their analysis suggests that our usual intuition about C_p, derived from selecting among projection estimates of different ranks, may be misleading in studying a nonlinear method like LARS that combines thresholding and shrinkage. They rewrite the LARS-C_p expression (4.5) in terms of a penalized criterion for selecting among orthogonal projections. Viewed in this unusual way (for the estimator to be used is *not* a projection!), they argue that *mult* in fact behaves like $\log(n/k)$ rather than 2 (in the case of a k-dimensional projection). It is indeed remarkable that this same model-dependent value of *mult*, which has emerged in several recent studies [Foster and Stine (1997), George and Foster (2000), Abramovich, Benjamini, Donoho and Johnstone (2000) and Birgé and Massart (2001)], should also appear as relevant for the analysis of LARS. We look

forward to the further extension of the Birgé–Massart approach to handling these nondeterministic penalties.

Cross-validation is a nearly unbiased estimator of prediction error and as such will perform similarly to C_p (with $mult = 2$). The differences between the two methods concern generality, efficiency and computational ease. Cross-validation, and nonparametric bootstrap methods such as the 632+ rule, can be applied to almost any prediction problem. C_p is more specialized, but when it does apply it gives more efficient estimates of prediction error [Efron (2004)] at almost no computational cost. It applies here to LAR, at least when $m < n$, as in David Madigan and Greg Ridgeway's example.

We agree with Madigan and Ridgeway that our new LARS algorithm may provide a boost for the Lasso, making it more useful and attractive for data analysts. Their suggested extension of LARS to generalized linear models is interesting. In logistic regression, the L_1-constrained solution is not piecewise linear and hence the pathwise optimization is more difficult. Madigan and Ridgeway also compare LAR and Lasso to least squares boosting for prediction accuracy on three real examples, with no one method prevailing.

Saharon Rosset and Ji Zhu characterize a class of problems for which the coefficient paths, like those in this paper, are piecewise linear. This is a useful advance, as demonstrated with their robust version of the Lasso, and the ℓ_1-regularized Support Vector Machine. The former addresses some of the robustness concerns of Weisberg. They also report on their work that strengthens the connections between ε-boosting and ℓ_1-regularized function fitting.

Berwin Turlach's example with uniform predictors surprised us as well. It turns out that 10-fold cross-validation selects the model with $|\beta_1| \approx 45$ in his Figure 3 (left panel), and by then the correct variables are active and the interactions have died down. However, the same problem with 10 times the noise variance does not recover in a similar way. For this example, if the X_j are uniform on $[-\frac{1}{2}, \frac{1}{2}]$ rather than $[0, 1]$, the problem goes away, strongly suggesting that proper centering of predictors (in this case the interactions, since the original variables are automatically centered by the algorithm) is important for LARS.

Turlach also suggests an interesting proposal for enforcing marginality, the hierarchical relationship between the main effects and interactions. In his notation, marginality says that $\beta_{i:j}$ can be nonzero only if β_i and β_j are nonzero. An alternative approach, more in the "continuous spirit" of the Lasso, would be to include constraints

$$|\beta_{i:j}| \le \min\{|\beta_i|, |\beta_j|\}.$$

This implies marginality but is stronger. These constraints are linear and, according to Rosset and Zhu above, a LARS-type algorithm should be available for its estimation. Leblanc and Tibshirani (1998) used constraints like these for shrinking classification and regression trees.

As Turlach suggests, there are various ways to restate the LAR algorithm, including the following nonalgebraic purely statistical statement in terms of repeated fitting of the residual vector $\mathbf{r}$:

1. Start with $\mathbf{r} = \mathbf{y}$ and $\widehat{\beta}_j = 0 \ \forall \, j$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Increase $\widehat{\beta}_j$ in the direction of the sign of $\operatorname{corr}(\mathbf{r}, \mathbf{x}_j)$ until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Update $\mathbf{r}$, and move $(\widehat{\beta}_j, \widehat{\beta}_k)$ in the joint least squares direction for the regression of $\mathbf{r}$ on $(\mathbf{x}_j, \mathbf{x}_k)$ until some other competitor $\mathbf{x}_\ell$ has as much correlation with the current residual.
5. Continue in this way until all predictors have been entered. Stop when $\operatorname{corr}(\mathbf{r}, \mathbf{x}_j) = 0 \ \forall \, j$, that is, the OLS solution.

Traditional forward stagewise would have completed the least-squares step at each stage; here it would go only a fraction of the way, until the next competitor joins in.

Keith Knight asks whether Forward Stagewise and LAR have implicit criteria that they are optimizing. In unpublished work with Trevor Hastie, Jonathan Taylor and Guenther Walther, we have made progress on that question. It can be shown that the Forward Stagewise procedure does a sequential minimization of the residual sum of squares, subject to

$$\sum_j \left| \int_0^t \beta_j'(s) \, ds \right| \le t.$$

This quantity is the total L_1 arc-length of the coefficient curve $\beta(t)$. If each component $\beta_j(t)$ is monotone nondecreasing or nonincreasing, then L_1 arc-length equals the L_1-norm $\sum_j |\beta_j|$. Otherwise, they are different and L_1 arc-length discourages sign changes in the derivative. That is why the Forward Stagewise solutions tend to have long flat plateaus. We are less sure of the criterion for LAR, but currently believe that it uses a constraint of the form $\sum_j |\int_0^k \beta_j(s) \, ds| \le A$.

Sandy Weisberg, as a ranking expert on the careful analysis of regression problems, has legitimate grounds for distrusting automatic methods. Only foolhardy statisticians dare to ignore a problem's context. (For instance it helps to know that diabetes progression behaves differently after menopause, implying strong age–sex interactions.) Nevertheless even for a "small" problem like the diabetes investigation there is a limit to how much context the investigator can provide. After that one is drawn to the use of automatic methods, even if the "automatic" part is not encapsulated in a single computer package.

In actual practice, or at least in good actual practice, there is a cycle of activity between the investigator, the statistician and the computer. For a multivariable prediction problem like the diabetes example, LARS-type programs are a good first step toward a solution, but hopefully not the last step. The statistician examines the output critically, as did several of our commentators, discussing the results with

the investigator, who may at this point suggest adding or removing explanatory variables, and so on, and so on.

Fully automatic regression algorithms have one notable advantage: they permit an honest evaluation of estimation error. For instance the C_p-selected LAR quadratic model estimates that a patient one standard deviation above average on BMI has an increased response expectation of 23.8 points. The bootstrap analysis (3.16) provided a standard error of 3.48 for this estimate. Bootstrapping, jackknifing and cross-validation require us to repeat the original estimation procedure for different data sets, which is easier to do if you know what the original procedure actually was.

Our thanks go to the discussants for their thoughtful remarks, and to the Editors for the formidable job of organizing this discussion.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19, Dept. Statistics, Stanford Univ.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268.

EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* To appear.

FOSTER, D. and STINE, R. (1997). An information theoretic comparison of model selection criteria. Technical report, Dept. Statistics, Univ. Pennsylvania.

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

LEBLANC, M. and TIBSHIRANI, R. (1998). Monotone shrinkage of trees. *J. Comput. Graph. Statist.* **7** 417–433.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: brad@stat.stanford.edu