Least Angle Regression

Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani
Statistics Department, Stanford University

January 9, 2003

Abstract

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 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 Cp 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 linear 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 Section 8.5 of Weisberg (1980): 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.

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	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

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.

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

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for} \quad j = 1, 2, \ldots m. \tag{1.1}$$

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 $\widehat{\boldsymbol{\mu}}$,

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

with total squared error

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

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

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

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

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

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 & Turlach (2000*a*).

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 & Friedman (2001).

The Lasso also has a parsimony property: for any given constraint value t, only a subset of the covariates have non-zero 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.

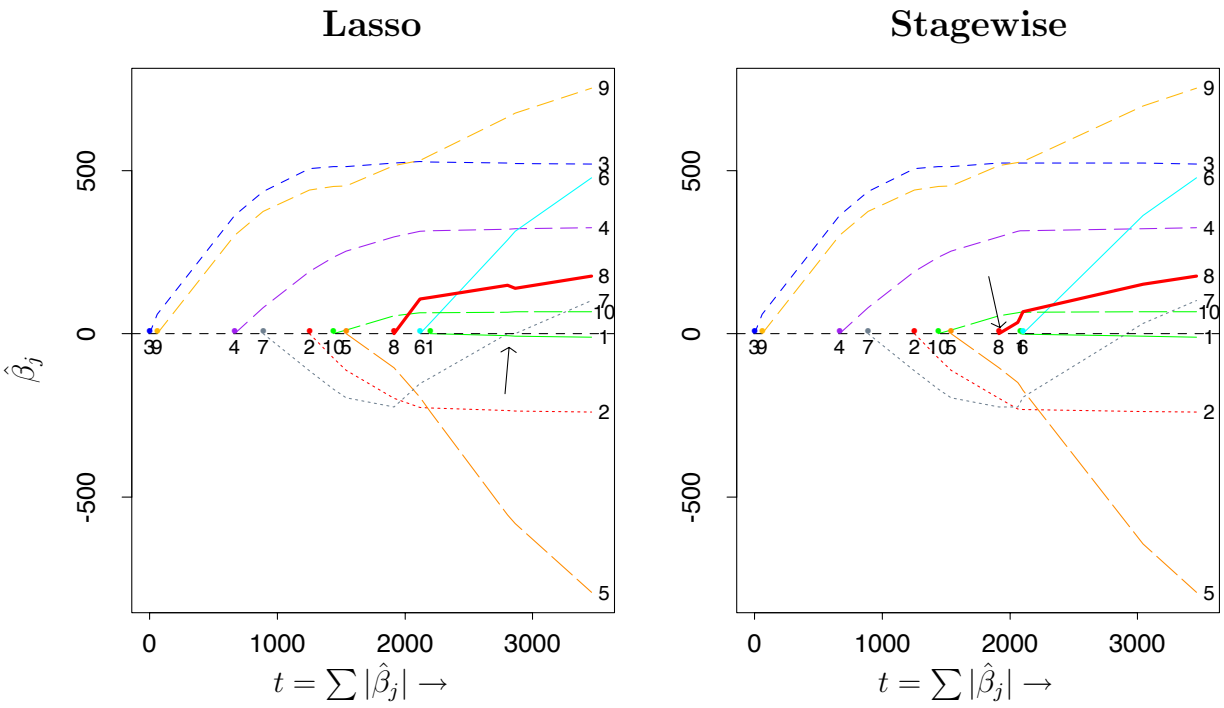

Figure 1. Estimates of regression coefficients $\widehat{\beta}_j, j = 1, 2, \ldots, 10$, for the diabetes study. *Left Panel* Lasso estimates, as a function of $t = \Sigma_j |\widehat{\beta}_j|$. The covariates enter the regression equation sequentially as t increases, in order $j = 3, 9, 4, 7, \ldots 1$. *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 track of covariate 8.

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,

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

with ϵ some small constant. "Small" is important here: the "big" choice $\epsilon = |\widehat{c}_{\widehat{j}}|$ 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 Chapter 10 and Algorithm 10.4 of Hastie et al. (2001).

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). 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 analyses 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}, i.e. along the "least angle direction", until a fourth variable enters, etc.

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 non-zero. 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 $\widehat{\boldsymbol{\mu}}_o = \mathbf{0}$ (remembering that the response has had its mean subtracted off, as in (1.1)). Figure 2 has $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_o$ making a smaller angle with $\mathbf{x}_1$ than $\mathbf{x}_2$, i.e. $c_1(\widehat{\boldsymbol{\mu}}_o) > c_2(\widehat{\boldsymbol{\mu}}_o)$. LARS then augments $\widehat{\boldsymbol{\mu}}_o$ in the direction of $\mathbf{x}_1$, to

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

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 $c_1(\widehat{\boldsymbol{\mu}}_1) = c_2(\widehat{\boldsymbol{\mu}}_1)$.

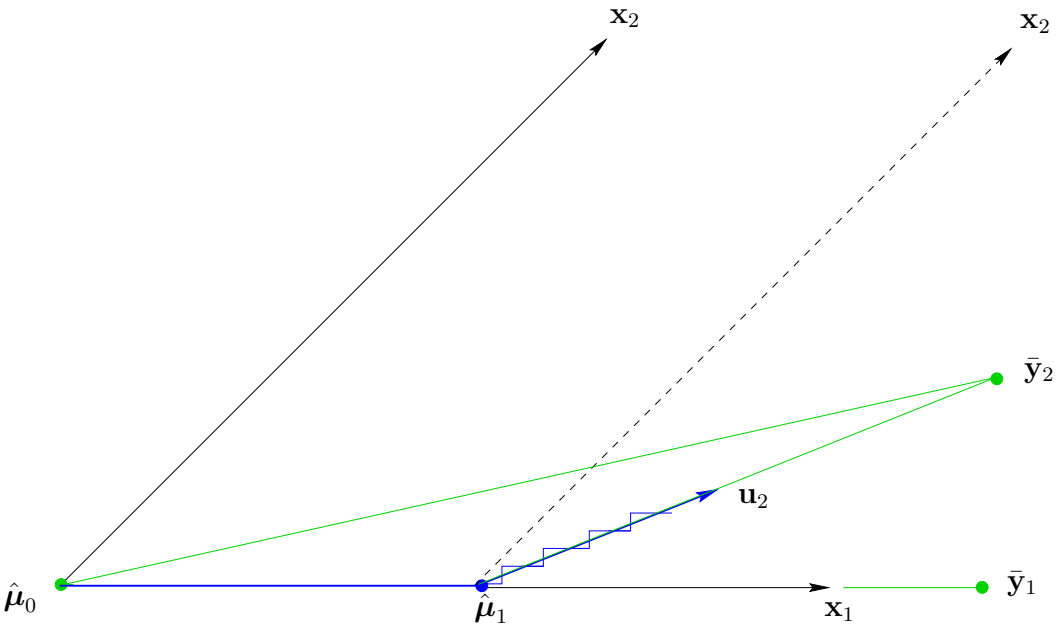

Figure 2. The LARS algorithm in the case of $m = 2$ covariates; $\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$. Beginning at $\widehat{\boldsymbol{\mu}}_o = \mathbf{0}$, the residual vector $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_o$ has greater correlation with $\mathbf{x}_1$ than $\mathbf{x}_2$; the next LARS estimate is $\widehat{\boldsymbol{\mu}}_1 = \widehat{\boldsymbol{\mu}}_o + \widehat{\gamma}_1 \mathbf{x}_1$, 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 $\epsilon \to 0$, 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 2 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

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

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

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

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

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

We can now fully describe the LARS algorithm. As with the Stagewise procedure we begin at $\widehat{\boldsymbol{\mu}}_o = \mathbf{0}$ 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

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

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)–(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 joining the active set $\mathcal{A}$ in the same order as for the Lasso: $3, 9, 4, 7, \ldots, 1$. 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 $\widehat{C}_k$,

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

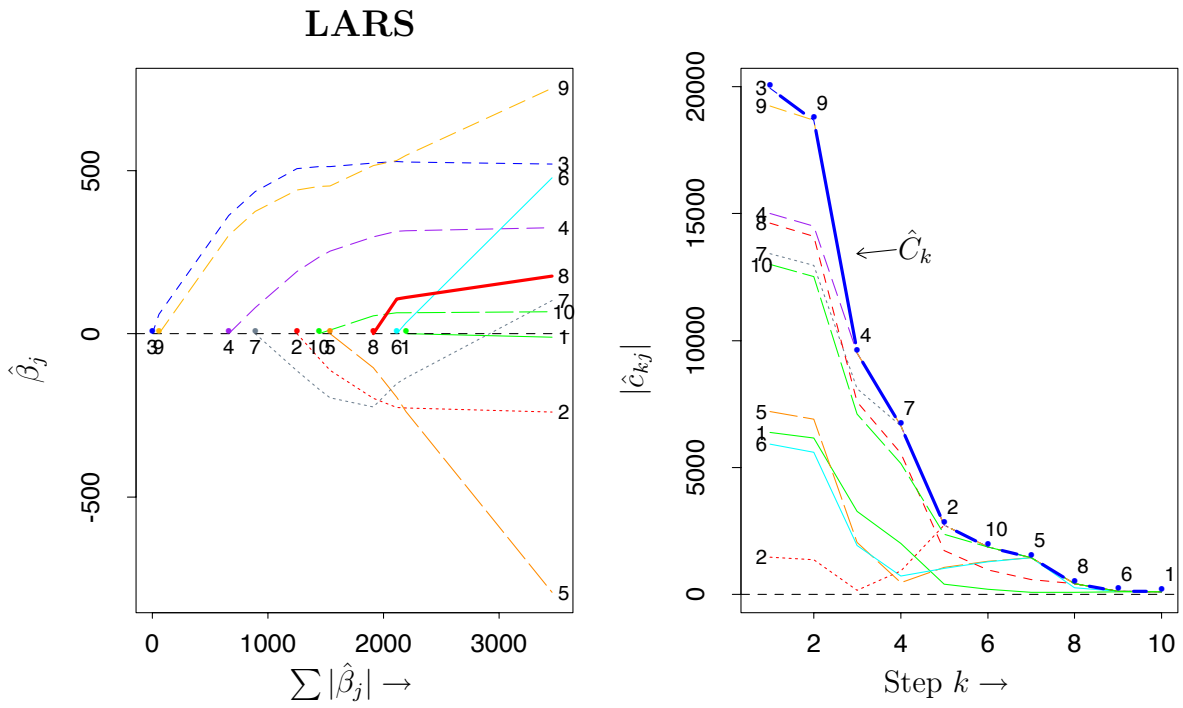

Figure 3. LARS analysis of the diabetes study. *Left*: estimates of regression coefficients $\widehat{\beta}_j$, $j = 1, 2, \ldots 10$; plotted versus $\Sigma|\widehat{\beta}_j|$; plot is slightly different than either Lasso or Stagewise, Figure 1. *Right*: Absolute current correlations as function of 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.

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 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.

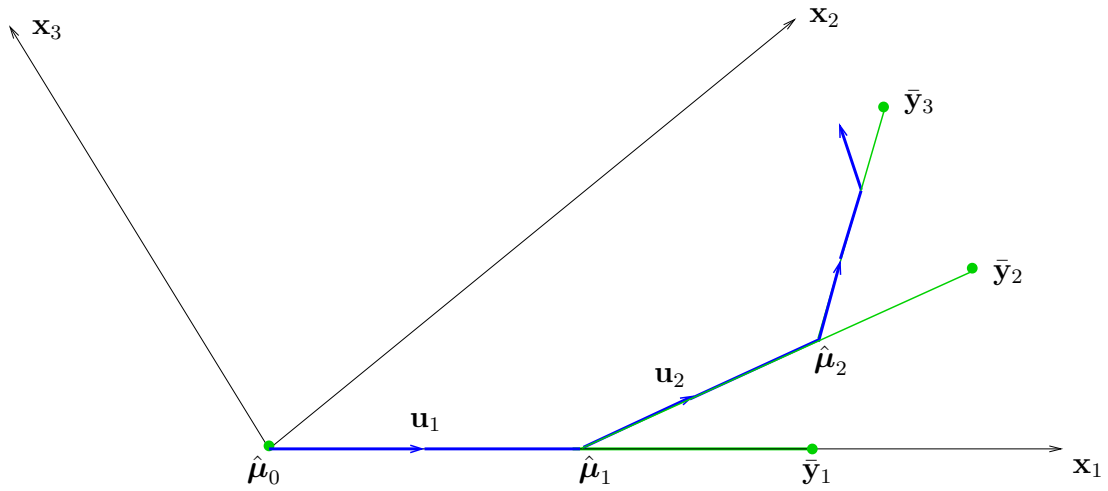

Figure 4. At each stage the LARS estimate $\widehat{\boldsymbol{\mu}}_k$ approaches, but does not reach, the corresponding OLS estimate $\bar{\mathbf{y}}_k$.

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, 2000b), 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 non-zero 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

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

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

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

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 $\widehat{\boldsymbol{\mu}}_1 + \epsilon \mathbf{x}_1$. Now variable 2 is more correlated,

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

forcing $j = 2$ to be the next Stagewise choice, etc.

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

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

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

$$\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}} + N\epsilon X_{\mathcal{A}} P_{\mathcal{A}} \tag{3.10}$$

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

The LARS algorithm (2.14) progresses along

$$\boldsymbol{\mu}_{\mathcal{A}} + \gamma X_{\mathcal{A}} w_{\mathcal{A}} \quad \text{where} \quad w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}}, \tag{3.11}$$

(2.6)–(3.2). 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 non-negative. 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 $\mathcal{C}_{\mathcal{A}}$, i.e. 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 non-negative 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:

- *Stagewise*: successive differences of $\widehat{\beta}_j$ agree in sign with the current correlation $\widehat{c}_j = \mathbf{x}_j'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$.
- *Lasso*: $\widehat{\beta}_j$ agrees in sign with $\widehat{c}_j$.
- *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 (3.19) 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:

$$\textit{Quadratic Model} \quad \text{10 main effects, 45 interactions, 9 squares,} \tag{3.15}$$

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 $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\mu}$ of $n = 442$ residuals. The "true R^2" for this model, $\|\boldsymbol{\mu}\|^2/(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\epsilon}\|^2)$, equaled 0.416.

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

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

with $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \epsilon_2^*, \ldots \epsilon_n^*)$ a random sample, with replacement, from the components of $\boldsymbol{\epsilon}$. 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* $pe(\widehat{\boldsymbol{\mu}})$ to be

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

so $pe(\mathbf{0}) = 0$ and $pe(\boldsymbol{\mu}) = 1$. The solid curve graphs the average of $pe(\widehat{\boldsymbol{\mu}}^{(k)*})$ 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 non-zero $\widehat{\beta}_j^*$ terms composing $\widehat{\boldsymbol{\mu}}^{(k)*}$. For example $\widehat{\boldsymbol{\mu}}^{(40)*}$ averaged 33.23 non-zero 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, 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 $pe(\widehat{\boldsymbol{\mu}}^{(k)*})$ 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

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

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

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

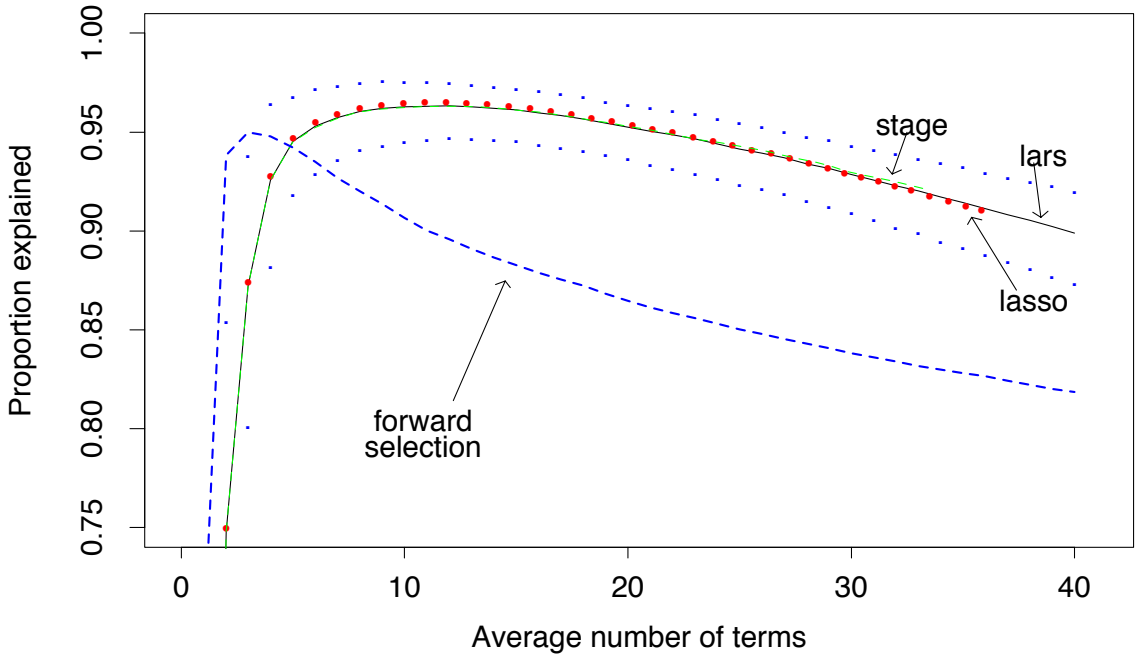

Figure 5. Simulation study comparing LARS, Lasso, and Stagewise algorithms; 100 replications of model (3.15)-(3.16). Solid curve shows average proportion explained, (3.17), for LARS estimates as function of number of steps $k = 1, 2, \ldots, 40$; Lasso and Stagewise give nearly identical results; small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

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 .01 for $k \geq 4$ compared with $R^2 \doteq .50$, 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 $\mathbf{x}_{3:9}, \mathbf{x}_{3:4}, \ldots, \mathbf{x}_{4:7}$.

Backwards Lasso The Lasso-modified LARS algorithm can be run backwards, starting from the full OLS solution $\bar{\boldsymbol{\beta}}_m$. Assuming that all the coordinates of $\bar{\boldsymbol{\beta}}_m$ are non-zero, 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 backwards 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 backwards, 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 backwards-moving algorithm. This is not the case for LARS or Stagewise, both of which are inherently forward-moving algorithms.

4. Degrees of Freedom and $\mathbf{C_p}$ 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

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

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

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

and a C_p-type risk estimation formula,

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

If σ^2 and df_{μ,σ^2} are known, $C_p(\widehat{\boldsymbol{\mu}})$ is an unbiased estimator of the true risk $E\{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/\sigma^2\}$. For linear estimators $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$, 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 & Tibshirani (1997) and Section 7 of Efron (1986) discuss formulas (4.4)–(4.5) and their role in C_p, AIC, and SURE estimation theory, a more recent reference being Ye (1998).

Practical use of C_p formula (4.5) requires preliminary estimates of $\boldsymbol{\mu}, \sigma^2$, 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 $\widehat{\boldsymbol{\mu}}^*$ were then generated according to

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

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

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

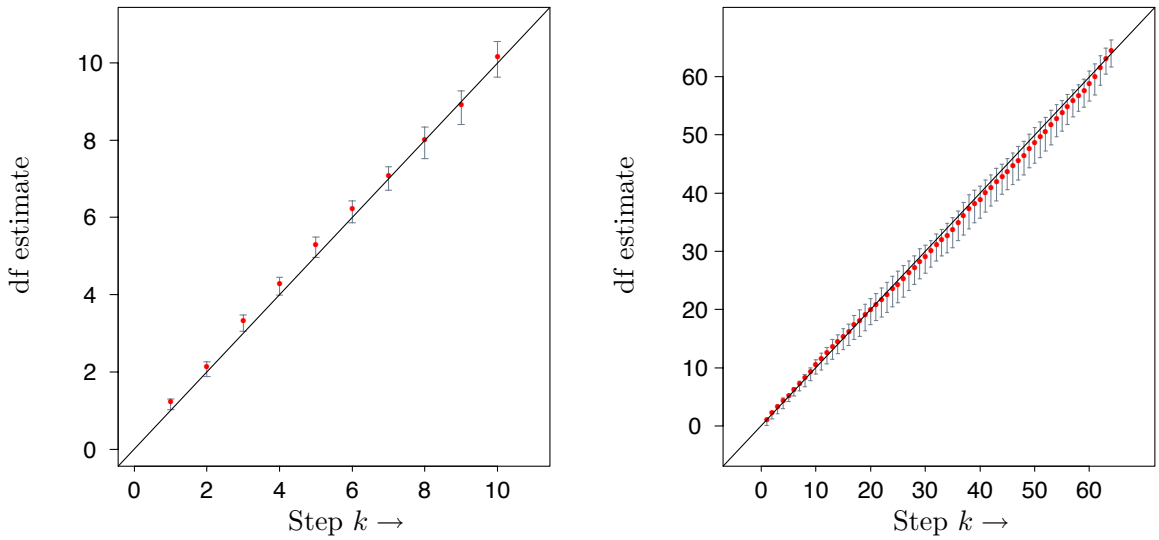

Figure 6. Degrees of freedom for LARS estimates $\widehat{\boldsymbol{\mu}}_k$. *Left Panel* diabetes study, Table 1, $k = 1, 2, \cdots, m = 10$; *Right Panel* quadratic model (3.15) for the diabetes data, $m = 64$. Solid line is simple approximation $df_k = k$. Dashed lines are approximate 95% confidence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

and then

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

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 $\widehat{\boldsymbol{\mu}}_k, k = 1, 2, \cdots, m = 10$. It portrays a startlingly simple situation that we will call the "simple approximation",

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

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 $\widehat{\boldsymbol{\mu}}_k$ by

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

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

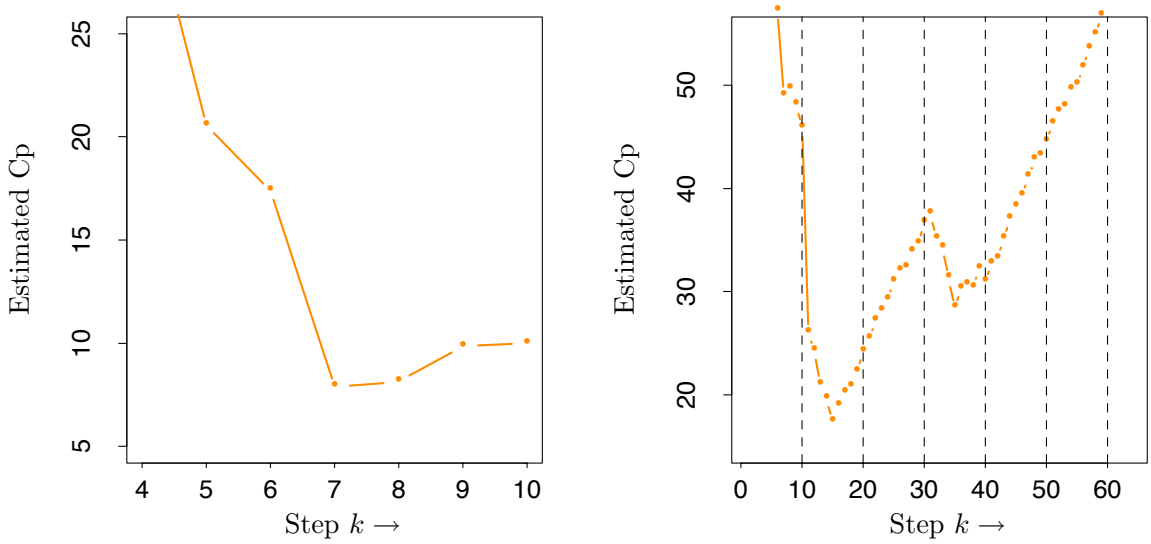

Figure 7. C_p estimates of risk (4.10) for the two situations of Figure 6. *Left Panel* $m = 10$ model has smallest C_p at $k = 7$; *Right Panel* $m = 64$ model has smallest C_p at $k = 16$.

further calculations beyond those for the original LARS estimates. The formula applies only to LARS, and not to Lasso or Stagewise.

Figure 7 displays $C_p(\widehat{\boldsymbol{\mu}}_k)$ 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 $x_1, x_2, \ldots, x_m$ 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

Positive Cone condition For all possible subsets $X_{\mathcal{A}}$ of the full design matrix X,

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

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 Appendix) and set $\nabla \cdot g = \sum_{i=1}^n \partial g_i/\partial x_i$. If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, then Stein's formula states that

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

The left 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 one, but that the continuity assumptions underlying formula (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, $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$. 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 re-express (4.13) as a locally linear transformation,

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

P_k being the usual projection matrix from 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 $\mathrm{trace}(M_k) = k$. 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 estimate (4.6)–(4.7).

It is clear that formula (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 non-zero 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 $\mathbf{x}_j = \mathbf{e}_j, j = 1, ..., n$, the kth LARS estimate ($0 \le k \le n$) is given by*

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

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

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:

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

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

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

and so

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

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

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

and

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

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 $|y|_{(k+1)} = \ldots = |y|_{(k+r)}$, then $\mathcal{A}_k(\mathbf{y})$ expands by r variables at step $k+1$ and $\hat{\boldsymbol{\mu}}_{k+\nu}(\mathbf{y}), \nu = 1, ..., r$ are all determined at the same time and are equal to $\hat{\boldsymbol{\mu}}_{k+1}(\mathbf{y})$. □

Proof of Theorem 4 in 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, i.e. 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 & Woodroof (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 Appendix Lemma 13, 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 non-zero, as argued in the proof of Lemma 13. As shown in (9.4) in the Appendix, formula (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

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

[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 $k \geq 1$,

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

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 et al. (2000*a*), 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 $w_k = A_k \mathcal{G}_k^{-1} 1_k$ for the equiangular vector $\mathbf{u}_k = X_k w_k$, (2.6), has its kth 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 kth 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. 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

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

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

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

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 $j \in \mathcal{A}_{k-1}$:

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

Thus

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

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

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 $A_{\mathcal{A}} = (1'\mathcal{G}_{\mathcal{A}}^{-1}1)^{-\frac{1}{2}}$, (2.4), (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*

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

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

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

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

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

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

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

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 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 derivative 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_o, within which the set $\mathcal{A}$ of non-zero Lasso coefficients $\widehat{\beta}_j(t)$ remains constant.

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

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

for $t \in \mathcal{T}$, *where* $\mathbf{u}_{\mathcal{A}}$ *is the equiangular vector* $X_{\mathcal{A}} w_{\mathcal{A}}, w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}}$, *(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 non-zero regression coefficients $\widehat{\beta}_{\mathcal{A}}(t)$,

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

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 non-zero 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

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

(The inequality in (1.5) can be replaced by $T(\widehat{\boldsymbol{\beta}}) = t$ as long as t is less than $\Sigma|\bar{\beta}_j|$ 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

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

Introducing a Lagrange multiplier (5.20) becomes

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

Differentiating we get

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

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

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

Hence

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

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

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

and

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

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

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 polytopes, 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*

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

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

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

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 $\lambda \cdot \mathrm{sign}(\widehat{\beta}_j)$ for $j \in \mathcal{A}$. Likewise $\lambda = |\widehat{c}_j| = \widehat{C}$.

□

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

$$\begin{aligned} &|c_j(t)| = \widehat{C}(t) \equiv \max\{|c_\ell(t)|\} \quad \mathit{for} \quad j \in \mathcal{A} \\ &\mathit{and} \\ &|c_j(t)| \le \widehat{C}(t) \quad \mathit{for} \quad j \notin \mathcal{A} \end{aligned} \tag{5.30}$$

Proof. (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 $T(\gamma) = \Sigma |\beta_j(\gamma)|$, 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

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

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

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

while if $\mathbf{d}$ has only component j non-zero 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 $t_o = \inf(\mathcal{T})$ 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 et al. 2000*a*); 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 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 $t = t_o$ indicate such a point, $t_o = \inf(\mathcal{T})$ as in Lemma 9, with Lasso regression vector $\widehat{\boldsymbol{\beta}}$, prediction estimate $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$, current correlations $\widehat{\mathbf{c}} = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$, $s_j = \text{sign}(\widehat{c}_j)$, and maximum absolute correlation $\widehat{C}$. Define

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

$\mathcal{A}_{10} = \mathcal{A}_1 \cup \mathcal{A}_o$ 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 $T(\gamma) = \Sigma|\beta_j(\gamma)|$.

Lemma 10. *The negative slope (5.31) at t_o 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

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

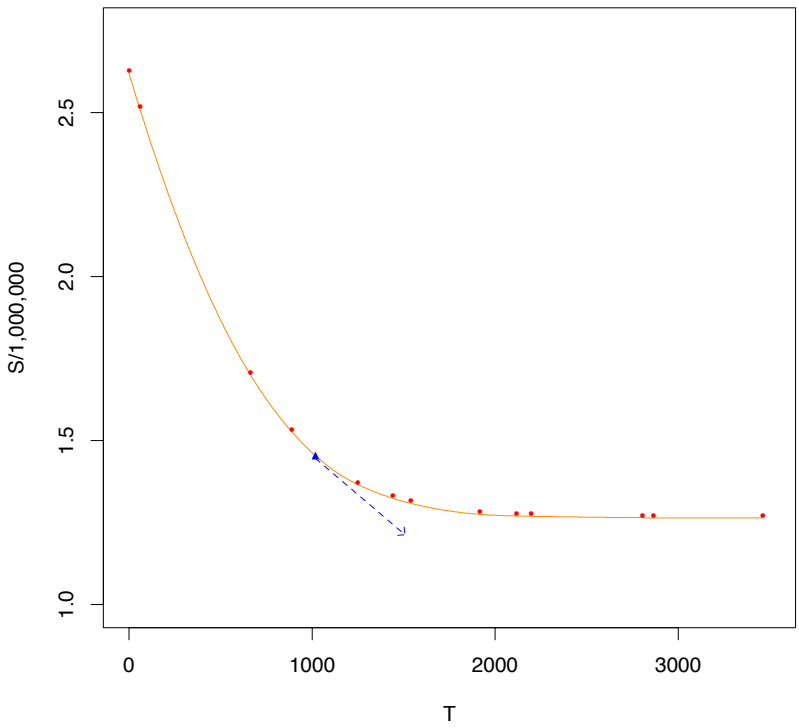

Figure 8. Plot of S versus T for Lasso applied to diabetes data; points indicate the 12 modified LARS steps of Figure 1; triangle is (T, S) boundary point at $t = 1000$; dashed arrow is tangent at $t = 1000$, negative slope R_t, (5.31). The (T, S) curve is a decreasing, convex, quadratic spline.

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)

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

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

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

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 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 $\mathcal{A}_1$, $\mathcal{A}_o$, and $\mathcal{A}_{10}$ as at (5.36).

Constraint I $\mathcal{A}_1 \subseteq \mathcal{A}$. 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 II $\mathcal{A} \subseteq \mathcal{A}_{10}$. 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 non-zero 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 III $w_{\mathcal{A}} = A_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}}$ cannot have $\text{sign}(w_j) \neq \text{sign}(\widehat{c}_j)$ for any coordinate $j \in \mathcal{A}_o$. If it does then $\text{sign}(\widehat{\beta}_j(\gamma)) \neq \text{sign}(\widehat{c}_j(\gamma))$ for sufficiently small γ, violating Lemma 8.

Constraint IV Subject to Constraints I-III, $\mathcal{A}$ must minimize $A_{\mathcal{A}}$. 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 $\widehat{\boldsymbol{\beta}}_o = 0$, 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 $\mathcal{A}_o$ is empty so that constraints I and II 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 break point, $\mathcal{A}_o$ has exactly one member, say j_o, so $\mathcal{A}$ must equal $\mathcal{A}_1$ or $\mathcal{A}_{10}$. There are two cases: if j_o

has just been *added* to the set $\{|\widehat{c}_j| = \widehat{C}\}$ then Lemma 4 says that $\text{sign}(w_{j_o}) = \text{sign}(\widehat{c}_{j_o})$, so that constraint III 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_o *deleted* from the active set as at (3.6). Now the choice $\mathcal{A} = \mathcal{A}_{10}$ is ruled out by Constraint III: it would keep $w_{\mathcal{A}}$ the same as in the previous LARS-Lasso step, and we know that that was stopped at (3.6) to prevent a sign contradiction at coordinate j_o. 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 $\mathcal{A}_o = \{j_1, j_2\}$. It is possible but not certain that $\mathcal{A}_{10}$ does not violate Constraint III, in which case $\mathcal{A} = \mathcal{A}_{10}$. However if it does violate III 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 $\Sigma|\widehat{\beta}_j| = 2000$ 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 $s_j = \text{sign}(\widehat{c}_j)$ are all non-negative.

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

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

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. □

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

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

$P_{\mathcal{A}}$ 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

Constraint I $\mathbf{v} \in \mathcal{S}_{\mathcal{A}}^+$, the non-negative simplex

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

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 non-zero 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 $\mathbf{v}$ must be proportional to the equiangular vector $\mathbf{u}_{\mathcal{B}}$, (2.6), that is $\mathbf{v} = \mathbf{v}_{\mathcal{B}}$, (5.8),

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

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 $\mathbf{v} = \lambda \mathcal{G}_{\mathcal{B}}^{-1} 1_{\mathcal{B}}$; 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 $X_{\mathcal{B}}'\mathbf{v}_{\mathcal{B}} = A_{\mathcal{B}}^2 1_{\mathcal{B}}$, or

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

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

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

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 re-enter (1.7).

To obtain strict inequality in (6.7), let $\mathcal{B}_o \subset \mathcal{A} - \mathcal{B}$ 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 $\epsilon \to 0$ 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),

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

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.

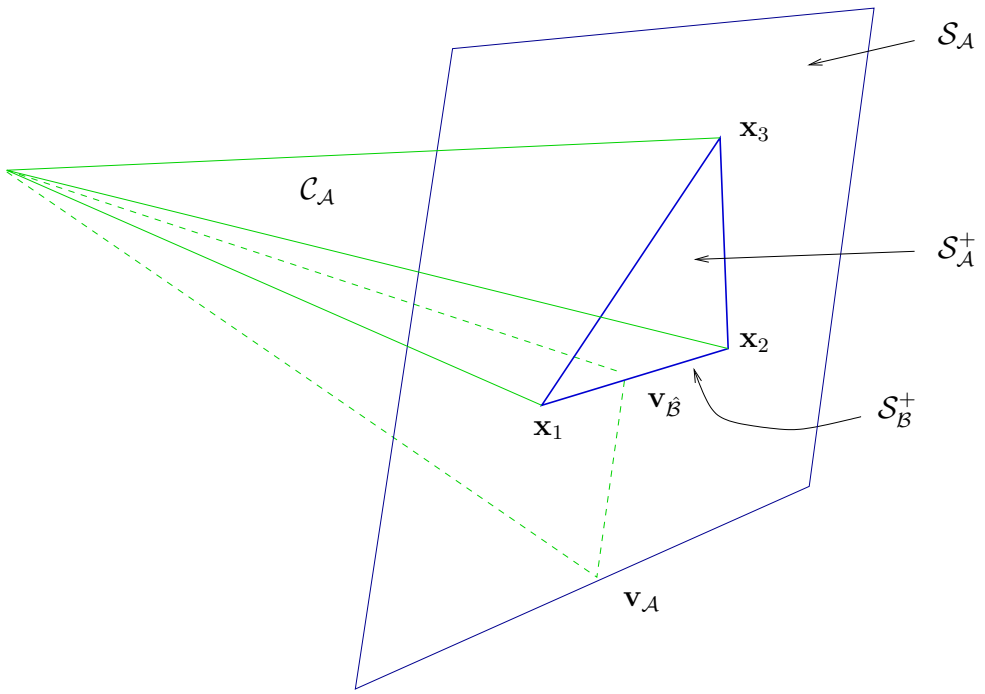

Figure 9. The geometry of the LARS-Stagewise modification

Lemma 12. $\mathbf{v}_{\hat{\mathcal{B}}}$ *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 $\mathrm{Sin} = [1 - \mathrm{Cos}^2]^{\frac{1}{2}}$, 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 $\mathcal{L}(X_{\mathcal{B}}) = \mathcal{L}(\mathbf{c}_1, \mathbf{c}_1, \ldots, \mathbf{c}_J)$, $J = |\mathcal{B}|$, the space of n-vectors with last $n - J$ coordinates zero, and also

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

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

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

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

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

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 $x_{\ell_3} \le 0$ (that is, $\mathbf{x}_\ell$ 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 non-zero 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 $\mathbf{v}_{\mathcal{B}} \neq \mathbf{v}_{\hat{\mathcal{B}}}$ 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 $\widehat{\mathbf{v}} = \mathbf{v}_{\widehat{\mathcal{B}}}$ 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 $\widehat{\mathbf{v}} = \mathbf{v}_{\widehat{\mathcal{B}}} = \phi(\widehat{\mathcal{A}})$. In the succeeding LARS-Stagewise interval $\widehat{\boldsymbol{\mu}}(\gamma) = \widehat{\boldsymbol{\mu}} + \gamma\mathbf{v}_{\widehat{\mathcal{B}}}$, the maximal set is immediately reduced to $\widehat{\mathcal{B}}$, according to properties (6.6), (6.7) of $\mathbf{v}_{\widehat{\mathcal{B}}}$; at which it stays during the entire interval. However $\phi(\widehat{\mathcal{B}}) = \phi(\widehat{\mathcal{A}}) = \mathbf{v}_{\widehat{\mathcal{B}}}$ since $\mathbf{v}_{\widehat{\mathcal{B}}} \in \mathcal{C}_{\widehat{\mathcal{B}}}$, 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 non-linear fitting problems, as in Chapter 10 of Hastie et al. (2001), which begins with a Stagewise analysis to "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 $\widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{B}}}$. 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

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

(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

$$\left.\frac{\partial S_{\text{Stage}}(\gamma)}{\partial\gamma}\right|_o \Big/ \left.\frac{\partial S_{\text{LARS}}(\gamma)}{\partial\gamma}\right|_o = \frac{A_{\hat{\mathcal{A}}}}{A_{\widehat{\mathcal{B}}}}, \tag{6.20}$$

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),

$$-\left.\frac{\partial\widehat{C}(\gamma)}{\partial\gamma}\right|_o = \min_{\hat{\mathcal{A}}}\{|x_j'\mathbf{u}|\}. \tag{6.21}$$

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

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

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 non-active $\mathbf{x}_j$ 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 & 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 counter-productive 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 & Hansen (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 factors 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:

- 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 non-zero coefficients.
- 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 $n - 1$.

- The model sequence, particularly near the saturated end, tends to be quite variable with respect to small changes in $\mathbf{y}$.
- The estimation of σ^2 may have to depend on an auxilliary 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 $m > n$.

A documented S-plus implementation of LARS and associated functions is 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 & 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 & Stone 1984).

Friedman, Hastie & 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 $\hat{\mathbf{y}} + \epsilon \cdot \mathbf{t}_1$, $\mathbf{r}$ 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 $\epsilon = 1$: 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 isϵ-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 fit 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 don't 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 et al. (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. But 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 stepsizes this procedure approximates LARS; its advantage is that it can be carried out on an infinite set of predictors such as trees.

9. Appendix

9.1. 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 $s_l \mathbf{x}_{j(l)} = \mathrm{sgn}(\hat{c}_{j(l)})\mathbf{x}_{j(l)}$. 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 $\Delta\mathbf{y}$,

$$\|\Delta g\| = \|g(\mathbf{y} + \Delta\mathbf{y}) - g(\mathbf{y})\| \le L\|\Delta\mathbf{y}\|. \tag{9.1}$$

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, differ by one, 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

$$w'(R\mathbf{y} - \gamma\mathbf{u}_k) = \mathbf{x}_{j_1}'(R\mathbf{y} - \gamma\mathbf{u}_k) = \mathbf{x}_{j_2}'(R\mathbf{y} - \gamma\mathbf{u}_k) \tag{9.2}$$

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 $\boldsymbol{\delta}_1'\mathbf{u}_k \neq 0$ 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

$$\boldsymbol{\delta}_2' R\mathbf{y} = \eta_1'\mathbf{y}\ \boldsymbol{\delta}_2'\mathbf{u}_k$$

In other words, setting $\boldsymbol{\eta}_2 = R'\boldsymbol{\delta}_2 - (\boldsymbol{\delta}_2'\mathbf{u}_k)\boldsymbol{\eta}_1$, we have

$$N(j_1, j_2) \subset \{\mathbf{y} : \boldsymbol{\eta}_2'\mathbf{y} = 0\}.$$

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

$$\hat{C}_k(\mathbf{y}_0) - \hat{\gamma}_k(\mathbf{y}_0)A_k \begin{Bmatrix} = \\ > \end{Bmatrix} \hat{c}_{k,j}(\mathbf{y}_0) - \hat{\gamma}_k(\mathbf{y}_0)a_{k,j} \quad \text{as} \quad \begin{Bmatrix} j = k+1 \\ j > k+1 \end{Bmatrix}. \tag{9.3}$$

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 (9.3) implies that

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

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

$$\hat{\gamma}_k(\mathbf{y}) = R_{k,k+1}(\mathbf{y}) = \frac{(\mathbf{x}_l - \mathbf{x}_{k+1})'(\mathbf{y} - \hat{\boldsymbol{\mu}}_{k-1}(\mathbf{y}))}{(\mathbf{x}_l - \mathbf{x}_{k+1})'\mathbf{u}_k} \tag{9.4}$$

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 (9.1),

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

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 (9.3)), 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

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

□

9.2. 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 span$(X_{\mathcal{A}})$, so that $a = \mathbf{x}_+'P_{\mathcal{A}}\mathbf{x}_+ < 1$. The the $+$-component of $G_{\mathcal{A}+}^{-1}1_{\mathcal{A}+}$ is*

$$(G_{\mathcal{A}+}^{-1}1_{\mathcal{A}+})_+ = (1-a)^{-1}\Big(1 - \frac{\mathbf{x}_+'\mathbf{u}_{\mathcal{A}}}{A_{\mathcal{A}}}\Big). \tag{9.5}$$

Consequently, under the positive cone condition (4.11),

$$\mathbf{x}_+'\mathbf{u}_{\mathcal{A}} < A_{\mathcal{A}}. \tag{9.6}$$

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)),

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

where

$$\begin{aligned} E = D - B'A^{-1}B &= 1 - \mathbf{x}'_+P_{\mathcal{A}}\mathbf{x}_+, \\ F = A^{-1}B &= G^{-1}_{\mathcal{A}}X'\mathbf{x}_+ \end{aligned}$$

from which (9.5) follows. The positive cone condition implies that $G^{-1}_{\mathcal{A}+}1_{\mathcal{A}+} > 0$, and so (9.6) is immediate. □

9.3. 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 sufficent to establish the same pattern of behavior at points of multiplicity three 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 $\mathbf{y}_0$ that 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, the property (9.3) 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 $B(\mathbf{y}_0)$, i.e. those for which $\mathcal{A}_{k+1}(\mathbf{y}) \setminus \mathcal{A}_k = \{k+1, k+2\}$. We make 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 co-ordinate 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 three 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. □

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association* **81**, 461–470.

Efron, B. & Tibshirani, R. (1997), 'Improvements on cross-validation: the 632+ bootstrap: method', *J. Amer. Statist. Assoc.* **92**, 548–560.

Freund, Y. & Schapire, R. (1997), 'A decision-theoretic generalization of online learning and an application to boosting', *Journal of Computer and System Sciences* **55**, 119–139.

Friedman, J. (2001), 'Greedy function approximation: the gradient boosting machine', *Annals of Statistics* . to appear.

Friedman, J., Hastie, T. & Tibshirani, R. (2000), 'Additive logistic regression: a statistical view of boosting (with discussion)', *Annals of Statistics* **28**, 337–307.

Golub, G. & Van Loan, C. (1983), *Matrix Computations*, Johns Hopkins University Press, Baltimore.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning; Data mining, Inference and Prediction*, Springer Verlag, New York.

Lawson, C. & Hansen, R. (1974), *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.

Mallows, C. (1973), 'Some comments on cp', *Technometrics* **15**, 661–675.

Meyer, M. & Woodroof, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Annals of Statistics* **28**, 1083–1104.

Osborne, M., Presnell, B. & Turlach, B. (2000*a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**, 389–404.

Osborne, M. R., Presnell, B. & Turlach, B. (2000*b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 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. Royal. Statist. Soc. 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', *Journal of the American Statistical Association* pp. 120–131.