IMA Journal of Numerical Analysis (2000) **20**, 389–404

A new approach to variable selection in least squares problems

M. R. OSBORNE
School of Mathematical Sciences, Australian National University, Australia

BRETT PRESNELL
Department of Statistics, University of Florida, USA

AND

B. A. TURLACH
Department of Mathematics and Statistics, University of Western Australia, Perth, Australia

[Received 20 October 1998 and in revised form 15 October 1999]

The title Lasso has been suggested by Tibshirani (1996) as a colourful name for a technique of variable selection which requires the minimization of a sum of squares subject to an l_1 bound κ on the solution. This forces zero components in the minimizing solution for small values of κ. Thus this bound can function as a selection parameter. This paper makes two contributions to computational problems associated with implementing the Lasso: (1) a compact descent method for solving the constrained problem for a particular value of κ is formulated, and (2) a homotopy method, in which the constraint bound κ becomes the homotopy parameter, is developed to completely describe the possible selection regimes. Both algorithms have a finite termination property. It is suggested that modified Gram–Schmidt orthogonalization applied to an augmented design matrix provides an effective basis for implementing the algorithms.

1. The problem

Exploratory data analysis involves a collection of techniques for variable selection (selection of a subset of the columns in the design matrix A) in the linear model

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{1}$$

The precise form of the model in this class of problems is not known *a priori* but must be selected from a class of possible models using information contained in the given data set. These selection methods are now finding important applications in areas such as data mining where there is a requirement for adequate and economical models for summarizing aspects of the information in very large data sets. The classical technique of exploratory data analysis is stepwise regression where a greedy algorithm adds a variable at a time to the model set using the criterion of best fitting the current residual vector $\mathbf{r}$ in the sense of making the most significant reduction in the sum of squares of residuals. Also, variables are removed from the model set if they satisfy a criterion of redundancy which basically reverses the addition criterion. In this approach the generated model with

k variables typically contains the model with $k-1$ variables, but this need not be true of the best subsets. Tibshirani's (1996) Lasso provides a new approach to the variable selection problem which makes use of the polyhedral structure of the l_1 norm. One reason for interest in the Lasso is that it provides a sense of optimality for the variable set selected not shared by stepwise regression. Implemented as a piecewise linear homotopy it also possesses many of the attributes of a stepwise procedure. Another advantage of the homotopy is that it provides an explanation of the variable selection attributes of the Lasso.

The basic problem of the Lasso is to minimize the sum of squares of residuals in (1),

$$\min \tfrac{1}{2}\mathbf{r}^T\mathbf{r}; \qquad \mathbf{r} = A\mathbf{x} - \mathbf{b}, \tag{2}$$

where $A : R^p \rightarrow R^n$, and where the columns of A provide the totality of variables available to the selection procedure, subject to the nonsmooth constraint expressed in terms of the l_1 norm of the solution vector

$$\kappa - \|\mathbf{x}\|_1 \geqslant 0. \tag{3}$$

This choice of norm provides a selection mechanism by forcing components of $\mathbf{x}$ to zero when κ is small.

The plan of the paper is as follows. In the next section, preliminary results on the Lasso are derived. These include necessary conditions for an optimum, duality, and a convenient partitioning of the variables. The descent and homotopy methods are derived in the following two sections. These have complementary roles. The descent algorithm solves the constrained optimization problem for a given κ by means of an active set algorithm that exploits a local constraint linearization, and the homotopy algorithm extends the optimum solution for a range of values of κ by making use of the result that the optimal trajectory is piecewise linear in κ. The emphasis here is on the explicit treatment of the norm constraint, and it is argued that this results in more compact problem formulations than possible alternatives which make use of families of linear inequalities. Brief concluding sections summarize the problem scaling strategy used, implementation details, and preliminary numerical results. Summary conclusions are presented.

2. Properties of the Lasso

It is the nonsmooth nature of the particular norm constraint (3), which treats zero values of $\mathbf{x}$ as special, that provides the mechanism for variable selection. It proves an interesting constraint set because it has a natural representation as a system of linear inequalities, but the number of these inequalities can be large. For example, if $\mathbf{x}$ has k zero components then there are 2^k possible representors $\mathbf{v}$ of $\|\mathbf{x}\|_1$, that is vectors with elements ± 1 such that

$$\mathbf{v}^T\mathbf{x} = \|\mathbf{x}\|_1 .$$

More compact representations which involve only $O(p)$ additional linear constraints are possible. For example, the constraint can be replaced by the inequalities

$$-t_i \leqslant x_i \leqslant t_i, \qquad t_i \geqslant 0, \qquad i = 1, 2, \ldots, p,$$

$$\kappa - \sum_{i=1}^{p} t_i \geqslant 0.$$

However, this system does involve increasing the number of variables, and it has the added disadvantage of leading to degenerate constraint sets in this application. For these reasons we are interested in developing compact algorithms which use the norm information directly. This will be done by means of a careful local linearization with a property we call *sign feasibility* being used to delimit its scope. The crucial dependence of the norm constraint is on the bound κ.

REMARK 1 Let κ_{LS} be the smallest value of κ in (3) for which the solution of (2), (3) also minimizes the sum of squares (2). Clearly, the least squares solution is obtained for all larger values of κ. Thus the interesting values, those which imply that the constraint (3) is active, satisfy $\kappa < \kappa_{LS}$.

The explicit use of the norm constraint leads to a compact form for the Kuhn–Tucker conditions (Osborne, 1985) for the problem (2), (3). These are:

$$\mathbf{r}^T A = -\mu \mathbf{v}^T, \qquad \mu \geqslant 0, \tag{4}$$

where

$$\mathbf{x} = P^T \begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix}, \qquad \mathbf{v} = P^T \begin{bmatrix} \theta_\sigma \\ \mathbf{v}_2 \end{bmatrix} \in \partial \|\mathbf{x}\|_1 , \tag{5}$$

P is the permutation matrix which collects together the nonzero components of $\mathbf{x}$ (these are pointed to by the index set σ), $\theta_{\sigma(j)} = \operatorname{sgn}(x_{\sigma(j)})$, $-1 \leqslant (\mathbf{v}_2)_j \leqslant 1$, $j \in \sigma^C$, and $\sigma \cup \sigma^C = \{1, 2, \ldots, p\}$. Note that $\mathbf{v}^T\mathbf{x} = \|\mathbf{x}\|_1$, $\|\mathbf{v}\|_\infty = 1$, and

$$\mu = -\frac{\mathbf{r}^T A\mathbf{x}}{\|\mathbf{x}\|_1} = \left\| \mathbf{r}^T A \right\|_\infty .$$

It follows from this equation that $\mu > 0$ if and only if $\kappa < \kappa_{LS}$.

The Lagrangian function is given by

$$\mathcal{L}(\mathbf{x}, \mu) = \tfrac{1}{2}\mathbf{r}^T\mathbf{r} - \mu (\kappa - \|\mathbf{x}\|_1)$$

and is convex in $\mathbf{x}$ for $\mu \geqslant 0$. The dual function is given by

$$w(\mu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu).$$

The condition for a minimum in $\mathbf{x}$ is $0 \in \partial_x \mathcal{L}(\mathbf{x}, \mu)$. This gives the necessary conditions (not quite the Kuhn–Tucker conditions (4) because of the dependence on μ)

$$\mathbf{r}^T A + \mu \mathbf{v}^T = 0, \qquad \mathbf{v}^T \in \partial \|\mathbf{x}\|_1 . \tag{6}$$

Taking the scalar product with $\mathbf{x}$ gives

$$\mathbf{r}^T A\mathbf{x} + \mu \|\mathbf{x}\|_1 = 0.$$

Substituting in the expression for $\mathcal{L}$ gives

$$w(\mu) = -\tfrac{1}{2}\mathbf{r}^T\mathbf{b} - \tfrac{1}{2}\kappa\mu$$

where the dependence on μ is through (6).

It is convenient to partition the Kuhn–Tucker conditions by introducing the (partial) factorization of AP^T into the product of an orthogonal times block upper triangular matrix (Clark & Osborne, 1988)

$$AP^T = Q\begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix}, \qquad Q^T\mathbf{b} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}. \tag{7}$$

Here U_1 is strictly upper triangular, but B need not be reduced. Substituting in (4) gives

$$\begin{bmatrix} U_1^T & \\ U_{12}^T & B^T \end{bmatrix}\left\{\begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix}\begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}\right\} + \mu P\mathbf{v} = 0.$$

This simplifies to

$$U_1\mathbf{x}_\sigma = \mathbf{c}_1 - \mu U_1^{-T}\theta_\sigma, \tag{8}$$

$$\mu\mathbf{v}_2 = B^T\mathbf{c}_2 + \mu U_{12}^T U_1^{-T}\theta_\sigma. \tag{9}$$

The optimal solution can be obtained from this pair of equations, which must be solved in conjunction with the constraint equation (3), provided the correct partitioning (5) is known *a priori*. Otherwise the problem reduces to one of finding this correct partition. Setting $\mathbf{w}_\sigma = U_1^{-T}\theta_\sigma$ in (8) gives another formula for the multiplier

$$\mu = \frac{\mathbf{w}_\sigma^T\mathbf{c}_1 - \kappa}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma} = \frac{\theta_\sigma^T\mathbf{x}_{LS}^\sigma - \kappa}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma} \tag{10}$$

where $\mathbf{x}_{LS}^\sigma = U_1^{-1}\mathbf{c}_1$ is the least squares solution corresponding to the variables referenced by σ.

3. The descent algorithm

This algorithm is a rather standard active set method in which a simple quadratic program is solved at each step in order to generate a descent direction. The interest lies in the treatment of the active constraint. The basic step involves computing a correction $\mathbf{h}$ to the current solution estimate using what amounts to a local linearization of (3) about the current $\mathbf{x}$:

$$\min_{\mathbf{h}} \tfrac{1}{2}\|\mathbf{r}(\mathbf{x}+\mathbf{h})\|_2^2; \qquad \theta_\sigma^T(\mathbf{x}_\sigma + \mathbf{h}_\sigma) \leqslant \kappa, \qquad \mathbf{h} = \begin{bmatrix} \mathbf{h}_\sigma \\ 0 \end{bmatrix} \tag{11}$$

where $\mathbf{x}$ is a feasible starting point and θ_σ is fixed by the local constraint condition

$$\|\mathbf{x}\|_1 = \theta_\sigma^T\mathbf{x}_\sigma \leqslant \kappa. \tag{12}$$

The Kuhn–Tucker conditions for the linearized system comprising the sum of squares and this linear constraint are

$$A_1^T \mathbf{r}(\mathbf{x}+\mathbf{h}) + \widetilde{\mu}\theta_\sigma = 0, \qquad \widetilde{\mu} \geqslant 0 \tag{13}$$

where A_1 is made up of the columns of A associated with the nonzero components of $\mathbf{x}$. The advantage of this 'linearized' problem is that the system (12) and (13) can be solved directly. An application of the factorization (7) gives the equations

$$\begin{aligned} U_1\mathbf{h}_\sigma &= -Q_1^T\mathbf{r}(\mathbf{x}) - \widetilde{\mu}U_1^{-T}\theta_\sigma, \\ \theta_\sigma^T(\mathbf{x}_\sigma+\mathbf{h}_\sigma) &= \kappa, \end{aligned}$$

assuming the constraint is active so $\widetilde{\mu} > 0$. Setting $\mathbf{w}_\sigma = U_1^{-T}\theta_\sigma$ gives

$$\widetilde{\mu} = \frac{\mathbf{w}_\sigma^T\mathbf{c}_1 - \kappa}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}.$$

Note that this agrees with (10) if the correct θ_σ is known. If the optimum satisfies

$$\mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x}_\sigma + \mathbf{h}_\sigma \\ 0 \end{bmatrix}, \qquad \operatorname{sgn}(\mathbf{x}_\sigma) = \theta_\sigma$$

then $\mathbf{x}$ is called *sign feasible* for (2) and satisfies (8). It can be tested for optimality in (2) by solving (9) for $\mathbf{v}_2$, and checking that

$$-1 \leqslant (\mathbf{v}_2)_i \leqslant 1.$$

If these constraints are satisfied then the solution has been obtained.

The following strategies can be used to develop an appropriate partition σ:

Case 1 $\mathbf{x}+\mathbf{h}$ given by (11) is not sign feasible for (2). Here the actions taken are (compare (Clark & Osborne, 1988)):

1. Move to first new zero component in the direction $\mathbf{h}$, $0 = x_k \leftarrow x_k + \gamma h_k$, $k \in \sigma$, $0 < \gamma < 1$.
2. Now there are two possibilities: either setting $\theta_k = -\theta_k$, $\mathbf{x}_\sigma \leftarrow \mathbf{x}_\sigma + \gamma\mathbf{h}$, and recomputing $\mathbf{h}$ yields a descent direction that is consistent with the revised θ_σ (this has the advantage that downdating of the factorization (7) is not necessary so the computation is relatively cheap), or it is necessary to update $\sigma \leftarrow \sigma \setminus \{k\}$, reset $\mathbf{x}_\sigma$, θ_σ (they are feasible for the restricted problem), downdate the factorization, and recompute $\mathbf{h}_\sigma$.
3. Iterate until a sign feasible $\mathbf{x}$ is obtained.

Case 2 $\mathbf{x}_\sigma$ is sign feasible for (11) but $\mathbf{v}_2$ is not feasible. Here the actions taken are:

1. Select an infeasible multiplier condition involving (say) $(\mathbf{v}_2)_s$.
2. Update

$$\sigma \leftarrow \sigma \cup \{s\}, \qquad \mathbf{x}_\sigma \leftarrow \begin{bmatrix} \mathbf{x}_\sigma \\ (x_s = 0) \end{bmatrix}, \qquad \theta_\sigma \leftarrow \begin{bmatrix} \theta_\sigma \\ \theta_s \end{bmatrix},$$

where θ_s is chosen so that $\operatorname{sgn}(\theta_s) = \operatorname{sgn}(h_s)$ in (11).

3. Solve (11) and iterate.

REMARK 2 To justify Case 1 note that $\mathbf{x}$ is either optimal for the restricted problem, in which case this stage of the iteration is terminated, or it is not optimal, in which case $\mathbf{h}$ is a descent direction so the objective is reduced in the next step. Thus there can be no cycling, and the procedure must converge. Further, the procedure must be finite as the total number of configurations is finite, and the final $\mathbf{x}$ must be sign feasible or the limiting process is contradicted. This argument extends to show that the full algorithm also has a finite termination property.

REMARK 3 Assuming that the norm constraint is active (otherwise there is nothing to do as (9) is trivially satisfied), then the correct choice of θ_s in Case 2 is

$$\theta_s = \operatorname{sgn}((\mathbf{v}_2)_s). \tag{14}$$

To show this note that, because the initial point for the augmented problem is not optimal, it follows that the solution of (11) gives a descent direction. Thus

$$\begin{aligned} 0 > \mathbf{r}\left(\begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} A_1 & \mathbf{a}_s \end{bmatrix} \begin{bmatrix} \mathbf{h}_\sigma \\ h \end{bmatrix} &= \mathbf{r}^T A_1 \mathbf{h}_\sigma + \mathbf{r}^T \mathbf{a}_s h \\ &= -\mu \left[\theta_\sigma^T \mathbf{h}_\sigma + (\mathbf{v}_2)_s\, h\right], \end{aligned}$$

where equations (8), (9) have been used in evaluating the right-hand side. Feasibility for (11) gives

$$\theta_\sigma^T \mathbf{h}_\sigma + \theta_s h \leqslant 0.$$

Adding μ times this to the previous inequality gives

$$0 > \mu \left(\theta_s - (\mathbf{v}_2)_s\right) h.$$

This shows that

$$\operatorname{sgn}(\theta_s) = \operatorname{sgn}\left((\mathbf{v}_2)_s\right) \Rightarrow \operatorname{sgn}(h) = \operatorname{sgn}(\theta_s).$$

It follows that the constraint

$$\begin{bmatrix} \mathbf{x}_\sigma^T & x_s \end{bmatrix} \begin{bmatrix} \theta_\sigma \\ \theta_s \end{bmatrix} \leqslant \kappa$$

is equivalent to the norm constraint for small enough displacements in the direction $\begin{bmatrix} \mathbf{h}_\sigma \\ h \end{bmatrix}$. Thus it gives the appropriate linearization.

REMARK 4 In Case 1, more than one fresh zero can be encountered, but this just causes the deletion of the corresponding components from σ. There is no analogue of degeneracy in linear programming here. This point is well made in Fletcher (1993).

REMARK 5 To detect and avoid dependences in the design matrix A, the test in Case 2 should be modified to one that selects the most infeasible of the $(\mathbf{v}_2)_i$ subject to the norm of the corresponding column of B being greater than a prescribed threshold. The implementation of this test requires that the problem has been scaled in a way which makes comparisons between columns of the design matrix meaningful.

The iteration can be started from $\mathbf{x} = 0$ by choosing an initial s to insert into σ and solving the resulting one-variable problem. Starting from this end of the problem has two advantages:

1. It puts the emphasis on building up the optimal σ by starting from a small base rather than by pruning a large one which could be ill-conditioned;
2. It permits the computation to proceed while at the same time building up the factorization (8), (9).

The solution of the unconstrained problem

$$\min_h \|h\mathbf{a}_s - \mathbf{b}\|_2^2$$

is

$$h = \frac{\mathbf{a}_s^T \mathbf{b}}{\|\mathbf{a}_s\|_2^2}.$$

Thus the norm bound κ must satisfy

$$\kappa < \max_s \frac{|\mathbf{a}_s^T \mathbf{b}|}{\|\mathbf{a}_s\|_2^2}$$

if a constrained solution is to be possible, and the appropriate choice of θ_s is

$$\theta_s = \operatorname{sgn}\left(\mathbf{a}_s^T \mathbf{b}\right)$$

where s is the maximizing index. The equations determining the solution of the constrained problem are

$$\begin{aligned} \mathbf{a}_s^T \mathbf{a}_s h - \mathbf{a}_s^T \mathbf{b} + \mu\theta_s &= 0, \\ h\theta_s &= \kappa. \end{aligned}$$

This gives

$$\begin{aligned} \mu &= \left|\mathbf{a}_s^T \mathbf{b}\right| - \kappa \mathbf{a}_s^T \mathbf{a}_s, \\ x_s &= \theta_s \kappa. \end{aligned}$$

It follows that the solution of this initial problem is sign feasible. Thus the next step increases $|\sigma|$ illustrating the building up of the approximation basis characteristic of this initialization.

4. Homotopy

The descent algorithm can only probe selection status information at a discrete set of values of κ. This could prove expensive if this information is needed in fine detail for $0 \leqslant \kappa \leqslant \kappa_{LS}$. In Osborne (1992) a piecewise linear homotopy is used to provide full and easily computed information on the optimal trajectory in a regression quantile problem. Based on this experience a homotopy approach to follow the trajectory of optimal variable selections as a function of κ is investigated. To develop this consider the basic necessary conditions (8), (9), (10):

$$\begin{aligned} U_1\mathbf{x}_\sigma &= \mathbf{c}_1 - \mu\mathbf{w}_\sigma, \\ \mu\mathbf{v}_2 &= B^T\mathbf{c}_2 + \mu U_{12}^T\mathbf{w}_\sigma, \\ \mu &= \frac{\mathbf{w}_\sigma^T\mathbf{c}_1 - \kappa}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}. \end{aligned}$$

The optimization problem minimizes a strictly convex objective over a convex feasible region and so has a unique minimizer for each κ. This implies that the trajectory must be continuous. Thus, if

$$-\mu\mathbf{e} < \mu\mathbf{v}_2 < \mu\mathbf{e}, \tag{15}$$

$$\left|x_{\sigma(i)}\right| > 0, \qquad i = 1, 2, \ldots, |\sigma|, \tag{16}$$

then this will be true also for points on the optimal trajectory corresponding to small enough perturbations in κ. Differentiating the above necessary conditions with respect to κ gives the equations

$$\frac{\mathrm{d}\mu}{\mathrm{d}\kappa} = -\frac{1}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}, \tag{17}$$

$$U_1\frac{\mathrm{d}\mathbf{x}_\sigma}{\mathrm{d}\kappa} = \frac{1}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}\mathbf{w}_\sigma, \tag{18}$$

$$\frac{\mathrm{d}(\mu\mathbf{v}_2)}{\mathrm{d}\kappa} = -\frac{1}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}U_{12}^T\mathbf{w}_\sigma. \tag{19}$$

The right-hand side of this system is constant provided the constraints (15) remain satisfied. It follows that there must be an interval of κ in which the solution variables $\mu, \mathbf{x}_\sigma, \mu\mathbf{v}_2$ are linear functions of κ, and the optimality conditions are satisfied pointwise. A key question is what happens at the end of such an interval when one of the constraints (15) becomes an equality. If a break occurs through a component of $\mathbf{x}_\sigma$ becoming zero then continuity requires that the corresponding component of θ_σ augment $\mathbf{v}_2$ and then remain feasible by moving away from its bound as κ increases. If a component of $\mathbf{v}_2$ reaches its bound of ± 1 then it augments θ_σ, and the corresponding component of $\mathbf{x}$ is added to $\mathbf{x}_\sigma$ and moves away from 0. Note that this updating corresponds to the rule (14) for ensuring sign feasibility in the descent method.

It is necessary to update or downdate the factorization (7) at each break point of the trajectory in order to actually carry out the homotopy calculations. The update step is considered here. It assumes that the break is caused by a component of $\mathbf{v}_2$ reaching a

bound and releasing the corresponding component of **x** from zero. If an **x** component reached 0 then the process must be reversed and the factorization is downdated. The specific assumptions are:

1. $|\sigma| = k < p$,
2. $|(\mathbf{v}_2)_1| \rightarrow 1$ as $\kappa \uparrow \kappa^*$, the critical value identifying the right-hand end of the interval of validity of the current piece of the homotopy, and
3. the homotopy is restarted with the new state component being $\overline{x}_{k+1}$ and corresponding sign ϕ.

REMARK 6 As with the descent algorithm, dependences can be removed from consideration by restricting this test to components of $\mathbf{v}_2$ that correspond to columns of B with norms exceeding a prescribed threshold.

The action here corresponds to freeing a component of **x** from zero and requires sweeping out the appropriate column of B in order to add another row and column to U_1. There is no real restriction in the choice of $(\mathbf{v}_2)_1$ to be the component reaching its bound, and it simplifies the notation in the subsequent calculation because the column added to U_1 is derived from the first column of the remainder of the matrix. The updating is done conveniently using a Householder transformation:

$$\left[I - 2 \begin{bmatrix} 0 \\ \mathbf{z} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{z}^T \end{bmatrix} \right] \begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} U_1 & [U_{12}]_1 \\ & \pi \|[B]_1\|_2 \end{bmatrix} & \begin{bmatrix} [U_{12}]_2 \\ \mathbf{u}^T \\ \overline{B} \end{bmatrix} \\ 0 & \end{bmatrix},$$

where the quantities defining the transformation are

$$\mathbf{z} = \gamma^{-1} \left([B]_1 - \pi \|[B]_1\|_2 \, \mathbf{e}_1 \right),$$
$$\gamma^2 = 2 \|[B]_1\|_2 \left(\|[B]_1\|_2 - \pi \, ([B]_1)_1 \right),$$

$\pi = \operatorname{sgn}(B_{11})$, and where the notation is used that $[G]_1$ denotes the first column of a matrix G, while $[G]_2$ is the remaining submatrix. Here $[B]_1$ maps into a multiple $\pi \|[B]_1\|_2$ of the first unit vector while the remainder of the transformation gives

$$[B]_2 \rightarrow \begin{bmatrix} \mathbf{u}^T \\ \overline{B} \end{bmatrix}.$$

The change in the right-hand side is

$$\mathbf{c} \rightarrow \begin{bmatrix} \mathbf{c}_1 \\ c_{11} \\ \overline{\mathbf{c}}_2 \end{bmatrix},$$

where

$$\begin{aligned} c_{11} &= \mathbf{e}_1^T \left(I - 2\mathbf{z}\mathbf{z}^T \right) \mathbf{c}_2, \\ &= (\mathbf{c}_2)_1 - 2 \frac{\left(([B]_1)_1 - \pi \|[B]_1\|_2 \right) \left([B]_1 - \pi \|[B]_1\|_2 \, \mathbf{e}_1 \right)^T \mathbf{c}_2}{2 \|[B]_1\|_2 \left(\|[B]_1\|_2 - \pi \, ([B]_1)_1 \right)}, \\ &= \frac{[B]_1^T \mathbf{c}_2}{\pi \|[B]_1\|_2} = \frac{\mu \left((\mathbf{v}_2)_1 - [U_{12}]^T \mathbf{w}_\sigma \right)}{\pi \|[B]_1\|_2}. \end{aligned} \tag{20}$$

The last equation follows from equation (9) and expresses the condition that $(\mathbf{v}_2)_1$ reaches its bound when $\kappa = \kappa^*$:

$$\mu\,(\mathbf{v}_2)_1 = [B]_1^T\,\mathbf{c}_2 + \mu\,[U_{12}]_1^T\,\mathbf{w}_\sigma .$$

The updated index set is

$$\overline{\sigma} = \sigma \cup \{k+1\}\,,$$

and bars are used to denote updated quantities. To update $\mathbf{w}_\sigma$ we have to select the correct sign ϕ for $\overline{x}_{k+1}$. Then

$$\overline{\mathbf{w}}_{\overline{\sigma}} = \begin{bmatrix} U_1^{-T} \\ -[U_{12}]_1^T\,U_1^{-T}/\pi\,\|[B]_1\|_2 & 1/\pi\,\|[B]_1\|_2 \end{bmatrix} \begin{bmatrix} \theta_\sigma \\ \phi \end{bmatrix}.$$

It follows that

$$\overline{w}_{k+1} = \left[-[U_{12}]_1^T\,\mathbf{w} + \phi\right]/\pi\,\|[B]_1\|_2\,. \tag{21}$$

Explicit calculations can now be made to verify continuity and sign feasibility. For example, consider the updated state value

$$\begin{aligned} \overline{x}_{k+1} &= \frac{c_{11}}{\pi\,\|[B]_1\|_2} - \frac{\mu}{\pi\,\|[B]_1\|_2}\mathbf{e}_{k+1}^T\overline{U}_1^{-T}\begin{bmatrix} \theta_\sigma \\ \phi \end{bmatrix}, \\ &= \frac{c_{11}}{\pi\,\|[B]_1\|_2} - \frac{\mu w_{k+1}}{\pi\,\|[B]_1\|_2}, \\ &= \frac{\mu}{\left(\pi\,\|[B]_1\|_2\right)^2}\left((\mathbf{v}_2)_1 - \phi\right). \end{aligned} \tag{22}$$

It follows that a necessary condition for $\overline{x}_{k+1}$ to be continuous with 0 when $\kappa = \kappa^*$ is

$$(\mathbf{v}_2)_1 - \phi = 0, \tag{23}$$

and this is identical to (14) in this case. It remains to verify that this choice of ϕ ensures sign feasibility when $\kappa > \kappa^*$. From (19) it follows that

$$\frac{\mathrm{d}\mu}{\mathrm{d}\kappa}\,(\mathbf{v}_2)_1 + \mu\frac{\mathrm{d}\,(\mathbf{v}_2)_1}{\mathrm{d}\kappa} = -\frac{1}{\mathbf{w}^T\mathbf{w}}\,[U_{12}]_1^T\,\mathbf{w}$$

whence

$$\mu\frac{\mathrm{d}\,(\mathbf{v}_2)_1}{\mathrm{d}\kappa} = \frac{1}{\mathbf{w}^T\mathbf{w}}\left((\mathbf{v}_2)_1 - [U_{12}]_1^T\,\mathbf{w}\right).$$

It follows from (22) and (23) that

$$\mathrm{sgn}\left(\frac{\mathrm{d}\overline{x}_{k+1}}{\mathrm{d}\kappa}\right) = \mathrm{sgn}(\overline{w}_{k+1}) = \mathrm{sgn}\left(\frac{\mathrm{d}\,(\mathbf{v}_2)_1}{\mathrm{d}\kappa}\right) = \mathrm{sgn}\,((\mathbf{v}_2)_1)\,.$$

Similar calculations can provide explicit verification of the continuity of μ and of the components of $\mathbf{v}_2$ not at a bound.

Two useful properties of the homotopy are noted:

PROPERTY 1 There are at most a finite number of steps in the homotopy. In particular, no index set σ can repeat in the homotopy. This follows on noting that if σ is appropriate at κ_1 and κ_2 then linearity forces it to be valid at all points in between.

PROPERTY 2 The sum of squares of residuals is monotonic decreasing as κ increases. Differentiating gives

$$\begin{aligned}\tfrac{1}{2}\frac{\mathrm{d}\,\|\mathbf{r}\|_2^2}{\mathrm{d}\kappa} &= \mathbf{r}^T A\frac{\mathrm{d}\mathbf{x}_\sigma}{\mathrm{d}\kappa},\\ &= -\frac{1}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma}\mu\theta_\sigma^T U_1^{-1}\mathbf{w}_\sigma,\\ &= -\mu < 0.\end{aligned}$$

This suggests that the typical action as κ increases through a breakpoint is variable addition.

REMARK 7 The following computation suggests a method for finding starting values for the homotopy. Assume the component of maximum modulus of $A^T\mathbf{b}$ is unique, and that the corresponding index is $s = 1$. Then there is an optimal solution for κ small enough in which only $x_1 \neq 0$. This is given by

$$\sigma = \{1\}, \qquad \mu = \left|\mathbf{a}_1^T\mathbf{b}\right| - \kappa\mathbf{a}_1^T\mathbf{a}_1, \qquad x = \theta_1\kappa.$$

To verify this, write the Kuhn–Tucker conditions as

$$\begin{bmatrix} \mathbf{a}_1^T \\ A_2^T \end{bmatrix}(x\mathbf{a}_1 - \mathbf{b}) = -\mu\begin{bmatrix} \theta_1 \\ \mathbf{v}_2 \end{bmatrix}.$$

The condition that the components of $\mathbf{v}_2$ lie in the correct range can be written

$$\max_{i>1}\left|\frac{\theta_1\kappa\mathbf{a}_i^T\mathbf{a}_1 - \mathbf{a}_i^T\mathbf{b}}{\left|\mathbf{a}_1^T\mathbf{b}\right| - \kappa\mathbf{a}_1^T\mathbf{a}_1}\right| < 1.$$

It is satisfied for small enough κ as a consequence of the assumption that there is a single component of $A^T\mathbf{b}$ of maximum modulus. The value of i for which equality holds for the smallest value of $\kappa > 0$ determines the range of validity of this solution.

REMARK 8 The above discussion has assumed that a breakpoint in the trajectory occurs by a single state or multiplier vector component changing status. However, the possibility of a multiple change, although unlikely, has not been excluded. If this was to cause cycling behaviour then a brute force remedy is to make a small increment in κ and then apply the descent algorithm using the current values as initial conditions. Use of an analogous procedure is discussed in Osborne (1992).

5. On problem scaling

Central to any consideration of variable selection is a requirement of scale independence in the criteria for making the necessary selection choices in the preceding algorithms. The

residual vector $\mathbf{r}$ is independent of the diagonal scaling of the columns of A for we can write

$$\mathbf{r} = A\mathbf{x} - \mathbf{b} = AD\left(D^{-1}\mathbf{x}\right) - \mathbf{b}.$$

However, scaling questions are important for $\mathbf{x}$ and hence necessarily have implications for the norm constraint (3). For example, is it legitimate to start by scaling the columns of A to have unit length in order to impose a form of comparability on the variables? Let $\mathbf{a}_i = A\mathbf{e}_i$ be the ith column of A. Use of the column scaled form of A corresponds to an initial scaling of the constraint given by

$$\kappa - \sum_{i=1}^{p} \|\mathbf{a}_i\|_2 \, |x_i| \geqslant 0.$$

This form is invariant under diagonal scaling, and has the further advantages:

1. The scaling depends only on the structure of the set of basis functions and so is independent of the particular vector of observations.
2. If κ is small then the first variable selected is the same as that chosen by the standard stepwise regression procedure.

Column norms of the design matrix are scaled to have length 1 in our implementations of both the descent and homotopy algorithms.

An alternative that could be considered is one that has been used to good effect in implementing a version of the Levenberg algorithm for nonlinear least squares (Moré, 1978). This scales each component x_i by $|\nabla F_i|$ where $F(\mathbf{x})$ is the objective function and the weight $|\nabla F_i|$ is evaluated at the beginning of the current step of the Levenberg iteration. Clearly it is invariant to diagonal matrix transformations. Also, in the Lasso, the weights can be evaluated at the initial point $\mathbf{x}_0$ because the design matrix is fixed. This gives the modified constraint

$$\kappa - \sum_{i=1}^{p} \left|\mathbf{r}\,(\mathbf{x}_0)^T \, A\mathbf{e}_i\right| |x_i| \geqslant 0.$$

It reduces to

$$\kappa - \sum_{i=1}^{p} \left|\mathbf{b}^T \mathbf{a}_i\right| |x_i| \geqslant 0, \tag{24}$$

in the case that the initial value is $\mathbf{x}_0 = 0$. However, this form has the interesting disadvantage that if the weights corresponding to (24) are chosen then the corresponding initial step is degenerate in the sense that all variables prove to meet the selection criteria of our algorithms.

6. Implementation

Modified Gram–Schmidt orthogonalization provides an elegant base for implementing the above algorithms. The basic idea is to consider the tableau array

$$W = \begin{bmatrix} A_1^T & A_1^T\mathbf{b} & I & 0 \\ A_2^T & A_2^T\mathbf{b} & 0 & I \\ -\mathbf{b}^T & 0 & 0 & 0 \end{bmatrix}.$$

At step i there will be $(i-1)$ orthonormalized rows corresponding to A_1^T in the above partition. The next sequence of operations is as follows.

1. Select a row of A_2^T to act as pivotal row. Let this be $(A_2)_1^T$. Compute
$$d_1^2 = \left\| (A_2)_1^T \right\|_2^2 .$$
2. Form the scalar products d_{1j} of the remaining rows of A_2^T (and including the last row of W) with the pivotal row. $d_{1j} = (A_2)_1^T (A_2)_j$.
3. Orthogonalize the pivotal row to each of the remaining rows (including the last row of W), but apply the operation to the full matrix W
$$(W_2)_j = (W_2)_j - \frac{d_{1j}}{d_1^2} (W_2)_1 .$$
4. Scale the pivotal row
$$(W_2)_1 \leftarrow \frac{1}{d_1} (W_2)_1 .$$
5. Repartition W by adding the pivotal row to W_1.

At this point the following identifications can be made using the partial QR factorization:

$$\begin{bmatrix} A_1 & A_2 \end{bmatrix} \rightarrow \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U_1 & U_{12} \\ & B \end{bmatrix}.$$

Modified Gram–Schmidt applied to A is equivalent to premultiplying A^T by

$$\begin{bmatrix} U_1^{-T} & 0 \\ -U_{12}^T U_1^{-T} & I \end{bmatrix}.$$

Applying this transformation, suitably augmented to take account of the last row, to W gives

$$W \rightarrow \begin{bmatrix} Q_1^T & \mathbf{c}_1 & U_1^{-T} & 0 \\ B^T Q_2^T & B^T \mathbf{c}_2 & -U_{12}^T U_1^{-T} & I \\ \mathbf{r}(\mathbf{x}_\sigma^{LS})^T & \mathbf{b}^T Q_1 Q_1^T \mathbf{b} & \left(\mathbf{x}_\sigma^{LS}\right)^T & 0 \end{bmatrix}.$$

Not only does the solution of the least squares problem defined by σ appear explicitly, but quantities needed to implement the descent and homotopy algorithms can be derived readily from the tableau. As an illustration note that

$$\begin{bmatrix} U_1^{-T} \\ -U_{12}^T U_1^{-T} \\ \left(\mathbf{x}_\sigma^{LS}\right)^T \end{bmatrix} \theta_\sigma = \begin{bmatrix} \mathbf{w}_\sigma \\ \left(\mathbf{w}_\sigma^T \mathbf{w}_\sigma\right) (\mathrm{d}(\mu \mathbf{v}_2)/\mathrm{d}\kappa) \\ \left(\mathbf{w}_\sigma^T \mathbf{w}_\sigma\right) \mu + \kappa \end{bmatrix}.$$

TABLE 1
Progress of homotopy, Hald data

κ	μ	x_1	x_2	x_3	x_4	x_5
0·00	0·989	0·0	0·0	0·0	0·0	0·0
0·18	0·806	17·635	0·0	0·0	0·0	0·0
0·73	0·270	44·072	0·0	0·52433	0·0	0·0
1·02	0·9–2	52·270	1·4152	0·65726	0·0	0·0
1·04	0·2–4	48·203	1·6952	0·65692	0·24972	0·0
1·13	0·00	62·405	1·5511	0·51017	0·10191	−0·14406

TABLE 2
Multiplier vectors, Hald data

v_1	v_2	v_3	v_4	v_5
1·0	0·8640	0·9917	0·8487	0·8226
1·0	0·8791	1·0	0·8401	0·8093
1·0	1·0	1·0	0·7684	0·8045
1·0	1·0	1·0	1·0	0·7077
1·0	1·0	1·0	1·0	−1·0

7. Numerical results

Results are presented for the homotopy algorithm applied to the Hald data set (Draper & Smith, 1998). It is considered here as a data set with 13 observations and 5 variables in order to show the intercept variable explicitly (it can be removed as part of a process of centring and scaling in standard least squares regression). Here the intercept corresponds to variable 1. This data set has been used as a good example to illustrate stepwise regression procedures because it shows both variable addition and deletion with the test being an F test for significance at the 10% level. The homotopy algorithm appears better organized. The progress towards the solution is illustrated in Table 1. This gives the value of the homotopy parameter κ, the multiplier μ, and the (unscaled) solution values $\mathbf{x}$ at the breakpoints. It shows that only addition steps are taken, that the intercept term is added first, and that the multiplier has already become small ($0{\cdot}9d - 2$) by the time the fourth variable to enter has been selected. If the computation had been stopped at this point then the $\mathbf{x}^{\sigma}_{LS}$ corresponds to the final solution of the stepwise regression. A check on the solution process is available because when $\mu = 0$ then $\mathbf{x} = \mathbf{x}_{LS}$ which is available from the tableau.

Table 2 gives the corresponding values of the multiplier vector $\mathbf{v}$. It shows how the values of θ_σ build up as the algorithm progresses.

An example that makes variable deletion steps in order to achieve sign feasibility is obtained by considering the descent algorithm applied to the Hald data with the constraint bound being given by $\kappa = 1{\cdot}03$. The results, including unscaled solution values, are given in Table 3. The initial steps correspond to stepwise regression steps which continue until the current $\mathbf{x}$ violates the constraint bound. The implementation of the descent algorithm allowed only for the second (downdating) option when a zero component of

TABLE 3

Descent algorithm, constraint bound $k = 103$

#	μ	x_1	x_2	x_3	x_4	x_5
1	0	95·423				0·0
5	0	117·57				−0·7382
2	2·661-2	95·924	0·0			−0·1023
3	2·281-2	81·874	1·7058	0·0		−0·0460
−5	5·529-4	54·4608	1·4848	0·6129		0·0
4	2·832-3	52·488	1·4529	0·6608	0·0	0·0
	5·823-3	50·620	1·5288	0·6571	0·1012	0·0

x is encountered. It is interesting that the steps actually mirror the stepwise regression computation reported in Draper & Smith (1998). In both cases variable 5 enters early (actually as a stepwise regression step) and is replaced at the penultimate step.

8. Conclusion

Two algorithms of complementary nature have been presented for computing the Lasso. Both appear capable of efficient implementation and prove effective in practice. The descent algorithm can operate as a probe for solving the selection problem for a particular value of κ, while the homotopy procedure gives global results. The descent algorithm can be used to generate initial values as a preliminary to exploring particular ranges of κ using the homotopy procedure. The functioning of the homotopy algorithm provides a satisfactory explanation of the variable selection mechanism. The usefulness of the Lasso depends on the successful development of an associated statistical theory. There is informal evidence that this problem is attracting a good deal of attention.

Acknowledgements

Helpful comments by a referee and by the Associate Editor have led to significant improvements in presentation.

REFERENCES

CLARK, D. I. & OSBORNE, M. R. 1988 On linear restricted and interval least-squares problems. *IMA J. Numer. Anal.* **8**, 23–36.

DRAPER, N. H. & SMITH, H. 1966 *Applied Regression Analysis*. [3rd edition, 1998], Chichester: Wiley.

FLETCHER, R. 1993 Resolving degeneracy in quadratic programming. *Ann. Op. Res.* **47**, 307–334.

MORÉ, J. J. 1978 The Levenberg-Marquardt algorithm: implementation and theory. *Proc. Numerical Analysis. (Dundee 1977)*, Lecture Notes in Mathematics No. 630, (G. A. Watson ed.). Springer, Berlin: pp. 105–116.

OSBORNE, M. R. 1985 *Finite Algorithms in Optimization and Data Analysis*. Chichester: Wiley.

OSBORNE, M. R. 1992 An effective method for computing regression quantiles. *IMA J. Numer. Anal.* **12**, 151–166.

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