

A homotopy algorithm for the quantile regression lasso and related piecewise linear problems

M.R. Osborne* B.A. Turlach†

October 13, 2009

Abstract

We present a proof of concept that the homotopy algorithm of Osborne, Presnell and Turlach [9] which has proved such an effective optimal path following method for implementing Tibshirani's "lasso" for variable selection can be extended to polyhedral objectives in examples such as the quantile regression lasso. The new algorithm introduces the novel feature that it requires two homotopy sequences involving continuation steps with respect to both the constraint bound and the Lagrange multiplier. Performance is illustrated by application to several standard data sets and these results are compared to calculations made with the original lasso homotopy program. This permits an assessment to be made of the computational complexity both of the new method and of the closely related linear programming post optimality procedures which generate essentially identical solution trajectories. This comparison is not favourable to the polyhedral objective selection methods. However, the new method still provides an effective computational procedure, plus it has distinct implementation advantages over the linear programming approaches to the polyhedral objective problem. The computational difficulty is explained and the problem that needs to be resolved identified.

*Mathematical Sciences Institute, Australian National University

†School of Mathematics and Statistics, University of Western Australia

1 Introduction

The problem motivating this investigation is the classic variable selection problem of determining a subset of the parameters $\boldsymbol{\beta} \in R^p$ in the linear model

$$\mathbf{r} = \mathbf{y} - X\boldsymbol{\beta}, \quad (1)$$

where $\mathbf{r} \in R^n$ is the residual vector, $\mathbf{y} \in R^n$ is the vector of observations, and $X : R^p \rightarrow R^n$ is the design matrix which is assumed to have full (typically column) rank, in order to provide an adequate but parsimonious representation of the problem data. If an l_1 inequality constraint

$$\sum_{i=1}^p |\beta_i| \leq \kappa \quad (2)$$

is imposed on $\boldsymbol{\beta}$ in the least squares regression formulation of the estimation problem then the number of nonzero components of $\boldsymbol{\beta}$ in the optimum solution increases as the constraint bound is increased away from zero. The problem eventually becomes unconstrained for large enough values of the bound so that all parameters can be expected to be involved in the solution. Tibshirani suggested in [11] that study of the sets of non-zero parameter estimates as a function of the constraint bound provided a basis for variable selection and gave the acronym “lasso” to the resulting process. Improved algorithms were developed in [9]. In particular, a “homotopy” algorithm which follows the piecewise linear optimal solution path as a function of the constraint bound was recommended. Subsequently significant evidence has accumulated that this algorithm is particularly efficient in many circumstances, frequently taking at most few more steps than the number of variables selected. When this is true then it rivals the cost of an unconstrained least squares algorithm for the full set of variables while providing considerable additional information. This has led to further applications. For example, it has proved distinctly effective in applications in compressed sensing [3]. Also, generalisations have been made to problems in which smoothness is preserved in multiple objectives and the form of constraint varied appropriately [12] while further applications have been mooted to piecewise quadratic objectives with continuous first derivative, and to the Huber-M estimator which involves a mixed piecewise linear, quadratic objective [10]. Both these problems have sufficient smoothness to ensure that the necessary conditions vary continuously at points where the objective function changes its structure. However, this property is not shared by the homotopy equations in either case so that an extra layer of algorithmic complexity is added. Different selection properties can be addressed by varying the form of constraint. For example, [1]

recommend using what is in effect the signed rank objective [5] as a constraint. They claim this has an advantage in suggesting the development of cluster variables as an aid to interpretation when the predictor variables are relatively highly correlated and make similar contributions to the response. A different suggestion for attacking similar problems is the so-called “elastic net” [15].

The other class of applications that has attracted recent attention is that corresponding to the relaxation of the smoothness of the objective. Examples include piecewise linear objectives for quantile regression [6], the training of support vector machines [14], [13], and [16] where the form of constraint used in [12] is used to develop a simultaneous variable selection procedure for simultaneous quantile regressions. A suggested compressed sensing application involves the minimization of the l_1 norm of the parameters subject to a maximum norm constraint on the components of the residual vector [2]. Here we develop a general homotopy type lasso algorithm for variable selection in a form which is applicable to piecewise linear objective functions of quite general type. There is one striking difference in the properties of the optimal homotopy trajectory between the case when the objective is at least once continuously differentiable and the non-smooth case when the objective is piecewise linear. In the former case the Lagrange multiplier for the l_1 constraint is a piecewise linear, continuous function of the constraint bound with the characteristic property that it decreases steadily from its initial positive value at $\kappa = 0$ to 0 for κ large enough. In contrast, the corresponding Lagrange multiplier associated with a piecewise linear objective is a decreasing step function with jumps at the points where the set of non-zero components of β changes, and it is necessary to include an explicit multiplier update phase into the computation. This can be developed quite generally given a generic form for the subdifferential of the polyhedral constraint. However, here we restrict ourselves to the classic l_1 form used in the lasso.

Initially the l_1 objective is considered. This is just the special case $\tau = .5$ of the lasso for the quantile regression problem

$$\min_{\beta} \sum_{i=1}^n (1 - \tau)(-r_i)_+ + \tau(r_i)_+, \quad 0 < \tau < 1, \quad (3)$$

where the r_i are residuals in a linear model fit. There is a hint that computational complexity is potentially an important consideration in designing algorithms for piecewise linear objectives in these problems. This arises because complexity is already a problem in simplicial algorithms for the l_1 estimation problem where it proves important to incorporate an effective line-search in order to make large correction steps especially in the initial

stages of the computation [8]. If this is not done then a sequence of small steps is made as the residual vector adjusts to adapt itself to the form required by the necessary conditions. But every residual changing sign triggers a point of nondifferentiability of the objective. Typically $O(n)$ changes at least are expected in this adaptive process given general initial conditions. A form of this problem occurs also in the homotopy algorithm. The reason for this is that frequently a succession of multiplier update steps occur with the successive parameter estimates contained exactly in the same orthant of a fixed subspace. The corresponding increments in these estimates correspond to descent steps for the objective function in this subspace. However, typically the objective minimum in this subspace is not on the homotopy path (does not satisfy the homotopy necessary conditions) so the basic l_1 strategy requires modification.

The novelty in our approach lies in the manner in which the non-smooth necessary conditions are used explicitly in the basic algorithm structure. This involves switching between the constrained and Lagrangian forms of the problem in order to continue with respect to the constraint bound (increase) and then to update (decrease) the Lagrangian parameter. In particular decisions about adding and subtracting variables are based on subgradient vector components attaining or departing from bounds. This basic pattern generalises to necessary conditions for more complicated piecewise linear structures. It is equivalent to the use of simplex algorithm post-optimality techniques applied to a linear programming formulation of the problem. It has significant advantages both in displaying problem structure and in avoiding the introduction of the slack variables required in the linear programming formulations. This enables the direct use of the design matrix as the basic data structure in the homotopy approach.

The lasso algorithm is developed in the next section. This is followed by a discussion of numerical experience together with an indication of the source of the additional work required in the non-smooth objective case. The detailed algebraic calculations to justify the homotopy algorithm have been postponed to the appendix. We describe this project as a “proof of concept” exercise. This means two separate issues have not been explored in detail.

1. The numerical work is thorough but has not been exhaustively optimised. For example, economised updating and downdating of matrix factorizations has not been implemented. There could be scope for this and it could involve, for example, keeping two lines of factorization for the two classes of continuation. Treatment of degeneracy is more important in the linear programming post optimality studies where it can have a structural role which is directly related to the use of a single

continuation variable [8], but it can and does occur in our approach. It is discussed here as one aspect of the computations.

2. There is no doubt of the utility of extracting an intercept term and standardising the design accordingly in ordinary least squares estimation. This is not quite so clear in the basic lasso as the l_1 norm does not respect orthogonality. It becomes less clear when the objective is polyhedral, and further complicated if a variable rather than its absolute value is entered into the constraint term. The calculations presented here took the simplest option that permitted direct comparison with the original lasso homotopy program [9]. This permitted the design to be augmented optionally by a column of 1's. Columns of the design and of the response were then scaled to have length 1. The augmented design has been used in the quoted numerical results.

2 Lasso for the l_1 objective

2.1 Necessary conditions

The basic idea of the homotopy algorithm is to follow the evolution of the parameters $\boldsymbol{\beta}$ as a piecewise linear function of the constraint bound κ . Two forms of the problem prove important in developing the homotopy trajectory. These are

Constrained form This is used to follow $\boldsymbol{\beta}$ as a piecewise linear function of the constraint bound κ . This problem is

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n |r_i|, \quad \sum_{i=1}^p |\beta_i| \leq \kappa. \quad (4)$$

If $|y_i| > 0$, $i = 1, 2, \dots, n$, then the problem has the solution $\boldsymbol{\beta} = 0$ when $\kappa = 0$. Increasing κ permits components of $\boldsymbol{\beta}$ to move from 0 (become active), typically one at a time. If κ is large enough then $\boldsymbol{\beta}$ is essentially unconstrained so the problem reduces to an unconstrained l_1 minimization problem. One consequence is that the associated Lagrange multiplier is $\lambda = 0$.

Lagrangian form This is used to update the subgradient vector components in the necessary conditions as a parametric function of the Lagrange multiplier λ with κ fixed. This update is performed whenever

there is a change in the set of active $\boldsymbol{\beta}$ components. The problem Lagrangian is

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^n |r_i| + \lambda \left\{ \sum_{i=1}^p |\beta_i| - \kappa \right\}.$$

It is convex if $\lambda \geq 0$. The form used here follows from the necessary condition on $\boldsymbol{\beta}$ expressed in subgradient form

$$0 \in \partial_{\boldsymbol{\beta}} \mathcal{L} \{ \boldsymbol{\beta}, \lambda \} = \partial_{\boldsymbol{\beta}} \sum_{i=1}^n |r_i| + \lambda \partial_{\boldsymbol{\beta}} \sum_{i=1}^p |\beta_i|. \quad (5)$$

The key to its utility is a consequence of κ not appearing explicitly in this expression. Setting the initial value of λ corresponding to $\kappa = 0$ is described in subsection 2.4.

In non-degenerate cases it will be possible to separate zero and non-zero values of both \mathbf{r} and $\boldsymbol{\beta}$ into complementary classes and to take uniquely specified actions in the key situations that occur when these classes have to be redefined. Let

$$\sigma = \{i : r_i = 0\}, \quad \psi = \{i : x_i \neq 0\}, \quad (6)$$

and denote the set complements by

$$\sigma^c = \{i : r_i \neq 0\}, \quad \psi^c = \{i : x_i = 0\}. \quad (7)$$

In addition define permutation matrices $P_\sigma : R^n \rightarrow R^n$ and $Q_\psi : R^p \rightarrow R^p$ by

$$P_\sigma \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}, \quad \begin{cases} (\mathbf{r}_1)_i = r_{\sigma^c(i)} \neq 0, & i = 1, 2, \dots, n - |\sigma|, \\ (\mathbf{r}_2)_i = r_{\sigma(i)} = 0, & i = 1, 2, \dots, |\sigma| \end{cases},$$

$$Q_\psi \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \quad \begin{cases} (\boldsymbol{\beta}_1)_i = \beta_{\psi(i)} \neq 0, & i = 1, 2, \dots, |\psi|, \\ (\boldsymbol{\beta}_2)_i = \beta_{\psi^c(i)} = 0, & i = 1, 2, \dots, p - |\psi| \end{cases}.$$

To specify necessary conditions set

$$P_\sigma X Q_\psi^T = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \quad (8)$$

$$P_\sigma \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}. \quad (9)$$

Let $[\boldsymbol{\theta}_\sigma^T \quad \mathbf{v}_\sigma^T] \in \partial\|P_\sigma \mathbf{r}\|_1$, and $[\boldsymbol{\theta}_\psi^T \quad \mathbf{u}_\psi^T] \in \partial\|Q_\psi \boldsymbol{\beta}\|_1$ be subgradient vectors associated with the corresponding l_1 norm functions. Then the necessary conditions (5) are

$$[\boldsymbol{\theta}_\sigma^T \quad \mathbf{v}_\sigma^T] \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} = \lambda [\boldsymbol{\theta}_\psi^T \quad \mathbf{u}_\psi^T], \quad \lambda \geq 0, \quad (10)$$

$$-1 \leq v_i \leq 1, \quad i = 1, 2, \dots, |\sigma|, \quad (11)$$

$$-1 \leq u_i \leq 1, \quad i = 1, 2, \dots, |\psi^c|, \quad (12)$$

$$\boldsymbol{\theta}_\sigma^T \mathbf{r}_1 = \|\mathbf{r}\|_1, \quad (13)$$

$$\|\boldsymbol{\beta}\|_1 = [\boldsymbol{\theta}_\psi^T \quad \mathbf{u}_\psi^T] Q_\psi \boldsymbol{\beta} = \boldsymbol{\theta}_\psi^T \boldsymbol{\beta}_1 \leq \kappa. \quad (14)$$

2.2 Varying κ

The evolution of the set of active parameter estimates as functions of κ follows standard homotopy procedures. Consider $\kappa > 0$, κ in an open interval such that the optimal $\boldsymbol{\beta}$ is determined by a fixed set of zero residuals $r_i = 0$, $i \in \sigma$. These can be identified as l_1 necessary conditions except that here the active bound constraint picks up a degree of freedom so that $|\sigma| = |\psi| - 1$. Then

$$\boldsymbol{\theta}_\psi^T \boldsymbol{\beta}_1 = \kappa, \quad l_1 \text{ norm condition}, \quad (15)$$

$$X_{21} \boldsymbol{\beta}_1 = \mathbf{y}_2, \quad \text{zero residual conditions}. \quad (16)$$

Differentiating with respect to κ gives

$$\boldsymbol{\theta}_\psi^T \frac{d\boldsymbol{\beta}_1}{d\kappa} = 1, \quad (17)$$

$$X_{21} \frac{d\boldsymbol{\beta}_1}{d\kappa} = 0, \quad (18)$$

$$\Rightarrow \frac{d\boldsymbol{\beta}_1}{d\kappa} = \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \mathbf{e}_1. \quad (19)$$

It follows from the partitioning of the necessary conditions (10) that

$$\boldsymbol{\theta}_\sigma^T X_{11} + \mathbf{v}_\sigma^T X_{21} = \lambda \boldsymbol{\theta}_\psi^T.$$

Differentiating with respect to κ gives

$$\frac{d\mathbf{v}_\sigma^T}{d\kappa} X_{21} = \frac{d\lambda}{d\kappa} \boldsymbol{\theta}_\psi^T. \quad (20)$$

To show $\frac{d\lambda}{d\kappa} = 0$ multiply by $\frac{d\boldsymbol{\beta}_1}{d\kappa}$ and use (18) and (17) to obtain

$$0 = \frac{d\mathbf{v}_\sigma^T}{d\kappa} X_{21} \frac{d\boldsymbol{\beta}_1}{d\kappa} = -\frac{d\lambda}{d\kappa}.$$

It now follows from (18) that $\frac{d\mathbf{v}_\sigma}{d\kappa} = 0$ given X_{21} has full rank. The second component of the partitioning of (10) gives

$$\boldsymbol{\theta}_\sigma^T X_{12} + \mathbf{v}_\sigma^T X_{22} = \lambda \mathbf{u}_\psi^T.$$

It follows directly from this that $\frac{d\mathbf{u}_\psi}{d\kappa} = 0$. Thus λ , \mathbf{v}_σ , and \mathbf{u}_ψ are constant in intervals of κ with fixed zero residual sets and can have jumps only when it is necessary to vary the partitioning of the optimality conditions. One connection between the above development and a direct optimisation approach is the following.

Lemma 1 $\frac{d\boldsymbol{\beta}}{d\kappa}$ is a direction of descent for minimising $\|\mathbf{r}\|_1$.

Proof. The calculation of the directional derivative can proceed as follows.

$$\begin{aligned} \|\mathbf{r}\|_1' \left(\boldsymbol{\beta} : \frac{d\boldsymbol{\beta}}{d\kappa} \right) &= \sup_{\mathbf{z} \in \partial \|\mathbf{r}\|_1} \mathbf{z}^T X \frac{d\boldsymbol{\beta}}{d\kappa}, \\ &= \boldsymbol{\theta}_\sigma^T X_{11} \frac{d\boldsymbol{\beta}_1}{d\kappa} = -\lambda \boldsymbol{\theta}_\psi^T \frac{d\boldsymbol{\beta}_1}{d\kappa}, \\ &= -\lambda \boldsymbol{\theta}_\psi^T \begin{bmatrix} X_{21} \\ \boldsymbol{\theta}_\psi \end{bmatrix}^{-1} \mathbf{e}_{|\psi|} = -\lambda < 0. \end{aligned}$$

where (18) has been used. Here $\mathbf{e}_{|\psi|}$ is the unit vector with 1 in the $|\psi|$ place.

■

A similar result is given in [7].

There are two possibilities for terminating the step in the direction $\frac{d\boldsymbol{\beta}_1}{d\kappa}$.

- 1 A new zero residual is generated. It is assumed that the change in the characteristic set corresponds to row $\sigma^c(k)$. Set $\mathbf{x}_k^T = \mathbf{e}_k^T X_{11}$, $y_k = \mathbf{e}_k^T \mathbf{y}_1$, and define r_k similarly. The update of the index set σ to take account of the new zero residual is $\sigma \leftarrow \sigma \cup \{\sigma^c(k)\}$. If σ is reordered such that $\sigma^c(k) \rightarrow \sigma(1)$ then the corresponding subgradient component is given by $v_1(\lambda_0) = \theta_k$.
- 2 A component of $\boldsymbol{\beta}_1$ vanishes before a new zero residual is reached. Let this component corresponds to index $\psi(j)$. To preserve the necessary conditions it is necessary to update the index set pointing to the $\boldsymbol{\beta}$ components, $(\psi \leftarrow \psi \setminus \{\psi(j)\})$. It is assumed that $\psi(j) \rightarrow \psi^c(1)$. The corresponding subgradient component is $u_1 = \theta_j$.

At this point the computation switches to consider continuation with respect to the Lagrange multiplier as parameter. Optimality as λ is reduced requires that the new subgradient components move from their bound values into the interior of their region of feasibility. This is guaranteed by the following results.

Lemma 2 *Optimality is preserved in a small enough reduction $\Delta\lambda$ in λ provided $\theta_k v_1(\lambda_0 - \Delta\lambda) < 1$. This is equivalent to the following.*

$$\theta_k \frac{dv_1(\lambda_0)}{d\lambda} \geq 0. \quad (21)$$

Lemma 3 *Optimality is preserved in a small enough reduction $\Delta\lambda$ in λ provided $\theta_j u_1(\lambda_0 - \Delta\lambda) \leq 1$. This result is equivalent to*

$$\theta_j \frac{du_1(\lambda_0)}{d\lambda} \geq 0. \quad (22)$$

These results are proved in appendix A.

2.3 Varying λ

Let the current value of the Lagrange multiplier be $\lambda = \lambda_0$. Also, let the current interval of linear dependence of the parameters $\boldsymbol{\beta}$ on κ be $[\kappa_0 \leq \kappa \leq \kappa_1]$. There are two cases which can interrupt this dependence:

1. Assume the κ homotopy step is terminated by the occurrence of the new zero residual r_k , $r_k(\kappa_1) = 0$, corresponding to the first termination possibility in the κ phase. In this case the updated quantities are given by

$$\bar{X}_{21} \leftarrow \begin{bmatrix} \mathbf{x}_k^T \\ X_{21} \end{bmatrix}, \quad \bar{\mathbf{y}}_2 \leftarrow \begin{bmatrix} y_k \\ \mathbf{y}_2 \end{bmatrix}, \quad (23)$$

\bar{X}_{21} is now square and invertible as the vanishing of r_k requires that the corresponding row \mathbf{x}_k from X_{11} be independent of the rows of X_{21} . Now both $\boldsymbol{\beta}$ and κ are fixed by the condition

$$\bar{X}_{21} \boldsymbol{\beta}_1 = \bar{\mathbf{y}}_2.$$

2. The break in the κ homotopy is triggered by the vanishing of a $\boldsymbol{\beta}$ component. In this case it is necessary to remove the column of X_{21} corresponding to the new zero component of $\boldsymbol{\beta}$ to produce \bar{X}_{21} . Again it is square and nonsingular but now has dimension $|\sigma|$. The vector θ_ψ has its j 'th component added to \mathbf{u}_ψ .

Note that λ does not appear explicitly in either possibility. Now differentiate the repartitioned necessary conditions with respect to λ to obtain

$$\begin{bmatrix} 0 & \cdots & 0 & \frac{d\mathbf{v}_\sigma^T}{d\lambda} \end{bmatrix} \begin{bmatrix} \bar{X}_{11} & \bar{X}_{12} \\ \bar{X}_{21} & \bar{X}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_\psi^T & \frac{d(\lambda \mathbf{u}_\psi^T)}{d\lambda} \end{bmatrix}.$$

Equating components gives

$$\frac{d\mathbf{v}_\sigma^T}{d\lambda} \bar{X}_{21} = \boldsymbol{\theta}_\psi^T, \quad (24)$$

$$\frac{d\mathbf{v}_\sigma^T}{d\lambda} \bar{X}_{22} = \frac{d(\lambda \mathbf{u}_\psi^T)}{d\lambda}. \quad (25)$$

Remark 4 It follows that $\frac{d\mathbf{v}_\sigma^T}{d\lambda}$ and $\frac{d(\lambda \mathbf{u}_\psi^T)}{d\lambda}$ are constant so the updates of \mathbf{v}_σ and \mathbf{u}_ψ are computed readily. However, \mathbf{u}_ψ is not itself piecewise linear in λ . A similar situation occurs in the lasso homotopy algorithm for the case of a quadratic objective function subject to an l_1 constraint [9]. Here this complicates the checking that the components of \mathbf{u}_ψ are within bound. The condition that $\lambda \mathbf{u}_\psi$ is piecewise linear can be expressed component-wise as

$$\frac{d}{d\lambda} (\lambda \mathbf{u}(\lambda)_i) = \gamma_i, \quad i = 1, 2, \dots, p,$$

where γ_i is constant independent of λ on each linear piece. It follows that

$$(\mathbf{u}_\psi)_i = \frac{\alpha_i}{\lambda} + \gamma_i,$$

where

$$\alpha_i = (\mathbf{u}_\psi(\lambda_0))_i - \gamma_i \lambda_0.$$

Thus

$$\mathbf{u}_\psi(\lambda_0 - \Delta)_i = \frac{\alpha_i}{\lambda_0 - \Delta} + \gamma_i$$

is an increasing function of $\Delta < \lambda_0$ provided $\alpha_i > 0$. This means that $\mathbf{u}_\psi(\lambda_0 - \Delta)_i$ can reach its upper bound as Δ is increased provided

$$1 = \frac{\alpha_i}{\lambda_0 - \Delta} + \gamma_i \Rightarrow \Delta = \lambda_0 - \frac{\alpha_i}{1 - \gamma_i}.$$

This requires $\alpha_i > 0 \Rightarrow \mathbf{u}_\psi(\lambda_0)_i > \gamma_i$. The lower bound is reached if

$$-1 = \frac{\alpha_i}{\lambda_0 - \Delta} + \gamma_i \Rightarrow \Delta = \lambda_0 + \frac{\alpha_i}{1 + \gamma_i}.$$

This requires $\alpha_i < 0 \Rightarrow \mathbf{u}_\psi(\lambda_0)_i < \gamma_i$.

A consequence is that the necessary conditions continue to hold as λ is reduced until either:

1. Let $u_q = \mathbf{e}_q^T \mathbf{u}_\psi$, and assume it is the component of \mathbf{u}_ψ to first reach a bound. Then $\psi \leftarrow \psi \cup \{\psi^c(q)\}$ and the increase κ phase is recommenced. The necessary conditions are maintained if the corresponding component of $\boldsymbol{\beta}$ moves away from zero with the sign of the bound. This is required in order to impose the norm condition correctly on the augmented $\boldsymbol{\beta}$.
2. Let $v_q = \mathbf{e}_q^T \mathbf{v}_\sigma$ and assume it is the component of \mathbf{v}_σ to first reach a bound. The appropriate action is to remove the corresponding index from σ , $\sigma \leftarrow \sigma \setminus \{\sigma(q)\}$, and restart the increase κ phase using the downdated X_{21} . Now $r_q = \mathbf{e}_q^T \mathbf{r}_2$ is no longer bound to zero. The necessary conditions will continue to hold provided its sign agrees with the sign of v_q as κ is increased.

The additional results required are as follows (see appendix A).

Lemma 5 *Let $v_q(\lambda)$ reach first a bound as λ is decreased. If $\sigma \leftarrow \sigma \setminus \{\sigma(q)\}$ and the κ phase restarted then*

$$\text{sign} \left(\frac{dr_q}{d\kappa} \right) = - \text{sign} \left(\frac{dv_q}{d\lambda} \right). \quad (26)$$

Lemma 6 *Let $u_q(\lambda)$ reach first a bound as λ is decreased. If $\psi \leftarrow \psi \cup \{\psi^c(q)\}$ and the κ phase restarted then*

$$\text{sign} \left(\frac{dx_q}{d\kappa} \right) = - \text{sign} \left(\frac{du_q}{d\lambda} \right). \quad (27)$$

2.4 Starting out

Assume $y_i \neq 0$, $i = 1, 2, \dots, n$. Then

$$\mathbf{r} = \mathbf{y}, \quad \sigma = \emptyset, \quad (\boldsymbol{\theta}_\sigma)_i = \text{sign}(y_i), \quad i = 1, 2, \dots, n,$$

when $\kappa = 0$. For simplicity, assume also there is a unique answer to

$$k = \arg \left\{ \max_i |\boldsymbol{\theta}_\sigma^T X_{*i}| \right\},$$

where X_{*i} denotes the i 'th column of X . Set

$$\lambda = |\boldsymbol{\theta}_\sigma^T X_{*k}|, \quad \psi = \{k\}, \quad \theta_\psi = \text{sign}(\boldsymbol{\theta}_\sigma^T X_{*k}), \quad (28)$$

$$\lambda \begin{bmatrix} \theta_\psi & \mathbf{u}_\psi^T \end{bmatrix} = \boldsymbol{\theta}_\sigma^T X Q_\psi^T. \quad (29)$$

Note that initially ψ is a singleton so that θ_ψ is a scalar. Equations (28) and (29) provide a set of quantities satisfying the necessary conditions for small enough κ . This permits the κ phase to be initiated from $\kappa = 0$.

| | p | n | SASD | SAXA | XDXA | XDSD |
|----------|----|-----|------|------|------|------|
| Hald | 4 | 13 | 8 | 4 | 1 | 0 |
| Iowa | 8 | 33 | 16 | 14 | 1 | 6 |
| diabetes | 10 | 442 | 514 | 23 | 1 | 13 |
| housing | 13 | 506 | 783 | 24 | 0 | 11 |

Table 1: Step counts for homotopy algorithm (R implementation)– l_1 objective

3 Numerical results

3.1 Computational complexity

Results of the homotopy algorithm for several well known data sets are displayed in Table 1. The data sets are

Hald data [4];

Iowa wheat data [4], <http://www.math.unm.edu/splus/node130.html>;

diabetes data <http://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt>;

Boston housing data <http://lib.stat.cmu.edu/datasets/boston> .

At the end of each step which varies κ either σ is increased or ψ is decreased, and at the end of each step that decreases λ either σ is decreased or ψ is increased. An iteration of our homotopy algorithm involves a step varying κ followed by a step varying λ . Table 1 gives the number of such iterations that were necessary in order to reach the unconstrained solution of the problem, and counts of how $|\sigma|$ and $|\psi|$ changed in each iteration. Specifically SA entries correspond to increments and SD to decrements in σ , XA to increments and XD to decrements in ψ . These results show that much of the time the basic step corresponds to entries in the SASD column. This counts descent steps that move one residual r_q away from zero in the κ step and adds a new residual r_k to the zero residual set in the λ step leaving $\beta(\kappa)$ in the same $|\psi|$ dimensional subspace. This is also illustrated for the diabetes data set in Figure 1. Because $\frac{d\beta}{d\kappa}$ is a descent direction (Lemma 1) it follows that the algorithm is making steps that approach the l_1 minimum in this current $|\psi|$ dimensional subspace (actually a fixed orthant as a sign change in β interrupts the SASD sequence).

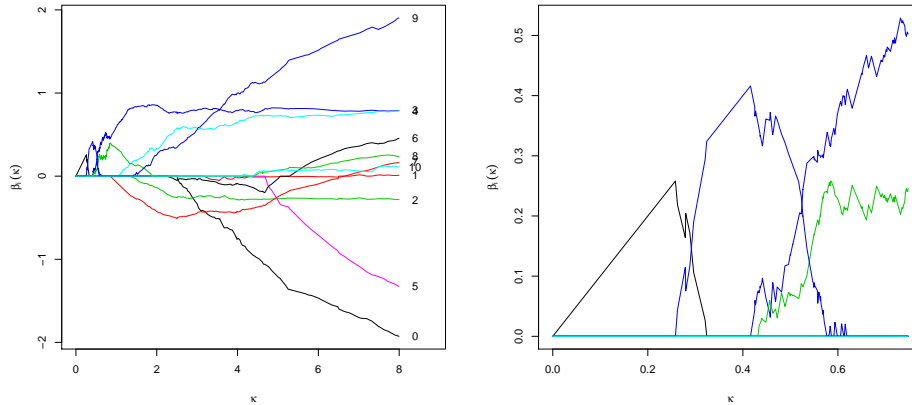


Figure 1: Homotopy algorithm illustrated on the diabetes data. The left panel shows the complete homotopy. The right panel is a magnification of the initial part of the homotopy illustrating the large number of SASD steps that are taken.

| | XA | XD |
|----------|----|----|
| Hald | 5 | 0 |
| Iowa | 9 | 0 |
| diabetes | 14 | 3 |
| housing | 16 | 2 |

Table 2: Step counts for homotopy algorithm – least squares objective

These SASD steps have no analog in the original lasso algorithm in which the usual step is one that adds a new variable - here corresponding most closely to the SAXA column in Table 1. The XDSD column entries correspond to back-track steps while the XDXA column indicates variables changing sign. Both these types of step are observed to occur infrequently in the original lasso formulation which has a complexity essentially dominated by SAXA steps. Results for the original lasso corresponding to those shown above are given in Table 2. These show an order of magnitude difference in the computational complexity of the l_1 homotopy algorithm which occurs because the large number of the SASD steps turns an $O(n)$ algorithm - the complexity of an update step when n is large and p typically fixed - into at least an $O(n^2)$ algorithm. This problem arises because the process is seeking improvements by taking many “small” SASD steps towards the l_1 minimum in the current subspace. However, this process terminates when-

ever the homotopy “escapes” corresponding to the occurrence of a step of type SAXA, XDXA, or XDSD, and this can happen before the minimum in the current subspace is reached. This is a consequence of the role of the currently zero-valued parameter estimates pointed to by ψ^c in the homotopy necessary conditions which has no counterpart in a straight l_1 minimization based on the non-zero parameter estimates alone. Thus large step methods based on using line search techniques developed for the l_1 minimization problem cannot be applied without modification.

The problem that the sequence of small steps solves can be formulated explicitly. The idea is to maximise κ or minimise λ in the orthant of the subspace determined by $\boldsymbol{\theta}_\psi$ subject to constraints which ensure that the necessary conditions remain satisfied. For example, the problem of determining the range of κ appropriate to the current ψ in a “large step” homotopy algorithm corresponds to the problem of designing a fast algorithm for solving the problem

$$\max_{\mathbf{v}, \mathbf{w}, \lambda, \boldsymbol{\beta}} \sum_{i \in \psi} (\boldsymbol{\theta}_\psi)_i x_i \quad (30)$$

subject to constraints which express a small reformulation of the necessary conditions (1) and (10) – (14)

$$\mathbf{v}^T \begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} + \lambda \boldsymbol{\theta}_\psi^T = 0, \quad (31)$$

$$-\mathbf{e}_\sigma \leq \mathbf{v} \leq \mathbf{e}_\sigma, \quad (32)$$

$$\mathbf{v}^T \mathbf{r} = \|\mathbf{r}\|_1, \quad (33)$$

$$\mathbf{v}^T \begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix} + \mathbf{w}_\psi^T = 0, \quad (34)$$

$$-\lambda \mathbf{e}_\psi \leq \mathbf{w}_\psi \leq \lambda \mathbf{e}_\psi, \quad (35)$$

$$\lambda \geq 0, \quad (36)$$

$$(\boldsymbol{\theta}_\psi)_i x_i \geq 0, \quad i \in \psi, \quad (37)$$

where $\mathbf{w} = \lambda \mathbf{u}$, and \mathbf{e}_σ and \mathbf{e}_ψ are vectors with all components 1 and having the indicated dimension. This is almost a linear program of the right structure. The catch is the single nonlinear constraint (33) which is required to restrict the components of \mathbf{v} to lie in the subdifferential of $\|\mathbf{r}\|_1$. This constraint is satisfied implicitly when the homotopy equations are solved exactly.

3.2 Dealing with degeneracies

The homotopy algorithm encounters degeneracy when the decision process at the end of either the κ or λ step requires a choice to be made between several possibilities. For example, several residuals could vanish identically to terminate a κ step. One cause encountered in the diabetes and Boston housing data corresponded to identical observations of the response and partial design in sub models when several potential model variables are omitted. This degeneracy could be avoided by repeated inspection at some computational cost. Two alternative approaches to resolving this problem have been employed successfully in our computations:

1. Small random perturbations can be added to the input data. This has been used in a Fortran 95 implementation of the algorithm.
2. An R implementation permits several indices to enter σ during the vary κ step. If this happens then the matrices in (19) and (25) are only square and invertible once duplicate rows are removed. These equations can be solved using generalised inverse technology. Typically, if several indices enter σ during a κ step then they are later simultaneously removed during a λ step. However, if the λ step ends with an increment in ψ then it may happen that, after the removal of duplicate rows, X_{21} does not have one more column than it has rows which is a requirement for (19) in the next κ step. In this case it is necessary to perform further λ steps, which should all involve decrementing σ , until the required condition on (19) is satisfied.

It is not easy to compare the computational performance of the two procedures. However, we have the impression that the procedure used in the R implementation appears more efficient.

Treatment of degeneracy is important also in post optimality methods used in conjunction with a linear programming formulation of the l_1 lasso. Typically these approaches start with the Lagrangian problem formulation and use the multiplier as the post optimality parameter. Here what corresponds to the κ step can be interpreted as a degeneracy [8]. This structural degeneracy is in addition to that discussed above.

A Proofs

Proofs of the Lemmas that govern the switching between the κ and λ phases are presented. Note that the conditions presented are necessary for the piecewise linear dependence of β on κ , and \mathbf{v}_σ and $\lambda \mathbf{u}_\psi$ on λ to be continuous.

However, the algebraic proofs presented here have some independent interest.

Remark 7 *It will be convenient to distinguish between the set of zero residuals in the κ phase and the augmented set corresponding to a new zero residual in the subsequent λ phase which starts with λ equal to its value λ_0 in the κ phase. The key components of the augmented system are written*

$$\bar{\mathbf{v}}_{\sigma}(\lambda_0) = \begin{bmatrix} \theta_k \\ \mathbf{v}_{\sigma} \end{bmatrix}, \quad \bar{X}_{21} = \begin{bmatrix} \mathbf{x}_k^T \\ X_{21} \end{bmatrix}.$$

As a result of augmenting X_{21} while holding ψ fixed it follows that \bar{X}_{21} is nonsingular. The key component of the necessary conditions becomes

$$\bar{\boldsymbol{\theta}}_{\sigma}^T \bar{X}_{11} + \bar{\mathbf{v}}_{\sigma}^T \bar{X}_{21} = -\lambda \boldsymbol{\theta}_{\psi}^T,$$

and

$$\frac{d\bar{v}_1}{d\lambda} = -\boldsymbol{\theta}_{\psi}^T \bar{X}_{21}^{-1} \mathbf{e}_1.$$

A similar notation is used in the case that ψ is downdated to take account of a component of $\boldsymbol{\beta}$ becoming zero. Let this component be $\beta_{\psi(j)}$ then it is assumed that it is swapped to position $|\psi|$ and then moved to ψ^c in position 1. The repositioning is summarised by

$$\begin{bmatrix} \boldsymbol{\theta}_{\psi}^T \\ X_{21} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{\theta}}_{\psi}^T & \theta_j \\ \bar{X}_{21} & \mathbf{x}^j \end{bmatrix} Q_j, \quad (38)$$

where the permutation matrix that expresses the operation that moves the j 'th component of $\boldsymbol{\theta}_{\psi}$ to the last position is written Q_j .

Proof of Lemma 2. It is necessary to show that the subgradient component associated with $r_k(\kappa_1) = 0$ satisfies equation (21).

Proof. It is convenient to characterise $\theta_k = \text{sign}(r_k)$ using the condition $r_k(\kappa_1) = 0$. This is

$$\mathbf{x}_k^T \left\{ \boldsymbol{\beta}_1(\kappa_0) + (\kappa_1 - \kappa_0) \frac{d\boldsymbol{\beta}_1}{d\kappa} \right\} - y_k = 0,$$

so that

$$\begin{aligned} \mathbf{x}_k^T \frac{d\boldsymbol{\beta}_1}{d\kappa} &= \mathbf{x}_k^T \left[\begin{bmatrix} \boldsymbol{\theta}_{\psi}^T \\ X_{21} \end{bmatrix} \right]^{-1} \mathbf{e}_1, \\ &= -\frac{1}{\kappa_1 - \kappa_0} r_k(\kappa_0). \end{aligned} \quad (39)$$

It follows that

$$\theta_k = -\text{sign} \left(\mathbf{x}_k^T \frac{d\beta_1}{d\kappa} \right). \quad (40)$$

Let \mathbf{w} be defined by

$$\mathbf{x}_k^T = \mathbf{w}^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix},$$

where $w_1 \neq 0$, and $\mathbf{w}^T = [w_1 \quad \mathbf{w}_2^T]$. Then

$$\begin{aligned} \bar{X}_{21} &= \begin{bmatrix} \mathbf{w}^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} \\ X_{21} \end{bmatrix}, \\ &= \begin{bmatrix} w_1 \boldsymbol{\theta}_\psi^T + \mathbf{w}_2^T X_{21} \\ (w_1 + (1 - w_1)) X_{21} \end{bmatrix}, \\ &= \begin{bmatrix} w_1 \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} + \begin{bmatrix} \mathbf{w}_2^T \\ (1 - w_1) I_{|\sigma|} \end{bmatrix} X_{21} \end{bmatrix}, \\ &= \begin{bmatrix} w_1 I_{|\psi|} + \begin{bmatrix} \mathbf{w}_2^T \\ (1 - w_1) I_{|\sigma|} \end{bmatrix} X_{21} \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \\ \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} \end{bmatrix}, \\ &= \begin{bmatrix} w_1 I_{|\psi|} + \begin{bmatrix} \mathbf{w}_2^T \\ (1 - w_1) I_{|\sigma|} \end{bmatrix} \begin{bmatrix} 0 & I_{|\sigma|} \end{bmatrix} \\ \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} \end{bmatrix}, \\ &= \begin{bmatrix} w_1 & \mathbf{w}_2^T \\ 0 & I_{|\sigma|} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}. \end{aligned}$$

Thus

$$\bar{X}_{21}^{-1} = \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{w_1} & -\frac{1}{w_1} \mathbf{w}_2^T \\ 0 & I_{|\sigma|} \end{bmatrix}. \quad (41)$$

It follows that

$$\begin{aligned} \frac{d\bar{v}_1}{d\lambda}(\lambda_0) &= -\boldsymbol{\theta}_\psi^T \bar{X}_{21}^{-1} \mathbf{e}_1, \\ &= -\mathbf{e}_1^T \begin{bmatrix} \frac{1}{w_1} & -\frac{1}{w_1} \mathbf{w}_2^T \\ 0 & I_{|\sigma|} \end{bmatrix} \mathbf{e}_1, \\ &= -\frac{1}{w_1}. \end{aligned} \quad (42)$$

But by (39)

$$\mathbf{x}_k^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \mathbf{e}_1 = \mathbf{w}^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \mathbf{e}_1 = w_1 = -\frac{1}{\kappa_1 - \kappa_0} r_k(\kappa_0). \quad (43)$$

Thus $\text{sign}(w_1) = -\theta_k$, so that

$$\text{sign} \left(\frac{d\bar{\mathbf{v}}_1}{d\lambda}(\lambda_0) \right) \theta_k = 1.$$

■

Proof of Lemma 3

A preliminary result is required.

Lemma 8 *The solution to the linear system*

$$\begin{bmatrix} \mathbf{a}^T & b \\ I & \mathbf{c} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \gamma \end{bmatrix} = \begin{bmatrix} \mu \\ \mathbf{z} \end{bmatrix}$$

is

$$\begin{bmatrix} \mathbf{x} \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbf{z} - \frac{\mu - \mathbf{a}^T \mathbf{z}}{b - \mathbf{a}^T \mathbf{c}} \mathbf{c} \\ \frac{\mu - \mathbf{a}^T \mathbf{z}}{b - \mathbf{a}^T \mathbf{c}} \end{bmatrix}. \quad (44)$$

The particular case of interest is

$$\begin{bmatrix} \mu \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which gives

$$\begin{bmatrix} \mathbf{x} \\ \gamma \end{bmatrix} = \frac{1}{b - \mathbf{a}^T \mathbf{c}} \begin{bmatrix} -\mathbf{c} \\ 1 \end{bmatrix}. \quad (45)$$

Proof. After the reordering the new component of \mathbf{u} will satisfy

$$u_1(\lambda_0) = (\boldsymbol{\theta}_\psi)_j = \theta_j$$

and is at its bound. It is required to move into its feasible region as λ is reduced. This gives the condition

$$\begin{aligned} \theta_j u_1(\lambda_0 - \Delta\lambda) &< 1, \\ \Rightarrow -\Delta\lambda \theta_j \frac{du_1(\lambda_0)}{d\lambda} &< 0, \\ \Rightarrow \theta_j \frac{du_1(\lambda_0)}{d\lambda} &\geq 0 \end{aligned}$$

The starting point is

$$\frac{d\beta_j}{d\kappa} = \mathbf{e}_j^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \mathbf{e}_1.$$

The next step is to highlight the updated quantities. This uses (38):

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix} &= \begin{bmatrix} \bar{\boldsymbol{\theta}}_\psi^T & \theta_j \\ \bar{X}_{21} & \mathbf{x}^j \end{bmatrix} P_j, \\
&= \begin{bmatrix} \bar{\boldsymbol{\theta}}_\psi^T \bar{X}_{21}^{-1} & \theta_j \\ I & \mathbf{x}^j \end{bmatrix} \begin{bmatrix} \bar{X}_{21} & \\ & 1 \end{bmatrix} P_j, \\
&= \begin{bmatrix} -\frac{d\bar{\mathbf{v}}^T}{d\lambda} & \theta_j \\ I & \mathbf{x}^j \end{bmatrix} \begin{bmatrix} \bar{X}_{21} & \\ & 1 \end{bmatrix} P_j.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{d\beta_j}{d\kappa} &= \mathbf{e}_{|\psi|}^T \begin{bmatrix} \bar{X}_{21} & \\ & 1 \end{bmatrix}^{-1} \begin{bmatrix} -\frac{d\bar{\mathbf{v}}^T}{d\lambda} & \theta_j \\ I & \mathbf{x}^j \end{bmatrix}^{-1} \mathbf{e}_1, \\
&= \mathbf{e}_{|\psi|}^T \begin{bmatrix} \bar{X}_{21} & \\ & 1 \end{bmatrix}^{-1} \frac{1}{\theta_j + \mathbf{x}^{jT} \frac{d\bar{\mathbf{v}}}{d\lambda}} \begin{bmatrix} -\mathbf{x}^j \\ 1 \end{bmatrix}, \\
&= \frac{\mathbf{e}_{|\psi|}^T}{\theta_j + \mathbf{x}^{jT} \frac{d\bar{\mathbf{v}}}{d\lambda}} \begin{bmatrix} \bar{X}_{21}^{-1} \mathbf{x}^j \\ 1 \end{bmatrix}, \\
&= \frac{1}{\theta_j + \mathbf{x}^{jT} \frac{d\bar{\mathbf{v}}}{d\lambda}}, \\
&= \frac{1}{\theta_j - \frac{d(\lambda u_1)}{d\lambda}}.
\end{aligned}$$

The final result is

$$\frac{dx_j}{d\kappa} \frac{du_1}{d\lambda} = -\frac{1}{\lambda} \tag{46}$$

showing that u_1 moves into its feasible region. ■

Proof of Lemma 5

Proof. Let the component v_q of \mathbf{v}_σ first reach a bound at $\lambda = \lambda_1 < \lambda_0$. Then

$$\begin{aligned}
\theta_q &= v_q(\lambda_1) = v_q(\lambda_0) + (\lambda_1 - \lambda_0) \frac{dv_q}{d\lambda}, \\
\Rightarrow \text{sign} \left(\frac{dv_q}{d\lambda} \right) &= -\theta_q, \text{ as } |v_q(\lambda_0)| < 1.
\end{aligned}$$

Also

$$\text{sign} \left(\frac{dr_q}{d\kappa} \right) = \text{sign}(\theta_q) \Rightarrow \theta_q r_q(\kappa) = \theta_q \left((\kappa - \kappa_0) \frac{dr_q}{d\kappa} \right) = |r_q|,$$

as $r_q(\kappa_0) = 0$. Now

$$\begin{aligned}\mathbf{x}_k^T &= \mathbf{w}^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}, \\ &= w_1 \boldsymbol{\theta}_\psi^T + w_q \mathbf{x}_q^T + \sum_{j \in \{1, \dots, |\sigma|\} \setminus \{q\}} w_j \mathbf{e}_j^T X_{21}.\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{x}_q^T &= \frac{1}{w_q} \mathbf{x}_k^T - \frac{w_1}{w_q} \boldsymbol{\theta}_\psi^T - \sum_{j \in \{1, \dots, |\sigma|\} \setminus \{q\}} \frac{w_j}{w_q} \mathbf{e}_j^T X_{21}, \\ &= \begin{bmatrix} -\frac{w_1}{w_q} & \frac{1}{w_q} & \dots & -\frac{w_j}{w_q} & \dots \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ \mathbf{x}_k^T \\ X_{21}^{-q} \end{bmatrix},\end{aligned}$$

where X_{21}^{-q} denotes X_{21} with row \mathbf{x}_q removed. This gives, using (19),

$$\frac{dr_q}{d\kappa} = \mathbf{x}_q^T \frac{d\boldsymbol{\beta}_1}{d\kappa} = \mathbf{x}_q^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ \mathbf{x}_k^T \\ X_{21}^{-q} \end{bmatrix}^{-1} \mathbf{e}_1 = -\frac{w_1}{w_q}.$$

On the other hand

$$\frac{dv_\sigma^T}{d\lambda} = -\boldsymbol{\theta}_\psi^T \bar{X}_{21}^{-1},$$

so that

$$\frac{dv_q}{d\lambda} = -\boldsymbol{\theta}_\psi^T X_{21}^{-1} \mathbf{e}_q = -\boldsymbol{\theta}_\psi^T \begin{bmatrix} \boldsymbol{\theta}_\psi^T \\ X_{21} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{w_1} & -\frac{\mathbf{w}_2^T}{w_1} \\ 0 & I_{|\sigma|} \end{bmatrix} \mathbf{e}_q = \frac{w_q}{w_1}.$$

Thus

$$\text{sign} \left(\frac{dr_q}{d\kappa} \right) = -\text{sign} \left(\frac{dv_q}{d\lambda} \right).$$

■

Proof of Lemma 6

Proof. Assume that u_q , $q \in \psi^c$ is the variable that reaches its bound as λ is decreased. Let $u_q(\lambda_1) = \theta_q$ then the equation determining $\frac{d\boldsymbol{\beta}}{d\kappa}$ is

$$\frac{d\boldsymbol{\beta}}{d\kappa} = \begin{bmatrix} \boldsymbol{\theta}_\psi^T & \theta_q \\ \bar{X}_{21} & \bar{X}_{22} \mathbf{e}_q \end{bmatrix}^{-1} \mathbf{e}_1.$$

From equations (24) and (25) it follows that

$$\begin{bmatrix} 1 & \frac{dv_\sigma^T}{d\lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\psi^T & \theta_q \\ \bar{X}_{21} & \bar{X}_{22} \mathbf{e}_q \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 & \theta_q - \frac{d(\lambda u_q)}{d\lambda} \end{bmatrix}.$$

Thus

$$\begin{bmatrix} 1 & \frac{d\mathbf{v}_q^T}{d\lambda} \end{bmatrix} \mathbf{e}_1 = \begin{bmatrix} 0 & \cdots & 0 & \theta_q - \frac{d(\lambda u_q)}{d\lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\psi^T & \theta_q \\ \bar{X}_{21} & \bar{X}_{22} \mathbf{e}_q \end{bmatrix}^{-1} \mathbf{e}_1,$$

giving

$$\begin{aligned} 1 &= \left(\theta_q - \frac{d(\lambda u_q)}{d\lambda} \right) \frac{d\beta_q}{d\lambda}, \\ &= -\lambda \frac{du_q}{d\lambda} \frac{d\beta_q}{d\kappa}. \end{aligned}$$

Equation (27) is a direct consequence as $\lambda > 0$. ■

References

- [1] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervising clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2007.
- [2] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics*, 35(6):2313–2351, 2007.
- [3] D. L. Donoho and Y. Tsaig. Fast solution of l_1 -norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54:4789–4812, 2008.
- [4] N. H. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, 3 edition, 1998.
- [5] T. P. Hettmansperger and J. W. McKean. *Robust Nonparametric Statistical Methods*. John Wiley & Sons, Chichester, 1998.
- [6] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- [7] Y. Li and J. Zhu. L_1 -norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.
- [8] M. R. Osborne. *Simplicial Algorithms for Minimizing Polyhedral Functions*. Cambridge University Press, 2001.
- [9] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–403, 2000.

- [10] S. Rosset and Ji Zhu. Piecewise linear regularised solution paths. *Annals of Statistics*, 35(3):1012–1030, 2007.
- [11] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [12] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.
- [13] Y. Yao and Y. Lee. Another look at linear programming for feature selection via methods of regularization. Technical Report 800, Department of Statistics, Ohio State University, 2007.
- [14] J. Zhu, T. Hastie, S. Rosset, and R. Tibshirani. l_1 -norm Support Vector Machines. *Advances in Neural Information Processing Systems*, 16:49–56, 2004.
- [15] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.
- [16] H. Zou and M. Yuan. Regularised simultaneous model selection in multiple quantiles regression. *Computational Statistics & Data Analysis*, 52:5296–5304, 2008.