On the LASSO and Its Dual

Michael R. Osborne* Brett Presnell† Berwin A. Turlach‡

June 23, 1999

Abstract

Proposed by Tibshirani (1996), the LASSO (least absolute shrinkage and selection operator) estimates a vector of regression coefficients by minimising the residual sum of squares subject to a constraint on the l^1-norm of coefficient vector. The LASSO estimator typically has one or more zero elements and thus shares characteristics of both shrinkage estimation and variable selection. In this paper we treat the LASSO as a convex programming problem and derive its dual. Consideration of the primal and dual problems together leads to important new insights into the characteristics of the LASSO estimator and to an improved method for estimating its covariance matrix. Using these results we also develop an efficient algorithm for computing LASSO estimates which is usable even in cases where the number of regressors exceeds the number of observations.

KEY WORDS AND PHRASES. Convex Programming, Dual Problem, Partial Least Squares, Quadratic Programming, Penalised Regression, Regression, Shrinkage, Subset Selection, Variable Selection.

*Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia

†Department of Statistics, University of Florida, Gainesville FL 32611–8545, USA; and Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia

‡Department of Statistics, University of Adelaide, Adelaide SA 5005, Australia; and Centre for Mathematics and its Applications and Cooperative Research Centre for Advanced Computational Systems, Australian National University, Canberra ACT 0200, Australia. The author wishes to acknowledge that part of this work was carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

1 Introduction

Consider the usual linear regression setting with data $(x_{i,1}, \ldots, x_{i,m}, y_i)$, $i = 1, \ldots, n$, where the x_{ij}s are the regressors and y_i the response for the ith observation. In this situation, ordinary least squares regression finds the linear combination of the x_{ij}s that minimises the residual sum of squares. However, if m is large or if the the regressor variables are highly correlated, then the variances of the least-squares coefficient estimates may be unacceptably high. Standard methods for addressing this difficulty include ridge regression and, particularly in cases where a more parsimonious model is desired, subset selection.

As an alternative to standard ridge regression and subset selection techniques, Tibshirani (1996) proposed the "least absolute shrinkage and selection operator" (LASSO), which minimises the residual sum of squares under a constraint on the l^1-norm of coefficient vector. Thus the LASSO estimator solves the optimisation problem

$$\underset{\beta_1,\ldots,\beta_m}{\text{minimise}} \qquad \tfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 \tag{1.1a}$$

$$\text{subject to} \qquad \sum_{j=1}^{m}|\beta_j| \leq t, \tag{1.1b}$$

for some $t > 0$. If t is greater than or equal to the l^1-norm of the ordinary least squares estimator, then that estimator is of course unchanged by the LASSO. For smaller values of t, the LASSO shrinks the estimated coefficient vector towards the origin (in the L_1 sense), typically setting some of the coefficients equal to zero. Thus the LASSO combines characteristics of ridge regression and subset selection and promises to be a useful tool for variable selection.

Though the optimisation problem (1.1) is easily stated, solving it numerically is not a trivial exercise. The algorithm proposed by Tibshirani (1996) is adequate for moderate values of m, but it is not the most efficient possible. Of course the effect of an inefficient algorithm is greatly magnified when techniques such as cross-validation and the bootstrap are used to choose an appropriate value of t (Tibshirani, 1996) or to estimate standard errors. Moreover, Tibshirani's algorithm is particularly inefficient when m is large and it is not usable at all when $m > n$. This can be a rather severe practical limitation, since problems in which the number of variables is of the same or larger order than the number of observations occur frequently in areas such as chemometrics, where *partial least squares* (Brown, 1993; Haagen *et al.*, 1993) is often employed. In fact our own interest in the LASSO was initially motivated by the problem of knot selection for regression splines (see Osborne *et al.*, 1998), which can be formulated as a variable selection problem with $m > n$.

In this paper, we treat (1.1) as a convex programming problem and derive the dual optimisation problem. By considering simultaneously the primal problem and its dual, we develop a highly efficient algorithm for calculating the LASSO estimator which is also applicable in the case that $m > n$. This approach also yields new insight into the LASSO by providing an exact characterisation of

the solution(s) of (1.1). In particular, in the case $m \leq n$, this characterisation suggests an estimator of the covariance matrix of the LASSO estimator different from the one proposed by Tibshirani (1996).

We shall concentrate on the optimisation problem (1.1). A closely related optimisation problem is

$$\underset{\beta_1,\ldots,\beta_m}{\text{minimise}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{m} |\beta_j|. \tag{1.2}$$

Problems (1.1) and (1.2) are equivalent; that is, for a given λ, $0 \leq \lambda < \infty$, there exists a $t \geq 0$ such that the two problems share the same solution, and vice versa. Optimisation problems like (1.1) are usually referred to as constrained regression problems while (1.2) would be called a penalised regression. Chen *et al.* (1999) propose to use penalised l^1-regression in the context of wavelet regression. They use a primal-dual log-barrier interior point algorithm to solve (1.2). Recently Sardy *et al.* (1999) propose another algorithm that is based on block coordinate relaxation techniques.

A possible generalisation of (1.1) is to change the constraint (1.1b) to

$$\sum_{j=1}^{m} |\beta_j|^\gamma \leq t \quad \text{for some} \quad \gamma \geq 1.$$

This was investigated by Fu (1998) (see also, Frank and Friedman, 1993). Fu (1998) also proposes an alternative algorithm to solve (1.2). However, his algorithm is again not applicable if $m > n$ as it starts from the unconstrained least-squares solution of (1.1a).

The rest of this paper is structured as follows. In Section 2 we derive the dual of (1.1) and discuss the relationship between the primal and dual problems. Section 3 discusses further theoretical properties of the LASSO estimator, including existence and uniqueness of solutions and the number of non-zero entries in the estimator. In Section 4 we discuss the estimation of standard errors and propose a new approach to this problem based on the duality results of Section 2. Further technical details concerning standard error estimation are given in the appendix. A new and efficient algorithm to calculate the LASSO estimator is developed in Section 5. In Section 6, this algorithm is applied to an example from Tibshirani (1996), and various standard error estimates are also compared in the context of this example.

2 Convex duality and the LASSO

To fix notation, let $\mathbf{y} = (y_1, \ldots, y_n)^T$ denote the vector of observed responses, let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be the $n \times m$-matrix with the vector $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})^T \in \mathbb{R}^n$ as jth column and let $\mathbf{A} = \mathbf{X}^T\mathbf{X}$. We assume that $\mathbf{X}$ has maximal rank. Let $\mathcal{N}(\mathbf{X}) \subset \mathbb{R}^m$ denote the null space of $\mathbf{X}$ and let β^0 be a solution to the unconstrained least squares problem. Of course if $m \leq n$, then $\mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$

and $\beta^0 = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}$ is unique. If $m > n$, then $\mathcal{N}(\mathbf{X})$ has dimension $m - n$, β^0 is not unique, and $\mathbf{X}(\beta^0 + \eta) = \mathbf{y}$ holds for any $\eta \in \mathcal{N}(\mathbf{X})$. But in either case we may define

$$t_0 = \min_{\eta \in \mathcal{N}(\mathbf{X})} \|\beta^0 + \eta\|_1,$$

where $\|\beta\|_1 = \sum_{i=1}^m |\beta_i|$ denotes the l^1 norm on $\mathbb{R}^m$. Note, that t_0 is unique even though in the case $m > n$ there may be several η's that attain t_0. Since the LASSO is equivalent to ordinary least squares when $t \geq t_0$, we assume in the sequel that $t < t_0$.

The optimisation problem (1.1) can be rewritten as

$$\underset{\beta}{\text{minimise}} \qquad f(\beta) \tag{2.1a}$$

$$\text{subject to} \qquad g(\beta) \geq 0, \tag{2.1b}$$

where

$$f(\beta) = \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \tfrac{1}{2}\mathbf{r}^T\mathbf{r} \tag{2.2}$$

and

$$g(\beta) = t - \sum_{i=1}^m |\beta_i|. \tag{2.3}$$

Here $\mathbf{r} = \mathbf{r}(\beta)$ is the vector of residuals corresponding to β and $g(\beta)$ is implicitly a function of t, which is treated as fixed in the present discussion.

Since f is continuous and the region of feasible β vectors is compact, a solution to (2.1) is guaranteed to exist. Further, since $t < t_0$, all the critical values of f occur outside the feasible region and any solution β^* of (2.1) must lie on its boundary, i.e., $\|\beta^*\|_1 = t$. Because g is a concave function, the region of feasible values defined by (2.1b) is convex, and since f is a convex function, it is clear that the solution set of (2.1) is convex. If $m \leq n$, then our assumption on $\mathbf{X}$ ensures that f is strictly convex, in which case the solution is unique. These facts are summarised in the following lemma.

Theorem 2.1 (Existence and Uniqueness). *If $t < t_0$, then the following hold:*

(a) If $m \leq n$, then a unique solution β^* of (1.1) exists and $\|\beta^*\|_1 = t$.

(b) If $m > n$, then a solution β^* of (1.1) exists and $\|\beta^*\|_1 = t$ for any solution. If β_1^* and β_2^* are both solutions of (1.1), then $\rho\beta_1^* + (1-\rho)\beta_2^*$ is also a solution for all $0 \leq \rho \leq 1$.

Treating (2.1) as a convex programming problem (Nash and Sofer, 1996, p. 21), the Lagrangian is

$$\mathcal{L}(\beta, \lambda) = f(\beta) - \lambda g(\beta). \tag{2.4}$$

If we define

$$\mathcal{L}^*(\beta) = \sup_{\lambda \geq 0} \mathcal{L}(\beta, \lambda), \tag{2.5}$$

then

$$\mathcal{L}^*(\beta) = \begin{cases} f(\beta) & \text{if } g(\beta) \geq 0, \\ \infty & \text{if } g(\beta) < 0. \end{cases}$$

Hence, minimising $\mathcal{L}^*(\beta)$ is equivalent to solving (2.1). In convex programming theory, (2.1) or the equivalent problem of minimising $\mathcal{L}^*(\beta)$ are called the primal problem and $f(\beta)$ is called the primal objective function.

For $\lambda \geq 0$ the dual objective function is defined to be

$$\mathcal{L}_*(\lambda) = \inf_{\beta} \mathcal{L}(\beta, \lambda), \tag{2.6}$$

and the dual problem is

$$\underset{\lambda \geq 0}{\text{maximise}}\, \mathcal{L}_*(\lambda). \tag{2.7}$$

If we fix $\lambda \geq 0$, then $\mathcal{L}(\beta, \lambda)$ is a convex function in β and $\mathcal{L}(\beta, \lambda) \to \infty$ as $\|\beta\|_1 \to \infty$. Hence, $\mathcal{L}(\cdot, \lambda)$ has at least one minimum and $\bar{\beta}$ minimises $\mathcal{L}(\beta, \lambda)$ if and only if the m-dimensional null-vector $\mathbf{0}$ is an element of the subdifferential $\partial_\beta \mathcal{L}(\bar{\beta}, \lambda)$ (Osborne, 1985, p. 23). In the current problem, the subdifferential is given by (Osborne, 1985, p. 20, Remark 4.4)

$$\partial_\beta \mathcal{L}(\beta, \lambda) = -\mathbf{X}^T \mathbf{r} + \lambda \mathbf{v},$$

where $\mathbf{v} = (v_1, \ldots, v_m)^T$ is of the following form: $v_i = 1$ if $\beta_i > 0$, $v_i = -1$ if $\beta_i < 0$ and $v_i \in [-1, 1]$ if $\beta_i = 0$. Thus if $\bar{\beta}$ minimises $\mathcal{L}(\beta, \lambda)$ for a given value of λ, then

$$\mathbf{0} = -\mathbf{X}^T \bar{\mathbf{r}} + \lambda \mathbf{v}, \tag{2.8}$$

for some $\mathbf{v}$ of the form described above and $\bar{\mathbf{r}} = \mathbf{r}(\bar{\beta}) = \mathbf{y} - \mathbf{X}\beta$ denotes the residual vector.

The form of $\mathbf{v}$ implies that $\mathbf{v}^T \bar{\beta} = \|\bar{\beta}\|_1$ and thus it follows from (2.8) that if $\bar{\beta}$ minimises $\mathcal{L}(\beta, \lambda)$, then $\lambda = \bar{\mathbf{r}}^T \mathbf{X} \bar{\beta} / \|\bar{\beta}\|_1$. Alternatively, if $\bar{\beta} \neq \mathbf{0}$, which is the case whenever $t > 0$ by Theorem 2.1, then $\|\mathbf{v}\|_\infty = 1$ and it follows, again from (2.8), that λ can also be calculated as $\lambda = \|\mathbf{X}^T \bar{\mathbf{r}}\|_\infty$. Using these two expressions for λ we find that

$$\begin{aligned} \mathcal{L}_*(\lambda) = \mathcal{L}(\bar{\beta}, \lambda) &= \tfrac{1}{2} \bar{\mathbf{r}}^T \bar{\mathbf{r}} - \frac{\bar{\mathbf{r}}^T \mathbf{X} \bar{\beta}}{\|\bar{\beta}\|_1} (t - \|\bar{\beta}\|_1) \\ &= \tfrac{1}{2} \bar{\mathbf{r}}^T \bar{\mathbf{r}} + \bar{\mathbf{r}}^T \mathbf{X} \bar{\beta} - t \frac{\bar{\mathbf{r}}^T \mathbf{X} \bar{\beta}}{\|\bar{\beta}\|_1} \end{aligned}$$

$$= \tfrac{1}{2}\mathbf{y}^T\mathbf{y} - \tfrac{1}{2}\bar{\beta}^T\mathbf{A}\bar{\beta} - t\frac{\bar{\mathbf{r}}^T\mathbf{X}\bar{\beta}}{\|\bar{\beta}\|_1}$$
$$= \tfrac{1}{2}\mathbf{y}^T\mathbf{y} - \tfrac{1}{2}\bar{\beta}^T\mathbf{A}\bar{\beta} - t\|\mathbf{X}^T\bar{\mathbf{r}}\|_\infty.$$

If we define

$$\begin{aligned}\tilde{h}(\beta) &= \tfrac{1}{2}\mathbf{y}^T\mathbf{y} - \tfrac{1}{2}\beta^T\mathbf{A}\beta - t\frac{\mathbf{r}^T\mathbf{X}\beta}{\|\beta\|_1},\\ \bar{h}(\beta) &= \tfrac{1}{2}\mathbf{y}^T\mathbf{y} - \tfrac{1}{2}\beta^T\mathbf{A}\beta - t\|\mathbf{X}^T\mathbf{r}\|_\infty,\end{aligned} \tag{2.9}$$

then $\tilde{h}(\bar{\beta}) = \bar{h}(\bar{\beta})$ for any $\bar{\beta}$ for which $\mathbf{0} \in \partial_\beta\mathcal{L}(\bar{\beta}, \lambda)$, and the dual function can be written as

$$\mathcal{L}_*(\lambda) = \mathcal{L}(\bar{\beta}, \lambda) = \bar{h}(\bar{\beta}) = \tilde{h}(\bar{\beta}) \text{ for any } \bar{\beta} \text{ for which } \mathbf{0} \in \partial_\beta\mathcal{L}(\bar{\beta}, \lambda). \tag{2.10}$$

In the sequel, we shall use $h(\beta)$ as a generic notation for either $\bar{h}(\beta)$ or $\tilde{h}(\beta)$, and statements about $h(\beta)$ will hold for both. However, note that in general $\tilde{h}(\beta) \neq \bar{h}(\beta)$.

Remark 2.2: Tibshirani (1996) notes that (1.1) can be written as a quadratic programming problem. The dual function $\bar{h}(\beta)$ was originally found by the authors while deriving the dual problem of this quadratic programming problem.

Remark 2.3: By the same arguments as above the solution of (1.2) must fulfil (2.8). This leads to an interesting observation concerning the choice of smoothing parameter in l^1-penalised regression versus l^2-penalised regression, i.e. ridge regression. In l^2-penalised regression, one typically observes that $\beta \to \mathbf{0}$ as $\lambda \to \infty$, but for any finite λ all entries in β are non-zero. By way of contrast, in l^1-penalised regression we see from (2.8) that as soon as $\lambda \geq \|\mathbf{X}^T\mathbf{y}\|_\infty$ is chosen, $\beta = \mathbf{0}$ is a solution of (1.2). To see this note that if $\beta = \mathbf{0}$, then $\mathbf{r} = \mathbf{y}$ and if we choose $\mathbf{v} = \mathbf{X}^T\mathbf{y}/\lambda$, then (2.8) holds and $\mathbf{v}$ is of the required form, i.e. each of its components has an absolute value less or equal to one. Thus, if in l^1-penalised regression the "smoothing" parameter λ is to be chosen adaptively, e.g. by cross-validation, then the search for the optimal parameter can be conveniently restricted to the interval from zero to $\|\mathbf{X}^T\mathbf{y}\|_\infty$.

The existence of a finite solution to the dual problem is guaranteed by Theorem 2.5 below. If $\bar{\lambda}$ is such a solution and $\mathcal{L}_*(\bar{\lambda}) = \mathcal{L}(\bar{\beta}, \bar{\lambda})$, then $\mathcal{L}_*(\bar{\lambda}) = \bar{h}(\bar{\beta}) = \tilde{h}(\bar{\beta})$. However, if β_* is a maximiser of $h(\beta)$, where the maximum is taken over all $\beta \in \mathbb{R}^m$, then β_* is not necessarily a feasible point for the primal problem, i.e. it is not guaranteed that $\|\beta_*\|_1 \leq t$. In fact, if $m > n$, it can be shown that a global maximiser $\bar{\beta}_*$ of $\bar{h}(\beta)$ exists with $\|\bar{\beta}_*\|_1 \leq t$ but that the l^1 norm of the global maximiser $\tilde{\beta}_*$ of $\tilde{h}(\beta)$ is strictly greater than t.

In many cases, efficient algorithms for solving optimisation problems can be developed by using the relationship between the primal and dual problems. This is also the case for solving (1.1). Below we give some results concerning the relationship between the primal problem (2.1) and its dual (2.7). The first result follows directly from the definitions of $\mathcal{L}^*$ and $\mathcal{L}_*$ and is known as *weak duality* (see, e.g. Nash and Sofer, 1996, Chapter 14.8).

Theorem 2.4. (Weak Duality) *If β^* is a solution of* (1.1) *and $\bar{\lambda}$ is a solution of the dual problem* (2.7)*, then $\mathcal{L}_*(\bar{\lambda}) \leq \mathcal{L}^*(\beta^*)$, i.e., $h(\bar{\beta}) \leq f(\beta^*)$, where $\bar{\beta}$ satisfies $\mathcal{L}_*(\bar{\lambda}) = \mathcal{L}(\bar{\beta}, \bar{\lambda})$.*

A direct consequence of this theorem is that $f(\beta^*) \geq h(\beta^*)$ for all solutions of (1.1). It is desirable that equality hold at solutions of (1.1), since this would allow us to use the *dual gap* $f(\beta) - h(\beta)$ to test for solution of (1.1). The discussion in Nash and Sofer (1996, Chapter 14.8) indicates that this is true if (and only if) there is some point (β^*, λ^*) that satisfies the saddle-point condition

$$\mathcal{L}(\beta^*, \lambda) \leq \mathcal{L}(\beta^*, \lambda^*) \leq \mathcal{L}(\beta, \lambda^*) \tag{2.11}$$

for all $\beta \in \mathbb{R}^m$ and $\lambda \geq 0$. The next theorem shows that for this problem such points exist. Here, the Lagrange multiplier λ corresponding to β is $\lambda = \mathbf{r}^T\mathbf{X}\beta/\|\beta\|_1$, as defined after (2.8).

Theorem 2.5. (Strong Duality) *If β^* is a solution of* (1.1) *and λ^* is the Lagrange multiplier corresponding to β^*, then λ^* is a solution of the dual problem* (2.7) *and $\mathcal{L}_*(\lambda^*) = \mathcal{L}(\beta^*, \lambda^*)$. It follows that the optimal primal and dual function values are equal, i.e., $h(\beta^*) = f(\beta^*)$.*

Proof. Following Osborne (1985, p. 34) we define the *perturbation function* to be

$$v(z) = \inf_{\beta \in \{\beta : g(\beta) \geq z\}} f(\beta).$$

By Lemma 1.6.2 of Osborne (1985), $v(z)$ is a convex function with effective domain $\mathrm{dom}(v) = (-\infty, t]$. Since 0 lies in the interior of $\mathrm{dom}(v)$, $v(z)$ is stable at $z = 0$ (Osborne, 1985, p. 15). The theorem now follows from Theorem 1.6.2(iv) of Osborne (1985). □

Since the dual function (2.6) is concave, any extreme in $(0, \infty)$ is a maximum and all maxima take the same value. Hence, as a consequence of Theorems 2.4 and 2.5, if $\bar{\lambda} > 0$ is a solution for the dual problem with $\mathcal{L}_*(\bar{\lambda}) = \mathcal{L}(\bar{\beta}, \bar{\lambda})$, and if $\bar{\beta}$ is primal feasible, then $\bar{\beta}$ is a solution to the primal problem.

3 Characteristics of solutions

3.1 Uniqueness

The following definitions are useful for proving properties of solutions of (1.1) when $m > n$. If β^* is a solution of (1.1), then we define $V(\beta^*)$ to be the collection of all vectors $\mathbf{e}$ of the form $e_i = 1$ if $\beta_i^* > 0$, $e_i = -1$ if $\beta_i^* < 0$ and e_i equals either -1 or 1 if $\beta_i^* = 0$. If β^* has $l < m$ entries equal to zero, then $V(\beta^*) = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is a set of $k = 2^l$ vectors, and we let $\mathbf{E}$ denote the $k \times m$ matrix whose ith row is $\mathbf{e}_i^T$. Note that $\|\mathbf{e}\|_1 = m$ and ${\beta^*}^T\mathbf{e} = t < t_0$ for each $\mathbf{e} \in V(\beta^*)$. Thus

$$\left(\beta^* - \tfrac{t}{m}\mathbf{e}\right)^T \mathbf{e} = 0 \quad \text{for all} \quad \mathbf{e} \in V(\beta^*),$$

i.e., β^* lies on the intersection of all the hyperplanes that have l^1-distance t from the origin and one of the $\mathbf{e} \in V(\beta^*)$ as normal vector.

Suppose now that $m > n$ and that $\beta^\dagger$ is also a solution of (1.1), and let $\eta = \beta^\dagger - \beta^*$. It follows from Theorem 2.1 that $\beta^* + \rho\eta$ is also a solution for all $0 \leq \rho \leq 1$ and that $\|\beta^* + \rho\eta\|_1 = t$. Since $\|\mathbf{y} - \mathbf{X}\beta\|_2$ must be constant across solutions, a standard argument then shows that $\eta \in \mathcal{N}(\mathbf{X})$. Moreover, the condition $\|\beta^* + \rho\eta\|_1 = t$ implies that $\eta^T\mathbf{e} \leq 0$ for all $\mathbf{e} \in V(\beta^*)$. To see this, first note that $\|\mathbf{b}\|_1 \geq \mathbf{b}^T\mathbf{w}$ for any m-vectors $\mathbf{b}$ and $\mathbf{w}$ with $\|\mathbf{w}\|_\infty \leq 1$. Thus if $\eta^T\mathbf{e} > 0$ for some $\mathbf{e} \in V(\beta^*)$, then $\|\beta^* + \rho\eta\|_1 \geq (\beta^* + \rho\eta)^T\mathbf{e} = t + \rho\eta^T\mathbf{e} > t$ for $0 < \rho \leq 1$, a contradiction. Geometrically, this argument reflects the fact that in moving from β^* to a new solution, we must either stay on all the hyperplanes in which β^* lies, or, if we move off any of these hyperplanes, we must move in a direction η for which $(\beta^* + \rho\eta)^T\mathbf{e} < t$ for those $\mathbf{e}$ that define the hyperplanes that we leave.

If we take $C(\beta^*)$ to be the convex cone defined by

$$C(\beta^*) = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^T\mathbf{e} \leq 0 \quad \text{for all} \quad \mathbf{e} \in V(\beta^*)\},$$

then the above discussion can be summarised by the following theorem.

Theorem 3.1. *β^* is a unique solution of (1.1) if and only if $C(\beta^*) \cap \mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$, where $\mathbf{0}$ is the m-dimensional null vector.*

Note that the condition of the theorem is trivially fulfilled if $m \leq n$ since we assume $\mathbf{X}$ to have full rank. Although this theorem gives a necessary and sufficient condition for the existence of a unique solution, the condition is difficult to verify in practice. The results of Section 2 enable us to develop a more useful condition.

Suppose again that β^* and $\beta^\dagger$ are both solutions of (1.1). Then arguing as before,

$$\eta = \beta^\dagger - \beta^* \in \mathcal{N}(\mathbf{X}) \quad \Longrightarrow \quad \mathbf{X}\beta^* = \mathbf{X}\beta^\dagger \quad \Longrightarrow \quad \mathbf{X}^T\mathbf{r}^* = \mathbf{X}^T\mathbf{r}^\dagger.$$

This means that the vector $\mathbf{X}^T\mathbf{r}$ is constant across solutions of (1.1). Let $\sigma = \{i_1, \ldots, i_p\}$ be the set of indices for which $|(\mathbf{X}^T\mathbf{r})_{i_j}| = \|\mathbf{X}^T\mathbf{r}\|_\infty$ for $j = 1, \ldots, p$. Since every solution of (1.1) must satisfy (2.8), it follows that if β is a solution then $\beta_i = 0$ for all $i \notin \sigma$. Hence, if β^* and $\beta^\dagger$ are solutions, then $\eta_i = 0$ for all $i \notin \sigma$. This leads to the following theorem.

Theorem 3.2. *Let β^* be a solution of (1.1). Denote by $\mathbf{X}_\sigma$ the $n \times p$-matrix whose jth column is the i_jth column of $\mathbf{X}$, $j = 1, \ldots, p$, and let $\mathbf{E}_\sigma$ denote the corresponding submatrix of $\mathbf{E}$, the matrix formed by the vectors in $V(\beta^*)$. Then β^* is a unique solution if and only if there exists no $\gamma \neq \mathbf{0}$ satisfying*

$$\mathbf{E}_\sigma\gamma \leq \mathbf{0}, \tag{3.1a}$$

$$\mathbf{X}_\sigma\gamma = \mathbf{0}. \tag{3.1b}$$

Given Theorem 3.2, it is often easy to verify whether a solution β^* is unique when $m > n$. Assume for example that any $n \times n$-submatrix of $\mathbf{X}$ has full rank. Then after calculating a solution to (1.1) with the algorithm to be proposed in Section 5, we determine σ. If the number of elements of σ is less than or equal to n, then the solution is unique, since in this case (3.1b) holds only for $\gamma = \mathbf{0}$. Otherwise we must check for a non-trivial solution of (3.1).

3.2 The number of non-zero coefficients

If $m > n$ and β^* is not a unique solution, then there exists a vector $\eta \in \mathcal{N}(\mathbf{X})$ such that $\eta^T \mathbf{e} \leq 0$ for all $\mathbf{e} \in V(\beta^*)$. But the "interesting" directions are those $\eta \in \mathcal{N}(\mathbf{X})$ for which $\eta^T \mathbf{e} = 0$ for all $\mathbf{e} \in V(\beta^*)$, i.e. $\eta \in \mathcal{N}(\mathbf{E})$. If we move along such a direction, from β^* to $\beta^\dagger$, until we reach (at least) one other hyperplane that defines the m-dimensional l^1-sphere of radius t, then $\beta^\dagger$ will have at least one fewer non-zero components than β^*, i.e.,$|V(\beta^*)| < |V(\beta^\dagger)|$. Hence, as long as there exists $\eta \in \mathcal{N}(\mathbf{E}) \cap \mathcal{N}(\mathbf{X})$, $\eta \neq \mathbf{0}$, we can move from β^* along η to a solution $\beta^\dagger$ which has fewer non-zero entries than β^*. From the results proven below it follows that $\mathbf{E}$ has full rank if β^* is a vertex of the m-dimensional l^1-sphere of radius t. Hence, this iterative process will end at the latest when a vertex is reached.

On the other hand, if $\eta \in \mathcal{N}(\mathbf{X})$ is such that $\eta \in C(\beta^*)$ and $\eta^T \mathbf{e} < 0$ for at least one $\mathbf{e} \in V(\beta^*)$, then the number of non-zero entries of β^* increases, at least initially, as we move along η. If we move along such a direction from β^* to $\beta^\dagger$ we cannot guarantee that $V(\beta^\dagger)$ will contain more vectors than $V(\beta^*)$. This discussion motivates the following definition.

Definition 3.3. A solution β^* of (1.1) is called *regular* if $\mathcal{N}(\mathbf{E}) \cap \mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$.

Note, that a unique solution of (1.1) is also a regular solution in the sense of this definition since $\mathcal{N}(\mathbf{E}) \subset C(\beta^*)$. From the discussion preceding the definition it is also clear that we can find at least one regular solution if (1.1) has multiple solutions. To obtain a bound for the number of non-zero elements a regular solution may have, we define $C(\beta^*)^\circ$ to be the polar cone of $C(\beta^*)$,

$$C(\beta^*)^\circ = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y}^T \mathbf{x} \leq 0 \quad \text{for all} \quad \mathbf{x} \in C(\beta^*)\}.$$

The polar cone $C(\beta^*)^{\circ\circ}$ of $C(\beta^*)^\circ$ is again $C(\beta^*)$ and the vectors $\mathbf{e} \in V(\beta^*)$ are *generators* of $C(\beta^*)^\circ$, i.e.

$$C(\beta^*)^\circ = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i, \quad \lambda_i \geq 0, \quad i = 1, \ldots, k\}$$

(Rockafellar, 1970, Ch. 14). Note that the dimension of a convex cone containing $\mathbf{0}$ is defined to be the dimension of the smallest subspace containing it.

Lemma 3.4. *The following results hold:*

(a) The rank of $\mathbf{E}$ *is* $l+1$*, i.e.* $V(\beta^*)$ *contains* $l+1$ *linearly independent vectors. Furthermore,* $\mathbf{E}$ *is* irreducible*, i.e. no row of* $\mathbf{E}$ *is a* positive *linear combination of other rows, and the origin is also not a* positive *linear combination of rows of* $\mathbf{E}$*.*

(b) The dimension of $C(\beta^*)^\circ$ *is* $l+1$*.*

(c) The dimension of $C(\beta^*)$ *is* m*.*

Proof. (a): Set $\mathcal{I} = \{i_1 \ldots, i_l\}$, where $\beta^*_{i_j} = 0$ for $j = 1, \ldots, l$. Take the following $l+1$ vectors: for $k = 1, \ldots, l$, let $\tilde{\mathbf{e}}_k \in V(\beta^*)$ have elements $e_{i_k,k} = -1$ and $e_{i_j,k} = 1$ for $j = 1, \ldots, k-1, k+1, \ldots, l$, and let $\tilde{\mathbf{e}}_{l+1} \in V(\beta^*)$ have elements $e_{i_j,l+1} = 1$ for $j = 1, \ldots, l$. It is easy to verify that these $l+1$ vectors are linearly independent and that every other vector in $V(\beta^*)$ can be written as a linear combination of these vectors. This proves that $\mathbf{E}$ has rank $l+1$. Assume that there exist $\lambda_j > 0$ and indices k_j, $j = 1, \ldots, p$, such that $\sum_{j=1}^p \lambda_j \mathbf{e}_{k_j}$ is equal either to $\mathbf{e}_k$ for some $k \notin \{k_1, \ldots, k_p\}$, or to the origin. Since $\|\beta^*\|_1 = t$, there exists at least one i_0 with $\beta^*_{i_0} \neq 0$ and the i_0-th components of all $\mathbf{e} \in V(\beta^*)$ are either all equal to 1 or all equal to -1. Thus $\sum_{j=1}^p \lambda_j$ is equal to 1 if $\sum_{j=1}^p \lambda_j \mathbf{e}_{k_j} = \mathbf{e}_k$ or 0 if $\sum_{j=1}^p \lambda_j \mathbf{e}_{k_j} = \mathbf{0}$. The latter case leads directly to a contradiction. In the former case, since all the components of both $\mathbf{e}_k$ and $\mathbf{e}_{k_j}$, $j = 1, \ldots, p$, have absolute value 1, it is clear that we must have $\mathbf{e}_{k_j i} = \mathbf{e}_{ki}$ for all $j = 1, \ldots, p$ and $i = 1, \ldots, m$. But this contradicts the fact that the $\mathbf{e}$ in $V(\beta^*)$ are all distinct. Hence, $\mathbf{E}$ is irreducible.
(b): The smallest subspace that contains $C(\beta^*)^\circ$ is (Rockafellar, 1970, p. 15)

$$C(\beta^*)^\circ - C(\beta^*)^\circ = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2 \quad \text{and} \quad \mathbf{x}_1, \mathbf{x}_2 \in C(\beta^*)^\circ\}.$$

Hence, the dimension of $C(\beta^*)^\circ$ is $l+1$, the number of linearly independent vectors in $V(\beta^*)$.
(c): The proof follows Meyer (1997) and is based on the fact that $\mathbf{E}$ is irreducible. If the dimension of $C(\beta^*)$ is less than m, then there exists a vector $\mathbf{v} \neq \mathbf{0}$ such that $\mathbf{v}^T\mathbf{x} = 0$ for all $\mathbf{x} \in C(\beta^*)$. But then both $\mathbf{v}$ and $-\mathbf{v}$ must be in $C(\beta^*)^\circ$, contradicting the fact that the origin is not a positive linear combination of rows of $\mathbf{E}$. □

Theorem 3.5. *If* $m > n$ *and* β^* *is a regular solution of* (1.1)*, then* β^* *has at most* n *non-zero entries.*

Proof. Let β^* be a regular solution with l zero and $m-l$ non-zero components. Then, by Lemma 3.4, $\mathcal{N}(\mathbf{E})$ is a $(m-l-1)$-dimensional subspace of $\mathbb{R}^m$ and $\mathcal{N}(\mathbf{E}) \cap \mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$ by definition. Of course $\mathbf{X}$ is assumed to have full rank, so that $\mathcal{N}(\mathbf{X})$ has dimension $m-n$.

Now consider the one-dimensional space $S = \{\eta : \eta = \lambda\bar{\mathbf{e}},\ \lambda \in \mathbb{R}\}$, where $\bar{\mathbf{e}} = \frac{1}{k}\sum_{j=1}^k \mathbf{e}_j$ has entries $\bar{e}_i = 1$ if $\beta^*_i > 0$, $\bar{e}_i = -1$ if $\beta^*_i < 0$ and $\bar{e}_i = 0$ if $\beta^*_i = 0$. It is clear that $\bar{\mathbf{e}}^T\mathbf{e} = m - l > 0$ for all $\mathbf{e} \in V(\beta^*)$, so that $S \cap \mathcal{N}(\mathbf{E}) = \{\mathbf{0}\}$. But also $S \cap \mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$, since otherwise, for $\lambda > 0$ sufficiently small, $\beta^* - \lambda\bar{\mathbf{e}}$ would be a solution of (1.1) with l^1 norm less than that of

β^*, contradicting Theorem 2.1. Since S, $\mathcal{N}(\mathbf{E})$ and $\mathcal{N}(\mathbf{X})$ can span at most $\mathbb{R}^m$, it follows that $(m-l-1)+(m-n)+1 \leq m$, or, equivalently, $m-l \leq n$. □

4 Standard errors of LASSO estimates

If β^* is a solution of (1.1) then it must satisfy (2.8). In Section 3.1 we showed that $\mathbf{X}^T\mathbf{r}$ does not depend on the particular solution β^* and hence the same is true for $\mathbf{v} = \mathbf{X}^T\mathbf{r}/\|\mathbf{X}^T\mathbf{r}\|_\infty$. in (2.8). Combining this with the fact that $\lambda = \mathbf{r}^T\mathbf{X}\beta^*/\|\beta^*\|_1$ if follows from (2.8) that

$$\mathbf{X}^T\mathbf{y} = \left(\mathbf{A} + \frac{1}{\|\beta^*\|_1\|\mathbf{X}^T\mathbf{r}\|_\infty}(\mathbf{X}^T\mathbf{r})(\mathbf{X}^T\mathbf{r})^T\right)\beta^* = (\mathbf{A}+\mathbf{W})\beta^*, \tag{4.1}$$

where $\mathbf{W}$ is a rank-1 matrix. Let $\mathbf{I}_n$ denote the $n \times n$ identity matrix and write

$$\mathbf{A}+\mathbf{W} = \mathbf{X}^T\left(\mathbf{I}_n + \frac{1}{\|\beta^*\|_1\|\mathbf{X}^T\mathbf{r}\|_\infty}\mathbf{r}\mathbf{r}^T\right)\mathbf{X}.$$

This shows that the rank of $\mathbf{A}+\mathbf{W}$ is equal to the rank of $\mathbf{X}$ and thus equal to the rank of $\mathbf{A}$. Hence, if $m \leq n$, then the covariance matrix of the estimates may be approximated by

$$\mathrm{Var}(\beta^*) = (\mathbf{A}+\mathbf{W})^{-1}\mathbf{A}(\mathbf{A}+\mathbf{W})^{-1}\hat{\sigma}^2, \tag{4.2}$$

where $\hat{\sigma}^2$ is an estimate of the error variance.

This should be contrasted with the suggestion of Tibshirani (1996, p. 272) that:

> An approximate closed form estimate may be derived by writing the penalty $\sum|\beta_j|$ as $\sum\beta_j^2/|\beta_j|$. Hence, at the lasso estimate β^*, we may approximate the solution by a ridge regression of the form $\beta^\dagger = (\mathbf{X}^T\mathbf{X}+\mu\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{y}$ where $\mathbf{W}$ is a diagonal matrix with diagonal elements $|\beta_j^*|$, $\mathbf{W}^-$ denotes the generalized inverse of $\mathbf{W}$ and μ is chosen so that $\sum|\beta_j^\dagger| = t$. The covariance matrix of the estimates may then be approximated by
>
> $$(\mathbf{X}^T\mathbf{X}+\mu\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\mu\mathbf{W}^-)^{-1}\hat{\sigma}^2, \tag{7}$$
>
> where $\hat{\sigma}^2$ is an estimate of the error variance.

(The quoted section has been slightly altered to fix a typographical error and to agree more closely with our notation.) We claim that this formula does not yield an appropriate estimate of the covariance matrix of β^*. To support this claim, in Appendix A we show that while both (4.2) and (7) can be motivated via sequences of "smooth" approximations to (1.1), the sequence of approximations leading to (7) breaks down as the original problem (1.1) is approached.

Tibshirani (1996) also notes that (7) "gives an estimated variance of 0 for predictors with $\hat{\beta}_j = 0$". In our view this is inappropriate, and we note that (4.1) yields a positive standard error for all coefficient estimates. Nevertheless, since the distribution of individual LASSO coefficient

estimates will typically have a condensation of probability at zero, they may be far from normally distributed. This suggests that summarising uncertainty by standard errors may not be appropriate and is a topic that deserves further investigation. In Section 6 these issues are examined further in the context of a reanalysis of the prostate cancer data from Tibshirani (1996).

5 Algorithms

In this Section we derive two algorithms for calculating solutions of (1.1). The first algorithm is based on the duality theory above and can be used to compute the LASSO estimator in any setting. The second algorithm is a simple one for the orthogonal design case only. Implementations of these algorithms are available upon request.

5.1 The general case

The iterative algorithm that we propose to solve (1.1) is based on local linearisation of (2.2) about the current value of β. At each step the ith component of β is non-zero if and only if $i \in \sigma$, where the index set σ is updated at various stages of the algorithm.

Let P represent the permutation matrix that collects the non-zero components of β in the first $|\sigma|$ components and write $\beta = P^T \left(\begin{smallmatrix} \beta_\sigma \\ \mathbf{0} \end{smallmatrix} \right)$. Let $\theta_\sigma = \text{sign}(\beta_\sigma)$ have entry 1 if the corresponding entry in β_σ is positive and -1 otherwise. At each step of the algorithm β must be feasible for (1.1), i.e., $\theta_\sigma^T \beta_\sigma \leq t$. This is ensured by our algorithm when the initial set σ and β_σ are chosen appropriately as below.

To obtain the next iterate from the current β, we solve what amounts to a local linearisation of (1.1) about the current β:

$$\underset{\mathbf{h}}{\text{minimise}} \qquad f(\beta + \mathbf{h}) \tag{5.1a}$$

$$\text{subject to} \qquad \theta_\sigma^T(\beta_\sigma + \mathbf{h}_\sigma) \leq t \quad \text{and} \quad \mathbf{h} = P^T \left(\begin{smallmatrix} \mathbf{h}_\sigma \\ \mathbf{0} \end{smallmatrix} \right) \tag{5.1b}$$

If the constraint is active, then the Karush–Kuhn–Tucker conditions (Nash and Sofer, 1996, p. 450) for this problem can be written as

$$\begin{pmatrix} \mathbf{X}_\sigma^T \mathbf{X}_\sigma & \theta_\sigma \\ \theta_\sigma^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{h}_\sigma \\ \mu \end{pmatrix} = \begin{pmatrix} \mathbf{X}_\sigma^T(\mathbf{Y} - \mathbf{X}_\sigma \beta_\sigma) \\ t - \theta_\sigma^T \beta_\sigma \end{pmatrix}, \tag{5.2}$$

whence the solution of (5.1) is

$$\mathbf{h}_\sigma = (\mathbf{X}_\sigma^T \mathbf{X}_\sigma)^{-1}(\mathbf{X}_\sigma^T(\mathbf{Y} - \mathbf{X}_\sigma \beta_\sigma) - \mu \theta_\sigma),$$

where

$$\mu = \max \left(0, \frac{\theta_\sigma^T (\mathbf{X}_\sigma^T \mathbf{X}_\sigma)^{-1} \mathbf{X}_\sigma^T \mathbf{Y} - t}{\theta_\sigma^T (\mathbf{X}_\sigma^T \mathbf{X}_\sigma)^{-1} \theta_\sigma} \right).$$

Let $\beta^\dagger = \beta + \mathbf{h}$. If $\text{sign}(\beta^\dagger_\sigma) = \theta_\sigma$, then we say that $\beta^\dagger$ is *sign feasible*. If $\beta^\dagger$ is not sign feasible, we proceed as follows (Clark and Osborne, 1988).

A1. Move to the first new zero component in direction $\mathbf{h}$, i.e. find the smallest γ, $0 < \gamma < 1$ and corresponding $k \in \sigma$ such that $0 = \beta_k + \gamma h_k$ and set $\beta = \beta + \gamma \mathbf{h}$.

A2. There are now two possibilities. (1) First set $\theta_k = -\theta_k$ and recompute $\mathbf{h}$ by again solving (5.1) with the new β and θ_σ. If $\mathbf{h}$ so computed is a descent direction compatible with the revised θ_σ, then let $\beta^\dagger = \beta + \mathbf{h}$ and proceed to the next stage of the algorithm. (2) Otherwise update σ by deleting k, resetting β_σ and θ_σ accordingly (they are both feasible) and recompute $\mathbf{h}$ for the revised problem (5.1).

A3. Iterate until a sign feasible $\beta^\dagger$ is obtained.

Once a sign feasible $\beta^\dagger$ is obtained, we can test optimality by verifying (2.8). Calculate

$$\mathbf{v}^\dagger = \mathbf{X}^T \mathbf{r}^\dagger / \|\mathbf{X}^T_\sigma \mathbf{r}^\dagger\|_\infty = P^T \begin{pmatrix} \mathbf{v}^\dagger_1 \\ \mathbf{v}^\dagger_2 \end{pmatrix}.$$

where $\mathbf{r}^\dagger = \mathbf{y} - \mathbf{X}\beta^\dagger$. By construction $(\mathbf{v}^\dagger_1)_i = \theta_i$ for $1 \le i \le |\sigma|$ and if $-1 \le (\mathbf{v}^\dagger_2)_i \le 1$ for $1 \le i \le m - |\sigma|$, then $\beta^\dagger$ is a solution of (1.1). Otherwise, we proceed as follows.

B1. Determine the most violated condition, i.e. find s such that $(\mathbf{v}^\dagger_2)_s$ has maximal absolute value.

B2. Update σ by adding s to it and update $\beta^\dagger_\sigma$ by appending a zero as its last element and θ_σ by appending $\text{sign}(\mathbf{v}^\dagger_2)_s$.

B3. Set $\beta = \beta^\dagger$, solve (5.1) and iterate.

Remark 5.1: Justification of the calculations in case that $\beta^\dagger$ is not sign feasible. First note that if the current β is optimal for the restricted problem (5.1), then this portion of the algorithm is skipped. Otherwise, $\mathbf{h}$ is a descent direction, so that the objective f is reduced in the next step. Thus there can be no cycling and the procedure must converge. This procedure must be finite, since there are only finitely many possible configurations of σ. Since convergence of the process would otherwise be contradicted, the final β must be sign feasible.

Remark 5.2: Justification of the calculations in the case that $\beta^\dagger$ is sign feasible. If $\beta^\dagger$ is not optimal for (1.1), then the augmented vector $\begin{pmatrix} \beta^\dagger_\sigma \\ 0 \end{pmatrix}$ is also suboptimal for the augmented problem (5.1) with σ updated by adding s and θ_σ augmented to $\begin{pmatrix} \theta_\sigma \\ \theta_s \end{pmatrix}$. Hence the solution, say $\begin{pmatrix} \tilde{\mathbf{h}}_\sigma \\ h_s \end{pmatrix}$, of the augmented problem will be a descent direction for the augmented problem and, as long as primal feasibility is maintained, for (1.1). The latter fact implies that the algorithm as a whole must converge and requires only that we choose θ_s properly.

To justify our choice of θ_s, note that (5.2) implies $\mu\theta_\sigma = \mathbf{X}_\sigma^T \mathbf{r}^\dagger$. Since $\begin{pmatrix} \tilde{\mathbf{h}}_\sigma \\ h_s \end{pmatrix}$ is a descent direction for the augmented problem, we have

$$0 > -(\mathbf{r}^\dagger)^T \begin{pmatrix} \mathbf{X}_\sigma & \mathbf{x}_s \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{h}}_\sigma \\ h_s \end{pmatrix} = -(\mathbf{r}^\dagger)^T \mathbf{X}_\sigma \tilde{\mathbf{h}}_\sigma - (\mathbf{r}^\dagger)^T \mathbf{x}_s h_s = -\mu(\theta_\sigma^T \tilde{\mathbf{h}}_\sigma + (\mathbf{v}_2^\dagger)_s h_s). \tag{5.3}$$

On the other hand, feasibility for (5.1) requires that

$$\theta_\sigma^T \tilde{\mathbf{h}}_\sigma + \theta_s h_s \leq 0. \tag{5.4}$$

Multiplying (5.4) by μ and adding (5.3) yields

$$0 > \mu(\theta_s - (\mathbf{v}_2^\dagger)_s) h_s$$

Now since $|(\mathbf{v}_2^\dagger)_s| > 1$, we see that choosing $\theta_s = \text{sign}((\mathbf{v}_2^\dagger)_s)$ yields $\text{sign}(h_s) = \theta_s$. But this implies that the linearised constraint for the augmented problem is equivalent to the norm constraint for small enough displacements in the direction $\begin{pmatrix} \tilde{\mathbf{h}} \\ h_s \end{pmatrix}$, i.e., $\theta_\sigma^T(\beta^\dagger + \rho\tilde{\mathbf{h}}_\sigma) + \rho\theta_s h_s \leq t$ is equivalent to $\|\beta^\dagger + \rho \begin{pmatrix} \tilde{\mathbf{h}} \\ h_s \end{pmatrix}\|_1 \leq t$ for small $\rho > 0$. This in turn insures that primal feasibility is maintained in the algorithm.

Remark 5.3: Solving (5.1). The solution to (5.1) is readily and efficiently computed at each stage of the algorithm by maintaining a QR factorisation of $\mathbf{X}_\sigma$. This factorisation can easily be updated and downdated whenever σ is changed.

Remark 5.4: Starting the iteration. The iteration can be started from $\beta = \mathbf{0}$ and $\sigma = \emptyset$, with the first component to add to σ being determined as in part B of the algorithm. Starting from this end of the problem has two advantages:

- It emphasises building up the optimal σ by starting from a small base rather than by pruning a large one which could be illconditioned;
- It permits the computation to proceed while at the same time building up the factorisations mentioned in Remark 5.3.

If the LASSO estimate is to be calculated for several values of t, say $t_1 < t_2 < \cdots < t_k$, then we first solve for t_1 starting with $\beta = \mathbf{0}$ and $\sigma = \emptyset$. For all further values of t_i, we take as starting point the solution for t_{i-1}. This situation occurs if t is to be chosen by, say, generalised cross-validation (Tibshirani, 1996).

Remark 5.5: Advantages over Tibshirani's algorithms. Remark 5.4 already suggests the primary advantages of the current algorithm over that proposed by Tibshirani (1996). Whereas our algorithm starts from a small base to build up the optimal solution, Tibshirani algorithm starts at the solution of the unconstrained problem. If $m > n$, then Tibshirani's approach is infeasible, and if m is large (but not larger than n), then it is inefficient for small to medium sized values of t, as

most of the LASSO coefficient estimates will typically be equal to zero. Similarly, if the LASSO estimate is to be calculated for several ordered values of t, then our algorithm allows the solution at t_{i-1} to be used as a starting point when calculating the solution at t_i.

Remark 5.6: Connexion to l^1-penalised regression. If $m \leq n$, then the algorithm proposed by Fu (1998) can be used to solve (1.2). Algorithms that can be used if m may be larger then n are discussed, within the specific context of wavelet regression, by Chen *et al.* (1999) and Sardy *et al.* (1999). In principle, our algorithm can also be used to solve (1.2). In this case it would be necessary to find that value of t for which the corresponding Lagrange multiplier is equal to the smoothing parameter λ in (1.2). This could be done within a further loop, either by performing a grid search or using a Newton–Raphson algorithm. Note, that since the solution of (1.1) for a value $t = t'$ is a convenient starting point for our algorithm if the bound is changed to $t'' > t'$, even a grid search can be implemented efficiently.

Remark 5.7: Connexion to other subset selection techniques. The way our algorithm calculates the solution to (1.1) illustrates interesting connexions between the LASSO and other well known subset selection techniques. To see this, assume that the regressor variables are centred and rescaled to have (sample) mean zero and (sample) variance one. With this standardisation, the ith entry of $\mathbf{X}^T\mathbf{r}$ is proportional to the (sample) correlation between the ith regressor variable and the vector of residuals. Thus, at each stage the index added to the set σ is the index of the variable that has maximal correlation with the residual vector (of the constrained subproblem (5.1)). This is not dissimilar to forward variable selection (Miller, 1990, Chapter 3.2). However, whenever an index is added to σ we have to solve a new subproblem (5.1) and while solving this new problem it may happen that the indices of some variables are deleted from σ. Thus backward deletion is, practically, "built into" the LASSO and one could argue that it rather behaves like stepwise regression (Miller, 1990, Chapter 3.3). However, the LASSO is driven by an overarching optimality criterion, while the more ad hoc stepwise regression procedure is not. Osborne *et al.* (1999) develop a homotopy method in which the constraint t becomes the homotopy parameter and which can be used to obtain a complete characterisation of all solutions for $0 \leq t \leq t_0$. This approach gives further insight into the relationship between the LASSO and stepwise regression.

5.2 The orthogonal design case

In the orthogonal design case $\mathbf{A}$ is a diagonal matrix and we assume without loss of generality that $\mathbf{A} = \mathbf{I}_m$, where of course $m \leq n$. Tibshirani (1996) notes that in this case the solution to (1.1) is given by

$$\hat{\beta}_i = \text{sign}\left(\hat{\beta}_i^0\right) \max\left(0, \left|\hat{\beta}_i^0\right| - \gamma\right), \qquad i = 1, \ldots, m, \tag{5.5}$$

where the $\hat{\beta}_i^0$'s are the solution to the unconstrained problem (1.1a) and γ is chosen so that $\sum |\hat{\beta}_i| = t$. Using this relationship the LASSO estimator can be calculated easily and we give a simple

algorithm for this purpose.

Letting $(|\hat{\beta}_i^0| - \gamma)_+$ be the positive part of $(|\hat{\beta}_i^0| - \gamma)$, we first note that

$$\begin{aligned} t_0 - t &= \sum_{i=1}^{m} |\hat{\beta}_i^0| - \sum_{i=1}^{m} |\hat{\beta}_i| = \sum_{i=1}^{m} \{ |\hat{\beta}_i^0| - \big(|\hat{\beta}_i^0| - \gamma\big)_+ \} \\ &= \sum_{i=1}^{m} |\hat{\beta}_i^0| I\big(|\hat{\beta}_i^0| \le \gamma\big) + \gamma \sum_{i=1}^{m} I\big(|\hat{\beta}_i^0| > \gamma\big) \\ &= \sum_{i=1}^{K} b_i + \gamma(m - K) \end{aligned}$$

where $b_1 \le \cdots \le b_m$ are the ordered values of $|\hat{\beta}_1^0|, \ldots, |\hat{\beta}_m^0|$ and $K = \max\{i : b_i \le \gamma\}$. Since $t < t_0$, clearly $K < m$ and $b_K \le \gamma < b_{K+1}$. Let $c_0 = 0$ and $c_j = \sum_{i=1}^{j} b_i + b_j(m - j)$ for $j = 1, \ldots, m$, so that $0 = c_0 \le c_1 \le \cdots \le c_m = t_0$. Then

$$0 \le K = \max\{i : c_i \le t_0 - t\},$$

which is easily computed, and

$$\gamma = \Big\{ (t_0 - t) - \sum_{i=1}^{K} b_i \Big\} \Big/ (m - K).$$

6 An example

In this section we reanalyse the prostate cancer data used by Tibshirani (1996). These data come from a study by Stamey *et al.* (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The regressor variables are log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45).

Following Tibshirani (1996), we standardised each regressor such that it had (sample) mean zero and (sample) variance one. This standardisation allows us to incorporate an intercept term whose parameter is not part of the penalty. That is, we are fitting the model

$$\underset{\beta}{\text{minimise}}\, \tfrac{1}{2}(\mathbf{y} - \alpha - \mathbf{X}^T\beta)^T(\mathbf{y} - \alpha - \mathbf{X}^T\beta) \quad \text{such that} \quad \|\beta\|_1 \le t. \tag{6.1}$$

Here $\mathbf{y}$ is the response variable, log(prostate specific antigen) (lpsa), and $\mathbf{X}$ is built from the (standardised) regressors mentioned above. Due to the standardisation we have immediately (see also Tibshirani, 1996)

$$\hat{\alpha} = \tfrac{1}{n} \sum_{i=1}^{n} y_i = \bar{\mathbf{y}}.$$

		Estimated Standard Errors			
			Using (7) with		
	Estimated		Tibshirani's $\mathbf{W}^-$		Moore–Penrose $\mathbf{W}^-$
Predictor	Coefficients	Using (4.2)	$\mu = 2$	$\mu = 17.892$	$\mu = 17.892$
intcpt	2.4784	0.0719	0.0719	0.0719	0.0719
lcavol	0.5588	0.1008	0.0789	0.0536	0.0610
lweight	0.0970	0.0812	0.0602	0.0245	0.0233
age	0.0000	0.0789	0.0000	0.0000	0.0812
lbph	0.0000	0.0801	0.0000	0.0000	0.0779
svi	0.1556	0.0969	0.0713	0.0312	0.0302
lcp	0.0000	0.1245	0.0000	0.0000	0.1044
gleason	0.0000	0.1136	0.0000	0.0000	0.1111
pgg45	0.0000	0.1226	0.0000	0.0000	0.1232

Table 6.1: Coefficient and standard error estimates for the prostate cancer example

Hence, after calculating $\hat{\alpha}$ we may standardise $\mathbf{y}$ such that $\bar{\mathbf{y}} = 0$ and thus transform problem (6.1) into (2.1).

Tibshirani (1996) presents results for the case $t = \hat{s}\|\beta^0\|_1 = 0.8114$, where β^0 is the result from an unconstrained least-squares fit and $\hat{s} = 0.44$ was chosen by generalised cross-validation. We used the algorithm described in Section 5 to fit (6.1) with $t = 0.8114$, obtaining the parameter estimates given in Table 6.1. These estimates are identical to those given by Tibshirani (1996) and reproduced by the software made available by Tibshirani at the Statlib archive at Carnegie Mellon University.

Remark 6.1: There is an error in Tibshirani's routine that evaluates the GCV. The routine only centres and standardises the regressors but neglects to centre the response variable. Hence the residual sum of squares for the GCV function is wrongly calculated. If this error is corrected, then the optimal $\hat{s}$ given by GCV is $\hat{s} = 0.78$. To find these optimal values the GCV function is evaluated on a grid of 10 values evenly spaced between 0 and 1.

We found it difficult to reproduce the standard error estimates given by Tibshirani (1996). By examining Tibshirani's programs we found firstly that he searches for μ only on the interval [0,2] and that for these data $\mu = 2$ was chosen. Secondly, before calculating $\mathbf{W}^- = \text{diag}(1/|\beta_i^*|)$, all the zero entries of β^* are somewhat arbitrarily set to 10^{-11}. Using these two facts, we were able to reproduce his standard errors, given in Table 6.1 in the column labelled $\mu = 2$. These standard errors were also reproduced by Tibshirani's software and, except for the standard error of lcavol, are the same as those given in Table 2 of Tibshirani (1996).

However, calculating $\beta^\dagger = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{y}$ with $\mu = 2$ we found that $\|\beta^\dagger\|_1 = 1.1075$, indicating, not surprisingly, that $\mu = 2$ is not the correct value of the Lagrange multiplier. Indeed, the correct value for enforcing the constraint $\|\beta^\dagger\|_1 = 0.8114$ is $\mu = \lambda = \|\mathbf{X}^T\mathbf{r}\|_\infty = 17.892$. Using this value and Tibshirani's method for estimating the covariance matrix of $\beta^\dagger$ yields the values given in the corresponding column of Table 6.1.

The corrected standard error estimates produced by (7) are quite small and are of course zero for those coefficients estimated to be zero. By comparison the standard errors calculated using (4.2) are all non-zero, and are considerably larger for all of the (constrained) non-zero coefficient estimates.

It is also interesting to note that though the matrix $\mathbf{W}^-$ chosen by Tibshirani is *a* generalised inverse of the matrix $\mathbf{W}$, one might also consider using, for example, the Moore–Penrose inverse (see, e.g., Rao, 1973, p. 26) in (7). The Moore–Penrose inverse places zero into those diagonal elements of $\mathbf{W}^-$ which correspond to parameters that are estimated to be zero. The estimated standard errors that one obtains from (7) using $\mu = 17.892$ and the Moore–Penrose inverse are given in the last column of Table 6.1. In this case, not surprisingly, $\|\beta^\dagger\|_1 = 1.2073 \neq t$.

The estimated standard errors for the parameter that are estimated to be zero yielded by (7) if the Moore–Penrose inverse is used are similar to those obtained from (4.2). However, the estimated standard errors for the non-zero parameter are much smaller. Those estimates are similar to those obtained by using Tibshirani's $\mathbf{W}^-$. Given these results and the discussion of Section 4 and the appendix, we believe that (4.2) is the preferred way of estimating the standard errors of the LASSO estimates.

Acknowledgements

The authors are grateful to Iain Johnstone for permission to use the prostate cancer data and to Wenjiang Fu for providing them. We are also grateful to Bill Venables for a number of helpful comments and discussions. We would further like to thank the Associate Editor and the referees whose comments let to significant improvements in presentation.

A Smooth approximations of the LASSO

In this section we concentrate on the case $m \leq n$ and show how the optimisation problem (2.1) can be approximated "smoothly". This is done by approximating the function $g(\beta)$ by smooth functions. Hence, we are changing the manifold onto which the (unconstrained) ordinary least-squares estimator is projected from the l^1-sphere to a smooth, differentiable manifold. In the following calculations we also assume that the contraint is always enforced. Note, that this is the case with probability arbitrarily close to one for large n since $t < t_0$. Otherwise, the distribution of β^* would clearly be a mixture distribution.

The following analysis also motivates the covariance matrices discussed in Section 4. It is shown that the two matrices stem from different approximations, i.e., from projections onto two different smooth manifolds.

The first approximation is obtained as follows. Consider the family of densities of the form $k(u) = c_\alpha(1-u^2)_+^\alpha$, where $\alpha \geq 0$, c_α is a normalisation constant such that $\int k(u)\,du = 1$ and $(\cdot)_+ = \max(\cdot, 0)$. With $k_c(u) = k(u/c)/c$, set $\psi_c(u) = 2\int_{-c}^u k_c(x)\,dx - 1$ and let $\rho_c(u) = \int_{-c}^u \psi_c(t)\,dt + c$ be

the primitive of ψ_c satisfying $\rho_c(u) = |u|$ for $|u| \geq c$. Then $|u|$ can be smoothly approximated by $\rho_c(u)$ for small c, with the smoothness of the approximation being controlled by α (for $\alpha = 0$ we obtain Huber's ψ and ρ functions). Hence, the first approximation that we consider is to minimise $f(\beta)$ subject to

$$g_c(\beta) = t - \sum_{j=1}^{m} \rho_c(\beta_j) \geq 0. \tag{I}$$

Another well-known approximation to the absolute function $|u|$ is $\sqrt{u^2 + c^2}$ (see, e.g., Koch, 1996). This leads to the second smooth approximation of the constraint (2.1b) by

$$g_c(\beta) = t - \sum_{j=1}^{m} \sqrt{\beta_j^2 + c^2} \geq 0. \tag{II}$$

In what follows, any quantity associated with one of the smooth optimisation problems will be indicated by the subscript c. We shall use additional subscripts I and II, respectively, only if a distinction between the two approximations is necessary. However, it should be noted that for any value of c the values of these quantities depend on the approximation used.

The function $g_c(\cdot)$ is again concave and thus the region over which we minimise is convex. Since we assume that $m \leq n$ and $\mathbf{X}$ has full rank, we are minimising a strictly convex function over a convex region and a unique solution β_c^* to the smooth problem must exist. Since $g_c(\cdot) \to g(\cdot)$ as $c \to 0$, it is easy to show that $\beta_c^* \to \beta^*$, where β^* is the solution of (2.1).

The Kuhn–Tucker conditions for the smooth problem are

$$\mathbf{0} = -\mathbf{X}^T \mathbf{r}_c^* + \lambda_c \mathbf{v}_c^*, \tag{A.1}$$

where the ith component of $\mathbf{v}_c^*$ is

$$v_{c,I,i}^* = \begin{cases} 1 & \text{if } \beta_{c,i}^* \geq c \\ \psi_c(\beta_{c,i}^*) & \text{if } |\beta_{c,i}^*| \leq c \\ -1 & \text{if } \beta_{c,i}^* \leq -c \end{cases} \qquad v_{c,II,i}^* = \frac{\beta_{c,i}^*}{\sqrt{{\beta_{c,i}^*}^2 + c^2}}. \tag{A.2}$$

Note that $\mathbf{v}_{c,I}$ has a form similar to $\mathbf{v}$ in (2.8), whereas all components of $\mathbf{v}_{c,II}$ are strictly between -1 and 1. From the fact that $\beta_c^* \to \beta^*$ as $c \to 0$, it follows that $\mathbf{r}_c^* \to \mathbf{r}^*$ and, using (A.2), ${\beta_c^*}^T \mathbf{v}_c^* \to \|\beta^*\|_1$. Thus it follows from (A.1) that $\lambda_c = {\beta_c^*}^T \mathbf{X}^T \mathbf{r}_c^* / ({\beta_c^*}^T \mathbf{v}_c^*) \to \lambda$ and $\mathbf{v}_c^* \to \mathbf{v}$. That is, all the quantities in (A.1) converge against their counterparts in (2.8) as $c \to 0$.

However, if we look at the matrix equation for β_c^* induced by each approximation another picture emerges. For the first approximation, using calculations similar to those in Section 4, we

obtain

$$\mathbf{X}^T\mathbf{y} = \left(\mathbf{A} + \frac{1}{{\beta^*_{c,I}}^T \mathbf{v}^*_c \, \|\mathbf{X}^T\mathbf{r}^*_c\|_\infty}(\mathbf{X}^T\mathbf{r}^*_c)(\mathbf{X}^T\mathbf{r}^*_c)^T\right)\beta^*_{c,I} = (\mathbf{A} + \mathbf{W}_{c,I})\,\beta^*_{c,I}. \tag{A.3}$$

Thus, we may approximate the variance matrix of $\beta^*_{c,I}$ by (see, among others, Gallant, 1987, Chapter 3.7)

$$(\mathbf{A} + \mathbf{W}_{c,I})^{-1}\,\mathbf{A}\,(\mathbf{A} + \mathbf{W}_{c,I})^{-1}\,\hat{\sigma}^2.$$

These two formulae are similar to (4.1) and (4.2) and converge against the quantities in those equations as c tends to zero.

If we use the second approximation we obtain

$$\mathbf{X}^T\mathbf{y} = \left(\mathbf{A} + \lambda_c \mathbf{W}^{-1}_{c,II}\right)\beta^*_{c,II}, \tag{A.4}$$

where $\mathbf{W}_{c,II} = \text{diag}\left(\sqrt{{\beta^*_{c,II,i}}^2 + c^2}\right)$. The resulting approximation for the covariance matrix of $\beta^*_{c,II}$ is

$$\left(\mathbf{A} + \lambda_c \mathbf{W}^{-1}_{c,II}\right)^{-1}\mathbf{A}\left(\mathbf{A} + \lambda_c \mathbf{W}^{-1}_{c,II}\right)^{-1}\hat{\sigma}^2,$$

which is of the form proposed in (7) of Tibshirani (1996). Note however that a problem arises in the approximation (A.4) as c tends to zero if the solution β^* has at least one zero entry. The matrix $\mathbf{W}_{c,II}$ is non-singular for all $c > 0$ but in this case its limit is singular and those elements on the diagonal of $\mathbf{W}^{-1}_{c,II}$ that correspond to the zero entries of β^* are tending to infinity. In this sense the approximation "breaks down". The more regular behaviour of the approximation (A.3) suggests the use of (4.2) in preference to (7) for estimating the covariance matrix of β^*.

References

Brown, P.J. (1993). *Measurement, regression, and calibration*, Clarendon Press, Oxford.

Chen, S.S., Donoho, D.L. and Saunders, M.A. (1999). Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing* **20**(1): 33–61.
URL: *http://www-stat.stanford.edu/~donoho/Reports/1995/30401.ps.Z*

Clark, D.I. and Osborne, M.R. (1988). On linear restricted and interval least-squares problems, *IMA Journal of Numerical Analysis* **8**: 23–36.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**: 109–148.

Fu, W.J. (1998). Penalized regression: The Bridge versus the Lasso, *Journal of Computational and Graphical Statistics* **7**(3): 397–416.

Gallant, A.R. (1987). *Nonlinear statistical models*, John Wiley & Sons, New York.

Haagen, K., Bartholomew, D.J. and Deistler, M. (eds) (1993). *Statistical modelling and latent variables*, Elsevier/North-Holland, New York; Amsterdam.

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression, *Annals of Statistics* **24**(4): 1648–1666.

Meyer, M.C. (1997). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. Unpublished manuscript.

Miller, A.J. (1990). *Subset Selection in Regression*, Vol. 40 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Nash, S.G. and Sofer, A. (1996). *Linear and Nonlinear Programming*, McGraw–Hill, New York.

Osborne, M.R. (1985). *Finite Algorithms in Optimization and Data Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Chichester.

Osborne, M.R., Presnell, B. and Turlach, B.A. (1998). Knot selection for regression splines via the lasso, *in* S. Weisberg (ed.), *Dimension Reduction, Computational Complexity, and Information*, Vol. 30 of *Computing Science and Statistics*, Interface Foundation of North America, Inc., Fairfax Station, VA 22039–7460, pp. 44–49.
URL: *http://www.stats.adelaide.edu.au/people/bturlach/psfiles/interface98.ps.gz*

Osborne, M.R., Presnell, B. and Turlach, B.A. (1999). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis*. Under consideration.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2 edn, John Wiley & Sons, New York.

Rockafellar, R.T. (1970). *Convex Analysis*, Vol. 28 of *Princeton Mathematical Series*, Princeton University Press, Princeton, New Jersey.

Sardy, S., Bruce, A.G. and Tseng, P. (1999). Block coordinate relaxation methods for nonparametric signal denoising with wavelet dictionaries, *Journal of Computational and Graphical Statistics*. To appear.

Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, *Journal of Urology* **141**(5): 1076–1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.