Applications of l_1 regularisation

M. R. Osborne[1] Tania Prvan[2]

(Received 23 December 2010; revised 27 September 2011)

Abstract

The lasso algorithm for variable selection in linear models, introduced by Tibshirani, works by imposing an l_1 norm bound constraint on the variables in a least squares model and then tuning the model estimation calculation using this bound. This introduction of the bound is interpreted as a form of regularisation step. It leads to a form of quadratic program which is solved by a straight-forward modification of a standard active set algorithm for each value of this bound. Considerable interest was generated by the discovery that the complete solution trajectory parametrised by this bound is piecewise linear and can be calculated very efficiently. Essentially it takes no more work than the solution of either the unconstrained least squares problem or the quadratic program at a single bound value. This has resulted in the study both of the selection problem for different objective and constraint choices and of applications to such areas as data compression and the generation of sparse solutions of very under-determined systems. One important class of generalisation is to quantile regression

estimation problems. The original continuation idea extends to these polyhedral objectives in an interesting two phase procedure which involves both the constrained and Lagrangian forms of the problem at each step. However, it is significantly less computationally effective than is the original algorithm for least squares objectives. In contrast, the piecewise linear estimation problem can be solved for each value of the l_1 bound by a relatively efficient simplicial descent algorithm, and that this can be used to explore trajectory information in a manner which is at least competitive with the homotopy algorithm in this context. The form of line search used in the descent steps has an important bearing on the effectiveness of the algorithm. A comparison is given between the relative performance of the simplicial descent algorithm used and an interior point method on the piecewise linear estimation problem.

Contents

1 Introduction

This article has two objectives. The first is to provide a review of applications of l_1 regularisation along with some insights not found in the literature. The second is to describe exploratory computations that provide interesting

information about the performance of the l_1 descent algorithm in the context of variable selection compared with the homotopy algorithms.

The lasso algorithm for variable selection in linear models, introduced by Tibshirani, works by imposing an l_1 norm bound constraint on the variables in a least squares model and then tuning the model estimation calculation using this bound [13]. This introduction of the bound can be interpreted as a form of regularisation step. It leads to a form of quadratic program which can be solved by a straight-forward modification of a standard active set algorithm for each value of this bound.

The selection problem that motivates Tibshirani's introduction of the lasso starts with a data vector $\mathbf{y} \in R^n$, which could be observations on a signal measured in the presence of noise, and a linear model $X\boldsymbol{\beta}$ where the design matrix $X : R^p \to R^n$ is assumed to have rank $\min(n, p)$. It seeks an economical representation of the data expressed by a close to minimal set of the non-zero components of $\boldsymbol{\beta}$ and corresponding columns of X such that the norm of the residual vector $\|\mathbf{r}\|_2$,

$$\mathbf{r} = \mathbf{y} - X\boldsymbol{\beta}\,, \tag{1}$$

is small in an appropriate sense. The lasso seeks to systematise the search for an economical representation by considering the constrained problem

$$\min_{\beta} \frac{1}{2}\|\mathbf{r}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leqslant \kappa\,. \tag{2}$$

Here κ has the role of a regularisation parameter and it certainly can be used also to control ill-conditioning. No columns of X are selected when κ equals zero which implies that $\boldsymbol{\beta}$ equals the zero vector. Increasing κ adds columns of X into the model, typically one at a time. When κ is large enough there is no constraint on the components of $\boldsymbol{\beta}$ so all are selected when $p \leqslant n$. One consequence is that the Lagrange multiplier μ for the l_1 constraint is zero for κ large enough.

The initial method for solving (2) transformed it into a standard quadratic program for each value of κ tested. Subsequently, it was shown that the

transformation step is unnecessary [9]. However, the big improvement came from the realisation that the optimal solution trajectory was piecewise linear in κ and that this observation could be made the basis of a remarkably efficient algorithm—the whole spectrum of solutions $\boldsymbol{\beta}(\kappa)$ can be calculated at a similar computational cost to the solution of the quadratic program for a single value of κ. This "homotopy" algorithm is discussed in the next section. It has been adapted to fit a range of applications. These include applications in compressed sensing [5], problems with multiple objectives where the form of constraint must be chosen appropriately [14], variable selection in generalised linear models where the likelihood is approximated by quadratic splines [16], and robust variable selection using the Huber-M estimator which involves a mixed piecewise linear, quadratic objective [12]. Different selection problems can be addressed by varying the form of constraint. Examples include those by Bondell and Reich [3] and Zou and Hastie [18]. The original homotopy algorithm for the lasso is summarised in the next section and applications which show its remarkable efficiency summarised.

What is common in all these applications is that the objective is strictly convex, have degree no more than two, and have continuous first derivatives. This has led to attempts to weaken these requirements. Most are related to the quantile regression objective

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

which is continuous, convex, and piecewise linear. Examples include the training of (so-called) l_1 norm support vector machines [17, 15] corresponding to $\tau = 1$ in (3), quantile regression [6, 7], and regularised simultaneous model selection in multiple quantile regressions [19] where the form of constraint used by Turlach, Venables and Wright [14] is used to develop a simultaneous variable selection procedure for simultaneous quantile regressions. The compressed sensing application suggested by Candes and Tao [4] involves the minimization of the l_1 norm of the parameters subject to a maximum norm constraint on the components of the residual vector. This fits the pattern developed here

because it follows from the necessary conditions that the roles of the constraint and objective can be reversed. The extension of the lasso to this class of problems is considered in Section 3. The presentation is specialised to the "generic" quantile regression case $\tau = 0.5$ corresponding to the l_1 objective. The l_1 lasso replaces $\frac{1}{2}\|\mathbf{r}\|_2^2$ in (2) with $\|\mathbf{r}\|_1$. The development of a homotopy algorithm is possible. However, it has important structural differences when compared with C^1 objectives. These have important ramifications for the computational efficiencey of the procedures. Numerical results are presented which allow direct comparison with those derived for the C^1 case.

The final section returns to the consideration of the direct solution of the optimization problem for fixed κ. For the l_1 lasso the necessary conditions show that the constrained problem actually reduces to an augmented l_1 minimization problem for which considerable computational experience is available. The comparisons made here are of the work involved in solving the basic optimization problem for a range of fixed values of κ using an efficient simplicial l_1 solver, and of the relative performance of the descent algorithm and of an interior point algorithm for the l_1 problem. The direct solution procedures are much more competitive for the l_1 case than they are for the least squares lasso. Also, although the simplicial methods appear more efficient for problems with up to several hundreds of observations, there is evidence of a computational penalty growing with n which suggests that the interior point methods would become methods of choice once the number of observations enter the thousands.

2 The least squares lasso

Results in this section are stated without proof. The necessary conditions for a minimum of (2) are [9]

$$\mathbf{r}^T X = \mu \mathbf{u}^T, \quad \mu \geqslant 0, \quad \mathbf{u}^T \in \partial\|\boldsymbol{\beta}\|_1, \quad \mu = \frac{\mathbf{r}^T X \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_1}, \tag{4}$$

where μ is the Lagrange multiplier for the l_1 constraint. Note $\mu = 0$ if $\kappa \geqslant \|\boldsymbol{\beta}_{LS}\|_1$ where $\boldsymbol{\beta}_{LS}$ is the solution of the unconstrained problem. Now introduce an index set ψ pointing to the nonzero components of $\boldsymbol{\beta}$ (the currently selected variables) and a permutation matrix Q_ψ which collects together these nonzero components. Then

$$\boldsymbol{\beta} = Q_\psi^T \begin{bmatrix} \boldsymbol{\beta}_\psi \\ 0 \end{bmatrix}, \quad \mathbf{u} = Q_\psi^T \begin{bmatrix} \theta_\psi \\ \mathbf{u}_2 \end{bmatrix} \in \partial\|\boldsymbol{\beta}\|_1 , \tag{5}$$

$$(\theta_\psi)_j = \operatorname{sign}(\beta_{\psi(j)}), \quad -1 \leqslant (\mathbf{u}_2)_k \leqslant 1 , \quad k \in \psi^c, \tag{6}$$

$$\psi \cup \psi^c = \{1, 2, \ldots, p\}, \quad \mathbf{u}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_1 , \quad \|\mathbf{u}\|_\infty = 1 . \tag{7}$$

For convenience introduce the partial orthogonal factorisation of the design matrix X,

$$XQ^T = S \begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix}, \tag{8}$$

where U_1 is upper triangular, but B need not be reduced, and the auxiliary vector $\mathbf{w}_\psi = U_1^{-T}\theta_\psi$. If the inequalities (6) are strict then differentiating the reduced necessary conditions with respect to the constraint bound gives the system

$$\frac{d\mu}{d\kappa} = -\frac{1}{\mathbf{w}_\psi^T\mathbf{w}_\psi} , \tag{9}$$

$$U_1 \frac{d\boldsymbol{\beta}_\psi}{d\kappa} = \frac{1}{\mathbf{w}_\psi^T\mathbf{w}_\psi}\mathbf{w}_\psi , \tag{10}$$

$$\frac{d(\mu\mathbf{u}_2)}{d\kappa} = -\frac{1}{\mathbf{w}_\psi^T\mathbf{w}_\psi} U_{12}^T\mathbf{w}_\psi . \tag{11}$$

Here the right hand side of the ODE system is independent of κ. It follows that the solution trajectory is locally linear on sub intervals of κ where the l_1 norm is smooth. This means it is a simple computation to follow the solution trajectory until smoothness breaks down! This corresponds either to a component of $\boldsymbol{\beta}_\psi$ becoming zero (a variable deletion step) or a component of $\mathbf{u}_2$ violating its bounds in (6) which corresponds to a variable selection

Table 1: Step counts for the homotopy algorithm—least squares objective.

	p	n	XA	XD
Hald	4	13	4	0
Iowa wheat	8	33	8	0
diabetes	10	442	11	1
Boston housing	13	506	13	0

step [9]. The continuity of the trajectory is guaranteed by the standard perturbation results which show how to restart at these breakpoints.

This observation is the basis for the homotopy algorithm of Osborne, Presnell, and Turlach [9]. It proves to be remarkably efficient, computing the entire solution trajectory in little more than the cost of solving the unconstrained problem and returning significant additional information. It links to the standard least squares solution algorithm based on orthogonal factorization by using stepwise updating techniques in the partial factorisation (8). Results for several classical Google accessible data sets are given in Table 1. Here XA counts homotopy steps while XD counts variable deletions. Variable addition is much the most common action, and this explains the observed efficiency. Tibshirani noted that addition is the only action when columns of the design matrix are orthogonal.

3 Non-smooth objectives

An important example of the kind to be considered is provided by the quantile regression objective (3) parametrised by the quantile parameter τ where the r_i are residuals in the linear model fit. This objective is of interest in econometrics and corresponds to one of the more popular applications of the lasso variable selection technology. The special case of $\tau = 1/2$ gives the l_1 fitting problem. The limiting case $\tau = 1$ of (3) gives, with

very minor modifications to take account of the unconstrained variable, a form of support vector machine ([17]). The idea is that given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$, find a rule so that given a new $\mathbf{x}$ we assign it to a class from $\{-1, 1\}$; this is achieved by solving

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j h_j(x_i) \right) \right]_+ \quad \text{subject to } |\beta_1| + \cdots + |\beta_p| \leqslant \kappa \,.$$

The fitted model is

$$\widehat{f}(x) = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j h_j(x) \,,$$

and the corresponding class assignment is given by $\operatorname{sign} \widehat{f}(x)$.

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 κ 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 subject to an l_1 bound constraint is a decreasing step function of κ with jumps at non-smooth points of both the objective and the constraint. It is necessary to include an explicit multiplier update phase as these jumps have to be determined as part of the computation. This phase uses λ as the homotopy parameter.

The l_1 lasso is a particular case of the quantile regression lasso with the quantile parameter set to 0.5; we solve

$$\min_{\boldsymbol{\beta}} \|\mathbf{r}\|_1 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leqslant \kappa \,. \tag{12}$$

The Lagrangian form with multiplier λ is also required,

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = \|\mathbf{r}\|_1 + \lambda \{ \|\boldsymbol{\beta}\|_1 - \kappa \} \,. \tag{13}$$

The Lagrangian is convex if $\lambda \geqslant 0$. Necessary conditions give

$$0 \in \partial_\beta \mathcal{L}(\boldsymbol{\beta}, \lambda) = \partial_\beta \|\mathbf{r}\|_1 + \lambda \partial_\beta \|\boldsymbol{\beta}\|_1 .$$

This is also the condition for the minimum of the l_1 minimization problem (λ fixed) which is

$$\min_{\boldsymbol{\beta}} \{\|\mathbf{r}\|_1 + \lambda \|\boldsymbol{\beta}\|_1\} . \tag{14}$$

An index set ψ is again used to follow the selected components of $\boldsymbol{\beta}$. However, a second index set σ is needed to follow the residual zeros which here play a significant role in the necessary conditions. Set

$$\sigma = \{i : r_i = 0\} \quad \text{and} \quad \psi = \{i : \beta_i \neq 0\} .$$

Define set complements by $\sigma \cup \sigma^c = \{1, 2, \ldots, n\}$ and $\psi \cup \psi^c = \{1, 2, \ldots, p\}$, and permutation matrices $P_\sigma : R^n \to R^n$ and $Q_\psi : R^p \to R^p$ by

$$P_\sigma \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}, \quad \begin{cases} (\mathbf{r}_1)_i = r_{\sigma^c(i)} \neq 0, & i = 1, 2, \ldots, n - |\sigma|, \\ (\mathbf{r}_2)_i = r_{\sigma(i)} = 0, & i = 1, 2, \ldots, |\sigma|. \end{cases} \tag{15}$$

$$Q_\psi \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \quad \begin{cases} (\boldsymbol{\beta}_1)_i = \beta_{\psi(i)} \neq 0, & i = 1, 2, \ldots, |\psi|, \\ (\boldsymbol{\beta}_2)_i = \beta_{\psi^c(i)} = 0, & i = 1, 2, \ldots, p - |\psi|. \end{cases} \tag{16}$$

$$P_\sigma X Q_\psi^T = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, P_\sigma \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} . \tag{17}$$

The subdifferential components for $\|\mathbf{r}\|_1$ and $\|\boldsymbol{\beta}\|_1$ in the permuted system are

$$\begin{bmatrix} \theta_\sigma^T & \mathbf{v}_\sigma^T \end{bmatrix} \in \partial_\mathbf{r} \|P_\sigma \mathbf{r}\|_1 \quad \text{and} \quad \begin{bmatrix} \theta_\psi^T & \mathbf{u}_\psi^T \end{bmatrix} \in \partial_\beta \|Q_\psi \boldsymbol{\beta}\|_1 .$$

These permit the necessary conditions to be written [10] as

$$\begin{bmatrix} \theta_\sigma^T & \mathbf{v}_\sigma^T \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} = \lambda \begin{bmatrix} \theta_\psi^T & \mathbf{u}_\psi^T \end{bmatrix}, \quad \lambda \geqslant 0 , \tag{18}$$

$$\|\mathbf{v}_\sigma\|_\infty \leqslant 1 , \tag{19}$$

$$\|\mathbf{u}_\psi\|_\infty \leqslant 1\,, \tag{20}$$

$$\|\mathbf{r}\|_1 = \begin{bmatrix} \theta_\sigma^T & \mathbf{v}_\sigma^T \end{bmatrix} P_\sigma \mathbf{r} = \theta_\sigma^T \mathbf{r}_1\,, \tag{21}$$

$$\|\boldsymbol{\beta}\|_1 = \begin{bmatrix} \theta_\psi^T & \mathbf{u}_\psi^T \end{bmatrix} Q_\psi \boldsymbol{\beta} = \theta_\psi^T \boldsymbol{\beta}_1 \leqslant \kappa\,. \tag{22}$$

A similar argument to that used in the least squares case is employed to generate the equations for the homotopy calculations [10]. For the κ-step, the necessary conditions take the form

$$\begin{aligned} \theta_\psi^T \boldsymbol{\beta}_1 &= \kappa\,, \quad \text{“}l_1 \text{ norm condition”}, \\ X_{21} \boldsymbol{\beta}_1 &= \mathbf{y}_2\,, \quad \text{“zero residual conditions”}. \end{aligned}$$

Differentiating with respect to κ gives the “κ phase” equations

$$\theta_\psi^T \frac{d\boldsymbol{\beta}_1}{d\kappa} = 1\,, \tag{23}$$

$$X_{21} \frac{d\boldsymbol{\beta}_1}{d\kappa} = 0\,. \tag{24}$$

This phase finishes when either a new component of $\mathbf{r}$ or a new component of $\boldsymbol{\beta}$ vanishes. Either way it is necessary to repartition the design matrix in (18). Differentiating (18) with respect to λ gives

$$\frac{d\mathbf{v}_\sigma^T}{d\lambda} X_{21} = \theta_\psi^T\,, \tag{25}$$

$$\frac{d\mathbf{v}_\sigma^T}{d\lambda} X_{22} = \frac{d\left(\lambda \mathbf{u}_\psi^T\right)}{d\lambda}\,. \tag{26}$$

Integration of these differential equations terminates when subdifferential components of either $\mathbf{v}_\sigma$ or $\mathbf{u}_\psi$ violate the bounds (19) and (20) respectively. This corresponds to either a residual zero becoming non-zero or a zero component of $\boldsymbol{\beta}$ being flagged to become non-zero.

Numerical results for the same data sets used to produce Table 1 using the least squares lasso are presented in Table 2. Clearly these results are

Table 2: Step counts for homotopy algorithm—l_1 objective

	p	n	SASD	SAXA	XDXA	XDSD
Hald	4	13	17	3	0	0
Iowa wheat	8	33	18	11	1	4
diabetes	10	442	546	12	0	3
Boston housing	13	506	872	28	1	16

less satisfactory. The new feature is the relative importance of the residual sign changes which here trigger update steps caused by these points of non-differentiability, and this causes the extra work as $\mathbf{r}$ adapts to the sign structure required by the necessary conditions. The new notation SA and SD is used to indicate addition and deletion of entries in σ corresponding to the vanishing of a new residual, or a residual becoming nonzero. Double entries (for example SA followed by SD) reflect the κ followed by the λ phases at each step of the computation.

4 Descent computations

The solution of the l_1 lasso for a fixed value of κ is equivalent to an l_1 minimization problem with an augmented design matrix (14). The formulation of this problem as a linear program has a long history [11]. Early implementations encountered similar problems to those reported above for the homotopy algorithm. Typically moving to the next zero residual in an LP formulation encounters the characteristic simplex single step problem illustrated in Figure 1. Barrodale and Roberts [1] discovered a route around this problem; however, simplicial algorithms can be developed directly. From the necessary conditions, a vertex corresponds to a particular set of p residual zeroes, relaxing off one zero in a generic representation of such a set can generate a descent direction, and this is used as a line search direction for the original l_1 objective. This is illustrated in Figure 1. Explicit use of fast

Table 3: l_1 descent calculations—diabetes data.

λ	l_1 iterations	solution zeros	variables selected
5000	12	9	1
2500	12	8	2
1000	18	7	3
500	31	4	6
250	25	5	5
100	35	4	6
50	27	3	7
25	34	2	8
5	44	2	8
0.0001	45	0	10

sorting related algorithms in this line search is described by Bloomfield and Steiger [2]. A modified secant algorithm appears to have a superior performance on systematically generated problems [8]. An alternative approach based on an adaptation of linear programming interior point methods is recommended for large problems [11].

The use of the l_1 descent approach on the diabetes data set is summarised in Table 3 for a range of values of λ. Random initialisation is used and the total number of iterations is 283. The procedure uses a simplicial algorithm and secant method based line-search. Independent starts for each value of λ mean that $p = 10$ iterations, each $O(np)$ operations, are needed to initialise the computation for each λ. The selection information provided by this exercise does not compare too badly with the results of the homotopy algorithm and costs significantly less.

The second table in this section compares the performance of the l_1 descent method with an interior point method specialised to the class of problems considered here. This interior point candidate considered is the "Frisch–Newton" implementation by Koenker and Portnoy for the R statistical programming

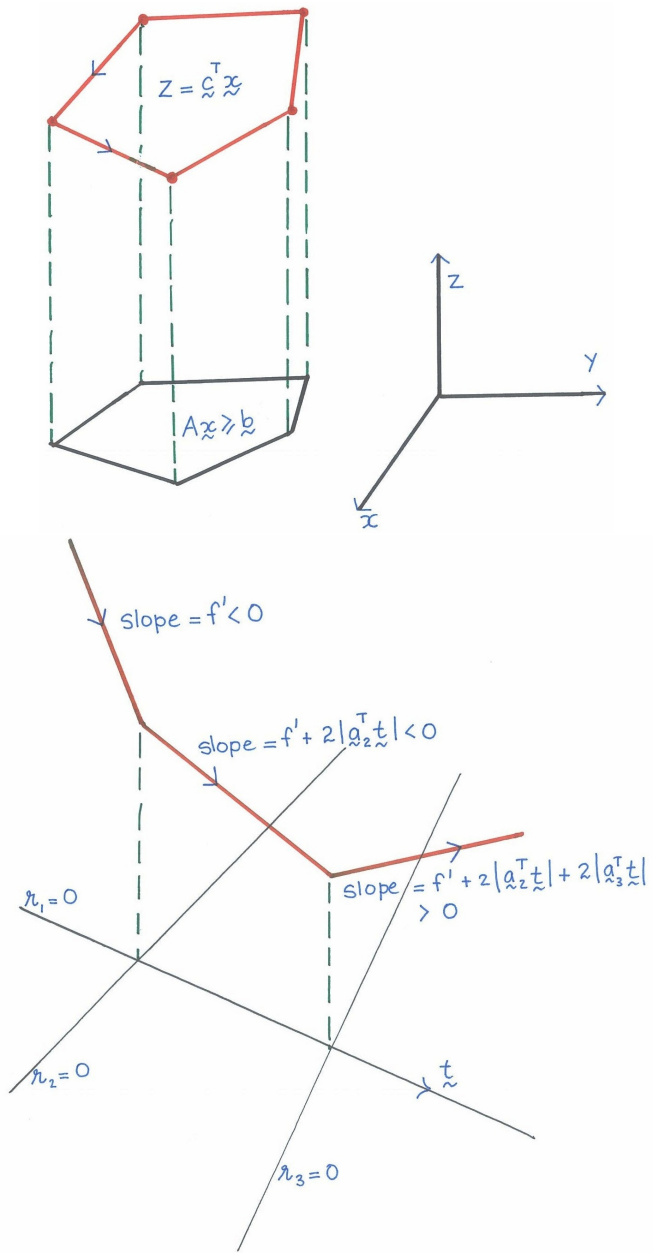

Figure 1: Standard linear programming does not support a line search.

Table 4: Simplicial descent versus interior point methods.

data set	n	p	descent	ls steps	FN IP
Hald	13	4	8	19	8
Iowa wheat	33	8	13	34	9
diabetes	442	10	45	182	10
Boston housing	506	13	56	251	13

language [6, 11]. The specific R package used was `quantreg` and the specific function `rq.fit.sfn`. To compare these figures it is necessary to note that the simplicial l_1 descent iterations cost $O(np)$ operations per recorded iteration, whereas the interior point methods are significantly more computer intense with each recorded iteration costing at least $O(np^2)$ operations. It appears that the simplicial descent method is distinctly competitive for the data sets used here. However, the reported results suggest it could suffer a growth term depending on n which is likely to reverse this conclusion for significantly larger data sets. This is in agreement with the recommendations accompanying the R software [6].

References

[1] I. Barrodale and F. D. K. Roberts. An improved algorithm for l_1 linear approximation. *SIAM J. Numer. Anal.*, 10:839–848, 1973. C876

[2] P. Bloomfield and W. L. Steiger. *Least Absolute Deviations*. Birkhauser, Boston, 1983. C877

[3] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervising clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2008. doi:10.1111/j.1541-0420.2007.00843.x C869

[4] E. Candes and T. Tao. The Dantzig selector: statistical estimation when $\mathfrak{p}$ is much larger than $\mathfrak{n}$. *Ann. Statist.*, 35(6):2313–2351, 2007. doi:10.1214/009053606000001523 C869

[5] D. L. Donoho and Y. Tsaig. Fast solution of $\mathfrak{l}_1$-norm minimization problems when the solution may be sparse. *IEEE Trans. Inf. Theory*, 54:4789–4812, 2008. doi:10.1109/TIT.2008.929958 C869

[6] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005. C869, C879

[7] Y. Li and J. Zhu. L_1-norm quantile regression. *J. Comp. Graph. Stat.*, 17(1):163–185, 2008. doi:10.1198/106186008X289155 C869

[8] M. R. Osborne. *Simplicial algorithms for minimizing polyhedral functions*. Cambridge University Press, 2001. C877

[9] M. R. Osborne, Brett Presnell, and B. A. Turlach. On the Lasso and its dual. *J. Comp. Graph. Stat.*, 9(2):319–337, 2000. http://www.jstor.org/stable/1390657 C869, C870, C872

[10] M. R. Osborne and B. A. Turlach. A homotopy algorithm for the quantile regression lasso and related piecewise linear problems. *J. Comp. Graph. Stat.*, 2010. Accepted for publication. doi:10.1198/jcgs.2011.09184 C874, C875

[11] S. Portnoy and R. Koenker. The Gaussian hare and the Laplacian tortoise: computability of squared error vs absolute error estimates. *Stat. Sci.*, 12:279–300, 1997. http://www.jstor.org/stable/2246216 C876, C877, C879

[12] S. Rosset and Ji Zhu. Piecewise linear regularised solution paths. *Ann. Stat.*, 35(3):1012–1030, 2007. doi:10.1214/009053606000001370 C869

[13] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B*, 58(1):267–288, 1996. http://www.jstor.org/stable/2346178 C868

[14] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005. doi:10.1198/004017005000000139 C869

[15] 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. C869

[16] M. Yuan and H. Zou. Efficient global approximation of generalised nonlinear ℓ_1 regularised solution paths and its applications. *J. Am. Stat. Assoc.*, 104:1562–1573, 2009. doi:10.1198/jasa.2009.tm08287 C869

[17] J. Zhu, T. Hastie, S. Rosset, and R. Tibshirani. $\mathfrak{l}_1$-norm support vector machines. *Adv. Neural Inf. Process. Syst.*, 16:49–56, 2004. http://books.nips.cc/papers/files/nips16/NIPS2003_AA07.pdf C869, C873

[18] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc., Ser. B*, 67:301–320, 2005. http://www.jstor.org/stable/3647580 C869

[19] H. Zou and M. Yuan. Regularised simultaneous model selection in multiple quantiles regression. *Comp. Stat. Data Anal.*, 52:5296–5304, 2008. doi:10.1016/j.csda.2008.05.013 C869

Author addresses

1. **M. R. Osborne**, Mathematical Sciences Institute, Australian National University, ACT 0200, Australia. mailto:Mike.Osborne@anu.edu.au

2. **Tania Prvan**, Department of Statistics, Macquarie University, NSW 2109, Australia.