Variable Selection and Control in Least Squares Problems

M.R. Osborne

Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University

Abstract. The classical technique of stepwise regression provides a paridigm for variable selection in the linear least squares problem. Trust region methods which control the size of the correction to the current solution estimate prove attractive for nonlinear least squares problems because of their good global convergence behaviour. Recently there has been a convergence of these techniques with the realisation that the l_1 trust region method also provides a form of variable selection. These results are reviewed here, and computational methods discussed.

AMS(MOS) classifications: 46N30, 46N40, 65U05.

1. Introduction

The basic linear least squares problem is

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{r}\|_2^2;\ \mathbf{r} = A\mathbf{x} - \mathbf{b}, \tag{1.1}$$

where $A : R^p \rightarrow R^n$, and it is assumed that $p \leq n$, and $\operatorname{rank}(A) = p$.

Two major application classes are considered:

- Variable selection. The computation here is essentially exploratory in nature. The idea is to choose the columns of the design matrix A from a potentially larger class of competitors in such a way as to provide an economical model for the vector of observations $\mathbf{b}$.
- Trust region computation. Here the aim is to estimate a vector of parameters $\mathbf{x}$ by minimizing a nonlinear sum of squares

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n} f_i(\mathbf{x})^2. \tag{1.2}$$

The norm constrained trust region method seeks a correction $\mathbf{h}_k$ to the current estimate $\mathbf{x}_k$ by solving the linear subproblem

$$\min_{\|\mathbf{h}\|\leq\kappa} \frac{1}{2}\|\mathbf{r}\|_2^2;\ \mathbf{r} = \mathbf{f}(\mathbf{x}_k) + \nabla\mathbf{f}(\mathbf{x}_k)\mathbf{h} \tag{LSP}$$

Here the parameter κ is a tuning parameter which is intended to provide a balance between good global convergence behaviour and rate of convergence by ensuring a satisfactory reduction in $F(\mathbf{x})$ at each step.

For our purposes, the connection between these problems is provided by the l_1 trust region method which has the property that solutions for an increasing sequence of values of κ contain, in general, an increasing number of variables. Thus the problem has a mechanism for variable selection. Further, the multiplier associated with the norm constraint decreases as κ increases, and the solution corresponding to $\mu = 0$ is the least squares solution for the currently selected variables.

Important questions concern the scaling of the variables in problem (1.1) and problem (LSP). The sum of squares of residuals is invariant under column rescaling, but this is not true in general of the norm constraint. Rescaling with weights w_i leads to a new norm given by

$$\left\|D(\mathbf{w})^{-1}\mathbf{x}\right\|_D = \left\|D(\mathbf{w})(D(\mathbf{w})^{-1}\mathbf{x})\right\| = \|\mathbf{x}\|$$

where $D(.)$ is a diagonal matrix. One important choice corresponds to $w_j = 1/\left\|A_{*j}\right\|_2$where A_{*j} is the j'th column of A. An important instance of this scaling corresponds to one in which the columns of A are scaled to have unit length in a preprocessing step. In many cases this step can be justified as establishing a basis for comparing the variables. Also, in the least squares regression case, it is usual for the design to have a column of 1's (an intercept term):

$$A = \begin{bmatrix} \mathbf{e} & A_- \end{bmatrix}.$$

In this case the necessary conditions for a minimum are

$$\mathbf{e}^T\mathbf{r} = 0,\ A_-^T\mathbf{r} = 0. \tag{1.3}$$

If the optimal solution to (1.1) is partitioned to highlight the intercept term then

$$\widehat{\mathbf{r}} = \left(I - \frac{\mathbf{e}\mathbf{e}^T}{n}\right)\widehat{\mathbf{r}} = \begin{bmatrix} 0 & \left(I - \frac{\mathbf{e}\mathbf{e}^T}{n}\right)A_- \end{bmatrix}\begin{bmatrix} \widehat{x}_1 \\ \widehat{\mathbf{x}}_1 \end{bmatrix} - \mathbf{b} + \overline{b}\mathbf{e},$$

where the least squares estimates are denoted by hats, and $\overline{b} = \mathbf{e}^T\mathbf{b}/n$. It follows from (1.3) that

$$\widehat{\mathbf{r}}\left(I - \frac{\mathbf{e}\mathbf{e}^T}{n}\right)A_- = 0.$$

Thus $\mathbf{x}_1$ is optimal for the reduced problem corresponding to the substitutions

$$\begin{aligned} A &\leftarrow \left(I - \frac{\mathbf{e}\mathbf{e}^T}{n}\right)A_-, \\ \mathbf{b} &\leftarrow \mathbf{b} - \overline{b}\mathbf{e}, \\ \widehat{x}_1 &= \overline{b} - \sum_{i=2}^{p} \left(\widehat{\mathbf{x}}_1\right)_i \frac{\mathbf{e}^T A_{*i}}{n}. \end{aligned} \tag{1.4}$$

This reduction is called centring the variables.

The plan of the paper is as follows. First the basic results in stepwise regression are summarised. A general trust region algorithm for which the linear subproblem is a norm constrained least squares problem is formulated in the next section, and a convergence proof provided for the nonlinear least squares problem. A novel feature is the form of the acceptance test for the trust region bound. Further insight into the structure of the algorithm is provided by considering the asymptotics for small κ. The results here show selection in the l_1 case, and an interesting robustness in the l_∞ case. The necessary conditions for the l_1 and l_∞ problems show a duality about a common structure which can be exploited in computation. Here, as κ increases, the set of active variables in the l_1 case tends to increase, while the set of extrema in the l_∞ case tends to decrease. An attempt is made in section 6 to develop a general form of algorithm for the case where the unit ball is polyhedral. This provides insight into the relation between the necessary conditions in the l_1 case and the variable selection property. It is noted that a homotopy approach can provide a complete solution to the l_1 constrained problem, and that an implementation based on a modified Gram-Schmidt tableau can have advantages for all methods considered. This material reports briefly on results given in [9]. Numerical results are reported in the final section. These make an interesting point regarding the solution sets that are reachable by the l_1 selection technique, and the form of normalisation of the variables used.

2. Stepwise regression

Stepwise regression has advantages as an exploratory tool both in simplicity of concept, and ease of access to software. As a consequence it has become a distinctly popular technique in exploratory data analysis. It makes the working assumption that an effective basis can be built up by adding columns to the design one at a time with the selection being made to maximise the improvement in the representation of the data vector. In addition, after each new variable is selected, the contributions of each of the current basis set is reviewed to ensure it remains significant. The process is heuristic in nature - there is no requirement for the best set of $p+1$ variables to contain the best set of p variables or to be reachable from it by the allowed strategy of additions and deletions. Nor indeed is there any guarantee that the process will not cycle for reasonable choices of the significance levels for acceptance and rejection. A full discussion of the statistical implications of the technique is given in [4].

It is convenient to begin by assuming that a partial basis has been established, and that the current step of the process sets out to augment this. To itemize this subset introduce an index set

$$\sigma = \{\sigma(1), \sigma(2), \cdots, \sigma(k)\}$$

pointing to the currently selected set of columns and let these be denoted by A_σ. It will be useful to define the orthogonal factorization of A_σ by

$$A_\sigma = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then the least squares problem augmented by the j'th column has the design

$$\mathbf{r}_j = \begin{bmatrix} A_\sigma & A_{*j} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ x_j \end{bmatrix} - \mathbf{b}.$$

The reduction in the sum of squares obtained by adding in the j'th column to the basis is given by

$$\|\mathbf{r}_j\|_2^2 = \|\mathbf{r}_\sigma\|_2^2 - \frac{\left(\mathbf{b}^T Q_2 Q_2^T A_{*j}\right)^2}{A_{*j}^T Q_2 Q_2^T A_{*j}}, \; j \in \sigma^C.$$

This proves a convenient way to compute variable addition as the terms involved are easy to update. Variable deletion also uses this identity, but reverses the process by considering the increases in the sum of squares corresponding to removing variables $j \in \sigma$. It is done more conveniently using an equivalent form based on the identity

$$\frac{\left(\mathbf{b}^T Q_2 Q_2^T A_{*j}\right)^2}{A_{*j}^T Q_2 Q_2^T A_{*j}} = \frac{x_j^2}{s_j \mathbf{e}_j^T \left(A_\sigma^T A_\sigma\right)^{-1} \mathbf{e}_j} \tag{2.1}$$

where

$$s_j = \frac{\|\mathbf{r}_\sigma\|_2^2}{n - k - 1},$$

and the degrees of freedom calculation assumes that the intercept term has been removed from the design and the remaining variables centred (1.4). The connection with a significance test is provided by (2.1) as the right hand side can be identified formally with an F statistic.

3. The norm constrained trust region algorithm

The problem that the trust region algorithm sets out to solve is to find $\mathbf{x}^*$ such that

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} F(\mathbf{x})$$

where $F(\mathbf{x})$ is given by (1.2), and where the problem data is assumed to be at least twice continuously differentiable. Let $\mathbf{r}_\kappa, \mathbf{h}_\kappa$ solve the linear subproblem (LSP)

$$\min_{\|\mathbf{h}\| \le \kappa} \frac{1}{2} \|\mathbf{r}\|_2^2 ; \; \mathbf{r} = \mathbf{f}(\mathbf{x}) + A\mathbf{h}$$

where $A = \nabla \mathbf{f}(\mathbf{x})$. It is assumed that A has full rank p, and this is a sufficient condition for (LSP) to have a unique solution. The essential step is encapsulated in the following result.

Lemma 1 *Let A have full rank p, and let $\mathbf{v} \in \partial \|\mathbf{h}\|$. Then the $(p+1) \times (p+1)$ matrix*

$$\begin{bmatrix} A^T A & \mathbf{v} \\ \mathbf{v}^T & 0 \end{bmatrix}$$

has full rank.

Proof. If the matrix is singular then there exists a vector $\mathbf{z}$ such that

$$\begin{bmatrix} \mathbf{z}^T & -1 \end{bmatrix} \begin{bmatrix} A^T A & \mathbf{v} \\ \mathbf{v}^T & 0 \end{bmatrix} = 0.$$

This gives

$$\begin{aligned} \mathbf{z}^T\mathbf{v} &= 0, \\ \mathbf{z}^T A^T A - \mathbf{v}^T &= \mathbf{0}, \end{aligned}$$

which implies $\mathbf{z}^T A^T A\mathbf{z} = 0$, and this gives a contradiction. ■

The basic trust region algorithm considered is as follows:

```
1. Select $\kappa > 0$, $0 < \sigma < .5$, $0 < \alpha < 1$, $\beta > 1$.
2. Set $step = 0$.
3. Solve (LSP) for $\mathbf{r}_\kappa, \mathbf{h}_\kappa$, set $step = step + 1$.
4. If
       $$\frac{1}{2} - \sigma \le \frac{F(\mathbf{x}) - F(\mathbf{x}+\mathbf{h}_\kappa)}{.5\left(\|\mathbf{f}(\mathbf{x})\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right)} \qquad (3.1)$$
   then  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}_\kappa$,
   else  $\kappa = \alpha * \kappa$,
         repeat 3.
5. If $step = 1$ then $\kappa = \beta * \kappa$.
6. If not converged repeat 2.
```

Discussion of the strong convergence results associated with variants of this algorithm go back to Moré [5] and Osborne [6] who discuss rather different aspects of the case of a Euclidean norm constraint. Discussions that apply to more general classes of problem have been given by Fletcher [3] who considers general (but smooth enough) unconstrained problems, and Osborne [7] who considers a convex composite objective subject to norm constraints which include l_2 and l_∞ but not l_1 norms. Here the main result considers the nonlinear least squares problem subject to a generic norm constraint. The direct attack on this general form of constraint would appear to be a new contribution.

The preliminary results required are as follows.

(i) Provided A has full (column) rank then problem (LSP) involves the minimization of a strictly convex objective function subject to a convex constraint. It has a unique solution, and the necessary conditions have the form

$$\begin{aligned} \mathbf{r}_\kappa^T A &= -\mu_\kappa \mathbf{v}, \\ \mu_\kappa &\ge 0,\ \mu_\kappa\left(\kappa - \|\mathbf{h}_\kappa\|\right) = 0, \\ \mathbf{v} &\in \partial\|\mathbf{h}_\kappa\|,\ \|\mathbf{v}\|^* = 1, \end{aligned} \tag{3.2}$$

where μ_κ is the multiplier, and $\|.\|^*$ denotes the norm dual to the constraint norm. If $\mathbf{v}$ is known then (LSP) can be solved directly. Algorithms for this problem develop the correct $\mathbf{v}$ iteratively. Two useful results follow from (3.2)

$$\mathbf{r}_\kappa^T A\mathbf{h}_\kappa = -\mu_\kappa\|\mathbf{h}_\kappa\| = -\mu_\kappa\kappa, \tag{3.3}$$

$$\mu_\kappa = \left\|\mathbf{r}_\kappa^T A\right\|^* \tag{3.4}$$

Note that if $\mathbf{h}^{LS}$ is the solution of the unconstrained least squares problem (1.1), and if $\left\|\mathbf{h}^{LS}\right\| \le \kappa$, then $\mu_\kappa = 0$, and $\mathbf{h}_\kappa = \mathbf{h}^{LS}$.

(ii) Selectively squaring (LSP) gives

$$\|\mathbf{r}_\kappa\|_2^2 - 2\mathbf{r}_\kappa^T A\mathbf{h}_\kappa + \mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa = \|\mathbf{f}\|_2^2 ,$$

and

$$\|\mathbf{r}_\kappa\|_2^2 = \|\mathbf{f}\|_2^2 + 2\mathbf{f}^T A\mathbf{h}_\kappa + \mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa .$$

Thus

$$\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa = \|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 - 2\mu_\kappa\kappa, \tag{3.5}$$

and

$$\begin{aligned}\mathbf{f}^T A\mathbf{h}_\kappa &= \nabla F(\mathbf{x})\mathbf{h}_\kappa = -\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa - \mu_\kappa\kappa, && (3.6)\\ &= -\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) + \mu_\kappa\kappa, \\ &= -\frac{1}{2}\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) - \frac{1}{2}\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa < 0. && (3.7)\end{aligned}$$

This result shows that (LSP) generates a direction of descent for minimizing $F(\mathbf{x})$. It follows from (3.7) that

$$-\nabla F(\mathbf{x})\mathbf{h}_\kappa \geq \frac{1}{2}\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right). \tag{3.8}$$

(iii) It follows from (3.5) that

$$0 < \mu_\kappa\kappa = \frac{1}{2}(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2) - \frac{1}{2}\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa .$$

A consequence is an order estimate as $\kappa \to 0$,

$$\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 \geq O(\kappa), \ \kappa \to 0, \tag{3.9}$$

at points for which $\mu_\kappa > 0$ as $\kappa \to 0$.
(iv) It is important to have an estimate of μ_κ at points $\mathbf{x}$ for which $\nabla F(\mathbf{x}) \neq 0$. From

$$\begin{aligned}\mu_\kappa &= \left\|\mathbf{r}_\kappa^T A\right\|^* , \\ &= \left\|A^T\mathbf{f} + A^T A\mathbf{h}_\kappa\right\|^* , \\ &\to \left\|\nabla F(\mathbf{x})\right\|^* , \kappa \to 0, && (3.10)\end{aligned}$$

as $\|\mathbf{h}_\kappa\| = \kappa$ for all $\kappa < \left\|\mathbf{h}^{LS}\right\|$. This is actually an upper bound for μ_κ because, from (3.6),

$$\begin{aligned}\mu_\kappa &\leq \left|\mathbf{f}^T A \frac{\mathbf{h}_\kappa}{\|\mathbf{h}_\kappa\|}\right| , \ \kappa \leq \left\|\mathbf{h}^{LS}\right\| , \\ &\leq \left\|\mathbf{f}^T A\right\|^* ,\end{aligned}$$

where the generalised Cauchy inequality is used in the last step.

(v) The observation that $\kappa \mathbf{h}^{LS} / \left\| \mathbf{h}^{LS} \right\|$ is feasible for the constrained least squares problem permits comparison between the constrained and least squares solutions. Let

$$\begin{aligned}
\mathbf{s} &= \mathbf{f} + \kappa A \frac{\mathbf{h}^{LS}}{\left\|\mathbf{h}^{LS}\right\|} \\
&= \mathbf{f} + A\mathbf{h}^{LS} - \left(1 - \frac{\kappa}{\left\|\mathbf{h}^{LS}\right\|}\right) A\mathbf{h}^{LS} \\
&= \mathbf{r}^{LS} - \left(1 - \frac{\kappa}{\left\|\mathbf{h}^{LS}\right\|}\right) A\mathbf{h}^{LS}.
\end{aligned}$$

Then

$$\begin{aligned}
\|\mathbf{r}_\kappa\|_2^2 \leq \|\mathbf{s}\|_2^2 &= \left\|\mathbf{r}^{LS}\right\|_2^2 + \left(1 - \frac{\kappa}{\left\|\mathbf{h}^{LS}\right\|}\right)^2 \left(\mathbf{h}^{LS}\right)^T A^T A\mathbf{h}^{LS}, \\
&= \left\|\mathbf{r}^{LS}\right\|_2^2 + \left(1 - \frac{\kappa}{\left\|\mathbf{h}^{LS}\right\|}\right)^2 \left(\|\mathbf{f}\|_2^2 - \left\|\mathbf{r}^{LS}\right\|_2^2\right).
\end{aligned}$$

It follows that if $\kappa < \left\|\mathbf{h}^{LS}\right\|$ (otherwise the problems have identical solutions) then

$$\|\mathbf{r}_\kappa\|_2^2 \leq \theta \|\mathbf{f}\|_2^2 + (1-\theta) \left\|\mathbf{r}^{LS}\right\|_2^2$$

where $\theta = \left(1 - \frac{\kappa}{\left\|\mathbf{h}^{LS}\right\|}\right)^2$. Rearrangement gives

$$\|\mathbf{f}\|_2^2 - \left\|\mathbf{r}^{LS}\right\|_2^2 \geq \|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 \geq (1-\theta)\left(\|\mathbf{f}\|_2^2 - \left\|\mathbf{r}^{LS}\right\|_2^2\right). \tag{3.11}$$

Note that $\theta \to 0$ as $\kappa \to \left\|\mathbf{h}^{LS}\right\|$.

Theorem 2 *At each point* $\mathbf{x}$ *at which* $\nabla F(\mathbf{x}) \neq 0$, *and* $\|\nabla F(\mathbf{x})\|$ *is bounded it is possible to find* $\kappa > 0$ *such that the solution of (LSP) satisfies (3.1). It follows that the sequence of values* $\{F(\mathbf{x}_j)\}$ *generated by the algorithm converges. The bounded limit points of the corresponding sequence* $\{\mathbf{x}_j\}$ *are stationary points of* $F(\mathbf{x})$ *provided* $\lim_{j\to\infty} \inf_j \{\kappa_j\} > 0$. *If this condition is not satisfied then the sequence* $\{\|\nabla^2 F(\mathbf{x}_j)\|\}$ *is unbounded at finite limit points which are not stationary points of* $F(\mathbf{x})$.

Proof. Taylor expansion plus an application of (3.7) gives

$$\begin{aligned}
F(\mathbf{x}+\mathbf{h}_\kappa) - F(\mathbf{x}) &= \nabla F(\mathbf{x})\mathbf{h}_\kappa + \frac{1}{2}\mathbf{h}_\kappa^T \nabla^2 F(\mathbf{x})\mathbf{h}_\kappa + o(\kappa^2), \\
&= -\frac{1}{2}\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) + \frac{1}{2}\sum_{i=1}^n f_i \mathbf{h}_\kappa^T \nabla^2 f_i(\mathbf{x})\mathbf{h}_\kappa + o(\kappa^2).
\end{aligned}$$

As the first term on the right hand side is $O(\kappa)$ by (3.9), while the second is $O(\kappa^2)$ as $\kappa \to 0$, it follows that (3.1) can be satisfied by taking κ small enough. Consequences are that

- The sequence $\{F(\mathbf{x})\}$ is decreasing. As the sequence is bounded below by assumption it is convergent.
- The sequence $\left\{\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right\} \to 0$.

If $\liminf\{\kappa\} > 0$ then an application of (3.11) shows that $\left\{\|\mathbf{f}\|_2^2 - \left\|\mathbf{r}^{LS}\right\|_2^2\right\} \to 0$. This suffices to establish that limit points are stationary points of $F(\mathbf{x})$ by the Osborne-Watson Lemma [10].

The final point that the sequence $\{\|\nabla^2 F(\mathbf{x}_j)\|\}$ is unbounded if $\lim_{j\to\infty}\inf_j\{\kappa_j\} = 0$ is readily established. For this limit to occur there must be an infinite sequence of steps for which (3.1) does not hold so that

$$\varepsilon > \frac{F(\mathbf{x}+\mathbf{h}_\kappa) - F(\mathbf{x})}{-.5\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right)} = \frac{|\nabla F(\mathbf{x})\mathbf{h}_\kappa| - \left|\frac{1}{2}\mathbf{h}_\kappa^T\overline{\nabla^2 F(\mathbf{x})}\mathbf{h}_\kappa\right|}{.5\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right)}$$

where $\varepsilon = 1/2 - \sigma > 0$, and the bar denotes a mean value. Using (3.6) and rearranging gives

$$\left|\frac{1}{2}\mathbf{h}_\kappa^T\overline{\nabla^2 F(\mathbf{x})}\mathbf{h}_\kappa\right| > \left(\left(1 - \frac{1}{2}\varepsilon\right)\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) - \mu\kappa\right).$$

It follows that

$$\begin{aligned}\left\|\overline{\nabla^2 F(\mathbf{x})}\right\|_2 &> \frac{(2-\varepsilon)\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) - 2\mu\kappa}{\|\mathbf{h}_\kappa\|_2^2}, \\ &> \frac{(2-\varepsilon)\left\{\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa + 2\mu\kappa\right\} - 2\mu\kappa}{\|\mathbf{h}_\kappa\|_2^2}, \\ &> 2(1-\varepsilon)\frac{\mathbf{h}_\kappa^T A^T A\mathbf{h}_\kappa + \mu\kappa}{\|\mathbf{h}_\kappa\|_2^2} = O(1/\kappa),\end{aligned}$$

provided μ is bounded away from 0. This condition is satisfied at finite limit points which are not stationary points of F by (3.10). ■

Remark 3 *This result gives an "almost global" convergence result if F is known to be smooth enough. For example, it works for exponential families, but not for approximation by rationals. A global result is not possible in general because of a lack of compactness results associated with nonlinear families.*

4. Trust region properties for small κ

Trust region properties for small values of the constraint bound κ follow from the multiplier conditions (3.2). In particular, It follows from

$$(\mathbf{f} + A\mathbf{h})^T A = -\mu\mathbf{v}^T,$$

and $\|\mathbf{h}\| \leq \kappa$, that for points at which $\nabla F \neq 0$ the subgradient vector is given by

$$\mathbf{v} \backsimeq -\frac{1}{\mu}\nabla F^T \backsimeq -\frac{\nabla F^T}{\|\nabla F\|^*}.$$

As $\mathbf{v}$ has to satisfy $\mathbf{v}^T\mathbf{h} = \|\mathbf{h}\|$, $\|\mathbf{v}\|^* = 1$, knowledge of $\mathbf{v}$ permits deductions to be made about $\mathbf{h}$.

Example 4 *Because the l_2 norm is self dual it follows that*

$$\mathbf{v} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2} \backsimeq -\frac{\nabla F^T}{\|\nabla F\|_2}.$$

It follows that $\|\mathbf{h}\|$ becomes parallel to the steepest descent direction as $\kappa \to 0$. This recovers a classical result.

Example 5 *In the l_∞ norm $\mathbf{v}$ is attained at the points of extremal deviation of $\mathbf{h}$, and*

$$\|\mathbf{v}\|^* = \sum_{i=1}^{p} |v_i| = 1.$$

Thus, as $\kappa \to 0$,

$$\mathbf{v} \backsimeq \frac{-1}{\sum_{i=1}^{p} |\nabla F_i|} \nabla F^T \Rightarrow \mathbf{h} \backsimeq -\kappa\theta,\ \theta_i = \operatorname{sgn}(\nabla F_i).$$

An interesting feature of this result is that $\mathbf{h}$ is invariant with respect to positive diagonal scaling of ∇F. This translates into invariance with respect to column scaling of A, and this translates directly to invariance with respect to positive diagonal transformation of $\mathbf{x}$. Because in general $\nabla F_i \neq 0$ this implies all components of $\mathbf{h}$ are nonzero as $\kappa \to 0$. This is not a situation that favours variable selection.

Example 6 *In the l_1 norm the small κ solution is given by $\mathbf{x} = \kappa \mathbf{e}_k$ in the case that*

$$k = \arg\max_i |\nabla F_i| \tag{4.1}$$

is uniquely determined. To verify this note that the multiplier equation becomes

$$\nabla F^T + A^T A (h_k \mathbf{e}_k) = -\mu \mathbf{v}$$

where the properties of the l_1 norm require $v_k = \theta = \operatorname{sgn}(h_k)$. Then

$$-\mu\kappa = \theta\kappa\nabla F_k + \kappa^2 \Rightarrow \theta = -\operatorname{sgn}(\nabla F_k),\ \mu = |\nabla F_k| + O(\kappa),$$

assuming the columns of A have been normalised to have length 1. To verify the solution note that

$$|v_i| = \frac{|\nabla F_i|}{|\nabla F_k|}(1 + O(\kappa)) < 1,\ i \neq k.$$

This shows the necessary conditions are satisfied for κ small enough.

Remark 7 *Two aspects of the l_1 norm result are of interest:*

- *Just one variable is selected in contrast to the other norms when , in general, all components of $\mathbf{h}$ are nonzero. This observation can be exploited to develop the l_1 norm constrained problem into a stepwise variable selection algorithm. In contrast, the number of components equal to the norm bound in the l_∞ norm decreases as κ increases. Generically it is just one when $\kappa = \left\|\mathbf{x}^{LS}\right\|$, and zero thereafter.*
- *The rule (4.1) for selecting k corresponds to the rule for entering the first variable in stepwise regression.*

5. Necessary conditions for l_1 and l_∞ constraints

Knowledge of the special form of $\mathbf{v} \in \partial \|\mathbf{h}\|$ can be exploited in designing algorithms for the norm constrained least squares problem. For example, in the case of the l_1 norm there is a permutation matrix P which selects the components of $\mathbf{x}$ at zero level such that:

$$\mathbf{v} \in \partial \|\mathbf{h}\| \Rightarrow P\mathbf{v} = \begin{bmatrix} \theta \\ \mathbf{v}_2 \end{bmatrix}, \; P\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ 0 \end{bmatrix}, \tag{5.1}$$

where $\mathbf{x}_1 \neq 0$, $\theta_i = \operatorname{sgn}((\mathbf{x}_1)_i)$, and $-1 \leq (\mathbf{v}_2)_i \leq 1$. The permutation matrix P is summarised conveniently by means of an index set σ pointing to the nonzero components of $\mathbf{x}$. In the l_∞ norm there is a similar decomposition in which P is determined by an index set σ pointing to the components of $\mathbf{x}$ of maximum modulus. This has the form

$$\mathbf{v} \in \partial \|\mathbf{h}\| \Rightarrow P\mathbf{v} = \begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}, \; P\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \kappa\theta \end{bmatrix}, \tag{5.2}$$

where $\operatorname{sgn}((\mathbf{v}_2)_i) = \theta_i$, $\sum_i |(\mathbf{v}_2)|_i = 1$, and $|(\mathbf{x}_1)_i| \leq \kappa$. The multiplier condition (3.2) can be written

$$PA^T\left(AP^TP\mathbf{x} - \mathbf{b}\right) = -\mu P\mathbf{v}.$$

Introducing the partial orthogonal factorization

$$Q^TAP^T = \begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix}, \; Q^T\mathbf{b} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}$$

reduces the necessary conditions to

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

Substituting the l_1 decomposition gives

$$U_1\mathbf{x}_1 = \mathbf{c}_1 - \mu U_1^{-T}\theta, \tag{5.3}$$

$$\mu\mathbf{v}_2 = B^T\mathbf{c}_2 + \mu U_{12}U_1^{-T}\theta. \tag{5.4}$$

The corresponding result for the l_∞ constraint is

$$U_1\mathbf{x}_1 = \mathbf{c}_1 - \kappa U_{12}\theta, \tag{5.5}$$

$$\mu\mathbf{v}_2 = B^T\mathbf{c}_2 - \kappa B^TB\theta. \tag{5.6}$$

In both cases the solution can be found once the partitioning determining P and the signs determining θ are known.

Remark 8 *One interesting aspect of these systems is that they suggest rather different conditioning properties. In the l_1 case the necessary conditions (5.3) suggest that an inversion of the normal matrix is inescapable forcing a condition number of order* $\operatorname{cond}(A)^2$. *However, in the l_∞ case a dependence on* $\operatorname{cond}(A)$ *is suggested by (5.5).*

6. Algorithms

A basic descent algorithm which includes the l_1 and l_∞ norms as special cases can be given in the case that the norm unit ball is polyhedral. The permutation matrix P then serves to split the components of $\mathbf{x}$ into those active components that enter the characterization conditions - zeros in l_1, components of maximum modulus in l_∞ - from the remainder. The algorithm is an active set method in the above sense, and aims to generate a descent direction by a form of local linearization.

1. Select $\mathbf{x}$ on the trust region boundary. Typically this would be done by shrinking the result of an initial computation which violates the norm bound. Select $\theta \in \partial \|\mathbf{x}_1\|$ where the subscript indicates restriction to the current active set. Select the value of the norm bound κ.
2. Determine a correction $\mathbf{h}$ to the current estimate by solving the linear sub problem

$$\min_{\mathbf{h}\in H} \|\mathbf{r}(\mathbf{x}+\mathbf{h})\|_2^2 ;\ H = \left\{\mathbf{h} : \theta^T (\mathbf{x}+\mathbf{h})_1 \leq \kappa\right\}. \tag{6.1}$$

3. If $\mathbf{x}^* = \mathbf{x}+\mathbf{h}$, and $\theta \in \partial \|\mathbf{x}_1^*\|$ then

   (i) test the multiplier conditions evaluated at the new point

$$A^T\mathbf{r}(\mathbf{x}^*) = -\mu^*\mathbf{v}^*.$$

   If satisfied then solution obtained, stop.

   (ii) else update θ and repeat 2. This step modifies θ to take account of the manner of violation of the multiplier conditions.

4. Else make a descent step in the direction determined by $\mathbf{h}$:

$$\max_{\gamma} \left\{\theta \in \partial \|(\mathbf{x}+\gamma\mathbf{h})_1\|\right\}. \tag{6.2}$$

   This step terminates at a point at which θ must be updated. The modification required to the active set here is complementary to that required by a violation of the multiplier conditions. Repeat step 2. The descent property associated with this step ensures that the other branch in 3. will be chosen eventually.

Remark 9 *In the l_1 case θ is given by the signs of the nonzero components of $\mathbf{x}$, and violation of the multiplier conditions (5.4) in step 3(i) of the algorithm corresponds to $|(\mathbf{v}_2)_s| > 1$ in (5.4). This changes θ by allowing a zero component of $\mathbf{x}$ (say x_k) to become nonzero. The appropriate choice of sign for x_k is given by the sign of the corresponding element of $\mathbf{v}_2$ [9]. In the descent step (6.2) the updating of θ is triggered by a component of $\mathbf{x}$ becoming zero.*

Remark 10 *In the l_∞ case θ is given by the signs of the components of maximum modulus of $\mathbf{x}$. The condition on $\mathbf{v}_2$ that is violated in the multiplier conditions (5.6) corresponds to $\operatorname{sgn}(\mathbf{v}_2)_s \neq \theta_s$, and the action taken is $\sigma \leftarrow \sigma\setminus\{\sigma(s)\}$. The descent step (6.2) terminates when a new component (say $\mathbf{x}_k$) reaches the norm bound. The action taken is $\sigma \leftarrow \sigma \cup \{k\}$. This algorithm plus refinements is discussed in [1].*

Variable selection in the l_1 constrained problem has its origin in the observation (Remark 9) that each step of variable addition occurs as a consequence of the violation of the multiplier condition $|(\mathbf{v}_2)_s| \leq 1$. Typically a strategy is employed that associates the component of $\mathbf{x}$ to relax from zero with the most violated (in an appropriate sense) of these conditions on the multiplier vector. What remains to be spelt out is the manner of selecting the constraint bound κ. As yet this problem does not have any very satisfactory solution, and it can be circumvented by noting that the selection record for $0 \leq \kappa \leq \infty$ can be computed by a piecewise linear homotopy involving only a finite number of linear pieces. This involves integrating the differential equations

$$\begin{aligned} \frac{d\mu}{d\kappa} &= -\frac{1}{\mathbf{w}^T\mathbf{w}}, \\ U_1\frac{d\mathbf{x}_1}{d\kappa} &= \frac{1}{\mathbf{w}^T\mathbf{w}}\mathbf{w}, \\ \frac{d\left(\mu\mathbf{v}_2\right)}{d\kappa} &= -\frac{1}{\mathbf{w}^T\mathbf{w}}U_{12}^T\mathbf{w}, \end{aligned}$$

where $\mathbf{w} = U_1^{-T}\theta$, and where the differential equations are valid provided $-\mu\mathbf{e} < \mu\mathbf{v}_2 < \mu\mathbf{e}$, $|x_i| > 0$, $i \in \sigma$. Continuity of the trajectory at break points corresponding to the equality case in these inequalities is shown in [9].

Implementation of descent and homotopy algorithms for l_1 and l_∞ norm constrained least squares problems can make use of the tableau based modified Gram-Schmidt algorithm considered in [8] for stepwise regression.

7. Numerical results

Numerical results are presented for the homotopy algorithm applied to the Iowa wheat data set [2]. This is displayed in Table 1. Progress of the algorithm is recorded for the two cases corresponding to:

- An explicit intercept variable is added to the data set as variable number 1, each column is scaled to have length 1; and
- The stepwise regression pattern is followed. That is each variable is centred first and then scaled to have length 1.

The point of the example is to show that the two forms of the data set are certainly not equivalent in the presence of the norm constraint. The results are displayed in Table 2 for the centred data, and Table 3 for the data with an explicit intercept added. The tables give the values of κ and μ and the signs of the variables currently selected. In the centred data variable #8 is added after two steps and dropped after a further 4 steps. It is then reinstated with the opposite sign in the final steps of the algorithm. In the case of the explicit intercept the iteration builds up a full complement of variables and then in the final steps of the algorithm drops and then resurrects variables #6 and #8. What is happening is that when the full complement first occurs the constraint is active causing these two to have opposite signs to those that obtain in the full least squares solution. A change in sign forces a drop followed by reentry after the corresponding multiplier moves from one bound to the other. In both cases this change occurs in consecutive steps.

A									b
1930	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	14.76	57.5	3.83	75.5	2.72	77.2	3.30	72.6	32.9
1932	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	16.76	60.5	1.64	77.8	3.45	76.4	3.01	70.5	40.0
1934	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	18.53	59.5	4.67	69.21	4.24	76.5	3.82	75.7	46.3
1939	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	1905	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	15.67	60.00	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

TABLE 1. Iowa wheat data

References

[1] D. I. CLARK AND M. R. OSBORNE, *On linear restricted and interval least-squares problems*, IMA J. Num. Anal., 8 (1988), pp. 23–36.

[2] N. H. DRAPER AND H. SMITH, *Applied Regression Analysis*, Wiley, 1966. 3rd Edition, 1998.

[3] R. FLETCHER, *Practical Methods of Optimization: Unconstrained Optimization*, vol. 1, Wiley, Chichester, 1980.

[4] A. J. MILLER, *Subset Selection in Regression*, Chapman and Hall, 1990.

[5] J. J. MORÉ, *The Levenberg-Marquardt algorithm: Implementation and theory*, in Numerical Analysis. Proceedings, Dundee 1977, G. A. Watson, ed., Springer-Verlag, 1978, pp. 105–116. Lecture Notes in Mathematics No. 630.

κ	μ	2	3	4	5	6	7	8	9	10
0.0	$7.506 - 1$	+								
.28	$4.676 - 1$	+					+			
.53	$2.971 - 1$	+					+	-		
.63	$2.310 - 1$	+					+	-		-
.75	$1.673 - 1$	+	+				+	-		-
.83	$1.433 - 1$	+	+				+	-	+	-
1.03	$8.428 - 2$	+	+				+		+	-
1.06	$7.251 - 2$	+	+	-			+		+	-
1.09	$6.510 - 2$	+	+	-	-		+		+	-
1.21	$4.724 - 2$	+	+	-	-	+	+		+	-
1.66	$2.260 - 2$	+	+	-	-	+	+	+	+	-
1.70	0.0	+	+	-	-	+	+	+	+	-

TABLE 2. Variable selection, centred data

κ	μ	1	2	3	4	5	6	7	8	9	10
0.0	$9.688 - 1$		+								
0.2	$7.652 - 1$		+					+			
0.55	$4.342 - 1$		+	+				+			
0.76	$2.313 - 1$		+	+				+		+	
1.01	$1.550 - 3$		+	+		-		+		+	
1.01	$1.387 - 3$		+	+		-		+	-	+	
1.48	$1.272 - 3$		+	+	-	-		+	-	+	
6.40	$2.747 - 4$		+	+	-	-	-	+	-	+	
6.46	$2.648 - 4$		+	+	-	-	-	+	-	+	-
7.07	$2.031 - 4$	-	+	+	-	-	-	+	-	+	-
14.99	$1.758 - 4$	-	+	+	-	-		+	-	+	-
30.44	$1.200 - 4$	-	+	+	-	-	+	+	-	+	-
65.30	$6.585 - 6$	-	+	+	-	-	+	+		+	-
65.30	$5.089 - 6$	-	+	+	-	-	+	+	+	+	-
67.47	0.0	-	+	+	-	-	+	+	+	+	-

TABLE 3. Variable selection, case of explicit intercept

[6] M. R. OSBORNE, *Nonlinear least squares - the Levenberg algorithm revisited*, J. Aust. Math. Soc., Series B, 19 (1977), pp. 343–357.

[7] ——, *Algorithms for nonlinear approximation*, in The Numerical Solution of Nonlinear Problems, C. T. H. Baker and C. Phillips, eds., Oxford University Press, 1981, pp. 270–286.

[8] ——, *Gram-Schmidt for least squares regression problems: A sweep based algorithm for orthogonal factorization*, Tech. Report SMS-015-90, School of Mathematical Sciences, Australian National University, 1990.

[9] M. R. OSBORNE, B. PRESNELL, AND B. A. TURLACH, *Berwin's problem*, tech. report, Australian National University, School of Mathematical Sciences, 1998.

[10] M. R. OSBORNE AND G. A. WATSON, *Nonlinear approximation problems in vector norms*, in

Numerical Analysis. Proceedings, Dundee 1977, G. A. Watson, ed., Springer Verlag, 1978, pp. 117–132. Lecture Notes in Mathematics No. 630.