Variable selection and control in the linear least squares problem

M.R.Osborne
Centre for Mathematics and its Applications
School of Mathematical Sciences
Australian National University

1 Introduction

1.1 The least squares problem

The basic problem is

$$\begin{aligned} \min_{\mathbf{x}} \|\mathbf{r}\|_2^2 ; \qquad & \mathbf{r} = A\mathbf{x} - \mathbf{b}, \\ A : R^p \rightarrow R^n, \qquad & \text{rank}(A) = p. \end{aligned}$$

Two main applications are considered

- Exploratory computations involving variable selection. These involve the making of a choice in selecting the columns of A that will be used in fitting to $\mathbf{b}$.

- A sub class of trust region computations. Here systematic restriction of the solution vector involving control of its length is used to force reduction in the objective function in nonlinear estimation problems.

1.2 Stepwise regression

Let A_σ be the submatrix with columns $\sigma(1), \cdots, \sigma(k)$, $\mathbf{r}_\sigma,\ \mathbf{x}_\sigma$ solutions to the partial regression

$$\begin{aligned} A_\sigma &= Q_\sigma U_\sigma = \left[\begin{array}{cc} A_{\sigma_j} & \mathbf{a}_{\sigma(j)} \end{array}\right] P_{\sigma_j},\ H_\sigma = Q_\sigma Q_\sigma^T, \\ \|\mathbf{r}_\sigma\|_2^2 &= \left\|\mathbf{r}_{\sigma_j}\right\|_2^2 - \frac{\left(\mathbf{b}^T\left(I - H_{\sigma_j}\right)\mathbf{a}_{\sigma(j)}\right)^2}{\mathbf{a}_{\sigma(j)}^T\left(I - H_{\sigma_j}\right)\mathbf{a}_{\sigma(j)}} \\ &= \left\|\mathbf{r}_{\sigma_j}\right\|_2^2 - \frac{(\mathbf{x}_\sigma)_j^2}{s_\sigma\left(A_\sigma^T A_\sigma\right)_{jj}^{-1}},\ s_\sigma = \frac{\|\mathbf{r}_\sigma\|_2^2}{\mathsf{ndf}}. \end{aligned}$$

Typically selection done by testing $H_0 : (\mathbf{x}_\sigma)_j = 0$ using statistic

$$t = \left|(\mathbf{x}_\sigma)_j\right| \Big/ \left(s_\sigma\left(A_\sigma^T A_\sigma\right)_{jj}^{-1}\right)^{1/2}.$$

1.3 Nonlinear least squares

$$\min_{\mathbf{x}} \sum_{i=1}^{n} f_i(\mathbf{x})^2, \ \mathbf{x} \in R^p.$$

Basic algorithm

$$\begin{array}{rcl} & & \mathbf{x} = \mathbf{x}_0 \\ 1 & : & \min_{\mathbf{h}} \|\mathbf{f} + \nabla \mathbf{f}\, \mathbf{h}\|_2^2 \\ & & \mathbf{x} \leftarrow \mathbf{x} + \mathbf{h} \\ & & \text{repeat 1 if no convergence} \end{array}$$

Need to control $\mathbf{h}$ to improve convergence. Consider

$$\min_{\mathbf{h} \in T(\mathbf{x}, \kappa)} \|\mathbf{f} + \nabla \mathbf{f}\, \mathbf{h}\|_2^2 . \qquad \text{(LSP)}$$

$T(\mathbf{x}, \kappa)$ is "trust region".κ is the control parameter. Most studied case is Levenberg algorithm

$$T(\mathbf{x}, \kappa) = \left\{ \mathbf{h};\ \|\mathbf{h}\|_2^2 \leq \kappa \right\}.$$

Obvious generalisation is to other norms

$$T(\mathbf{x}, \kappa) = \left\{ \mathbf{h};\ \|\mathbf{h}\|^2 \leq \kappa \right\}.$$

1.4 Basic trust region algorithm

1. Select $\kappa > 0, 0 < \sigma < .5, 0 < \alpha < 1, \beta > 1$.

2. Set $step = 0$.

3. Solve (LSP) for $\mathbf{r}_\kappa, \mathbf{h}_\kappa$, $step \leftarrow step + 1$

4. If

$$\frac{1}{2} - \sigma \leq \frac{F(\mathbf{x}) - F(\mathbf{x} + \mathbf{h}_\kappa)}{.5\left(\|\mathbf{f}(\mathbf{x})\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right)}$$

then $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}_\kappa$

else $\{\kappa \leftarrow \alpha * \kappa$;

repeat 3$\}$

5. If $step = 1$ then $\kappa \leftarrow \beta * \kappa$

6. If not converged repeat 2.

1.5 Basic properties

$$\begin{aligned} F &= \frac{1}{2}\sum_{i=1}^{n} f_i^2, \\ \mathbf{g}^T &= \nabla F = \mathbf{f}^T \nabla \mathbf{f}, \\ \nabla^2 F &= \nabla \mathbf{f}^T \nabla \mathbf{f} + \sum_{i=1}^{n} f_i \nabla^2 f_i \end{aligned}$$

To compute $\mathbf{h}$

$$\min_{\|\mathbf{h}\| \le \kappa} \|\mathbf{f} + \nabla \mathbf{f}\, \mathbf{h}\|_2^2 .$$

Kuhn-Tucker conditions are

$$\begin{aligned} & (\mathbf{f} + \nabla \mathbf{f}\, \mathbf{h})^T \nabla \mathbf{f} = -\boldsymbol{\mu} \mathbf{v}^T, \\ \mu \ge 0, \qquad & \mathbf{v} \in \partial \|\mathbf{h}\|, \ \mu (\kappa - \|\mathbf{h}\|) = 0. \end{aligned}$$

Using $\|\mathbf{v}\|_* = 1$ in dual norm gives

$$\mu = \left\| (\mathbf{f} + \nabla \mathbf{f}\, \mathbf{h})^T \nabla \mathbf{f} \right\|_* .$$

Using $\mathbf{v}^T \mathbf{h} = \|\mathbf{h}\|$ gives

$$\begin{aligned} \mu\kappa &= -\nabla F \mathbf{h} - \mathbf{h}^T \nabla \mathbf{f}^T \nabla \mathbf{f} \mathbf{h} \ge 0 \\ \Longrightarrow & -\nabla F \mathbf{h} \ge \mathbf{h}^T \nabla \mathbf{f}^T \nabla \mathbf{f} \mathbf{h} \ge 0. \end{aligned}$$

1.6 Estimate of linear decrease

Selectively squaring (LSP) gives

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

and

$$\nabla F(\mathbf{x})\,\mathbf{h}_\kappa = -\frac{1}{2}\left(\|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) - \frac{1}{2}\mathbf{h}_\kappa^T A^T A \mathbf{h}_\kappa.$$

Have following estimates

$$\begin{aligned} \mu_\kappa \le \left\|\mathbf{f}^T A\right\|^* &\rightarrow \left\|\mathbf{f}^T A\right\|^*,\ \kappa \to 0, \\ \|\mathbf{f}\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 &\geq O(\kappa),\ \kappa \to 0. \end{aligned}$$

Comparison with least squares solution

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

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

1.7 Bit about convergence

$$F(\mathbf{x}) - F(\mathbf{x}+\mathbf{h}_\kappa) = -\nabla F \mathbf{h}_\kappa - \frac{1}{2}\mathbf{h}_\kappa^T \nabla^2 F \mathbf{h}_\kappa - \cdots$$

$$= \mu\kappa + \mathbf{h}_\kappa^T \nabla \mathbf{f}^T \nabla \mathbf{f} \mathbf{h}_\kappa - \frac{1}{2}\mathbf{h}_\kappa^T \nabla^2 F \mathbf{h}_\kappa - \cdots$$

$$= \frac{1}{2}\left(\|\mathbf{f}(\mathbf{x})\|_2^2 - \|\mathbf{r}_\kappa\|_2^2\right) - \frac{1}{2}\sum_{i=1}^{n} f_i \mathbf{h}_\kappa^T \nabla^2 f_i \mathbf{h}_\kappa - \cdots$$

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

rhs is close to 1 for fixed $\mathbf{x}$ and κ small enough. Thus algorithm generates decreasing sequence of values $F(\mathbf{x}_j)$. This implies $\|\mathbf{f}(\mathbf{x})\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 \to 0$. That limit points are stationary points follows from $\|\mathbf{f}(\mathbf{x})\|_2^2 - \left\|\mathbf{r}^{LS}\right\|_2^2 \to 0$ which is a consequence of the least squares inequalities provided $\lim_{j\to\infty} \inf_j \kappa_j > 0$. Also, as $\|\mathbf{f}(\mathbf{x})\|_2^2 - \|\mathbf{r}_\kappa\|_2^2 \geq \mu_j \kappa_j$,

$$\lim_{j\to\infty} \inf_j \kappa_j > 0 \Longrightarrow \mu_j \to 0.$$

1.8 Unbounded Hessians

This just about says that the unconstrained algorithm will work eventually.

However, if $\lim_{j\to\infty} \inf_j \kappa_j = 0$ then there is an unbounded sequence of steps in which κ_j is reduced. If this sequence is associated with a finite limit point which is not a stationary point then the Hessian is unbounded there. The condition for decrease is

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

From this it follows that

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

provided the limit point is not a stationary point.This is close to a global convergence result.

1.9 Properties for small κ

$$\mathbf{g} + \nabla \mathbf{f}^T \nabla \mathbf{f} \mathbf{h} = -\boldsymbol{\mu} \mathbf{v}$$

For small κ, $\mathbf{x} \neq \widehat{\mathbf{x}}$, $-\boldsymbol{\mu}\mathbf{v} \approx \mathbf{g}$

$$l_2 : \mathbf{v} = \mathbf{h} / \left\| \mathbf{h} \right\| \Longrightarrow \mathbf{h} \approx -\gamma \mathbf{g}.$$

For small κ algorithm is close to steepest descent.

$$\begin{aligned} l_\infty \quad : \quad & \sum |v_i| = 1, \ \mathbf{v} \approx \frac{1}{\sum |g_i|} \mathbf{g} \\ \Longrightarrow \quad & \mathbf{h} \approx -\gamma \boldsymbol{\theta}, \ \theta_i = \operatorname{sgn}(g_i) \end{aligned}$$

This direction is invariant to positive diagonal scaling of $\mathbf{g}$.

$$\begin{aligned} l_1 \quad : \quad & k = \arg\max_i (|g_i|) \\ & \text{if } \mu \approx |g_k| \text{ then} \\ & |v_k| = 1, \ \left| v_j \right| < 1, \ j \neq k \\ \Longrightarrow \quad & \mathbf{h} = \gamma \mathbf{e}_k \text{ optimal } \kappa \text{ small enough.} \end{aligned}$$

For small κ this is relaxation, but also an example of variable selection.

2 The l_1 trust region

2.1 The problem

$$\min_{\mathbf{x}\in T} \frac{1}{2}\mathbf{r}^T\mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b},\ \mathbf{T} = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq \kappa\}$$

To reflect selection introduce σ pointing to nonzero components of $\mathbf{x}$. Make this explicit

$$\mathbf{x} = P^T \begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix}, \mathbf{v} = P^T \begin{bmatrix} \boldsymbol{\theta}_\sigma \\ \mathbf{v}_2 \end{bmatrix} \in \partial \|\mathbf{x}\|_1 .$$

Introduce partial orthogonal factorization of AP^T

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

Kuhn-Tucker conditions simplify to

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

plus constraint equation. Assumes partitioning.

2.2 Algorithm - develop correction

'Local linearization' about current $\mathbf{x}$, $\|\mathbf{x}\|_1 = \boldsymbol{\theta}_\sigma^T \mathbf{x}_\sigma$

$$\min_{\mathbf{h}} \frac{1}{2} \|\mathbf{r}(\mathbf{x}+\mathbf{h})\|_2^2 \,;\; \boldsymbol{\theta}_\sigma^T(\mathbf{x}_\sigma + \mathbf{h}_\sigma) \leq \kappa$$

The Kuhn-Tucker conditions are

$$A_1^T \mathbf{r}(\mathbf{x}+\mathbf{h}) + \widetilde{\mu}\boldsymbol{\theta}_\sigma = 0,\; \widetilde{\mu} \geq 0$$

Assuming the constraint is satisfied ($\mu > 0$)

$$\begin{aligned} U_1 \mathbf{h}_\sigma &= Q_1^T \mathbf{r}(\mathbf{x}) - \widetilde{\mu} U_1^{-T} \boldsymbol{\theta}_\sigma, \\ \boldsymbol{\theta}_\sigma^T \mathbf{h}_\sigma &= \kappa - \|\mathbf{x}\|_1 . \end{aligned}$$

This can be solved directly ($\mathbf{w} = U_1^{-T}\boldsymbol{\theta}_\sigma$)

$$\widetilde{\mu} = \frac{\mathbf{w}^T \mathbf{c}_1 - \kappa}{\mathbf{w}^T \mathbf{w}} \text{ (same formula).}$$

$\mathbf{x}$ *sign feasible* if the optimum satisfies

$$\mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x}_\sigma + \mathbf{h}_\sigma \\ 0 \end{bmatrix}, \text{sgn}(\mathbf{x}_\sigma) = \boldsymbol{\theta}_\sigma.$$

It can now be checked for optimality by testing

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

2.3 $\mathbf{x} + \mathbf{h}$ not sign feasible

1. Move to first new zero

$$0 = x_k \leftarrow x_k + \gamma h_k,\ k \in \sigma,\ 0 < \gamma < 1.$$

2. Now two possibilities. Either

$$\theta_k \leftarrow -\theta_k,\ \text{recompute } \mathbf{h}$$

gives direction compatible with new $\boldsymbol{\theta}_\sigma$ (saves updating factorization), or

$$\sigma \leftarrow \sigma \setminus \{k\},\ \text{recompute } \mathbf{h}$$

3. Iterate until $\mathbf{x}$ sign feasible.

If $\mathbf{x}$ not optimal for 'linearised problem' then $\mathbf{h}$ is a descent direction. Implies objective reduced in descent step. Thus no cycling and convergence in finite number of steps. Result must be sign feasible.

2.4 $\mathbf{v}_2$ not feasible

1. Determine the most violated multiplier condition - say involving $(\mathbf{v}_2)_s$. That is find s so that either $(\mathbf{v}_2)_s < -1$, or $(\mathbf{v}_2)_s > 1$.

2. Update

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

where θ_s chosen so that $\mathrm{sgn}(\theta_s) = \mathrm{sgn}(h_s)$. Correct choice is

$$\theta_s = \mathrm{sgn}((\mathbf{v}_2)_s).$$

3. Iterate to find next sign feasible $\mathbf{x}$.

2.5 Verification of sign rule

First step gives a descent direction

$$\begin{aligned} 0 &> \mathbf{r}(\begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix})^T \begin{bmatrix} A_1 & \mathbf{a}_s \end{bmatrix} \begin{bmatrix} \mathbf{h}_\sigma \\ h \end{bmatrix} \\ &= \mathbf{r}^T A_1 \mathbf{h}_\sigma + \mathbf{r}^T \mathbf{a}_s h = -\mu \boldsymbol{\theta}_\sigma^T \mathbf{h}_\sigma - \mu(\mathbf{v}_2)_s h. \end{aligned}$$

Feasibility gives

$$\boldsymbol{\theta}_\sigma^T \mathbf{h}_\sigma + \theta_s h \leq 0.$$

Adding μ times this inequality to the previous gives

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

It follows that the constraint is satisfied

$$\|\mathbf{x}_\sigma\|_1 + |x_s| = \begin{bmatrix} \mathbf{x}_\sigma & x_s \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\sigma \\ \theta_s \end{bmatrix} \leq \kappa.$$

This shows we have the correct linearization.

2.6 l_∞ case as contrast

With $\mathbf{x}$ suitably permuted the K-T conditions are

$$A^T \left(A \begin{bmatrix} \mathbf{x}_\sigma \\ \kappa \boldsymbol{\theta}_\sigma \end{bmatrix} - \mathbf{b} \right) = -\mu \begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}$$

where $\mathbf{v}_2^T \boldsymbol{\theta}_\sigma = 1$, $\|\mathbf{v}_2\|_1 = 1$, $\|\mathbf{x}_\sigma\|_\infty < \kappa$. The governing equations are

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

Similar algorithm is possible!

3 Homotopy

3.1 Basic equations

A complete characterization of solutions for $0 < \kappa < \infty$ can be found via a piecewise linear homotopy. If the optimal solution for a particular value of κ satisfies

$$-\mu\mathbf{e} < \mu\mathbf{v}_2 < \mu\mathbf{e}, \ |\mathbf{x}_\sigma| > 0$$

then same is true for small perturbations about the current values. Differentiating the K-T conditions gives

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

Right hand side constant implies piecewise linearity.

3.2 Critical Points

What happens when one of the constraints becomes an equality ($\kappa = \kappa^*$)? To answer assume

1. $|\sigma| = k < p$,

2. $|(\mathbf{v}_2)_1| \rightarrow 1, \kappa \uparrow \kappa^*$, and

3. the homotopy is restarted with new state component $\overline{x}_{k+1}$, sgn $(\overline{x}_{k+1}) = \phi$.

Need change in the (partial) factorization

$$\left[I - 2 \begin{bmatrix} 0 \\ \mathbf{z} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{z}^T \end{bmatrix} \right] \begin{bmatrix} U_1 & U_{12} \\ 0 & B \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} U_1 & \lfloor U_{12} \rfloor_1 \\ & \pi \| \lfloor B \rfloor_1 \| \end{bmatrix} & \begin{bmatrix} \lfloor U_{12} \rfloor_2 \\ \mathbf{u}^T \\ \overline{B} \end{bmatrix} \end{bmatrix}$$

3.3 Factorization details

$$
\begin{aligned}
\mathbf{z} &= \gamma\left(\lfloor B\rfloor_1 - \pi \left\|\lfloor B\rfloor_1\right\| \mathbf{e}_1\right),\ \pi = -\,\mathsf{sgn}\left(B_{11}\right) \\
\gamma^2 &= 2\left\|\lfloor B\rfloor_1\right\|\left(\left\|\lfloor B\rfloor_1\right\| - \pi B_{11}\right) \\
\mathbf{c} &\rightarrow \begin{bmatrix} \mathbf{c}_1 \\ c_{11} \\ \overline{\mathbf{c}}_2 \end{bmatrix}; c_{11} = \mathbf{e}_1^T\left(I - 2\mathbf{z}\mathbf{z}^T\right)\mathbf{c}_2 \\
c_{11} &= \frac{\lfloor B\rfloor_1^T \mathbf{c}_2}{\pi\left\|\lfloor B\rfloor_1\right\|} = \mu\frac{\left((\mathbf{v}_2)_1 - \lfloor U_{12}\rfloor_1^T \mathbf{w}\right)}{\pi\left\|\lfloor B\rfloor_1\right\|} \\
\overline{\mathbf{w}} &= \begin{bmatrix} U_1^{-T} \\ -\frac{\lfloor U_{12}\rfloor_1^T U_1^{-T}}{\pi\left\|\lfloor B\rfloor_1\right\|} & \frac{1}{\pi\left\|\lfloor B\rfloor_1\right\|} \end{bmatrix}\begin{bmatrix} \boldsymbol{\theta}_\sigma \\ \phi \end{bmatrix} \\
\overline{w}_{k+1} &= \frac{\left(\phi - \lfloor U_{12}\rfloor_1^T \mathbf{w}\right)}{\pi\left\|\lfloor B\rfloor_1\right\|}
\end{aligned}
$$

The hint is that the sign condition should be

$$\phi = (\mathbf{v}_2)_1\,.$$

3.4 Continuity at critical points

The quantities of concern are μ, $\mathbf{x}$ - this requires $\overline{x}_{k+1}$ continuous with 0 and sign feasible, and $\mu\mathbf{v}_2$.

$$
\begin{aligned}
\overline{\mu} &= \frac{\overline{\mathbf{w}}^T \begin{bmatrix} \mathbf{c}_1 \\ c_{11} \end{bmatrix} - \kappa}{\overline{\mathbf{w}}^T \overline{\mathbf{w}}} = \frac{\mu \mathbf{w}^T \mathbf{w} + w_{k+1} c_{11}}{\overline{\mathbf{w}}^T \overline{\mathbf{w}}} = \mu \\
B^T \mathbf{c}_2 &= B^T \left(I - 2\mathbf{z}\mathbf{z}^T\right)\left(I - 2\mathbf{z}\mathbf{z}^T\right)\mathbf{c}_2 \\
&= \begin{bmatrix} \pi \left\| \lfloor B \rfloor_1 \right\| c_{11} \\ c_{11}\mathbf{u} + \overline{B}\overline{\mathbf{c}}_2 \end{bmatrix} \\
\mu\mathbf{v}_2 &= \begin{bmatrix} \pi \left\| \lfloor B \rfloor_1 \right\| c_{11} \\ c_{11}\mathbf{u} + \overline{B^T}\overline{\mathbf{c}}_2 \end{bmatrix} + \mu U_{12}\mathbf{w} \\
&= \begin{bmatrix} \pi \left\| \lfloor B \rfloor_1 \right\| c_{11} + \mu \lfloor U_{12} \rfloor_1^T \mathbf{w} \\ \overline{B^T}\overline{\mathbf{c}}_2 + \mu \begin{bmatrix} \lfloor U_{12} \rfloor_2^T & \mathbf{u} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ c_{11}/\mu \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \pi \left\| \lfloor B \rfloor_1 \right\| c_{11} + \mu \lfloor U_{12} \rfloor_1^T \mathbf{w} \\ \mu\overline{\mathbf{v}}_2 \end{bmatrix}
\end{aligned}
$$

3.5 Properties

The number of pieces in the homotopy is finite. No index set can repeat because if σ is right at both κ_1 and κ_2 then it is valid for $\kappa_1 \leq \kappa \leq \kappa_2$ by linearity.

$\|\mathbf{r}\|_2^2$ is monotonic decreasing as a function of κ.

$$\begin{aligned}\frac{1}{2}\frac{d\,\|\mathbf{r}\|_2^2}{d\kappa} &= \mathbf{r}^T A_1 \frac{d\mathbf{x}_\sigma}{d\kappa} = \frac{\mathbf{r}^T A_1 U_1^{-1}\mathbf{w}_\sigma}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma} \\ &= \frac{-\mu\boldsymbol{\theta}_\sigma^T U_1^{-1}\mathbf{w}_\sigma}{\mathbf{w}_\sigma^T\mathbf{w}_\sigma} = -\mu < 0.\end{aligned}$$

Problem soluble for small enough κ. Assume component of max modulus in $A^T\mathbf{b}$ is first. Then

$$\sigma = \{1\}, \mu = \left|\mathbf{a}_1^T\mathbf{b}\right| - \kappa\,\|\mathbf{a}_1\|_2^2,\ x_1 = \theta_1\kappa.$$

is the required solution.

4 Implementation

A common base for implementation is provided by the modified Gram-Schmidt algorithm in the case of variable selection, and l_1 and l_∞ trust regions.

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

The result of orthonormalising the first block of rows plus orthogonalising them to the remaining rows is:

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

Updating: To this stage all rows corresponding to A_2 have been treated equally. Just select pivot row.

4.1 Downdating

Start by moving deleted variable to last position.

$$\begin{aligned} P\mathbf{x} &= PU^{-1}P^T T^T T P Q_1^T \mathbf{b} \\ Q_1^T &\leftarrow TPQ_1^T\mathbf{b},\ \mathbf{U}^{-T} \leftarrow TPU^{-T}P^T. \end{aligned}$$

T is the orthogoal matrix which applies the surgery.

$$\begin{bmatrix} x & & & \\ x & x & & \\ x & x & x & \\ x & x & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & & & \\ x & x & & x \\ x & x & x & x \\ x & & & x \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} x & & & \\ x & x & & 0_1 \\ x & x & x & 0_2 \\ x & 1 & 2 & x \end{bmatrix}$$

Finally, eliminate elements below pivotal row in final column, scale last element in pivotal row to 1.

4.2 Hald data (Draper and Smith)

13 observations, 5 variables

κ	μ	$x(1)$	$x(2)$	$x(3)$	$x(4)$	$x(5)$
0.0	.989	0.	0.	0.	0.	0.
0.18	.806	.1837	0.	0.	0.	0.
0.73	.270	.4567	0.	.2739	0.	0.
1.02	.9-2	.5416	.1373	.3434	0.	0.
1.04	.2-4	.4995	.1644	.3432	.0343	0.
1.13	0.	.6466	.1504	.2665	.0140	-.051

l_1 homotopy results

$v(1)$	$v(2)$	$v(3)$	$v(4)$	$v(5)$
1.	.8640	.9917	.8487	.8226
1.	.8791	1.	.8401	.8093
1.	1.	1.	.7684	.8045
1.	1.	1.	1.	.7077
1.	1.	1.	1.	-1.

mutipliers

(1)	5			
(1)	2	5		
(1)	2	3	5	
(1)	2	3		

Variable selection order in stepwise regression
-variables centred, intercept computed separately.

μ	$x(1)$	$x(2)$	$x(3)$	$x(4)$	$x(5)$
0	95.423				0.
0	117.57				-.7382
2.66-2	95.924	0.			-.1023
2.28-2	81.874	1.7058	0.		-.0460
5.53-4	54.461	1.485	.6129		
2.83-3	52.488	1.453	.6608	0.	
5.82-3	50.620	1.528	.6571	.1012	

Descent algorithm, constraint bound = 1.03

4.3 Iowa wheat data(Draper and Smith)

33 observations on 10 variables (including intercept)

κ	μ	2	3	4	5	6	7	8	9	10
0.0	7.51-1	+								
.28	4.68-1	+					+			
.53	2.97-1	+					+	-		
.63	2.31-1	+					+	-		-
.75	1.67-1	+	+				+	-		-
.83	1.43-1	+	+				+	-	+	-
1.03	8.43-2	+	+				+		+	-
1.06	7.25-2	+	+	-			+		+	-
1.09	6.51-2	+	+	-	-		+		+	-
1.21	4.72-2	+	+	-	-	+	+		+	-
1.66	2.26-2	+	+	-	-	+	+	+	+	-
1.70	0.0	+	+	-	-	+	+	+	+	-

Homotopy selection, centred data

κ	μ	1	2	3	4	5	6	7	8	9	10
0.0	9.69-1		+								
0.2	7.65-1		+					+			
0.55	4.34-1		+	+				+			
0.76	2.31-1		+	+				+		+	
1.01	1.55-3		+	+		-		+		+	
1.01	1.38-3		+	+		-		+	-	+	
1.48	1.27-3		+	+	-	-		+	-	+	
6.40	2.75-4		+	+	-	-	-	+	-	+	
6.46	2.65-4		+	+	-	-	-	+	-	+	-
7.07	2.03-4	-	+	+	-	-	-	+	-	+	-
15.0	1.76-4	-	+	+	-	-		+	-	+	-
30.4	1.20-4	-	+	+	-	-	+	+	-	+	-
65.3	6.56-6	-	+	+	-	-	+	+		+	-
65.3	5.09-6	-	+	+	-	-	+	+	+	+	-
67.5	0.0	-	+	+	-	-	+	+	+	+	-

Homotopy selection, explicit intercept.