Computer Centre Publication CC 2/69.

Least Squares Methods in Data Analysis

Edited by R.S. Anderssen and M.R. Osborne

The Australian National University

TABLE OF CONTENTS

* Discussants

Something old,something new, and .....

M.R. Osborne

Mathematical Sciences Institute
Australian National University

Anderssen 70
Canberra, January 30, 2009

Outline

linear least squares

$$\min_{\mathbf{x}} \|\mathbf{r}\|^2, \quad \mathbf{r} = A\mathbf{x} - \mathbf{b}.$$

$A : R^p \rightarrow R^n$, p fixed, n 'large'.

normal equations–Choleski

$$\widehat{\mathbf{x}} = \left(A^T A\right)^{-1} A^T \mathbf{b}$$

Householder (1958), Golub (1965)–Orthogonal transformation

$$A = Q \begin{bmatrix} U \\ 0 \end{bmatrix}, \quad \mathbf{c} = Q^T \mathbf{b},$$

$$\|\mathbf{r}\|^2 = \begin{bmatrix} U\mathbf{x} - \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}^T \begin{bmatrix} U\mathbf{x} - \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix},$$

$$\widehat{\mathbf{x}} = U^{-1}\mathbf{c}_1, \quad \|\widehat{\mathbf{r}}\|^2 = \|\mathbf{c}_2\|^2.$$

Sensitivity

Let $\text{cond}\{A\} = \sigma_{max}/\sigma_{min}$ and ε be an appropriate precision scale then (worst case) algorithm sensitivities proportional to

normal equations $\varepsilon\,\text{cond}\{A\}^2$

Householder Golub, Wilkinson (1966). Two terms appear in their error estimate, the first proportional to $\varepsilon\,\text{cond}\{A\}$, and the second to $\varepsilon\,\|\mathbf{r}\|\,\text{cond}\{A\}^2$.

It was argued that if $\|\mathbf{r}\|$ is small then the orthogonal factorization algorithm would be favoured. This argument has been generally accepted.

There may well be overkill in the $\|\mathbf{r}\|$ term. This term does not allow for the kind of cancellation that can occur when the perturbations are independent random variables in a structured modelling environment. Such effects strengthen the case for orthogonal factorization.

How do rounding errors accumulate in large scale computations?

Gauss-Newton iteration

Start with

$$y_i = f(t, \beta^*) + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $\mathcal{E}\{\varepsilon_i \varepsilon_j\} = 0$, and *errors are additive*.
Problem is to estimate β^* by solving

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})\|^2 .$$

Gauss-Newton method solves sequence of linear problems

$$\mathbf{h}_s = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathbf{f}(\boldsymbol{\beta}_s) - \nabla\mathbf{f}(\boldsymbol{\beta}_s)\mathbf{h}\|^2 ,$$
$$= \left(\nabla\mathbf{f}_s^T \nabla\mathbf{f}_s\right)^{-1} \nabla\mathbf{f}_s^T (\mathbf{y} - \mathbf{f}_s),$$
$$\boldsymbol{\beta}_{s+1} = \boldsymbol{\beta}_s + \mathbf{h}_s .$$

This is a fixed point iteration of the form $\boldsymbol{\beta}_{s+1} = F(\boldsymbol{\beta}_s)$.

rate of convergence

Let $\widehat{\beta}$ minimize the sum of squares. This is a point of attraction of the iteration provided

$$\lambda = \varpi\left\{F'\left(\widehat{\beta}\right)\right\} < 1$$

where $\varpi\left\{F'\right\}$ indicates the spectral radius and the variation $F'\left(\widehat{\beta}\right)$ is given by

$$F'\left(\widehat{\beta}\right) = \left(\nabla\mathbf{f}^T\nabla\mathbf{f}\right)^{-1}\sum_{i=1}^{n}\left(y_i - f_i\right)\nabla^2 f_i.$$

The iteration is first order convergent if $0 < \lambda < 1$. A higher order convergence rate requires $\lambda = 0$.

In 1968

The LSQMIDA paper makes no strong recommendations in a context in which these problems were our largest class of applications, and the software used was usually obtained informally. It made some attempt to describe the convergence behaviour. Basically, it recognised the generic first order convergence rate, stressed that this was potentially a problem, and noted that the Davidon method could be tried if problems were encountered. It mentioned there was also the Levenberg method.

40 years later Gauss-Newton is still going strong while the Levenberg modification is also widely used. The scaling properties of Gauss-Newton are superior No quasi-Newton type implementation seems to have made a big impact in this specific problem context.

large sample behaviour

Gauss-Newton has good large sample behaviour in important cases.

$$F_n'\left(\widehat{\beta}_n\right) = \left(\frac{1}{n}\nabla \mathbf{f}^{(n)T}\nabla \mathbf{f}^{(n)}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}$$
$$\left(\varepsilon_i^n + f_i^{(n)}\left(\beta^*\right) - f_i^{(n)}\left(\widehat{\beta}_n\right)\right)\nabla^2 f_i^{(n)},$$
$$\to 0, \quad n \to \infty,$$
$$\Rightarrow \lambda_n \to 0.$$

Sufficient conditions:

1. $\left(\frac{1}{n}\nabla \mathbf{f}^{(n)T}\nabla \mathbf{f}^{(n)}\right)^{-1} \to$ bounded, positive definite, design;
2. $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^n\nabla^2 f_i^{(n)} \to 0$, law of large numbers;
3. $\frac{1}{n}\sum_{i=1}^{n}\left(f_i^{(n)}\left(\beta^*\right) - f_i^{(n)}\left(\widehat{\beta}_n\right)\right)\nabla^2 f_i^{(n)} \to 0$, consistency.

This is not an isolated example

additive errors

- Fisher scoring
- Wedderburn quasi–likelihoods

separable regression

- Golub and Pereyra variable projection
- Modified Prony algorithms

simultaneous method for ODE estimation Bock iteration

Fisher scoring

$$\mathcal{L}(\beta, \mathbf{y}) = -\sum_{i=1}^{n} \log\left(f\left(t_i, \beta, y_i\right)\right)$$

then scoring iteration is

$$\mathbf{h}_s = -\mathcal{E}\left\{\nabla^2 \mathcal{L}\right\}^{-1} \nabla \mathcal{L}^T,$$
$$\beta_{s+1} = \beta_s + \mathbf{h}_s.$$

Aspects to notice include:

1. $\mathcal{L}$ is a sum of independent random variables.
2. $\mathcal{E}\left\{\nabla^2 \mathcal{L}\right\} = -\mathcal{E}\left\{\nabla \mathcal{L}^T \nabla \mathcal{L}\right\}$. Note second derivatives disappear.
3. orthogonal methods apply.

Variable projection

separable regression.

$$\mathbf{r} = \mathbf{y} - A(\beta)\,\boldsymbol{\alpha}$$

Define the orthogonal matrix $Q(\beta)$ by

$$A(\beta) = \begin{bmatrix} Q_1(\beta) & Q_2(\beta) \end{bmatrix} \begin{bmatrix} U(\beta) \\ 0 \end{bmatrix}.$$

Then

$$\min_{\boldsymbol{\alpha}} \|\mathbf{r}\|^2 = \left\| Q_2(\beta)^T \mathbf{y} \right\|^2,$$

$$\widehat{\boldsymbol{\alpha}} = U^{-1} Q_1^T \mathbf{y}.$$

Variable projection uses Gauss-Newton applied to sum of squares of the transformed residuals. Algorithm is asymptotically second order.

Constrained problems

$$\min_{\mathbf{x}} \Phi(\mathbf{x}), \ \mathbf{c}(\mathbf{x}) = 0, \ \mathbf{c} \in R^m, \ m < p.$$

Introduce *Lagrangian*: $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \Phi(\mathbf{x}) + \Sigma_{i=1}^{m} \lambda_i c_i(\mathbf{x})$.
Necessary conditions are:

$$\nabla_x \mathcal{L} = 0, \ \nabla_\lambda \mathcal{L} = \mathbf{c}^T = 0.$$

Newton correction $(\triangle \mathbf{x}, \ \triangle \boldsymbol{\lambda})$ is given by

$$\begin{aligned} \nabla^2_{xx} \mathcal{L} \triangle \mathbf{x} + \nabla_x \mathbf{c}^T \triangle \boldsymbol{\lambda} &= -\nabla_x \mathcal{L}^T, \\ \nabla_x \mathbf{c} \triangle \mathbf{x} &= -\mathbf{c}. \end{aligned}$$

Let

$$C^T = \nabla_x \mathbf{c}^T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then a sufficient condition for local convergence at $\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}}$ is

$$U \text{ full rank}, \ Q_2^T \nabla^2_{xx} \mathcal{L} Q_2 \succ 0$$

Iterations formulated

Set $\nabla^2_{xx}\Phi = A$, $\sum_{i=1}^{m} \lambda_i \nabla^2_{xx} c_i = B$, $\nabla_x \mathbf{c} = C$. Then basic iteration calculations solve

Newton $\begin{bmatrix} A+B & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \triangle \mathbf{x} \\ \triangle \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L} \\ -\mathbf{c} \end{bmatrix}$.

Bock $\begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \triangle \mathbf{x} \\ \triangle \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L} \\ -\mathbf{c} \end{bmatrix}$.

Write Bock as a fixed point iteration

$$\begin{bmatrix} \mathbf{x}_{i+1} \\ \boldsymbol{\lambda}_{i+1} \end{bmatrix} = \mathbf{F}\left(\begin{bmatrix} \mathbf{x}_i \\ \boldsymbol{\lambda}_i \end{bmatrix} \right),$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} - \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_x \mathcal{L} \\ \mathbf{c} \end{bmatrix}.$$

Condition for an attractive fixed point is $\varpi(\mathbf{F}') < 1$.