Least squares methods in maximum likelihood problems

M.R.Osborne
School of Mathematical Sciences, ANU

September 22, 2004

Abstract: It is well known that the Gauss-Newton algorithm for solving nonlinear least squares problems is a special case of the scoring algorithm for maximizing log likelihoods. What has received less attention is that the computation of the current correction in the scoring algorithm in both its line search and trust region forms can be cast as a linear least squares problem. This is an important observation both because it provides likelihood methods with a general framework which accords with computational orthodoxy, and because it can be seen as underpinning computational procedures which have been developed for particular classes of likelihood problems (for example, generalised linear models). Aspects of this orthodoxy as it affects considerations such as convergence and effectiveness will be reviewed.

1. Point of paper has changed somewhat with more emphasis on what do we know about Gauss-Newton.
2.Will not consider estimation of DE's per se. However,

1. Results apply to stable simple shooting estimation problems, and

2. to suitably imbedded multiple shooting formulations - and we know how to do this.

3. Strictly results do not apply to (our work on) the simultaneous approach to ODE estimation. Problem is technically a mixed problem as limiting multipliers satisfy a stochastic DE.
4. Will not say anything about optimum observation points. However, the information matrix will be much in evidence,and its determinant is a quantity to maximize as a function of these.

Start with independent event outcomes $\mathbf{y}_t \in R^q, \quad t = 1, 2, \cdots, n,$
associated pdf $g(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t})$ indexed by "points" $\mathbf{t} \in T \subset R^l,$
and structural information provided by a known parametric model

$$\boldsymbol{\theta}_t = \boldsymbol{\eta}(\mathbf{t}, \mathbf{x})$$

where $\boldsymbol{\theta} \in R^s$, and $\mathbf{x} \in R^p$.
A priori information is the experimental design T_n, $|T_n| = n$. For asymptotics require

$$\frac{1}{n} \sum_{\mathbf{t} \in T_n} f(\mathbf{t}) \to \int_{S(T)} f(\mathbf{t}) \rho(\mathbf{t}) d\mathbf{t}$$

The problem is given the event outcomes $\mathbf{y}_t$ it is required to estimate $\mathbf{x}$. It is not assumed that the individual components of $\mathbf{y}_t$ are independent.

Example Exponential family:

$$g\left(\mathbf{y};\left[\begin{array}{c}\boldsymbol{\theta}\\ \boldsymbol{\phi}\end{array}\right]\right)=c\left(\mathbf{y},\boldsymbol{\phi}\right)\exp\left[\left\{\mathbf{y}^{T}\boldsymbol{\theta}-b\left(\boldsymbol{\theta}\right)\right\}/a\left(\boldsymbol{\phi}\right)\right]$$

$$\begin{aligned}\mathcal{E}^{*}\{\mathbf{y}\}&=\mu\left(\mathbf{x}^{*},\mathbf{t}\right)=\nabla b\left(\boldsymbol{\theta}\right)^{T},\\ \mathcal{V}^{*}\{\mathbf{y}\}&=a\left(\boldsymbol{\phi}\right)\nabla^{2}b\left(\boldsymbol{\theta}\right).\end{aligned}$$

"signal in noise" model $\theta=\mu$.

(i)normal density

$$\begin{aligned}g&=\frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{1}{2\sigma^{2}}\left(y-\mu\right)^{2}\\ c(y,\phi)&=\frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{y^{2}}{2\sigma^{2}},\quad a(\phi)=\sigma^{2}\\ \theta&=\mu,\quad b(\theta)=\mu^{2}.\end{aligned}$$

(ii) multinomial (discrete) distribution

$$g(\mathbf{n};\omega) = \frac{n!}{\prod_{j=1}^{p} n_j!} \prod_{j=1}^{p} \omega_j^{n_j} = \frac{n!}{\prod_{j=1}^{p} n_j!} e^{\sum_{j=1}^{p} n_j \log \omega_j},$$

where $\sum_{j=1}^{p} n_j = n$, and the frequencies must satisfy the condition $\sum_{j=1}^{p} \omega_j = 1$. Eliminating ω_p gives

$$\sum_{j=1}^{p} n_j \log \omega_j = \sum_{j=1}^{p-1} n_j \log \frac{\omega_j}{1 - \sum_{i=1}^{p-1} \omega_i} + n \log \left(1 - \sum_{j=1}^{p-1}$$

It follows that

$$\theta_j = \log \frac{\omega_j}{1 - \sum_{i=1}^{p-1} \omega_i}, \; b(\boldsymbol{\theta}) = n \log \left(1 - \sum_{j=1}^{p-1} e^{\theta_j} \right).$$

Likelihood: $\mathcal{G}(\mathbf{y}; \mathbf{x}, T) = \prod_{\mathbf{t} \in T} g(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t})$

Estimation principle: $\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{G}(\mathbf{y}; \mathbf{x}, T)$.

Log likelihood

$$\begin{aligned} \mathcal{L}(\mathbf{y}; \mathbf{x}, T) &= \sum_{\mathbf{t} \in T} \log g(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}) \\ &= \sum_{\mathbf{t} \in T} L_t(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}) \end{aligned}$$

Assume:

- $\exists$ true model $\boldsymbol{\eta}$, parameter vector $\mathbf{x}^*$;
- $\mathbf{x}^*$ properly in interior of region in which $\mathcal{L}$ is well behaved;
- boundedness of integrals (computing expectations etc.).

Properties:

- $\mathcal{E}\left\{\nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)\right\} = 0$;

- $\mathcal{E}\left\{\nabla_x^2 \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)\right\} = -\mathcal{E}\left\{\nabla_x \mathcal{L}^T \nabla_x \mathcal{L}\right\}$.

Fisher information

$$\begin{aligned}\mathcal{I}_n &= \frac{1}{n}\mathcal{E}\left\{\nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)^T \nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)\right\} \\ &= \mathcal{V}\left\{\frac{1}{\sqrt{n}}\nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)\right\}.\end{aligned}$$

Maximum likelihood estimates are consistent and asymptotically minimum variance.

Algorithms: Describe step $\mathbf{h}$, $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}$

Newton

$$\mathcal{J}_n = \frac{1}{n}\nabla_x^2 \mathcal{L}\left(\mathbf{y}; \mathbf{x}, T_n\right)$$
$$\mathbf{h} = -\mathcal{J}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}\left(\mathbf{y}; \mathbf{x}, T\right)^T$$

Scoring

$$\mathbf{h} = \mathcal{I}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}\left(\mathbf{y}; \mathbf{x}, T\right)^T$$

Sample

$$\mathcal{S}_n = \frac{1}{n}\sum_{\mathbf{t}\in T_n} \nabla_x L_t\left(\mathbf{y}_t; \mathbf{x}, \mathbf{t}\right)^T \nabla_x L_t\left(\mathbf{y}_t; \mathbf{x}, \mathbf{t}\right)$$
$$\mathbf{h} = \mathcal{S}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}\left(\mathbf{y}; \mathbf{x}, T\right)^T$$

Equivalence

$$\lim_{n\to\infty} \mathcal{I}_n(\mathbf{x}^*) = \lim_{n\to\infty} \mathcal{S}_n(\mathbf{x}^*) = -\lim_{n\to\infty} \mathcal{J}_n(\mathbf{x}^*) = \mathcal{I}$$

where

$$\mathcal{I} = \int_{S(T)} \mathcal{E}^*\{\nabla_x^2 L\left(\mathbf{y}; \boldsymbol{\theta_t}, \mathbf{t}\right)\}\rho(\mathbf{t})d\mathbf{t}.$$

Transformation invariance - scoring, sample, but not Newton: Let $\mathbf{u} = \mathbf{u}(\mathbf{x})$, $W = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ then

$$\begin{aligned}
\nabla_x \mathcal{L} &= \nabla_u \mathcal{L} W, \quad \mathcal{I}^x = W^T \mathcal{I}^u W \\
\mathbf{h}_x &= \left(W^T \mathcal{I}^u W\right)^{-1} \frac{1}{n} W^T \nabla_u \mathcal{L}^T \\
&= W^{-1} (\mathcal{I}^u)^{-1} \frac{1}{n} \nabla_u \mathcal{L}^T
\end{aligned}$$

$\Rightarrow \mathbf{h}_u = W\mathbf{h}_x.$

Implementation: This requires:
(1) A method for generating $\mathbf{h}$, a direction in which the objective is increasing.
(2) A method for estimating progress. A full step need not be satisfactory. This is especially true of initial steps.
To measure progress introduce a monitor $\Phi(\mathbf{x})$ - needs same local stationary points and to be increasing when objective F is increasing

$$\nabla F\mathbf{h} \geq 0 \Rightarrow \nabla\Phi\mathbf{h} \geq 0$$

Two approaches:
(1) **Line search**: effective search direction computed, monitor used to gauge a profitable step in this direction.
(2) **Trust region**: step required to lie in an adaptively defined control region - typically one in which linearization of F does not depart too far from true behaviour.

Direction of search is required to be one which increases the objective. This is the case here for both scoring and sample algorithms

$$\nabla_x \mathcal{L}\mathbf{h} = \frac{1}{n}\nabla_x \mathcal{L}\mathcal{I}_n^{-1}\nabla_x \mathcal{L}^T > 0, \quad \mathbf{x} \neq \widehat{\mathbf{x}}.$$

Note this shows that $\mathcal{L}$ is a suitable monitor. $\nabla_x \mathcal{L}\mathbf{h}$ is invariant:

$$\begin{aligned} n\nabla_x \mathcal{L}\mathbf{h}_x &= \nabla_x \mathcal{L}\mathcal{I}_n^{-1}\nabla_x \mathcal{L}^T \\ &= \nabla_u \mathcal{L}W(W^T\mathcal{I}_n W)^{-1}W^T\nabla_u \mathcal{L}^T \\ &= \nabla_u \mathcal{L}(\mathcal{I}_n^u)^{-1}\nabla_x \mathcal{L}^T \\ &= n\nabla_u \mathcal{L}\mathbf{h}_u \end{aligned}$$

Will see there are good ways to compute this.

Measuring effectiveness. *Goldstein: accept step* $\mathbf{x} \to \mathbf{x} + \lambda \mathbf{h}$ *if*

$$\rho \leq \Psi(\lambda, \mathbf{x}, \mathbf{h}) \leq 1 - \rho, \quad 0 < \rho < .5$$
$$\Psi = \frac{\Phi(\mathbf{x} + \lambda \mathbf{h}) - \Phi(\mathbf{x})}{\lambda \nabla_x \Phi(\mathbf{x}) \mathbf{h}}$$

Can always choose λ *to satisfy this test. Armijo: Let* $0 < \rho < 1$, *and* $\lambda = \rho^k$ *where* k *is smallest integer such that*

$$\Phi(\mathbf{x} + \rho^{k-1} \mathbf{h}) \leq \Phi(\mathbf{x}) < \Phi(\mathbf{x} + \rho^k \mathbf{h}).$$

Goldstein fine for proving results. No direct link in sense of small step asymptotic equivalence relating λ *and* ρ^k. *It does not tell how to find* λ *while Armijo does. Choice* $\lambda = 1$ *favoured for scale invariant, Newton type methods. Also* $\Psi \to .5$ *for scoring and sample methods provided* n *large enough.* $.5 \geq \rho \geq .1$.

Convergence: nice inequality (Kantorovitch)

$$\frac{\nabla_x \mathcal{L}\mathbf{h}}{\|\nabla_x \mathcal{L}\|\|\mathbf{h}\|} \geq \frac{1}{\mathrm{cond}_S\left(\mathcal{I}_n\right)^{1/2}}.$$

If $\exists$ region $R \ni \mathcal{L}$ bounded and $\mathcal{I}_n$ positive definite then ascent direction exists uniformly $\Rightarrow$ lower bound λ_R exists for linesearch step multiplier. Effectiveness means $\mathcal{L}$ can be used as a monitor (Not true for pure Newton). Goldstein gives

$$\nabla_x \mathcal{L}\mathbf{h} \leq \frac{\mathcal{L}(\mathbf{x}+\lambda\mathbf{h}) - \mathcal{L}(\mathbf{x})}{\rho\lambda_R} \to 0$$

$\mathcal{L}(\mathbf{x})$ increasing and bounded $\Rightarrow$ convergence. Also

$$\|\mathbf{h}\| = \|(\mathcal{I}_n)^{-1}\frac{1}{n}\nabla_x \mathcal{L}^T\| \geq \frac{\|\nabla_x \mathcal{L}\|}{n\|\mathcal{I}_n\|}.$$

Putting it all together gives

$$\|\nabla_x \mathcal{L}\|^2 \leq K\left(\mathcal{L}(\mathbf{x}+\lambda\mathbf{h}) - \mathcal{L}(\mathbf{x})\right) \to 0$$

What happens if $\inf\{\lambda_i\} = 0$? Must be sequence $\{\widetilde{\lambda}_i\}$, $\inf \widetilde{\lambda}_i = 0 \ni$

$$\rho > \frac{\mathcal{L}(\mathbf{x_i} + \widetilde{\lambda}_i \mathbf{h_i}) - \mathcal{L}(\mathbf{x_i})}{\widetilde{\lambda}_i \nabla_x \mathcal{L}(\mathbf{x}_i)\mathbf{h_i}}$$

Here, Taylor plus mean value theorem gives

$$\|\frac{1}{n}\nabla_x^2 \mathcal{L}(\overline{\mathbf{x}})\| > \frac{2(1-\rho)}{\widetilde{\lambda}_i}\sigma_{min}(\mathcal{I}_n) \uparrow \infty.$$

Almost a global result! Set of allowed approximations need not be closed:

$$\eta = x(1) + x(2)\exp -x(3)t$$
$$t = \lim_{n\to\infty}(n - n\exp -\frac{1}{n}t).$$

Least squares. Consider sample estimate:

$$\begin{aligned}\mathcal{S}_n &= \frac{1}{n}\sum_{\mathbf{t}\in T_n} \nabla_x L_t^T \nabla_x L_t \\ &= \frac{1}{n} S_n^T S_n \\ S_n &= \begin{bmatrix} \vdots \\ \nabla_x L_t \\ \vdots \end{bmatrix}\end{aligned}$$

Correction satisfies the least squares problem

$$\min_h \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = S_n\mathbf{h} - \mathbf{e}$$

Scoring works pretty much the same:

$$\mathcal{I}_n = \frac{1}{n} \sum_{\mathbf{t} \in T_n} \nabla_x \boldsymbol{\eta}^T \mathcal{E}\{\nabla_\eta L_t^T \nabla_\eta L_t\} \nabla_x \boldsymbol{\eta}$$

Set $V_t^{-1} = \mathcal{E}\{\nabla_\eta L_t^T \nabla_\eta L_t\}$ then obtain least squares problem

$$\min_h \mathbf{r}^T \mathbf{r}; \quad \mathbf{r} = I_n^L \mathbf{h} - \mathbf{b}$$

where

$$I_n^L = \begin{bmatrix} \vdots \\ V_t^{-1/2} \nabla_x \boldsymbol{\eta} \\ \vdots \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \vdots \\ V_t^{1/2} \nabla_\eta L_t \\ \vdots \end{bmatrix}.$$

Example: normal distribution

$$L_t = -\frac{1}{2\sigma^2}(y_t - \mu(\mathbf{x}, t))^2$$

$$\nabla_x L_t = \frac{1}{\sigma^2}(y_t - \mu(\mathbf{x}, t))\nabla_x \mu_t$$

$$\nabla_x^2 L_t = -\frac{1}{\sigma^2}\left(-(y_t - \mu(\mathbf{x}, t))\,\nabla_x^2 \mu_t\right.$$

$$\mathcal{I}_n = \frac{1}{n\sigma^2}\sum_{t \in T_n} \nabla_x \mu_t^T \nabla_x \mu_t$$

$$I_n^L = \begin{bmatrix} \vdots \\ \nabla_x \mu_t \\ \vdots \end{bmatrix}$$

As σ cancels the result is the Gauss-Newton method. In contrast

$$\mathcal{S}_n = \frac{1}{n\sigma^2}\sum_{t \in T_n} \frac{(y_t - \mu(\mathbf{x}, t))^2}{\sigma^2}\nabla_x \mu_t^T \nabla_x \mu_t.$$

Here the scale does not exit so agreeably.

Computation:

$$I_n^L = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}$$
$$\mathbf{h} = U^{-1} Q_1^T \mathbf{b}$$

Can get more value from factorization

$$\begin{aligned} \nabla_x \mathcal{L}\mathbf{h} &= \left(\mathbf{b}^T I_n^L\right) \mathbf{h} \\ &= \mathbf{b}^T Q \begin{bmatrix} U \\ 0 \end{bmatrix} U^{-1} Q_1^T \mathbf{b} \\ &= \|Q_1^T \mathbf{b}\|^2 \geq 0 \end{aligned}$$

This is needed for Goldstein test, Also $\rightarrow 0$ as $\nabla_x \mathcal{L}\mathbf{h} \rightarrow 0$ so provides a scale invariant quantity for convergence testing.
Typically would scale columns of I_n^L - see Higham's book.

Rate of convergence: Consider the unit step scoring iteration in fixed point form:

$$\mathbf{x}_{i+1} = F_n(\mathbf{x}_i),$$

where

$$F_n(\mathbf{x}) = \mathbf{x} + \mathcal{I}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})^T.$$

The condition for convergence is

$$\varpi\left(F_n'(\mathbf{x}_n)\right) < 1,$$

where $\varpi(F_n'(\mathbf{x}_n))$ is the spectral radius of the variation $F_n' = \nabla_x F_n$. Then

$$\varpi\left(F_n'(\mathbf{x}_n)\right) \to 0, \text{ a.s., } n \to \infty.$$

$\varpi(F_n'(\mathbf{x}_n))$ is a (Newton-like) invariant of the likelihood surface, is a measure of the quality of the modelling, and can be estimated by a modification of the power method.

To calculate $\varpi\left(F_n'\left(\mathbf{x}_n\right)\right)$ note that $\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{x}_n\right) = 0$. thus

$$F_n'\left(\mathbf{x}_n\right) = I + \mathcal{I}_n\left(\mathbf{x}_n\right)^{-1}\frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L}\left(\mathbf{x}_n\right),$$
$$= \mathcal{I}_n\left(\mathbf{x}_n\right)^{-1}\left(\mathcal{I}_n\left(\mathbf{x}_n\right) + \frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L}\left(\mathbf{x}_n\right)\right).$$

If the right hand side were evaluated at $\mathbf{x}^*$ then the result would follow from the strong law of large numbers which shows that the matrix gets small (hence ϖ gets small) almost surely as $n \to \infty$. But, by consistency of the estimates, we have

$$\varpi\left(F_n'\left(\mathbf{x}_n\right)\right) = \varpi\left(F_n'\left(\mathbf{x}^*\right)\right) + O\left(\left\|\mathbf{x}_n - \mathbf{x}^*\right\|\right), \text{ a.s.,}$$

and the desired result follows.

Trust region: Idea is to limit scope of the linear subproblem to a closed control region containing the current estimate:

$$\min_{\mathbf{h}, \|\mathbf{h}\|_D^2 \le \gamma} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = I_n^L\mathbf{h} - \mathbf{b}$$

$$\|\mathbf{h}\|_D^2 = \mathbf{h}^T D^2\mathbf{h}, \quad D > 0 \text{ diagonal.}$$

Necessary conditions give

$$\left[\begin{array}{cc} \mathbf{r}^T & 0 \end{array}\right] = \boldsymbol{\lambda}^T \left[\begin{array}{cc} I & -I_n^L \end{array}\right] - \pi \left[\begin{array}{cc} 0 & \mathbf{h}^T D^2 \end{array}\right]$$

so $\mathbf{h}$ satisfies the perturbed scoring equations

$$\left(\mathcal{I}_n + \frac{\pi}{n} D^2\right)\mathbf{h} = \frac{1}{n}\nabla_x \mathcal{L}^T$$

π can be used to control the trust region by controlling the size of $\mathbf{h}$. Differentiating wrt π gives

$$\left(\mathcal{I}_n + \frac{\pi}{n} D^2\right)\frac{d\mathbf{h}}{d\pi} = -\frac{1}{n} D^2\mathbf{h}$$

$$\frac{d\mathbf{h}}{d\pi}^T D^2\mathbf{h} = -n\frac{d\mathbf{h}}{d\pi}^T \left(\mathcal{I}_n + \frac{\pi}{n} D^2\right)\frac{d\mathbf{h}}{d\pi} < 0$$

$$\Rightarrow -\frac{d}{d\pi}\|\mathbf{h}\|_D^2 < 0.$$

The classical form of the algorithm goes back to Levenberg 1944. Here two parameters α, β are kept, and the basic sequence of operations is:

```
count=1: do while F(x+h(\pi))<F(x)
    count=count+1
    \pi=\alpha*\pi
    loop
x\leftarrow x+h(\pi)
if count=1 then \pi=\beta*\pi
```

Successful steps will be taken eventually as

$$\mathbf{h} \to \frac{1}{\pi} D^{-2} \nabla_x \mathcal{L}^T, \quad \pi \to \infty$$

Experience suggests choices of α, β are not critical. Typically $\alpha\beta < 1$ to approach Newton like methods.

The scoring and sample algorithms are not exact Newton methods. Both can be regarded as regularised methods because of their generic positive definiteness. Not obvious that setting $\pi = 0$ will increase rate of convergence.

Basic theorems mirror line search results.
Convergence: Let $\{\mathbf{x}_i\}$ produced by the α, β procedure be contained in compact region R in which $\{\pi_i\} < \infty$ then $\{\mathcal{L}(\mathbf{y}; \mathbf{x}_i, T_n)\}$ converges, and limit points of $\{\mathbf{x}_i\}$ are stationary points of $\mathcal{L}$.
Boundedness: If sequence $\{\pi_i\}$ determined by α, β procedure is unbounded in R then the norm of $\nabla_x^2 \mathcal{L}$ is also unbounded.

Remember iterations have the potential to become unbounded for smooth sets of nonlinear approximating functions as closure of these sets of functions cannot be guaranteed without more work.

Computation: The trust region problem is:

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = \begin{bmatrix} X_n^L \\ \sqrt{\pi}I \end{bmatrix}\mathbf{h} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$$

Let $X_n^L = Q\begin{bmatrix} U \\ 0 \end{bmatrix}$ then $\mathbf{h}(\pi)$ can be found by solving the typically much smaller problem

$$\min_{\mathbf{h}} \mathbf{s}^T\mathbf{s}; \quad \mathbf{s} = \begin{bmatrix} U \\ \sqrt{\pi}I \end{bmatrix}\mathbf{h} - \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}$$

where $\mathbf{c}_1 = Q_1^T\mathbf{b}$. Make further factorization $\begin{bmatrix} U \\ \sqrt{\pi}I \end{bmatrix} = Q'\begin{bmatrix} U' \\ 0 \end{bmatrix}$, $Q'^T\begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{bmatrix}$ then

$$\begin{aligned} \mathbf{h}(\pi) &= (U')^{-1}\mathbf{c}_1' \\ \nabla_x \mathcal{L}\mathbf{h}(\pi) &= \mathbf{c}_1^T U\mathbf{h}(\pi) \\ &= \begin{bmatrix} \mathbf{c}_1^T & 0 \end{bmatrix}\begin{bmatrix} U \\ \sqrt{\pi}I \end{bmatrix} \\ &\quad \left(U^TU + \pi I\right)^{-1}\begin{bmatrix} U^T & \sqrt{\pi}I \end{bmatrix}\begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix} \\ &= \|\mathbf{c}_1'\|^2 \end{aligned}$$

The linear subproblem has a useful invariance property with respect to diagonal scaling. Introduce the new variables $\mathbf{u} = T\mathbf{x}$ where T is diagonal.

$$T^{-1}\left(S_x^T S_x + \pi D^2\right) T^{-1} T\mathbf{h}_x = T^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T.$$

This is equivalent to

$$\left(S_u^T S_u + \pi T^{-1} D^2 T^{-1}\right)\mathbf{h}_u = \nabla_{\mathbf{u}}\mathcal{L}^T.$$

Thus if D_i transforms with $\frac{\partial}{\partial x_i}$ then $T_i^{-1} D_i$ transforms in the same way with respect to $\frac{\partial}{\partial u_i}$. This requirement is satisfied by

$$D_i = \|(S_n)_{*i}\|.$$

This transformation effects a rescaling of the least squares problem. We have

$$\mathbf{h} = \left(S_n^T S_n + \pi D^2\right)^{-1} S_n^T \mathbf{e},$$

$$\Rightarrow D\mathbf{h} = \left(D^{-1} S_n^T S_n D^{-1} + \pi I\right)^{-1} D^{-1} S_n^T \mathbf{e}.$$

The effect of this choice is to rescale the columns of S_n to have unit length. It is often sufficient to set $\pi = 1$, and $D = \text{diag}\left\{\|(S_n)_{*i}\|, i = 1, 2, \cdots, p\right\}$ initially .

One catch is the initial choice $\pi = \pi_0$. One example is provided by models containing exponential terms such as $e^{-x_k t}$ which should have negative exponents ($x_k > 0$) but which can become positive ($x_k + h_k < 0$) by too large a step. To fix introduce a damping factor τ such that the critical components of $\mathbf{x} + \tau \mathbf{h}(\pi_0) \geq 0$. $\|\tau \mathbf{h}(\pi_0)\|$ provides a possible choice for a revised trust region bound γ. This involves solving for π by eg Newton's method the equation

$$\|\mathbf{h}(\pi)\| = \gamma.$$

$\frac{d\mathbf{h}}{d\pi}$ can be found by solving

$$\left(U^T U + \pi I\right) \frac{d\mathbf{h}}{d\pi} = -\mathbf{h}$$

This can be written in least squares form:

$$\min_{\mathbf{h}} \mathbf{s}^T \mathbf{s}; \quad \mathbf{s} = \begin{bmatrix} U \\ \pi^{1/2} I \end{bmatrix} \frac{d\mathbf{h}}{d\pi} - \begin{bmatrix} U^{-T} \mathbf{h} \\ 0 \end{bmatrix}.$$

Newton's method extrapolates the function as a linear. An alternative strategy could be preferable here. This is based on the observation that, if

$$A = W\Sigma V^T$$

is the singular value decomposition of the matrix in the least squares formulation, then

$$\mathbf{h} = \sum_{i=1}^{p} \frac{\sigma_i(\mathbf{w}_i^T\mathbf{b})}{\sigma_i^2 + \pi}\mathbf{v}_i$$

is a rational function of π. To mirror this set

$$\|\mathbf{h}\| \approx \frac{a}{b + (\pi - \pi_0)}.$$

To identify the parameters a and b use the values $\|\mathbf{h}(\pi_0)\|$ and $\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|$ - this corresponds to the same information as used in Newton's method.

To solve for the parameters we have the equations:

$$\|\mathbf{h}(\pi_0)\| = \frac{a}{b},$$
$$\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\| = \frac{\mathbf{h}^T \frac{d\mathbf{h}}{d\pi}}{\|\mathbf{h}\|}(\pi_0) = -\frac{a}{b^2}.$$

The result is

$$b = -\frac{\|\mathbf{h}(\pi_0)\|}{\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|}, \quad a = \|\mathbf{h}(\pi_0)\|b,$$

giving the correction

$$\pi = \pi_0 - \frac{\|\mathbf{h}(\pi_0)\| - \gamma}{\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|}\frac{\|\mathbf{h}(\pi_0)\|}{\gamma}.$$

The effect of the rational extrapolation is just the Newton step modulated by the term $\frac{\|\mathbf{h}(\pi_0)\|}{\gamma}$. As this term $\rightarrow 1$ this is also a second order convergent process.

Simple exponential model: Here the model used is

$$\mu(t, \mathbf{x}) = x(1) + x(2)\exp(-x(3)t). \quad (1)$$

The values chosen for the parameters are $x(1) = 1$, $x(2) = 5$, and $x(3) = 10$. Initial values are generated using

$$x(i)_0 = x(i) + (1 + x(i))(.5 - Rnd)$$

where Rnd indicates a call to a uniform random number generator giving values in $[0, 1]$.

This model is not difficult in the sense that the graph has features that depend on each of the parameters. Thus the interest is in the effects of simulated data errors.

Two types of random numbers are used to simulate the experimental data.

Normal data: The data is generated by evaluating $\mu(t, \mathbf{x})$ on a uniform grid with spacing $\Delta = 1/(n+1)$ and then perturbing these values using normally distributed random numbers to give values

$$z_i = \mu(i\Delta, \mathbf{x}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, 2), \quad i = 1, 2, \cdots, n.$$

The choice of standard deviation was made to make small sample problems ($n = 32$) relatively difficult. The log likelihood is taken as

$$\mathcal{L}(\mathbf{x}) = -\frac{1}{2} \sum_{i=1}^{n} \left(z_i - \mu(i\Delta, \mathbf{x}) \right)^2 .$$

While the scale is not evident here it resurfaces in its effects on the generated data.

Poisson data: A Poisson random number generator is used to generate random counts z_i corresponding to $\mu(i\Delta, \mathbf{x})$ as the mean model. The log likelihood used is

$$\mathcal{L}(\mathbf{x}) = \sum_{i=1}^{n} z_i \log\left(\frac{\mu(i\Delta, \mathbf{x})}{z_i}\right) + (z_i - \mu(i\Delta, \mathbf{x}))\,.$$

Note that if $z_i = 0$ then the contribution from the logarithm term to the log likelihood is zero. The rows of the least squares problem design matrix are given by

$$\mathbf{e}_i^T A = \frac{1}{s_i}, \frac{\exp(-x(3)t_i)}{s_i}, \frac{-x(2)t_i \exp(-x(3)t_i)}{s_i},$$
$$i = 1, 2, \cdots, n$$

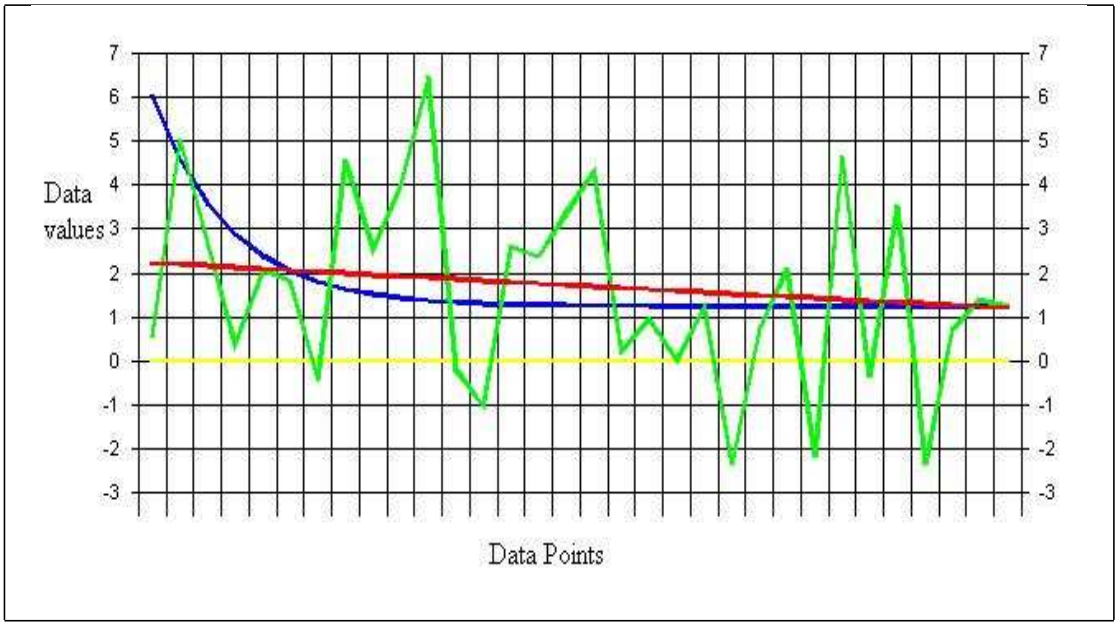
where $s_i = \sqrt{\mu(i\Delta, \mathbf{x})}$. The corresponding components of the right hand side are

$$b_i = \frac{z_i - \mu(i\Delta, \mathbf{x})}{s_i}.$$

Numerical experiments comparing the performance of the line search (LS) and trust region (TR) methods are summarised below. For each n the computations were initiated with 10 different seeds for the basic random number generator, and the average number of iterations is reported as a guide to algorithm performance. The parameter settings used are $\alpha = 2.5$, $\beta = .1$ for the trust region method and $\rho = .25$ for the Armijo parameter used in the line search. Experimenting with these values (for example, the choice $\alpha = 1.5$, $\beta = .5$) made very little difference in the trust region results. Convergence is assumed if $\nabla_x \mathcal{L}\mathbf{h} < 1.0e^{-8}$. This corresponds to final values of $\|\mathbf{h}\|$ in the range $1.e^{-4}$ to $1.e^{-6}$.

	Normal		Poisson	
n	LS	TR	LS	TR
32	10.3*	14*	11	12.3
128	9.3	11.9	7.6	7.9
512	7.3	7.3	7.1	6.9
2048	6.7	6.1	6.3	5.8

The starred entries in the table correspond to two cases of nonconvergence. The figure shows (in red) the current estimate after 50 iterations together with the data and the starting estimate. This gives an illustration that the set of approximations is not closed.



Result shows a straight line fit

Second example - Gaussians with an exponential background.

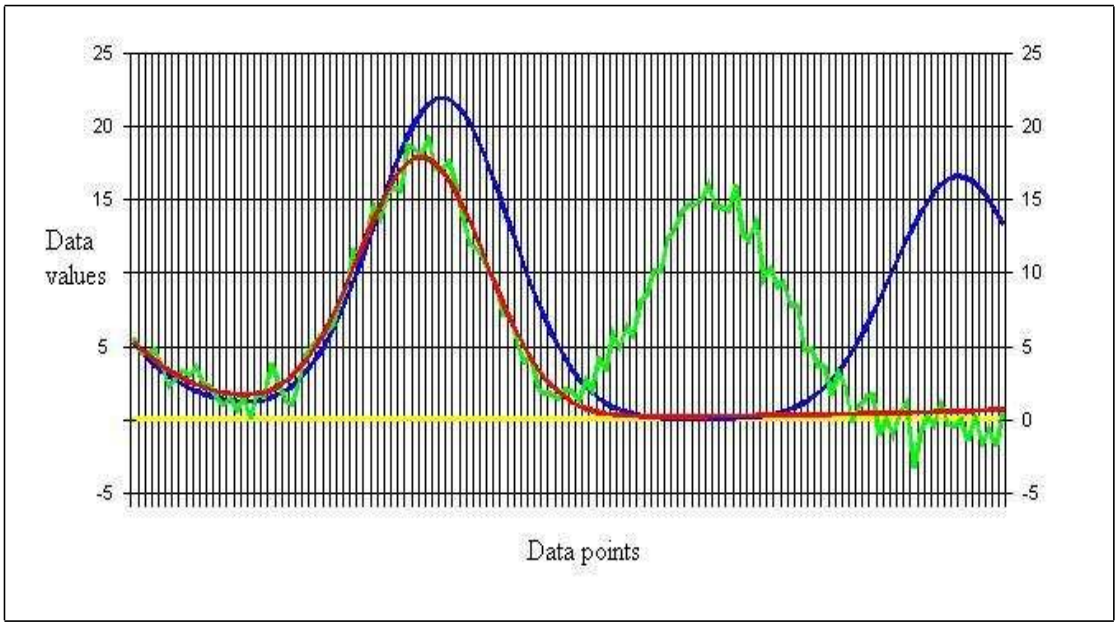
$$\mu = x(1)\exp -x(2)t + x(3)\exp -(t - x(4))^2/x(5) + x(6)\exp -(t - x(7))^2/x(8)$$

Line search case:

$$x(1) = 5,\ x(2) = 10,\ x(3) = 18,\ x(4) = .3333,$$
$$x(5) = .05,\ x(6) = 15,\ x(7) = .6667,\ x(8) = .05$$

.05	32	128	512	2048
1	-	10	8	10
2	17	41	42	24
3	-	64	11	6
4	84	11	-	53
5	27	15	8	142
6	20	13	11	8
7	6	7	26	8
8	-	40	15	8
9	137	10	9	66
10	11	6	14	23

The results here are much more of a mixed bag. The problems are closely related to the choice of starting values. The figure is produced using peak widths of .01 which makes it easier to see what is going on. In this case the starting values do not see the second peak. Both the background and the first peak are picked up well.



Initial conditions miss the second peak

Third example - trinomial example

Data for a trinomial example ($m = 3$) came from a consulting exercise with CSIRO. It is derived from a study of the effects of a cattle virus on chicken embryos.

$\log_{10}(titre)$	dead	deformed	normal
-0.42	0	0	18
0.58	1	2	13
1.58	5	6	4
2.58	12	6	1
3.58	18	1	0
4.58	16	0	0

Cattle virus data

The model suggested fits the frequencies explicitly

$$\omega_1 = \frac{1}{1 + \exp(-\beta_1 - \beta_3 \log(t))},$$
$$1 - \omega_2 = \frac{1}{1 + \exp(-\beta_2 - \beta_3 \log(t))},$$
$$\omega_3 = 1 - \omega_1 - \omega_2.$$

Makes sense to develop the algorithm in terms of the frequencies. Numerical results show an impressive rate of convergence for a relatively small data set. Suggests the model analysis is good.

its	$\mathcal{L}$	$\nabla\mathcal{L}\mathbf{h}$	β_1	β_2	β_3
0	49.31		-4.597	-3.145	.7405
1	47.53	.1353+2	-3.937	-2.369	.7834
2	47.00	.9778+0	-4.419	-2.584	.8882
3	46.99	.2145-1	-4.405	-2.620	.9060
4	46.99	.7370-5	-4.405	-2.619	.9060
5	46.99	.8861-8	-4.405	-2.619	.9060

Results of computations for the trinomial data

In conclusion Work on the Levenberg algorithm in the 1970's was responsible for at least some of the encouragement for the shift from linesearch to trust region methods in optimization problems. It can now be argued that this component of the move had somewhat dubious validity.

1. Use of the expected Hessian is already a "regularising" step. Improved conditioning derived from the trust region parameter could be illusory if the aim is small values for rapid convergence. If significant values of π are required then in the data analytic context the modelling could well be suspect.

2. The old papers relied on a small residual argument to explain good convergence rates. Again in our context this is not satisfactory. It is not completely obvious what the effect of non zero trust region parameters is in any particular case.

3. The trust region algorithm does not scale as well as the linesearch scoring algorithm.

4. Global convergence results of similar power appear available for both approaches.