Least squares methods in maximum likelihood problems

M.R.Osborne
School of Mathematical Sciences, ANU

Abstract

The Gauss-Newton algorithm for solving nonlinear least squares problems proves particularly efficient for solving parameter estimation problems when the number of independent observations is large and the fitted model is appropriate. In this context the conventional assumption that the residuals are small is not needed. The Gauss-Newton method is a special case of the Fisher scoring algorithm for maximizing log likelihoods and shares with this a number of desirable properties. The formal structural correspondence is striking with the linear subproblem for the general scoring algorithm having the form of a linear least squares problem. This is an important observation because it provides likelihood methods with a computational framework which accords with computational orthodoxy. Both line search and trust region algorithms are available and these are compared and contrasted here. It is shown that the types of theoretical results which have led to the wide acceptance of trust region methods have direct equivalents in the line search case, while the latter have better transformation invariance properties. Computational experiments for both continuous and discrete distributions show no advantage for the trust region approach.

1 Introduction

This paper has two main aims:

1. To show that a least squares framework is appropriate for the implementation of the Fisher scoring method for estimating parameters by maximizing an important class of log likelihoods. These methods include the Gauss-Newton algorithm for solving non-linear least squares

problems, but they also include parameter estimation problems for likelihoods associated with discrete probability distributions.

2. To compare and contrast the properties of line search and trust region methods in the implementation of these methods.

There is even historical interest in the comparison of the line search and trust region methods. The first trust region algorithm is probably Levenberg's modification of the Gauss-Newton algorithm [4]. Good convergence properties publicized for variants of this method ([5], [6]) served to draw attention to the potential advantages of the general trust region approach. The advantages in general are not denied, but it is suggested that the advantages in this particular context are not as great as might have been thought originally.

To establish the emphasis on a data analytic context in the algorithmic developments it is convenient to start with the following assumptions which serve to outline the problem context.

1. The experimental data consists of independent event outcomes $\mathbf{y}_j \in R^q, \quad j = 1, 2, \cdots, n$, but it is not assumed that the individual components of $\mathbf{y}_j$ are independent;

2. there is an associated probability density function (probability mass function for discrete distributions) $g(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t})$ indexed by "points" $\mathbf{t} \in T_n \subset R^l$ where $|T_n| = n$; and

3. the structural information is provided by a known parametric model

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

where $\boldsymbol{\theta} \in R^s$, and $\mathbf{x} \in R^p$. The true parameter vector, which is assumed to exist, is denoted by $\mathbf{x}^*$.

Expectations with respect to the density g are written

$$\mathcal{E}\{\bullet\}(\mathbf{t}) = \int_Y (\bullet)\, g(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t})\, d\mathbf{y},$$

where $Y = \text{range}(\mathbf{y})$ is assumed independent of $\mathbf{x}$. If the correct density associated with the observed event is needed then the expectation is indicated by $\mathcal{E}^*$.

A priori information is provided by the experimental design T_n. Here the subscript n is included for the purpose of asymptotic analysis (so that it is

typically assumed large), and the experimental design is required to satisfy the condition for a designed experiment:

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

Here $S(T)$ is an appropriately measurable set which is filled out as $n \to \infty$ by the sets of sample points T_n for each n. This really says no more than that there exists an hypothesised mechanism for conducting experiments for large values of n which is captured asymptotically by the limiting density $\rho(\mathbf{t})$. Here this is used typically in conjunction with the law of large numbers in applications such as the following:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(y_i) &= \frac{1}{n}\sum_{i=1}^{n} f(y_i) - \frac{1}{n}\mathcal{E}^*\left\{\sum_{i=1}^{n} f(y_i)\right\} + \frac{1}{n}\mathcal{E}^*\left\{\sum_{i=1}^{n} f(y_i)\right\}, \\ &\to \int_{S(T)} \mathcal{E}^*\left\{f(y)\right\}(\mathbf{t})\rho(\mathbf{t})d\mathbf{t}, n \to \infty. \end{aligned} \tag{1}$$

The parameter estimation problem is: "given the event outcomes $\mathbf{y}_t$ it is required to estimate $\mathbf{x}^*$".

Example 1 *All the densities to be considered belong to the exponential family:*

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

Here expectation and variance follow directly from the form of the density:

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

Particular cases include

1. *normal density (continuous distribution):*

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

Cases in which $\theta = \mu$ are called "signal in noise" models.

2. *Poisson density (discrete distribution):*

$$g(i,\lambda) = \frac{\exp(-\lambda)\,\lambda^i}{i!} = \frac{1}{i!}\exp(-\lambda)\exp(i\log(\lambda)) \tag{2}$$

so that $\theta = \log(\lambda)$, $b(\theta) = \exp(\theta)$, *and* $\mu = \frac{db}{d\theta} = \lambda$.

3. *multinomial (discrete distribution):*

$$\begin{aligned} g(\mathbf{m};\omega) &= \frac{m!}{\prod_{j=1}^{p} m_j!}\prod_{j=1}^{p}\omega_j^{m_j}, \\ &= \frac{m!}{\prod_{j=1}^{p} m_j!}e^{\sum_{j=1}^{p} m_j \log\omega_j}, \end{aligned} \tag{3}$$

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

$$\sum_{j=1}^{p} m_j \log\omega_j = \sum_{j=1}^{p-1} m_j \log\frac{\omega_j}{1-\sum_{i=1}^{p-1}\omega_i} + m\log\left(1-\sum_{j=1}^{p-1}\omega_j\right). \tag{4}$$

It follows that

$$\theta_j = \log\frac{\omega_j}{1-\sum_{i=1}^{p-1}\omega_i}, \tag{5}$$

$$b(\boldsymbol{\theta}) = m\log\left(1+\sum_{j=1}^{p-1} e^{\theta_j}\right). \tag{6}$$

Parameter estimation by the method of maximum likelihood starts with the likelihood

$$\mathcal{G}(\mathbf{y};\mathbf{x},T) = \prod_{\mathbf{t}\in T} g(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}). \tag{7}$$

This can be expected to be relatively large at the true parameter values in the sense that the actual event outcomes observed will contain at most a small proportion corresponding to events of low probability, and this can be expected to be true, in particular, for large data sets. This leads to the estimation principle:

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{G}(\mathbf{y};\mathbf{x},T). \tag{8}$$

In practice it is more convenient to use the log likelihood

$$\begin{aligned} \mathcal{L}_n(\mathbf{y};\mathbf{x},T) &= \sum_{\mathbf{t}\in T}\log g(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}), \\ &= \sum_{\mathbf{t}\in T} L_t(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}). \end{aligned} \tag{9}$$

The method of estimating parameters by maximizing $\mathcal{L}_n(\mathbf{y};\mathbf{x},T)$ is called the method of maximum likelihood. The following assumptions simplify the application of this method.

- There exists both a true model $\boldsymbol{\eta}$, and a unique correct parameter vector $\mathbf{x}^*$;
- $\mathbf{x}^*$ is properly in the interior of a compact region in which $\mathcal{L}_n$ is well behaved; and
- any required boundedness of integrals needed for computing expectations etc is available.

The important theoretical results that follows from all this [3] include the consistency of the maximum likelihood estimator in the sense of almost sure convergence

$$\widehat{\mathbf{x}} \overset{as}{\to} \mathbf{x}^*, \quad n \to \infty,$$

and the attractive property that the estimator asymptotically attains the minimum variance lower bound given by the inverse of the Fisher information:

$$\begin{aligned}\mathcal{I}_n &= \frac{1}{n}\mathcal{E}\left\{\nabla_x \mathcal{L}_n(\mathbf{y};\mathbf{x},T_n)^T \nabla_x \mathcal{L}_n(\mathbf{y};\mathbf{x},T_n)\right\}, &(10)\\ &\to \int_{S(T)} \mathcal{E}^*\left\{\nabla_x L_t(\mathbf{y};\mathbf{x}^*,\mathbf{t})^T \nabla_x L_t(\mathbf{y};\mathbf{x}^*,\mathbf{t})\right\}\rho(\mathbf{t})\,d\mathbf{t}, \quad n\to\infty,\\ &= \mathcal{I}. &(11)\end{aligned}$$

where L_t is defined in (9). These results are summarised in the form of the limiting normal distribution for the parameter estimates:

$$\sqrt{n}\,(\widehat{\mathbf{x}} - \mathbf{x}^*) \sim N\left(0, \mathcal{I}^{-1}\right)$$

This shows also that asymptotically the error in the computed estimate gets small thanks to a factor $1/\sqrt{n}$. The computational significance of this term is that it indicates what is in practice a slow rate of convergence.

Key properties of the log likelihood needed for the connection to least squares problems are:

- $$\mathcal{E}\left\{\nabla_x \mathcal{L}_n(\mathbf{y};\mathbf{x},T)\right\} = 0; \tag{12}$$
- $$\mathcal{E}\left\{\nabla_x^2 \mathcal{L}_n(\mathbf{y};\mathbf{x},T)\right\} = -\mathcal{E}\left\{\nabla_x \mathcal{L}_n^T \nabla_x \mathcal{L}_n\right\}. \tag{13}$$

Under the above assumptions they follow directly by reversing the order of differentiation and integration.

2 Algorithms

The types of method considered are all modifications of the basic Newton algorithm. Here, at the current point $\mathbf{x}$, a correction $\mathbf{h}$ is computed by linearizing the problem. This leads to the algorithms:

Newton

$$\mathcal{J}_n = \frac{1}{n}\nabla_x^2 \mathcal{L}_n\left(\mathbf{y};\mathbf{x},T_n\right), \tag{14}$$

$$\mathbf{h} = -\mathcal{J}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}_n\left(\mathbf{y};\mathbf{x},T\right)^T, \tag{15}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}.$$

Scoring

$$\mathbf{h} = \mathcal{I}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)^T, \tag{16}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}.$$

Sample

$$\mathcal{S}_n = \frac{1}{n}\sum_{\mathbf{t}\in T_n} \nabla_x L_t\left(\mathbf{y}_t;\mathbf{x},\mathbf{t}\right)^T \nabla_x L_t\left(\mathbf{y}_t;\mathbf{x},\mathbf{t}\right), \tag{17}$$

$$\mathbf{h} = \mathcal{S}_n^{-1}\frac{1}{n}\nabla_x \mathcal{L}\left(\mathbf{y};\mathbf{x},T\right)^T, \tag{18}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}.$$

Remark 2 *The key feature of these modified Newton methods is the replacement of the Hessian of the log likelihood objective by its expectation in the scoring algorithm and by a sample estimate in the sample algorithm. These replacements have the attractive structural property of being generically positive definite. An important result here is the asymptotic equivalence of the Hessian and its modifications when evaluated at the true solution. The proof employs the law of large numbers which is distinctly useful in this context.*

Theorem 3 *In the sense of almost sure convergence:*

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

Proof. Only the equivalences relating $\mathcal{J}_n$ and $\mathcal{I}_n$ are considered as the same basic form of argument is used in all cases. We have by (10), (13)

$$\begin{aligned}\mathcal{J}_n &= \mathcal{J}_n - \mathcal{E}\left\{\mathcal{J}_n\right\} + \mathcal{E}\left\{\mathcal{J}_n\right\}, \\ &= \mathcal{J}_n + \mathcal{I}_n - \mathcal{I}_n.\end{aligned}$$

Now $\mathcal{J}_n + \mathcal{I}_n = \mathcal{J}_n - \mathcal{E}\left\{\mathcal{J}_n\right\} \stackrel{as}{\to} 0,\ n \to \infty$ by the law of large numbers, and $\mathcal{I}_n \to \mathcal{I}$ by the designed experiment condition (compare (1)). ■

This result has immediate computational implications for if one of the modified Hessians is strictly positive definite for n large enough then in each of the cases (15), (16), and (18), the step estimation problem is associated with a bounded condition number for n large enough. In the Newton iteration there is the added requirement that $\mathbf{x}$ is sufficiently close to $\widehat{\mathbf{x}}$.

An important feature of the scoring and sample algorithms not shared by the Newton algorithm in general is a strong transformation invariance property. Let $\mathbf{w} = \mathbf{w}(\mathbf{x}),\ W = \frac{\partial \mathbf{w}}{\partial \mathbf{x}}$ then scoring gives:

$$\nabla_x \mathcal{L} = \nabla_w \mathcal{L} W, \quad \mathcal{I}^x = W^T \mathcal{I}^w W \tag{19}$$

$$\mathbf{h}_x = \left(W^T \mathcal{I}^w W\right)^{-1} \frac{1}{n} W^T \nabla_w \mathcal{L}^T, \tag{20}$$

$$= W^{-1} (\mathcal{I}^w)^{-1} \frac{1}{n} \nabla_w \mathcal{L}^T. \tag{21}$$

Thus $\mathbf{h}_w = W \mathbf{h}_x$. This result only applies to the Newton algorithm when the transformation is linear because the transformed Hessian involves derivatives of the transformation in general.

Implementation requires:

1. A method for computing the current correction $\mathbf{h}$. Both the scoring and sample algorithms are guaranteed to generate a direction in which the objective is increasing provided the weak condition that the modified Hessian is positive definite is assumed. In this sense they determine effective search directions.

$$\nabla_x \mathcal{L}_n \mathbf{h} = \frac{1}{n} \nabla_x \mathcal{L}_n \mathcal{I}_n^{-1} \nabla_x \mathcal{L}_n^T > 0, \quad \mathbf{x} \neq \widehat{\mathbf{x}}. \tag{22}$$

2. A method for estimating progress. A full step need not be satisfactory, and a more conservative approach is needed if the implementation is to have good global properties . This is needed especially for initial steps when good initial parameter estimates are not available.

To measure progress introduce a monitor function $\Phi(\mathbf{x})$. To be satisfactory this needs to have both the same local stationary points and to be increasing when the objective $\mathcal{L}_n$ is increasing. Thus it must satisfy

$$\nabla \mathcal{L}_n \mathbf{h} \geq 0 \Rightarrow \nabla \Phi \mathbf{h} \geq 0.$$

It is also very desirable that it reflect the good transformation properties of the basic algorithms.

Remark 4 *Both the scoring and sample algorithms have the important property that the objective function $\mathcal{L}_n$ provides a suitable monitor. This follows from (22). Transformation invariance follows from (20) and (21). In general, $\mathcal{L}_n$ is not a suitable monitor for the Newton iteration because it cannot be assumed that $\mathcal{J}_n$ is globally positive definite.*

We consider two different strategies for using the monitor.

Line search Assume that an effective search direction has been computed using (16) or (18). The monitor is used to gauge a profitable length of step in this direction.

Trust region This strategy modifies the basic step by requiring it to lie in an adaptively defined control region - typically one in which the linearization of $\mathcal{L}_n$ does not depart too far from true nonlinear behaviour. Here the monitor is used to control the adaptation.

These are presented in sections 3 and 5. Section 4 introduces the least squares formulation and some of its properties. The final sections presents numerical results and conclusions.

3 Properties of the line search methods

In the line search based computation two approaches are considered in determining a suitable step λ in the computed search direction . Both seek to choose λ relatively large in the set of values that permit Φ to increase.

Goldstein The new point $\mathbf{x} \rightarrow \mathbf{x} + \lambda \mathbf{h}$ is accepted provided

$$\varrho \leq \Psi(\lambda, \mathbf{x}, \mathbf{h}) \leq 1 - \varrho, \quad 0 < \rho < .5, \tag{23}$$

where

$$\Psi = \frac{\Phi(\mathbf{x} + \lambda \mathbf{h}) - \Phi(\mathbf{x})}{\lambda \nabla_x \Phi(\mathbf{x}) \mathbf{h}}. \tag{24}$$

Can always choose λ to satisfy this test under modest conditions on Φ. There are effective methods for computing λ to satisfy (23) [1].

Simple Let $0 < \varpi < 1$, and $\lambda = \varpi^k$ where $k \in \{0, 1, 2, \cdots\}$ is the smallest value such that

$$\Phi(\mathbf{x} + \varpi^{k-1}\mathbf{h}) \leq \Phi(\mathbf{x}) < \Phi(\mathbf{x} + \varpi^k \mathbf{h}). \tag{25}$$

There are some useful connections between the two line search strategies. Typically, if $\Phi(\mathbf{x} + \mathbf{h}) > \Phi(\mathbf{x})$ then $\lambda = 1$ is accepted corresponding to the Simple test being satisfied with $k = 0$. The argument uses that this step can yield a fast rate of convergence for the transformation invariant algorithms under appropriate conditions so that it makes sense to use it to set the search scale. Equation (23) appears to express a somewhat more stringent condition. However, if successive iterates are contained in a bounded region R in which $\mathcal{I}_n$ is positive definite and $\widehat{k}$ is an upper bound to the values of k computed in the Simple steps then a simple Taylor series analysis shows that $\varpi^{\widehat{k}} \approx \varrho$ satisfies the left hand inequality in (23) for each step provided the Hessian is negative definite. Also, if $\lambda = \varpi^k$ is accepted for a Simple step with $k > 1$ then at the previous step it follows that in (24)

$$\Psi\left(\varpi^{k-1}, \mathbf{x}, \mathbf{h}\right) < 0 < \varrho \tag{26}$$

for any $\varrho > 0$ which satisfies the requirements in the Goldstein test as the numerator in Ψ is ≤ 0 for every failed Simple step.

Discussion of convergence for the scoring (alternatively sample) algorithm in the case that the sequence of line search steps $\{\lambda_i\}$ generated by either strategy is bounded away from 0 is now routine. The sequence $\{\mathcal{L}_n(\mathbf{y}; \mathbf{x}_i, T_n)\}$ is increasing and bounded above and therefore converges. Consider the Goldstein test. This shows that the numerator in (24) tends to zero. Now (23) can be inverted to give an upper bound for $\nabla_x \mathcal{L}_n(\mathbf{y}; \mathbf{x}_i, T_n)\,\mathbf{h}_i$ that tends to 0 as $i \to \infty$. Because $\mathcal{I}_n$ (alternatively $\mathcal{S}_n$) is positive definite in R it follows that $\|\nabla_x \mathcal{L}_n(\mathbf{y}; \mathbf{x_i}, T_n)\| \to 0$. Thus limit points of the sequence of iterates $\{\mathbf{x}_i\}$ are stationary points of $\mathcal{L}_n$.

What happens if $\inf\{\lambda_i\} = 0$? If this occurs with the Goldstein line search then, because the right hand inequality in (23) must be satisfied at each iteration, it follows that mean values of the Hessian must be becoming unbounded. The corresponding result for the Simple line search makes for a closer comparison with the trust region methods.

Theorem 5 *Let the sequence of iterates $\{\mathbf{x}_i\}$ produced by the scoring algorithm implemented using the Simple line search be contained in a bounded region R in which $\mathcal{I}_n$ has full rank and $\inf\{\lambda_i\} = 0$. Then $\frac{1}{n}\nabla_x^2 \mathcal{L}_n$ is unbounded in R.*

Proof. For $\inf\{\lambda_i\} = 0$ to hold there must be an infinite sequence of points $\widetilde{\mathbf{x}}_i = \mathbf{x}_i + \widetilde{\lambda}_i \mathbf{h}_i$ where the Simple test fails so that, by (26) and any ϱ compatible with the requirements of the Goldstein test,

$$\varrho > \frac{\mathcal{L}_n\left(\widetilde{\mathbf{x}}_i\right) - \mathcal{L}_n\left(\mathbf{x}_i\right)}{\widetilde{\lambda}_i \nabla_x \mathcal{L}_n\left(\mathbf{x}_i\right) \mathbf{h}_i}, \quad \inf_i \left\{\widetilde{\lambda}_i\right\} = 0.$$

Using the mean value theorem and the property that $\mathbf{h}_i$ is a direction of ascent gives

$$\varrho > \frac{\widetilde{\lambda}_i \nabla_x \mathcal{L}_n(\mathbf{x}_i)\mathbf{h}_i - \frac{\widetilde{\lambda}_i^2}{2} |\mathbf{h}_i^T \nabla_x^2 \mathcal{L}_n(\overline{\mathbf{x}})\mathbf{h}_i|}{\widetilde{\lambda}_i \nabla_x \mathcal{L}_n(\mathbf{x}_i)\mathbf{h}_i},$$

where the bar denotes that a mean value is appropriate. Thus

$$\widetilde{\lambda}_i |\mathbf{h}_i^T \frac{1}{n} \nabla_x^2 \mathcal{L}_n(\overline{\mathbf{x}})\mathbf{h}_i| > 2(1-\varrho)\frac{1}{n}\nabla_x \mathcal{L}_n(\mathbf{x}_i)\mathbf{h}_i,$$

so that

$$\begin{aligned} \|\frac{1}{n}\nabla_x^2 \mathcal{L}_n(\overline{\mathbf{x}})\| &> \frac{2(1-\varrho)}{\widetilde{\lambda}_i} \frac{\frac{1}{n}\nabla_x \mathcal{L}_n(\mathbf{x}_i)\mathbf{h}_i}{\|\mathbf{h}_i\|^2} \\ &= \frac{2(1-\varrho)}{\widetilde{\lambda}_i} \frac{\mathbf{h}_i^T \mathcal{I}_n \mathbf{h}_i}{\|\mathbf{h}_i\|^2} \\ &> \frac{2(1-\varrho)}{\widetilde{\lambda}_i}\, \sigma_{min}(\mathcal{I}_n), \end{aligned} \tag{27}$$

where σ_{min} is the smallest eigenvalue. The result now follows from the definition of $\widetilde{\lambda}_i$. ■

These convergence results amount almost to a global result for if $\mathcal{L}_n$ is at least twice continuously differentiable in R then it follows that necessarily $\inf\{\lambda_i\} > 0$ so that limit points must be stationary points of the objective function. The key assumption is that R is bounded. If this is relaxed then the set of allowed approximations need not be closed. Consider the simple exponential model

$$\eta = x(1) + x(2)\exp\left(-x(3)t\right). \tag{28}$$

Let data be given by sampling the independent variable t. For large n the parameter estimates can be deduced from the relation

$$t = \lim_{n\to\infty}(n - n\exp\left(-\frac{1}{n}t\right)).$$

This shows both the closure problem and its relationship with R unbounded.

In practical algorithms rate of convergence ranks in importance with actual convergence. There is a good story [7], and it is summarized here for completeness. Consider the unit step scoring iteration in fixed point form:

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

where

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

The condition for convergence is

$$\pi(F_n'(\widehat{\mathbf{x}})) < 1,$$

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

To calculate $\pi(F_n'(\widehat{\mathbf{x}}))$ note that $\nabla_x \mathcal{L}_n(\widehat{\mathbf{x}}) = 0$. Thus

$$\begin{aligned} F_n'(\widehat{\mathbf{x}}) &= I + \mathcal{I}_n(\widehat{\mathbf{x}})^{-1} \frac{1}{n} \nabla_x^2 \mathcal{L}_n(\widehat{\mathbf{x}}), \\ &= \mathcal{I}_n(\widehat{\mathbf{x}})^{-1} \left(\mathcal{I}_n(\widehat{\mathbf{x}}) + \frac{1}{n} \nabla_x^2 \mathcal{L}_n(\widehat{\mathbf{x}}) \right). \end{aligned}$$

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

$$\pi(F_n'(\widehat{\mathbf{x}})) = \pi(F_n'(\mathbf{x}^*)) + O(\|\widehat{\mathbf{x}} - \mathbf{x}^*\|), \text{ a.s.},$$

and the desired result follows. Asymptotically, both the scoring and sample algorithms approach a second order rate.

$\pi(F_n'(\widehat{\mathbf{x}}))$ is a curvature invariant of the likelihood surface. It is a measure of the quality of the modelling, and can be estimated by a modification of the power method.

4 Least squares formulation of the step computation

In this section the least squares formulation of the step computation for the scoring method is presented. This is adequate for present purposes because needed expectations can be computed explicitly in the numerical examples considered. However, there is an exactly parallel development for the sample method which has the advantage of avoiding the computation of these expectations when analytic results are not available.

The basic idea behind the least squares formulation is pretty simple. It uses (13) to transform the expected Hessian and then notes the result has the form of a normal matrix. This then permits (16) to be written as a linear least squares problem. The key quantities in (16) are

$$\mathcal{I}_n = \frac{1}{n}\sum_{\mathbf{t}\in T_n} \nabla_x \boldsymbol{\eta}^T \mathcal{E}\{\nabla_\eta L_t^T \nabla_\eta L_t\}\nabla_x \boldsymbol{\eta}, \quad \nabla_x L_t^T = \nabla_x \boldsymbol{\eta}^T \nabla_\eta L_t^T. \tag{29}$$

Now set $V_t = \mathcal{E}\{\nabla_\eta L_t^T \nabla_\eta L_t\} = V^{T/2}V^{1/2}$. This reveals the required structure in (29) and leads to the least squares formulation

$$\min_h \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = I_n^L\mathbf{h} - \mathbf{b}, \tag{30}$$

where

$$I_n^L = \begin{bmatrix} \vdots \\ V_t^{1/2}\nabla_x \boldsymbol{\eta} \\ \vdots \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \vdots \\ V_t^{-T/2}\nabla_\eta L_t \\ \vdots \end{bmatrix}. \tag{31}$$

Example 6 *The best known example corresponds to the normal distribution. Here*

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

In matrix terms this gives:

$$I_n^L = \begin{bmatrix} \vdots \\ \nabla_x \mu_t \\ \vdots \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \vdots \\ y_t - \mu(\mathbf{x},t) \\ \vdots \end{bmatrix}.$$

Here σ cancels. The result is the Gauss-Newton method linear sub problem.

However, note that the approximation $\mathcal{E}\left\{(y_t - \mu(\mathbf{x},t))^2\right\} = \sigma^2$ has been used. This simplification is characteristic of the scoring algorithm.

Example 7 *The multinomial distribution provides an example where the component blocks in (31) are less trivial. Here L_t is given by (4), and $b(\boldsymbol{\theta})$ by (6).*

$$\frac{\partial L_t}{\partial \theta_i} = m_i - \frac{\partial b(\boldsymbol{\theta})}{\partial \theta_i} = m_i - m\omega_i, \tag{32}$$

$$V_{ij} = -\mathcal{E}\left\{\frac{\partial^2 L_t}{\partial\theta_i \partial\theta_j}\right\} = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = m\left\{\omega_i\delta_{ij} - \omega_i\omega_j\right\}. \tag{33}$$

To solve (30) it is convenient to make an orthogonal factorization of I_n^L:

$$I_n^L = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}, \tag{34}$$
$$\mathbf{h} = U^{-1} Q_1^T \mathbf{b}.$$

Can get more value from this factorization:

$$\nabla_x \mathcal{L}\mathbf{h} = \left(\mathbf{b}^T I_n^L\right)\mathbf{h}, \tag{35}$$
$$= \mathbf{b}^T Q \begin{bmatrix} U \\ 0 \end{bmatrix} U^{-1} Q_1^T \mathbf{b},$$
$$= \|Q_1^T \mathbf{b}\|^2 \geq 0. \tag{36}$$

This quantity $\rightarrow 0$ as the iteration proceeds and so provides a scale invariant quantity for testing convergence. It is expressed as a sum of squares and so is necessarily non negative. If the Goldstein test is being used then this is the way to evaluate the denominator.

In implementing the orthogonal factorization there are good arguments for scaling the columns of I_n^L to have unit norm [2].

5 Properties of trust region methods

The trust region methods accept a full step but control the length of the step by requiring the next iterate to lie in a closed, adaptively defined control region containing the current estimate. This is achieved by imposing a length constraint on the least squares form of the linear subproblem (30). A typical form for the resulting problem is

$$\min_{\mathbf{h}, \|\mathbf{h}\|_D^2 \leq \gamma} \mathbf{r}^T \mathbf{r}; \quad \mathbf{r} = I_n^L \mathbf{h} - \mathbf{b}, \tag{37}$$
$$\|\mathbf{h}\|_D^2 = \mathbf{h}^T D^2 \mathbf{h}, \quad D > 0 \text{ diagonal}. \tag{38}$$

Necessary conditions give

$$\begin{bmatrix} \mathbf{r}^T & 0 \end{bmatrix} = \mathbf{u}^T \begin{bmatrix} I & -I_n^L \end{bmatrix} - \lambda \begin{bmatrix} 0 & \mathbf{h}^T D^2 \end{bmatrix},$$

where $\left[\mathbf{u}^T, \lambda\right]$ are the Lagrange multipliers, so $\mathbf{h}$ satisfies the perturbed scoring equations

$$\left(\mathcal{I}_n + \frac{\lambda}{n} D^2\right) \mathbf{h} = \frac{1}{n} \nabla_x \mathcal{L}^T. \tag{39}$$

It is necessary to solve the constraint inequality $\|\mathbf{h}(\lambda)\|_D^2 \leq \gamma$ for λ in order to use γ as the control variable. An efficient method is given in [5]. However,

it is also possible to use the multiplier λ for this purpose directly. The key result is that $\mathbf{h}$ is a decreasing function of λ. This follows by differentiating equation (39) with respect to λ. This gives

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

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

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

The classical form of the algorithm goes back to [4]. Here two parameters α, β are kept to adjust λ, and the basic sequence of operations is:

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

Successful steps will be taken eventually as

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

Experience suggests the choices of α, β are not critical. Typically $\alpha\beta < 1$ so the iteration approaches the Newton like methods as convergence is achieved. The scoring and sample algorithms are not exact Newton methods, and both can be regarded as regularized methods with λ as regularization parameter because of their generic positive definiteness. In this context it is not completely obvious that setting $\lambda = 0$ will increase the rate of convergence over that for λ small for given n in the data analysis context.

Basic theorems mirror the line search results [6].

Convergence Let $\{\mathbf{x}_i\}$ produced by the α, β procedure be contained in a bounded region R in which $\{\lambda_i\} < \infty$ then $\{\mathcal{L}_n(\mathbf{y}; \mathbf{x}_i, T_n)\}$ converges, and limit points of $\{\mathbf{x}_i\}$ are stationary points of $\mathcal{L}_n$.

Boundedness If the sequence $\{\lambda_i\}$ determined by the α, β procedure is unbounded while $\{\mathbf{x}_i\} \subset R$ then the norm of $\nabla_x^2\mathcal{L}_n$ is also unbounded in R.

The necessary conditions (39) for the trust region method can be written as the linear least squares problem:

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = \begin{bmatrix} X_n^L \\ \sqrt{\lambda}D \end{bmatrix} \mathbf{h} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}. \tag{40}$$

It is convenient to make the a preliminary factorization $X_n^L = Q \begin{bmatrix} U \\ 0 \end{bmatrix}$ independent of λ. Then $\mathbf{h}(\lambda)$ can be found by solving the typically much smaller problem:

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

where $\mathbf{c}_1 = Q_1^T\mathbf{b}$. This is a considerable advantage in general as an iteration of the α, β method may require the solution of (39) for several values of λ. Solution of the reduced problem requires making the further factorization

$$\begin{bmatrix} U \\ \sqrt{\lambda}D \end{bmatrix} = Q' \begin{bmatrix} U' \\ 0 \end{bmatrix},$$
$$Q'^T \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix} \to \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{bmatrix}.$$

The results corresponding to (34), (36) in the line search case are

$$\begin{aligned} \mathbf{h}(\lambda) &= (U')^{-1}\mathbf{c}_1', \\ \nabla_x \mathcal{L}_n \mathbf{h}(\lambda) &= \mathbf{c}_1^T U \mathbf{h}(\lambda), \\ &= \begin{cases} \begin{bmatrix} \mathbf{c}_1^T & 0 \end{bmatrix} \begin{bmatrix} U \\ \sqrt{\lambda}D \end{bmatrix} \left(U^TU + \lambda D^2\right)^{-1} \\ \begin{bmatrix} U^T & \sqrt{\lambda}D \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}, \end{cases} \\ &= \|\mathbf{c}_1'\|^2. \end{aligned}$$

The form of the trust region constraint interferes with the good scaling properties of the scoring algorithm. The best that can be hoped for in practical terms is that the linear subproblem has a useful invariance property with respect to diagonal scaling. Introduce the new variables $\mathbf{w} = W\mathbf{x}$ where W is diagonal. Then, using subscripts to distinguish $\mathbf{x}$, $\mathbf{w}$ variables

$$W^{-1}\left(U_x^TU_x + \lambda D^2\right)W^{-1}W\mathbf{h}_x = W^{-1}\nabla_{\mathbf{x}}\mathcal{L}_n^T.$$

This is equivalent to

$$\left(U_w^TU_w + \lambda W^{-1}D^2W^{-1}\right)\mathbf{h}_w = \nabla_{\mathbf{w}}\mathcal{L}_n^T.$$

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

$$D_i = \left\| \left(I_n^L \right)_{*i} \right\|.$$

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

$$\begin{aligned} \mathbf{h} &= \left(\left(I_n^L \right)^T I_n^L + \lambda D^2 \right)^{-1} \left(I_n^L \right)^T \mathbf{b}, \\ \Rightarrow D\mathbf{h} &= \left(D^{-1} \left(I_n^L \right)^T I_n^L D^{-1} + \lambda I \right)^{-1} D^{-1} \left(I_n^L \right)^T \mathbf{b}. \end{aligned}$$

The effect of this choice is to rescale the columns of I_n^L to have unit length. It is often sufficient to set $\lambda = 1$, and $D = \text{diag}\left\{ \left\| \left(I_n^L \right)_{*i} \right\|, \, i = 1, 2, \cdots, p \right\}$ initially . However, if there are significant fluctuations in the size of the elements of I_n^L then [5] recommends updating D by

$$D_i = \max \left\{ D_i, \, \| \left(I_n^L \right)_{*i} \| \right\}.$$

6 Numerical Results

The first example is based on the simple exponential model

$$\mu(t, \mathbf{x}) = x(1) + x(2) \exp\left(-x(3) t \right). \tag{41}$$

The values chosen for the parameters are $x^*(1) = 1$, $x^*(2) = 5$, and $x^*(3) = 10$. The model with this choice of parameters is not difficult in the sense that dependence on each of the parameters is reflected strongly in different features of the graph. Thus the interest is in the effects of simulated data errors. 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 $[0, 1]$.

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 ($\sigma^2 = 2$) was made so that small sample problems ($n = 32$) are relatively difficult. The log likelihood omits constant terms and 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 σ 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 according to the probability distribution (2). The log likelihood used is

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

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

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

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 in table 1. 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 Simple 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 $1.e^{-6}$.

The starred entries in the table correspond to two cases of nonconvergence. Figure 1 shows (in red) the current estimate after 50 iterations together with the data and the starting estimate. This shows clearly that the

	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

Table 1: Algorithm performance, mean of 10 runs

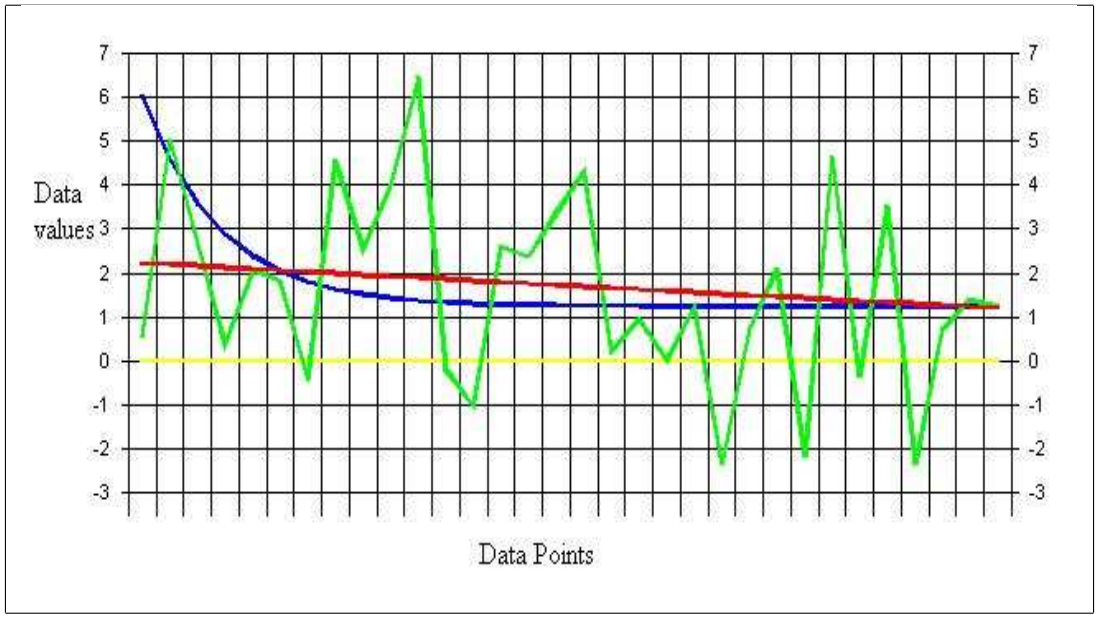

Figure 1: Result shows a straight line fit

.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

Table 2: Results for the peaks data

iteration is trying to approximate a straight line so that the set of approximations is not closed.

The second example uses as a model a sum of Gaussian peaks together with an exponential background.

$$\mu = \begin{cases} x(1)\exp\left(-x(2)t\right) + x(3)\exp\left(-(t-x(4))^2/x(5)\right) \\ +x(6)\exp\left(-(t-x(7))^2/x(8)\right) \end{cases}$$

To generate the data the following parameter values are used:

$$\begin{cases} x^*(1) = 5,\ x^*(2) = 10,\ x^*(3) = 18,\ x^*(4) = .3333, \\ x^*(5) = .05,\ x^*(6) = 15,\ x^*(7) = .6667,\ x^*(8) = .05 \end{cases}$$

Starting values are perturbed by multiplying by $1 + .5 * RND$. Note that this shifts the peaks in the initial approximation to the right. Results for the line search algorithm are given in table 2. These are much more of a mixed bag. The problems are closely related to the choice of starting values, in particular, the values chosen for the initial peak locations. The problem is illustrated in figure 2. This 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 which is badly positioned by the initial values. Both the background and the first peak are picked up well.

The third example makes use of a trinomial distribution (multinomial with $m = 3$). The data for this example is given in Table 3. It has its origin in a consulting exercise. It is derived from a study of the effects of a cattle virus on chicken embryos. The model suggested fits the frequencies explicitly ((42), (43), (44)). Thus it makes sense to develop the algorithm in terms of

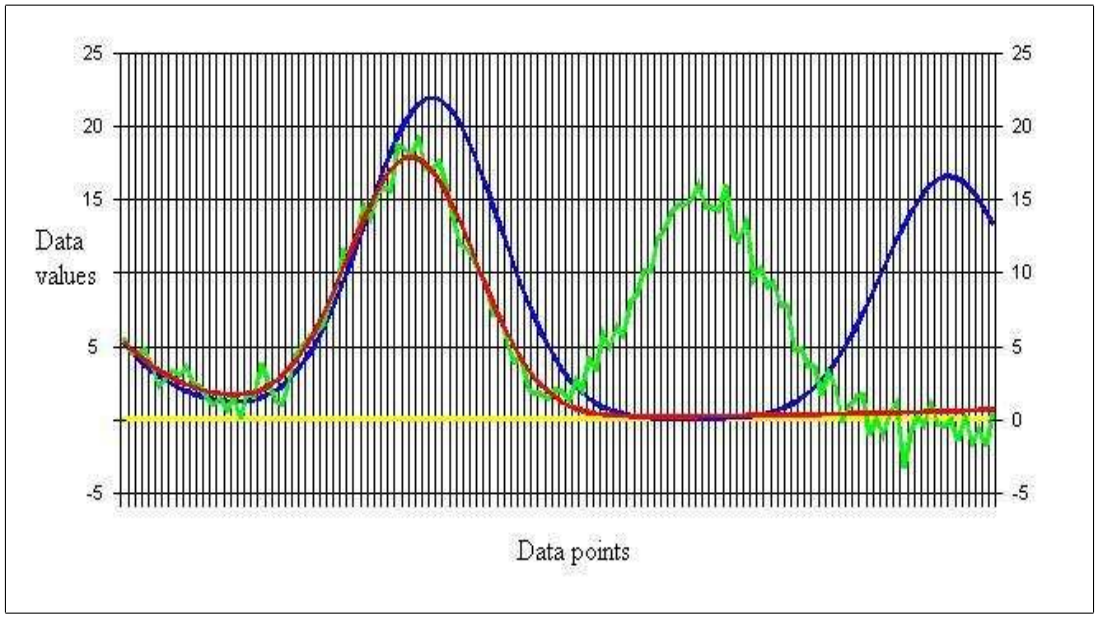

Figure 2: Initial conditions miss the second peak

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

Table 3: Cattle virus data

its	$\mathcal{L}$	$\nabla\mathcal{L}\mathbf{h}$	β_1	β_2	β_3
0	-54.86		-4.597	-3.145	.7405
1	-47.70	.1401+2	-3.737	-2.200	.7555
2	-47.01	.1277+1	-4.373	-2.551	.8803
3	-46.99	.3829-1	-4.503	-2.618	.9056
4	-46.99	.1234-4	-4.505	-2.619	.9061
5	-46.99	.3085-8	-4.505	-2.619	.9061

Table 4: Results of computations for the trinomial data

these.

$$\omega_1 = \frac{1}{1 + \exp\left(-\beta_1 - \beta_3 \log\left(t\right)\right)}, \tag{42}$$

$$1 - \omega_2 = \frac{1}{1 + \exp\left(-\beta_2 - \beta_3 \log\left(t\right)\right)}, \tag{43}$$

$$\omega_3 = 1 - \omega_1 - \omega_2. \tag{44}$$

Numerical results given in table 4 show an impressive rate of convergence for a relatively small data set. This suggests the model chosen is good.

7 Conclusion

The use of least squares methods in implementing scoring and sample algorithms has been exemplified. Numerical results have been presented illustrating algorithmic aspects such as the effects of initial values, the non closure of sets of approximating functions, and the importance of asymptotic convergence rates. Possibilities include both line search and trust region methods. Work on the Levenberg algorithm in the 1970's was responsible for at least some of the encouragement for the shift from line search to trust region methods in optimization problems. However, evidence has been presented here that conclusions derived from the nonlinear least squares problem area had a 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 λ are required then in the data analytic context the modelling could well be suspect.

2. The early papers relied on a small residual argument to explain good convergence rates. Again, in our context, this is not satisfactory. Here

the mechanism has to do with cancellation in sums of independent random variables. It is not completely obvious what the effect of small, non zero trust region parameters are in any particular case.

3. The trust region algorithms do not scale as well as the linesearch algorithms.

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

8 Acknowledgement

I am indebted to Hubert Schwetlick for his careful reading of the paper and for a number of suggestions which have significantly improved presentation.

References

[1] R. Fletcher, *Practical methods of optimization: Unconstrained optimization*, vol. 1, Wiley, Chichester, 1980.

[2] N.J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 1996.

[3] M. G. Kendall and A. Stuart, *The advanced theory of statistics*, vol. 2: Inference and Relationship, Charles Griffin and Company Limited, London, 1967.

[4] K. Levenberg, *A method for the solution of certain nonlinear problems in least squares*, Quart. Appl. Math. **2** (1944), 164–168.

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

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

[7] M.R. Osborne, *Fisher's method of scoring*, Int. Stat. Rev. **86** (1992), 271–286.