International Statistical Review (1992), **60**, 1, pp. 99–117. Printed in Great Britain
© International Statistical Institute

Fisher's Method of Scoring

M.R. Osborne

Statistics Research Section, Mathematical Sciences School, Australian National University

Summary

An analysis is given of the computational properties of Fisher's method of scoring for maximizing likelihoods and solving estimating equations based on quasi-likelihoods. Consistent estimation of the true parameter vector is shown to be important if a fast rate of convergence is to be achieved, but if this condition is met then the algorithm is very attractive. This link between the performance of the scoring algorithm and the adequacy of the underlying problem modelling is stressed. The effect of linear constraints on performance is discussed, and examples of likelihood and quasi-likelihood calculations are presented.

Key words and phrases: Maximum likelihood; Newtons method; Law of large numbers; Consistency; Quasi-likelihood; Method of scoring; Line search; Rate of convergence; Bound constraints; Multinomial likelihood.

1 Introduction

Two basic paradigms play important roles in the material developed in this paper. These are:

(1) Newton's method for function minimization, and
(2) the method of maximum likelihood for parameter estimation in data analysis problems.

The main aim is to examine aspects of the structure and performance of Fisher's method of scoring, a minimization technique based on the first paradigm, which provides a general approach to minimizing objective functions formulated using estimation strategies based on the second. The context is one in which the structure of the objective function depends on assumptions of a stochastic nature made about the problem, and it will be important to show how these affect the performance of the minimization algorithm.

The main assumptions made about the estimation problem are:

(1) The data set on which the estimation problem is based consists of event outcomes $\mathbf{y}_t \in R^m$, $t \in T$ where T contains the labelling information which identifies the particular events. Typically, $\mathbf{y}_t$ is a vector of a measurements on a system, while T could be an index set, the times at which particular observations are made, or a set of coordinates describing the possible configurations of a piece of apparatus. Events associated with distinct elements of T are assumed to be independent. The restriction to a single event type is made for simplicity.

(2) A quantity N, the effective sample size, is associated with the data set. It is assumed there is a systematic procedure which associates the particular data set with a realization of one of a family of experiments (each characterized by a distinct value of N) in such a way that the limiting behaviour as $N \rightarrow \infty$ can be described. Let T_N label the particular event set. Then 'systematic' means that

there is a measureable set $E \subset R^k$, $T_N \subset E$, and a increasing weight function $\omega(u)$ defined on E such that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t\in T_N} \phi(t) = \left[\int_0^1\right]^k \phi(t(u))\, d\omega(u) \tag{1.1}$$

where ϕ is any continuous functional defined on E and $t(u)$ maps $[0, 1]^k \to E$. For example, consider N observations taken at equispaced intervals of time such that $t\left(\frac{i}{N}\right) = \frac{i}{N}$, $\omega(u) = u$, and $E = [0, 1]$. Note that just the mechanism of sampling is being discussed here. That the mechanism is appropriate, so that information in the system increases without limit as $N \to \infty$, involves further assumptions.

(3) The 'true' probability density (mass for discrete distributions) associated with the event $\mathbf{y}_t$ is $g(\mathbf{y}_t, \boldsymbol{\eta}(t, \boldsymbol{\beta}^*), t)$. Here $\boldsymbol{\beta}^* \in R^p$ is a vector of parameters which is to be estimated from the given data, and $\boldsymbol{\eta}(t, \boldsymbol{\beta}) : R^p \times T \to R^q$ expresses the dependence on parametric and covariate data. That is $\boldsymbol{\eta}$ provides a model for the underlying data. It will always be assumed that quantities are smooth enough and that $\nabla_\beta \boldsymbol{\eta}$ has its maximum rank (this serves to get obvious regularity conditions out of the way).

(4) The negative of the log likelihood function used for the estimation of $\boldsymbol{\beta}$ is

$$\mathcal{K}(\boldsymbol{\beta}) = \sum_{t\in T} -\mathcal{L}_t; \quad \mathcal{L}_t = \log(f(\mathbf{y}_t, \boldsymbol{\eta}(t, \boldsymbol{\beta}), t)) \tag{1.2}$$

where f is the density hypothesized for the event $\mathbf{y}_t$, and may not be the same as g (but it will have similar smoothness properties). Note that the same $\boldsymbol{\eta}$ appears in both f and g so that the belief that a 'true model' belongs to a particular parametric class of possible models is not questioned explicitly. However, such cases as $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^* \times \boldsymbol{\beta}_2^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}_1 \times 0$ are not excluded.

Remark 1.1. Our development of scoring is essentially from the same point of view as that presented in [1], [11], [12] in considering generalized linear models and the statistical language GLIM. In this setting the implicit assumption is made that the appropriate limiting behaviour corresponds to $m/N \to 0$ as $N \to \infty$. Together with the assumption of the independence of the events $\mathbf{y}_t$, this makes practical an important connection between the basic scoring step (4.1) and a linear least squares problem (4.19). *This provides an effective method for implementing the scoring algorithm with distinct numerical advantages in terms of scaling and stability.* It also puts an emphasis on the form of the sampling strategy defined by T_N, and in subsequent argument involving appeals to the law of large numbers it will be assumed that appropriate regularity conditions are satisfied. A minimum condition on T is that it serves to identify $\boldsymbol{\beta}$ in the case that $\mathbf{y}_t$, $t \in T$ is given without error.

Example 1.1. One important model for the event data $\mathbf{y}_t$ is that of a signal observed in the presence of noise. In this case it is assumed that there is a parametric model for the signal given by

$$\mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\{\mathbf{y}_t\} = \boldsymbol{\mu}(t, \boldsymbol{\beta}^*) \tag{1.3}$$

where $\mathscr{E}$ is the expectation operator (for the density specified), and the mean vector $\boldsymbol{\mu}(t, \boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathscr{B}$, describes the parametric family of possible signals. Typically, $\boldsymbol{\mu} = \boldsymbol{\eta}$ in (1.2).

The purpose of this paper is to show not only that scoring provides a very satisfactory computing algorithm, but also that the performance of the algorithm provides important insights into the underlying problem modelling. This is done by establishing a connection between the two questions of the convergence of the computed estimates $\boldsymbol{\beta}_N^*$ as $N \to \infty$, and the asymptotic rate of convergence of the scoring algorithm. This link is provided by Newton's method.

The plan of the paper is as follows. The necessary properties of Newton's method for function minimization are sketched in the next section. It is used to show consistency of maximum likelihood and quasi-likelihood estimates in Section 3. The case for scoring is presented in Section 4. The argument in [16], which shows that the natural extension of the Levenberg algorithm for nonlinear least squares to scoring possesses both good global convergence properties and a fast rate of convergence in the case of likelihoods based on the exponential family of distributions provided the problem data is adequate, extends to the more general case (1.2) with only minor modification. For this reason, discussion is concentrated on the numerical properties of quasi-likelihood based estimation procedures. These have the novel property that it is the derivative of the objective function which is directly available rather than the function values themselves, and this consideration affects the numerical detail of the method. The emphasis on the rate of convergence of the scoring method complements the discussion of quasi-likelihoods in [11] and exponential dispersion models in [8]. It does this by stressing the nexus between efficient performance of the scoring algorithm and adequacy of the problem modelling. In other words, if scoring does not work well then the results are likely not of interest. The treatment of linear constraints on $\boldsymbol{\beta}$ is considered in Section 5, and some numerical results are presented in Section 6.

2 Newton's Method

In this section properties of Newton's method for function minimization are summarized as a preliminary to a comparison with Fisher's method of scoring. The method seeks to estimate a stationary point of the log likelihood by making a quadratic approximation to the objective function at the current point. Let

$$\mathcal{J} = \nabla^2 \mathcal{K}(\boldsymbol{\beta}), \tag{2.1}$$

then a step of the iteration is given by

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \mathbf{h}$$

where

$$\mathbf{h} = -\mathcal{J}(\boldsymbol{\beta})^{-1} \nabla \mathcal{K}(\boldsymbol{\beta})^T. \tag{2.2}$$

Also, it is usual to stabilize Newton's method by making a linesearch step using a merit function $Q(\boldsymbol{\beta})$ which has the same minimum as $\mathcal{K}(\boldsymbol{\beta})$, and which satisfies the descent condition

$$\nabla Q(\boldsymbol{\beta})\mathbf{h} < 0 \tag{2.3}$$

whenever $\mathbf{h}$ and $\nabla Q(\boldsymbol{\beta}) \neq 0$. The idea is that the step actually achieved in $\boldsymbol{\beta}$ should reduce Q and also satisfy some further criterion of effectiveness. The criterion discussed in [16] will serve the purpose here. It sets

$$\psi(\boldsymbol{\beta}, \mathbf{h}, \lambda) = \frac{Q(\boldsymbol{\beta} + \lambda \mathbf{h}) - Q(\boldsymbol{\beta})}{\lambda \nabla Q(\boldsymbol{\beta})\mathbf{h}} \tag{2.4}$$

and updates $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \lambda \mathbf{h} \tag{2.5}$$

where λ is chosen to satisfy the conditions

$$0 < \sigma < \psi(\boldsymbol{\beta}, \mathbf{h}, \lambda) < 1 - \sigma, \tag{2.6}$$

where typically σ is chosen small (say 10^{-4}).

Newton's method has the particular advantages:

(1) It has a fast (second order) rate of ultimate convergence to a zero $\hat{\boldsymbol{\beta}}$ of $\nabla\mathcal{K}$ provided $\mathcal{J}$ is nonsingular (but this does not imply convergence to a minimum of $\mathcal{K}$).
(2) It is invariant under constant, nonsingular, linear transformations. Close enough to the solution it is invariant to first order under general, nonsingular transformations.

However, there are also potential problems with Newton's method:

(1) Selection of a suitable monitor may not be straightforward. For example, it is not necessarily true that $\mathcal{K}$ is reduced in making a small enough step in the direction determined by $\mathbf{h}$ (a sufficient condition is that $\nabla^2\mathcal{K}$ be positive definite). Thus the simple tactic of setting $Q = \mathcal{K}$ in (2.4) may not be available for the Newton correction. Also, while $\mathbf{h}$ is downhill in the sense of (2.3) for minimizing $Q = \|\nabla\mathcal{K}\|^2$ which is minimized at all stationary points of $\mathcal{K}$, the use of this modified objective function can introduce further difficulties (the scaling may not be appropriate [2]). A possibility that needs to be explored further is that negative curvature in $\mathcal{K}$ is informative, and that the appropriate action needs to take this into account.
(2) Calculation of $\mathbf{h}$ requires a knowledge of second derivatives of $\mathcal{K}$. But often it is believed to be either uneconomical or inconvenient to compute these. For this reason it is conventional wisdom to seek methods which use at most first derivatives.

Newton's method has an advantage of a different kind in that it can be used to prove existence theorems. Consider, for example, the iteration started from an initial point $\boldsymbol{\beta}_0$ contained within some ball $B = \{\boldsymbol{\beta};\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \rho\}$ in which something is known (or is going to be assumed) about the regularity properties of the problem data. If

(1) $\mathcal{K}$ is twice continuously differentiable, and $\|\mathcal{J}(\boldsymbol{\beta}) - \mathcal{J}(\boldsymbol{\beta}')\| \leq K_1\, \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$ for all $\boldsymbol{\beta}, \boldsymbol{\beta}' \in B$.
(2) $\|\mathcal{J}(\boldsymbol{\beta}_0)^{-1}\| \leq K_2$, $\|\mathcal{J}(\boldsymbol{\beta}_0)^{-1}\nabla\mathcal{K}(\boldsymbol{\beta}_0)^T\| \leq K_3$,
(3) $\xi = K_1K_2K_3$ satisfies $\xi < \frac{1}{2}$,
(4) $\rho > \tau = (1/\xi)(1 - \sqrt{1 - 2\xi})\, K_3$,

then the full step Newton's method converges to a point $\hat{\boldsymbol{\beta}} \in B$, and $\hat{\boldsymbol{\beta}}$ is the only root of $\nabla\mathcal{K}(\boldsymbol{\beta}) = 0$ in B provided ρ is near enough to τ.

Remark 2.1. Note that τ can be chosen small if K_3 is small and the other assumptions hold with moderate values for the constants. Thus, if $\boldsymbol{\beta}_0$ gives a small value of K_3 then $\hat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}_0$ (certainly inside a ball of radius $2K_3$). An application of this result is made in the next section in discussing the convergence of the minima of $\mathcal{K}_N$ as $N \to \infty$. The idea is to use the stochastic properties to identify a limiting value $\boldsymbol{\beta}_L$. This is then used as $\boldsymbol{\beta}_0$ in showing that τ_N and ξ_N are small provided N is large enough.

3 Consistency

Consistency has to do with the question of how the parameter estimates $\hat{\boldsymbol{\beta}}_N$ obtained by minimizing (1.2) behave as $N \to \infty$. Newton's method finds stationary points so the assumption of a minimizing sequence involves a further condition. Here it is essential because the standard consistency results for maximum likelihood are results about global minima. A further question concerns the manner in which convergence behaviour is affected if the true probability density of the events $\mathbf{y}_t$ is unknown and guessed incorrectly; and this leads in turn to a third question which asks what kind of procedure can be employed to ensure that correct limiting behaviour of the computed estimates is not too critically dependent on guessing the right probability structure.

It is convenient to consider these questions here:

(1) because the convergence result for Newton's method permits a simple discussion of consistency (but one limited by quite strong smoothness requirements). A discussion under weaker conditions is given in Huber [6].

(2) because this approach provides the results in a form which permits them to be related directly to the numerical performance of the scoring method.

Remark 3.1. To discuss consistency it is necessary to identify first an appropriate limiting value $\boldsymbol{\beta}_L$ (assumed to exist). We associate the two ideas $\hat{\boldsymbol{\beta}}_N \to \boldsymbol{\beta}_L$ in the sense of almost sure convergence, and $\hat{\boldsymbol{\beta}}_N$ is a *consistent estimator of* $\boldsymbol{\beta}_L$ *under the assumed probability model.* Also it is important to characterize the case $\boldsymbol{\beta}_L = \boldsymbol{\beta}^*$. Here we will say that $\hat{\boldsymbol{\beta}}_N$ is *consistent* without further qualification. Assume that the true density is g. Then, indicating dependence on the current sample by subscript N and on the labelling within the sample by subscript t,

$$\begin{aligned}\nabla_\beta \mathscr{H}_N(\boldsymbol{\beta}) &= -\sum_{t\in T_N} \frac{1}{f_t} \nabla_\eta f_t \nabla_\beta \boldsymbol{\eta}_t \\ &= -\sum_{t\in T_N} \left(\frac{1}{f_t}\nabla_\eta f_t - \mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\left\{\frac{1}{f_t}\nabla_\eta f_t\right\}\right)\nabla_\beta \boldsymbol{\eta}_t - \sum_{t\in T_N} \mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\left\{\frac{1}{f_t}\nabla_\eta f_t\right\}\nabla_\beta \boldsymbol{\eta}_t. \end{aligned} \tag{3.1}$$

Now, assuming that the law of large numbers can be applied to the first summation in (3.1), this gives (compare (1.1))

$$\frac{1}{N}\nabla_\beta \mathscr{H}_N \to -\left[\int_0^1\right]^k \mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\left\{\frac{1}{f_t}\nabla_\eta f_t\right\}\nabla_\beta \boldsymbol{\eta}_t \, d\omega(u) \quad \text{a.s.} \tag{3.2}$$

where $\omega(u)$ is defined in (1.1) and describes the limiting distribution of data events and $t = t(u)$. Thus the relevant limiting value $\boldsymbol{\beta}_L$ for the minimizers of $\mathscr{H}_N$ for finite N solves the system of equations

$$\left[\int_0^1\right]^k \mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\left\{\frac{1}{f_t}\nabla_\eta f_t\right\}\nabla_\beta \boldsymbol{\eta}_t \, d\omega(u) = 0. \tag{3.3}$$

Similar expressions can be found in [10], [20]. Two important cases in which it is possible to deduce from (3.3) that $\boldsymbol{\beta}_L = \boldsymbol{\beta}^*$, the true vector of parameters, are:

(1) Let

$$\nabla_\eta \mathscr{L}_t = V_t(\boldsymbol{\beta})^{-1}\{\mathbf{y}_t - \boldsymbol{\mu}_t(\boldsymbol{\beta})\} \tag{3.4}$$

where V_t is positive definite. An important class of examples corresponds to f a member of the exponential family of distributions, and in this case $V_t(\boldsymbol{\beta}) = V(\boldsymbol{\mu}_t)$.

(2) Let $f = g$. Then the standard identity

$$\mathscr{E}_f\{\nabla_\eta \mathscr{L}_t\} = 0 \tag{3.5}$$

shows that $\boldsymbol{\beta}_L = \boldsymbol{\beta}^*$ provided $\boldsymbol{\beta}^*$ is an isolated solution of (3.3) and we recover the usual result.

To examine conditions under which $\hat{\boldsymbol{\beta}}_N \to \boldsymbol{\beta}_L$ a.s., assume the smoothness conditions needed for the Newton's method convergence theorem and set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_L$ in the first step for finding a root of $\frac{1}{N}\nabla\mathscr{K}_N(\boldsymbol{\beta})$. Then it follows (compare Remark 2.1) that a sufficient condition for

$$\tau_N \to 0, \quad \text{a.s.,} \quad N \to \infty \quad \Rightarrow \hat{\boldsymbol{\beta}}_N \to \boldsymbol{\beta}_L \quad \text{a.s.}$$

is that

$$\frac{1}{N}\nabla\mathscr{K}_N(\boldsymbol{\beta}_L) \to 0, \quad \frac{1}{N}\mathscr{I}_N(\boldsymbol{\beta}_L) \to \text{bounded, positive definite,} \quad \text{a.s.,} \quad N \to \infty.$$

The limit of the first term is (3.3). To analyse the second term note that the strong law now gives

$$\frac{1}{N}\mathscr{I}_N(\boldsymbol{\beta}_L) - \frac{1}{N}\mathscr{E}_{g(\boldsymbol{\beta}^*)}\{\mathscr{I}_N(\boldsymbol{\beta}_L)\} \to 0 \quad \text{a.s.} \tag{3.6}$$

But

$$\begin{aligned}\frac{1}{N}\mathscr{E}_{g(\boldsymbol{\beta}^*)}\{\mathscr{I}_N\} &= \frac{1}{N}\mathscr{E}_{f(\boldsymbol{\beta}_L)}\{\mathscr{I}_N\} - \frac{1}{N}\sum_{t\in T_N}\int_{\text{range }\mathbf{y}}(g(t, \boldsymbol{\beta}^*) - f(t, \boldsymbol{\beta}_L))\nabla^2\mathscr{L}_t\, d\mathbf{y} \\ &= \frac{1}{N}\sum_t \nabla_\beta \boldsymbol{\eta}_t^T \mathscr{E}_{f(t,\boldsymbol{\beta}_L)}\{\nabla_\eta \mathscr{L}_t(\boldsymbol{\beta}_L)^T \nabla_\eta \mathscr{L}_t(\boldsymbol{\beta}_L)\}\nabla_\beta \boldsymbol{\eta}_t \\ &\quad - \frac{1}{N}\sum_{t\in T_N}\int_{\text{range }\mathbf{y}}(g(t, \boldsymbol{\beta}^*) - f(t, \boldsymbol{\beta}_L))\nabla_\beta^2\mathscr{L}(\boldsymbol{\beta}_L)\, d\mathbf{y}\end{aligned} \tag{3.7}$$

where use has been made of the identity

$$-\mathscr{E}\left\{\frac{\partial^2 \mathscr{L}_t}{\partial\eta_i\,\partial\eta_j}\right\} = \mathscr{E}\left\{\frac{\partial\mathscr{L}_t}{\partial\eta_i}\frac{\partial\mathscr{L}_t}{\partial\eta_j}\right\}. \tag{3.8}$$

The first term on the right-hand side of (3.7) has the limiting value

$$\left[\int_0^1\right]^k \nabla_\beta \boldsymbol{\eta}_t^T V_t(\boldsymbol{\beta}_L)^{-1}\nabla_\beta \boldsymbol{\eta}_t\, d\omega(u) \tag{3.9}$$

where

$$V_t(\boldsymbol{\beta})^{-1} = \mathscr{E}_{f(t,\boldsymbol{\beta})}\{\nabla_\eta \mathscr{L}_t(\boldsymbol{\beta})^T \nabla_\eta \mathscr{L}_t(\boldsymbol{\beta})\}. \tag{3.10}$$

The second term has the limiting value

$$-\left[\int_0^1\right]^k \left\{\int_{\text{range }\mathbf{y}}(g(t, \boldsymbol{\beta}^*) - f(t, \boldsymbol{\beta}_L))\nabla_\beta^2\mathscr{L}_t(\boldsymbol{\beta}_L)\, d\mathbf{y}\right\} d\omega(u). \tag{3.11}$$

It is clear that (3.9) will fail to meet the bounded, positive definite requirement only under rather unusual circumstances so that $\hat{\boldsymbol{\beta}}_N$ being a consistent estimate of $\boldsymbol{\beta}_L$ under the hypothesized model will be ensured provided the relative size of the contribution of (3.11) is sufficiently small. That is provided the averaged contribution of the difference $g(\boldsymbol{\beta}^*) - f(\boldsymbol{\beta}_L)$ is sufficiently small. Of course, if the Hessian is indefinite a.s. for N large enough then the corresponding stationary points cannot be minima and the whole discussion loses its point. In practice, positive definiteness of $\mathscr{I}$ will likely be recognized from the performance of the minimization algorithm (this point is made explicit in the next section).

In many cases it is not difficult to construct consistent estimators of $\boldsymbol{\beta}^*$. The idea is to copy the form of $\mathscr{K}_n$ and prescribe the *quasilikelihood function* $\mathscr{W}_N = \sum_{t\in T} - \mathscr{W}_t$ by starting with a gradient specification analogous to (3.4)

$$\nabla_\mu \mathcal{W}_t^T = W_t(\boldsymbol{\mu})^{-1}(\mathbf{y}_t - \boldsymbol{\mu}(\boldsymbol{\beta}, t)) \tag{3.12}$$

where $W_t(\boldsymbol{\mu})$ is assumed to be positive definite. Note that this means that it is derivative rather than function values of $\mathcal{W}_N$ that are immediately available. Provided (1.3) holds, it follows from (3.12) that

$$\mathcal{E}_{g(t,\boldsymbol{\beta}^*)}\{\nabla_\mu \mathcal{W}_t(\boldsymbol{\beta}^*)\} = 0. \tag{3.13}$$

Thus, if $\hat{\boldsymbol{\beta}}_N$ satisfies the *estimating equation*

$$\nabla_\beta \mathcal{W}_N = \sum_{t \in T} \nabla_\mu \mathcal{W}_t \nabla_\beta \boldsymbol{\mu}(\boldsymbol{\beta}, t) = 0 \tag{3.14}$$

then the preceding argument goes through essentially unchanged to show that $\hat{\boldsymbol{\beta}}_N \rightarrow \boldsymbol{\beta}^*$ a.s., $N \rightarrow \infty$, where $\boldsymbol{\beta}^*$ satisfies (1.3) provided the law of large numbers applied to show that

$$\frac{1}{N}\nabla_\beta \mathcal{W}_N(\boldsymbol{\beta}^*) \rightarrow 0 \quad \text{a.s.,} \quad N \rightarrow \infty, \tag{3.15}$$

and

$$\begin{aligned}\frac{1}{N}\nabla_\beta^2 \mathcal{W}_N &= \frac{1}{N}\sum_t \{\nabla_\beta \boldsymbol{\mu}_t^T W_t^{-1} \nabla_\beta \boldsymbol{\mu}_t - (\nabla_\beta \boldsymbol{\mu}_t^T \nabla_\mu W_t^{-1} + \nabla_\beta^2 \boldsymbol{\mu}_t^T)(\mathbf{y}_t - \boldsymbol{\mu}_t)\} \\ &\rightarrow \left[\int_0^1\right]^k \nabla_\beta \boldsymbol{\mu}_t^T W_t^{-1} \nabla_\beta \boldsymbol{\mu}_t \, d\omega(u) \\ &= \text{bounded, positive definite.}\end{aligned} \tag{3.16}$$

Remarks 3.2. It is not difficult to satisfy these conditions for reasonably general distributions of errors. In particular, it is not necessary for W_t to be a consistent estimate of variance. One familiar and important example corresponds to $W_t = I$, the case of (nonlinear) least squares. It follows that this estimating procedure is consistent under typical conditions (adequate smoothness, density with bounded second moments, and sampling distribution compatible with (3.15), (3.16)). It is these latter conditions which contain the real substance of this result which is due to Jennrich [7].

Remark 3.3. The original suggestion of starting with the estimating equation (3.14) is due to Wedderburn [19]. A detailed discussion is given in [12]. One consequence of the easy access to a consistent estimator is a loss of efficiency. This can be minimized by an appropriate choice of W_t, and this question has received some attention. Wedderburn recommended $W_t(\mu) = V_t(\mu)$ as providing a generalization of maximum likelihood estimation for the exponential family in the case of scalar $\mathbf{y}$. Extension to the multivariate case is considered in [11]. Zeger & Liang discuss the use of a consistent estimate of V_t in quasilikelihood modelling of longitudinal data in several papers (for example [21]).

Remark 3.4. An alternative is to start with $\nabla \mathcal{W}_t$ satisfying (3.5). This is the basis for a discussion of optimal estimating equations in, for example, Godambe & Heyde [4] and McLeish & Small [13].

4 Fisher's Method of Scoring

The characteristic feature of the scoring algorithm is its replacement of the exact Hessian $\nabla^2\mathcal{K}$ in the Newton step (2.1) by its expectation. Thus the scoring step has the particular form

$$\mathbf{h} = -\mathcal{I}(\boldsymbol{\beta})^{-1}\nabla\mathcal{K}(\boldsymbol{\beta})^T \tag{4.1}$$

where

$$\nabla_\beta\mathcal{K}(\boldsymbol{\beta}) = \sum_t -\nabla_\eta\mathcal{L}_t\nabla_\beta\boldsymbol{\eta}_t,$$

$$\nabla_\beta^2\mathcal{K}(\boldsymbol{\beta}) = \sum_t \{-\nabla_\beta\boldsymbol{\eta}_t^T\nabla_\eta^2\mathcal{L}_t\nabla_\beta\boldsymbol{\eta}_t - \nabla_\beta^2\boldsymbol{\eta}_t\nabla_\eta\mathcal{L}_t\},$$

and

$$\mathcal{I} = \mathcal{E}_f\{\nabla_\beta^2\mathcal{K}(\boldsymbol{\beta})\} = \sum_t \nabla_\beta\boldsymbol{\eta}_t^T V_t(\boldsymbol{\beta})^{-1}\nabla_\beta\boldsymbol{\eta}_t, \tag{4.2}$$

where $V_t(\boldsymbol{\beta})$ is given by (3.10), and use has been made of the standard identities (3.4) and (3.8).

It follows immediately that:

(1) $\mathcal{I}$ is positive (semi) definite so that $\nabla\mathcal{K}\mathbf{h} < (=)0$ showing that the scoring step is necessarily downhill for minimizing $\mathcal{K}$ when $\mathcal{I}$ is nonsingular, and that the choice of monitor function $Q = \mathcal{K}$ is available in (2.4)—a distinct advantage over Newton's method,

(2) scoring has the same transformation invariance properties as Newton's method, and

(3) scoring requires only first derivative information for its implementation.

On the other hand it is necessary to compute expectations, and this could prove a difficulty because numerical evaluation of the integrals required in (4.2) does not appear practical for continuous distributions in general. However, in many cases the results needed are known exactly (the exponential family of distributions and their generalizations [8] providing an important class of examples).

But the potential problems of computing expectations does provide another reason for considering estimating equations based on the modified score function (3.12). In this case we set

$$\mathcal{I}^s(\boldsymbol{\beta}) = \sum_t \nabla_\beta\boldsymbol{\mu}_t^T W_t^{-1}\nabla_\beta\boldsymbol{\mu}_t \tag{4.3}$$

so that

$$\mathcal{I}^s(\boldsymbol{\beta}^*) = \mathcal{E}_{g(\boldsymbol{\beta}^*)}\{\nabla_\beta^2\mathcal{W}(\boldsymbol{\beta}^*)\} \tag{4.4}$$

and define the scoring step by

$$\mathbf{h} = -\mathcal{I}^s(\boldsymbol{\beta})^{-1}\nabla_\beta\mathcal{W}(\boldsymbol{\beta})^T. \tag{4.5}$$

Most of the good features of the scoring algorithm recorded above extend to this modified version. However, the line search is complicated by the lack of a monitor function in general. An alternative is to determine the step length parameter λ by finding the closest zero of the directional derivative by solving

$$\mathcal{W}'(\boldsymbol{\beta} + \lambda\mathbf{h} : \mathbf{h}) = 0. \tag{4.6}$$

This equation may have to be solved fairly accurately to ensure that the stability condition that $\mathcal{W}$ be reduced in the current step is satisfied.

The above discussion shows that scoring inherits most of the good properties of Newton's method, and the most serious remaining questions concern its convergence behaviour. Here too there are satisfactory answers. To summarize these assume that (4.1) is used to generate the scoring step, the monitor function is $Q = \mathcal{K}$, the step λ is chosen to satisfy (2.4), and successive iterates are confined to a compact region $D \subset R^p$ (for example, D could be bounded by a closed equipotential) in which $\mathcal{I}$ has full rank and $\nabla^3\mathcal{K}$ is continuous. Then the standard arguments [16] show that

(1)
$$\nabla\mathcal{K}\mathbf{h}/\|\nabla\mathcal{K}\|\,\|\mathbf{h}\| < -1/\operatorname{cond}(\mathcal{I}) \tag{4.8}$$
(where $\operatorname{cond}(\mathcal{I}) = \|\mathcal{I}\|\,\|\mathcal{I}^{-1}\|$ is the condition number of $\mathcal{I}$) so that the descent direction given by the scoring algorithm is uniformly downhill for minimizing $\mathcal{K}$, and

(2) limit points of the iteration are stationary points of $\mathcal{K}$.

Similar results hold for the scoring step based on (4.5) with $Q = \mathcal{W}$ as an (implicitly defined) monitor corresponding to the strategy based on finding a zero of (4.6).

Remark 4.1. Note that $\mathcal{I}$ having full rank in D is not too difficult a global condition. It does not imply that g has full rank so that it does not suffice to guarantee a single stationary point in D. It may suggest superior global convergence properties for the scoring algorithm. To make further progress it is convenient to assume convergence to a local minimum $\hat{\boldsymbol{\beta}}$ which is a consistent estimate of the true parameter vector $\boldsymbol{\beta}^*$. This need not be too extreme an assumption in the likelihood case when the objective function can have good convexity properties [18]. The rate of convergence can now be estimated using essentially the same argument as in [16], [17]. For this reason only the estimating equation case is considered here. The argument goes in two stages. In the first it is shown that $\lambda = 1$ will satisfy the step choice criterion when the current iterate $\boldsymbol{\beta}$ is close enough to $\hat{\boldsymbol{\beta}}_N$ and N is large enough. In the second stage it is assumed that $\lambda = 1$ is allowable, and the rate of convergence of the full step method is discussed.

In the first stage note that λ satisfies (4.6) when $\|\mathbf{h}\|$ is small provided

$$\begin{aligned} 0 &= \nabla\mathcal{W}_N(\boldsymbol{\beta})\mathbf{h} + \lambda\mathbf{h}^T\nabla^2\mathcal{W}_N\mathbf{h} + O(\|\mathbf{h}\|^3), \\ &= \mathbf{h}^T((\lambda-1)\nabla^2\mathcal{W}_N + \nabla^2\mathcal{W}_N - \mathcal{I}_N^s)\mathbf{h} + O(\|\mathbf{h}\|^3). \end{aligned} \tag{4.9}$$

We assume that

$$\frac{1}{N}\nabla^2\mathcal{W}_N(\boldsymbol{\beta}) \to \text{bounded, positive definite},$$

while the standard argument invoking the law of large numbers now shows that (almost surely as $N \to \infty$)

$$\frac{1}{N}(\nabla^2\mathcal{W}_N(\boldsymbol{\beta}) - \mathcal{I}_N^s(\boldsymbol{\beta})) = o(1) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|)$$

so that the term involving $(\lambda - 1)$ dominates the bracketed term in (4.9). It follows that the minimizing λ is close to 1.

Remark 4.2. There is a corresponding result in the likelihood case—that $\lambda = 1$ can be used eventually when N is large enough. But now (2.4) provides a criterion which permits us to test if $\lambda = 1$ is acceptable. We do not have access to a corresponding test when values of $\mathcal{W}$ are not available.

In the second stage of the rate of convergence result the iteration (4.5) is written in fixed point form as

$$\boldsymbol{\beta}_{i+1} = F(\boldsymbol{\beta}_i); \qquad F(\boldsymbol{\beta}) = \boldsymbol{\beta} - \mathcal{I}^s(\boldsymbol{\beta})^{-1}\nabla_\beta\mathcal{W}(\boldsymbol{\beta})^T. \tag{4.10}$$

It is a standard result that $\hat{\boldsymbol{\beta}}_N$ is a point of attraction for this iteration provided the spectral radius of F' (written $\varpi(F')$) satisfies

$$\varpi(F'(\hat{\boldsymbol{\beta}}_N)) < 1. \tag{4.11}$$

Because $\nabla_\beta \mathscr{W}(\hat{\boldsymbol{\beta}}_N) = 0$ we have (using (4.4) and the familiar argument)

$$\begin{aligned} F'_N(\hat{\boldsymbol{\beta}}_N) &= I - \mathscr{I}^s_N(\hat{\boldsymbol{\beta}}_N)^{-1}\nabla^2_x \mathscr{W}_N(\hat{\boldsymbol{\beta}}_N) \\ &= \left(\frac{1}{N}\mathscr{I}^s(\hat{\boldsymbol{\beta}}_N)\right)^{-1}\left(\frac{1}{N}(\mathscr{I}^s_N(\hat{\boldsymbol{\beta}}_N) - \nabla^2_x \mathscr{W}_N(\hat{\boldsymbol{\beta}}_N))\right) \\ &= F'_N(\boldsymbol{\beta}^*) + O(\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\|) \quad \text{a.s.,} \quad N \to \infty \end{aligned} \tag{4.12}$$

where

$$F'_N(\boldsymbol{\beta}^*) = o(1) \quad \text{a.s.,} \quad N \to \infty. \tag{4.13}$$

Thus

$$\varpi(F'_N(\hat{\boldsymbol{\beta}}_N)) \to 0 \quad \text{a.s.,} \quad N \to \infty. \tag{4.14}$$

This shows that the rate of convergence of the scoring algorithm increases as the effective sample size N increases. The consequence is that scoring is an attractive algorithm, in the sense that it has a fast rate of ultimate convergence, provided the measure N of the effective sample size is large enough.

Remark 4.3. When the choice $f \neq g$ is made in the definition of $\mathscr{H}$ then the scoring algorithm uses as Hessian estimate

$$\mathscr{I}(\boldsymbol{\beta}) = \mathscr{E}_{f(\boldsymbol{\beta})}\{\mathscr{I}(\boldsymbol{\beta})\}. \tag{4.15}$$

Arguing as in the derivation of (3.6), (3.7), the effect of the misidentification is to give the almost sure limiting value

$$\begin{aligned} F'_\infty(\boldsymbol{\beta}_L) = &\left\{\left[\int_0^1\right]^k \nabla_\beta \boldsymbol{\eta}_t^T V_t(\boldsymbol{\beta}_L)^{-1} \nabla_\beta \boldsymbol{\eta}_t \, d\omega(u)\right\}^{-1} \\ &\times \left\{\left[\int_0^1\right]^k \left(\int_{\text{range } \mathbf{y}} (g(t, \boldsymbol{\beta}^*) - f(t, \boldsymbol{\beta}_L)) \nabla^2 \mathscr{L}_t(\boldsymbol{\beta}_L) \, d\mathbf{y}\right) d\omega(u)\right\}. \end{aligned} \tag{4.16}$$

Comparing (4.16) *with* (3.6), (3.7) *it will be seen that the condition for* $\boldsymbol{\beta}_L$ *to be an attractive fixed point for the scoring algorithm implies the condition for* $\mathscr{I}(\boldsymbol{\beta}_L)$*, the sample information, to be positive definite.* This follows from the result that if A is positive definite and B is symmetric then

$$\varpi\{A^{-1}B\} < 1 \Rightarrow A \pm B \quad \text{positive definite.}$$

Thus there is a very close connection between the ultimate convergence of the scoring algorithm with a unit step in the line search, and the condition for $\hat{\boldsymbol{\beta}}_N$ to be a consistent estimator for $\boldsymbol{\beta}_L$. It follows from (4.16) that misidentification of the density can explain poor performance of the scoring algorithm for large N, in particular, when $\hat{\boldsymbol{\beta}}_N \to \boldsymbol{\beta}_L \neq \boldsymbol{\beta}^*$ a.s. $N \to \infty$.

Remark 4.4. An alternative way to look at the above results is that if the scoring algorithm converges rapidly then there is direct evidence that the problem modelling is adequate. It is known that $\varpi(F'(\boldsymbol{\beta}^*))$ is an invariant characterizing the local nonlinearity of the likelihood surface (see, for example, [9]). If it is small then confidence intervals based on linear theory will be adequate for the parameter estimates. *Clearly knowledge of* $\varpi(F'(\beta^*))$ *is of value, and this can frequently be estimated from the ratio* $\|\mathbf{h}_{i+1}\|/\|\mathbf{h}_i\|$ *in the case that the largest eigenvalue in modulus of* $F'(\boldsymbol{\beta}^*)$ *is isolated.* This just corresponds to an application of the classical power method for determining eigenvalues and will work best when the convergence rate is slow—but in the contrary case it will return a small number as required.

Remark 4.5. Both the forms (4.1) and (4.5) of the scoring algorithm possess the property that the calculation of the scoring step can be cast in the form of a linear least squares problem. This point is useful in implementation. Consider (4.5), define

$$A_t = \nabla_\beta \boldsymbol{\mu}_t^T W_t^{-\frac{1}{2}}, \tag{4.17}$$

and set

$$A = \begin{bmatrix} A_1^T \\ \vdots \\ A_{|T|}^T \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (W_1^{-\frac{1}{2}})^T(\mathbf{y}_1 - \boldsymbol{\mu}_1) \\ \vdots \\ (W_{|T|}^{-\frac{1}{2}})^T(\mathbf{y}_{|T|} - \boldsymbol{\mu}_{|T|}) \end{bmatrix}. \tag{4.18}$$

Then (4.5) can be written

$$\min \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = A\mathbf{h} - \mathbf{b}. \tag{4.19}$$

Note that this system has a standardized right-hand side at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ if W_t is the correct covariance matrix. This suggests that this formulation corresponds to an efficacious scaling.

5 Scoring Under Constraints

Discussion of general constrained problems is not attempted. But, as a general rule, local behaviour can be predicted on the basis of quadratic approximation of the objective function and linear approximation of the active constraints; and linear constraints of one form or another are never too far away from the class of problems we are considering. The point to be made here is that scoring usually maintains its good properties when applied to constrained problems. Examples of constraints include:

(1) constraints defining discrete probability distributions. For example:

$$\pi_i \geq 0, \quad i = 1, 2, \ldots, m, \quad \sum_{i=1}^{m} \pi_i = 1; \tag{5.1}$$

(2) constraints to ensure identifiability in analysis of variance calculations. For example: let

$$g(\mu_{ij}) = m + a_i + b_j, \quad i = 1, 2, \ldots, n_a, \quad j = 1, 2, \ldots, n_b, \tag{5.2}$$

then, typically the conditions

$$\sum_{i=1}^{n_a} a_i = 0, \quad \sum_{j=1}^{n_b} b_j = 0 \tag{5.3}$$

are imposed to remove the obvious ambiguities.

(3) Constraints can express additional information about the problem. For example, information that some parameters must be non-negative could be the expression of some physical law. Constraints of this kind are called bound constraints.

(4) Sometimes it is convenient to fix certain components of the parameter vector $\boldsymbol{\beta}$ in exploratory calculations involving controlling and setting parameter values in testing a range of models. Scoring may not be too robust with respect to this kind of activity which could introduce inconsistency if the complexity is understated, and exaggerate any tendency towards ill-conditioning in $\mathcal{I}$ if it is overstated.

Frequently problems of the first kind are eliminated by introducing a suitable parametrization (this point is illustrated in the next section where we consider an application involving a multi-nomial likelihood). Linear equality constraints (for example (5.3)) can be used to eliminate variables, but this can cause fill-in in sparse matrix structures such as (5.2) and so may not be an optimal strategy. To discuss the effect of linear equality constraints on the performance of the scoring algorithm consider the constrained maximum likelihood problem (there is a corresponding quasi-likelihood development)

$$\min_{\boldsymbol{\beta}} \mathscr{K}_N(\boldsymbol{\beta}) \quad \text{subject to } C\boldsymbol{\beta} = \mathbf{d}. \tag{5.4}$$

The necessary conditions for a minimum give the equations

$$\begin{gathered} \nabla_\beta \mathscr{K}_N(\boldsymbol{\beta}) = \mathbf{z}^T C \\ C\boldsymbol{\beta} = \mathbf{d} \end{gathered} \tag{5.5}$$

where $\mathbf{z}$ is the vector of multipliers. Setting $\boldsymbol{\zeta} = \mathbf{z}/N$, then the scoring step is given by

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\zeta} \end{bmatrix} \leftarrow \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\zeta} \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{h}_\beta \\ \mathbf{h}_\zeta \end{bmatrix}$$

where the descent direction satisfies the system

$$\mathscr{I}_N \mathbf{h}_\beta - NC^T \mathbf{h}_\zeta = -\nabla \mathscr{K}_N^T(\boldsymbol{\beta}) + NC^T \boldsymbol{\zeta} \tag{5.6a}$$

$$-NC\mathbf{h}_\beta = N(C\boldsymbol{\beta} - \mathbf{d}). \tag{5.6b}$$

To discuss the rate of convergence note that the appropriate limiting system corresponding to the $N \rightarrow \infty$ limit in (5.5) is

$$\begin{gathered} \left[\int_0^1\right]^k \mathscr{E}_{g(t,\boldsymbol{\beta}^*)}\{\nabla_\beta \mathscr{L}_t\}\, d\omega(u) = \boldsymbol{\zeta}^{*T} C \\ C\boldsymbol{\beta}^* = \mathbf{d}. \end{gathered} \tag{5.7}$$

Arguing as in (4.10)–(4.14), the rate of convergence depends on

$$\varpi\left(\begin{bmatrix} \frac{1}{N}\mathscr{I}_N & -C^T \\ -C & 0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{N}(\mathscr{J}_N - \mathscr{I}_N) & 0 \\ 0 & 0 \end{bmatrix}\right).$$

It gets small as $N \rightarrow \infty$ under similar conditions. Now consider the factorization

$$C^T = Q\begin{bmatrix} U \\ 0 \end{bmatrix} \tag{5.8}$$

where $Q = [Q_1 \;\; Q_2]$ is orthogonal and U is upper triangular. Substituting in (5.6) gives, in particular,

$$Q_2^T \mathscr{I}_N Q_2 \mathbf{h}_c = -Q_2^T \nabla_\beta \mathscr{K}_n \tag{5.9}$$

provided $\boldsymbol{\beta}$ satisfies the equality constraints, where

$$\mathbf{h}_c = Q_2^T \mathbf{h}_\beta \qquad (\mathbf{h}_\beta = Q_2 \mathbf{h}_c) \tag{5.10}$$

and this can readily be put into the form (4.19). Again, a similar development holds for quasi-likelihoods.

Similar results can be obtained for bound constraints. Major changes are not needed in the consistency arguments (they are based on the distribution of $\mathbf{y}$ rather than of $\boldsymbol{\beta}$)

although an extension of the Newton result is needed and second order sufficiency becomes the right property in showing $\boldsymbol{\beta}^*$ is isolated. A discussion of consistency for bound constraints under much weaker conditions is given in Moran [14]. He derives the distribution of $\boldsymbol{\beta}$ in Moran [15].

There is no real restriction in assuming lower bounds only, and in this case the maximum likelihood problem becomes

$$\min_{\boldsymbol{\beta}\in\mathcal{B}} \mathcal{K}(\beta); \quad \mathcal{B}=\{\boldsymbol{\beta};\boldsymbol{\beta}\geqslant\mathbf{b}\}. \tag{5.11}$$

The Kuhn–Tucker conditions characterizing the optimum are

$$\nabla\mathcal{K}(\boldsymbol{\beta})=\mathbf{u}^T, \tag{5.12a}$$

$$\mathbf{u}\geqslant 0, \quad u_i(\hat{\beta}_i-b_i)=0, \quad i=1,2,\ldots,p. \tag{5.12b}$$

Let $\sigma(\boldsymbol{\beta})=\{i;\beta_i>b_i\}$ be an index set pointing to the variables not at their bounds, and let $P=\begin{bmatrix}H_\sigma\\K_\sigma\end{bmatrix}$ be a permutation matrix with the property that

$$H_\sigma\boldsymbol{\beta}=\boldsymbol{\beta}_\sigma, \quad K_\sigma\boldsymbol{\beta}=\boldsymbol{\beta}_{\sigma^c} \tag{5.13}$$

separates $\boldsymbol{\beta}$ into its fixed and free components. Then the scoring step constrained by the condition $K_\sigma\boldsymbol{\beta}=0$ is given by the solution of the system of equations (compare (5.6))

$$\mathcal{I}\mathbf{h}+\nabla\mathcal{K}^T-K_\sigma^T\boldsymbol{\zeta}=0 \tag{5.14a}$$

$$K_\sigma\mathbf{h}=0; \tag{5.14b}$$

and this can be put in the equivalent form

$$\mathbf{h}=-H_\sigma^T(H_\sigma\mathcal{I}H_\sigma^T)^{-1}H_\sigma\nabla\mathcal{K}^T. \tag{5.15}$$

Here $\boldsymbol{\zeta}$ can be interpreted as a set of tentative Kuhn–Tucker multipliers, and optimality corresponds to

$$\nabla\mathcal{K}^T=K_\sigma^T\boldsymbol{\zeta}, \quad \boldsymbol{\zeta}\geqslant 0. \tag{5.16}$$

The algorithm could proceed as follows:

```
(1) compute $\mathbf{h}$, $\boldsymbol{\zeta}$ at the current point
(2) if $\zeta_q<0$ where $q \leftarrow arg\ \min_i \zeta_i$
            then $\sigma\leftarrow\sigma\cup\{q\}$
                recompute $\mathbf{h}$
                (it can be shown that this defines a feasible direction)
            else stop if $\|H_\sigma\nabla\mathcal{K}^T\|$ is small enough.
(3) Let $\lambda_1$ satisfy the step criterion (2.4)
            $q\leftarrow arg\ \min_i(\lambda;\beta_i+\lambda h_i=0,\lambda>0,i\in\sigma)$
            $\lambda_2\leftarrow-\beta_q/h_q$
            if $\lambda_2<\lambda_1$ then $\lambda\leftarrow\lambda_2$, $\sigma\leftarrow\sigma\backslash\{q\}$
                          else $\lambda\leftarrow\lambda_1$
            $\boldsymbol{\beta}\leftarrow\boldsymbol{\beta}+\lambda\mathbf{h}$
            repeat (1).
```

To analyse the rate of convergence of the algorithm assume that $\hat{\boldsymbol{\beta}} \in S \subset \mathcal{B}$ where S possesses the properties

(1) S is small so that $\boldsymbol{\beta} \in S \Rightarrow \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\| \leq \delta$.
(2) S is bounded by an equipotential, $S = \{\boldsymbol{\beta};\, \mathcal{K}(\boldsymbol{\beta}) \leq \bar{\mathcal{K}}\} \cap \mathcal{B}$.
(3) S intersects only constraints active at $\hat{\boldsymbol{\beta}}$ so that

$$(\beta_i = b_i, \quad \boldsymbol{\beta} \in S) \rightarrow \hat{\beta}_i = b_i.$$

(4) A form of strict complementarity implying a condition on S:

$$\hat{\beta}_i = b_i \Rightarrow (\nabla\mathcal{K}(\boldsymbol{\beta}))_i > 0, \quad \forall \boldsymbol{\beta} \in S. \tag{5.17}$$

It follows from (3), (4) that

$$H_\sigma \nabla\mathcal{K}(\boldsymbol{\beta})^T = 0 \Rightarrow \boldsymbol{\beta} = \hat{\boldsymbol{\beta}} \text{ because } \boldsymbol{\beta} \in S \Rightarrow \sigma(\boldsymbol{\beta}) \subseteq \sigma(\hat{\boldsymbol{\beta}}).$$

Lemma 5.1. *If $\sigma(\hat{\boldsymbol{\beta}})\backslash\sigma(\boldsymbol{\beta}) \neq \emptyset$ then* **h** *is downhill for minimizing $\mathcal{K}$, and the minimum with respect to λ in S of $\mathcal{K}(\boldsymbol{\beta} + \lambda\mathbf{h})$ is attained at a point where at least one more component of* $\boldsymbol{\beta}$ *is at a bound.*

Proof. From the above assumptions

$$H_\sigma \nabla\mathcal{K}(\boldsymbol{\beta}) = H_\sigma \mathbf{u} + O(\delta) = O(1), \quad \boldsymbol{\beta} \in S, \quad \sigma(\hat{\boldsymbol{\beta}})\backslash\sigma(\boldsymbol{\beta}) \neq \emptyset,$$

so that

$$\begin{aligned} \nabla\mathcal{K}(\boldsymbol{\beta} + \lambda\mathbf{h})\mathbf{h} &= -\mathbf{u}^T H_\sigma^T (H_\sigma \mathcal{I} H_\sigma^T)^{-1} H_\sigma \mathbf{u} + O(\delta) \\ &< 0, \quad \lambda \leq \lambda_2. \end{aligned} \tag{5.18}$$

It follows from assumption (2) above that $\mathcal{K}$ is minimized in S for $\lambda = \lambda_2$ so that the constraint $\beta_q = b_q$ becomes active.

Corollary 5.1. *If the scoring algorithm is started from a point $\boldsymbol{\beta} \in S$ then in a number of steps bounded by the number of constraints active at $\hat{\boldsymbol{\beta}}$ the correct active set is located.*

This result is reported in [17]. A similar result has been given in Burke & Moré [3].

Remark 5.1. It now follows from assumption (d) that constraints are never deleted once $\boldsymbol{\beta} \in S$. The constraint set is built up to $\sigma(\hat{\boldsymbol{\beta}})$ in a finite number of steps and the iteration then proceeds as an unconstrained problem in the reduced variables $\mathbf{z} = H_{\sigma(\hat{\boldsymbol{\beta}})}\boldsymbol{\beta}$. The rate of convergence of the scoring method is now given by

$$\varpi(F') = \varpi\{(H_\sigma \mathcal{I} H_\sigma^T)^{-1}(H_\sigma(\mathcal{I} - \mathcal{J})H_\sigma^T)\}.$$

This quantity is small under the same conditions that obtain in the unconstrained case.

6 Numerical Examples

Here numerical results illustrating a number of the points made in previous sections are given. Two problems are considered. The first is an example of a multinomial likelihood and illustrates a successful application of the scoring method. The second example is a quasi-likelihood calculation taken from Wedderburn's original paper [19]. Here the scoring method is only moderately successful and gives rise to some doubts about the modelling of the problem. Both examples involve constraints of the types previously considered. In the likelihood problem they are removed by reparametrization. In the quasi-likelihood problem which is solved iteratively it is convenient to use the transformed subproblem (5.9).

The multinomial likelihood. This distribution occurs when a sequence of n independent and identical experiments is performed. If each experiment results in one of m possible outcomes with respective probabilities $\pi_i \geqslant 0$, $i = 1, 2, \ldots, m$, $\sum_{i=1}^{m} \pi_i = 1$, then

$$P\{Y_i = y_i, i = 1, 2, \ldots, m\} = \frac{n!}{y_1! \ldots y_m!} \pi_1^{y_1} \ldots \pi_m^{y_m} \tag{6.1}$$

where y_i is the number of times that the i'th outcome occurred and $\sum_{i=1}^{m} y_i = n$. Consider now the case in which the multinomial distribution governs the outcome of $n(t)$ trials for each t, $t \in T$. Here the $y_i(t)$ are observed and the quantities to be modelled are the π_i. Assuming that the outcomes of each group of trials are mutually independent then the log likelihood is additive, and the contribution from each group of trials is given by

$$\mathscr{L}_t = \sum_{i=1}^{m} y_i(t) \log (\pi_i(t)) + \text{const.}, \tag{6.2}$$

and this can be put in the form (1.2) by setting

$$(\boldsymbol{\eta}_t)_i = \pi_i(\boldsymbol{\beta}, t), \quad i = 1, 2, \ldots, m-1, \quad \pi_m(\boldsymbol{\beta}, t) = 1 - \mathbf{e}^T \boldsymbol{\eta}_t. \tag{6.3}$$

The reduction of a scoring step to a linear least squares problem is straightforward in this case. Differentiating (6.2) and taking expectations where appropriate gives

$$\frac{\partial \mathscr{L}_t}{\partial \eta_i} = \frac{y_i(t)}{(\boldsymbol{\eta}_t)_i} - \frac{y_m(t)}{1 - \mathbf{e}^T \boldsymbol{\eta}_t} = \frac{y_i(t)}{\pi_i(t)} - \frac{y_m(t)}{\pi_m(t)}, \quad i = 1, 2, \ldots, m-1, \tag{6.4}$$

$$\frac{\partial^2 \mathscr{L}_t}{\partial \eta_i \, \partial \eta_j} = -\frac{y_i(t)}{\pi_i(t)^2} \delta_{ij} - \frac{y_m(t)}{\pi_m(t)^2}, \quad i, j = 1, 2, \ldots, m-1, \tag{6.5}$$

$$\mathscr{E}\left\{\frac{\partial^2 \mathscr{L}_t}{\partial \eta_i \, \partial \eta_j}\right\} = -\frac{n(t)}{\pi_i(t)} \delta_{ij} - \frac{n(t)}{\pi_m(t)}, \quad i, j = 1, 2, \ldots, m-1. \tag{6.6}$$

Thus V_t^{-1} in (3.10) is given by

$$\begin{aligned} V_t^{-1} &= n(t)\{\operatorname{diag}(1/\pi_i(t)) + (1/\pi_m(t))\mathbf{e}\mathbf{e}^T\} \\ &= n(t) D_t^{\frac{1}{2}} (I + (1/\pi_m(t)) \mathbf{v}\mathbf{v}^T) D_t^{\frac{1}{2}} \end{aligned} \tag{6.7}$$

where

$$D_t = \operatorname{diag}(1/\pi_i(t); i = 1, 2, \ldots, m-1), \quad \mathbf{v}_t = D_t^{-\frac{1}{2}} \mathbf{e}. \tag{6.8}$$

Then one form for $V_t^{-\frac{1}{2}}$ is

$$V_t^{-\frac{1}{2}} = n(t)^{\frac{1}{2}} D_t^{\frac{1}{2}} (I + \rho_t \mathbf{v}_t \mathbf{v}_t^T) \tag{6.9}$$

where

$$\rho_t = \frac{1}{\pi_m(t) + \pi_m(t)^{\frac{1}{2}}}.$$

The quantities defining the linear least squares form of the scoring step (4.19) can now be evaluated. This gives

$$\begin{aligned} A_t &= n(t)^{\frac{1}{2}} (\nabla_\beta \boldsymbol{\eta}_t^T D_t^{\frac{1}{2}} + \rho_t (\nabla_\beta \boldsymbol{\eta}_t^T \mathbf{e}) \mathbf{v}_t^T) \\ &= n(t)^{\frac{1}{2}} (\nabla_\beta \boldsymbol{\eta}_t^T D_t^{\frac{1}{2}} - \rho_t \nabla_\beta \pi_m(t)^T \mathbf{v}_t^T), \end{aligned} \tag{6.10}$$

and, noting that

$$(I + \rho_t \mathbf{v}_t \mathbf{v}_t^T)^{-1} (D_t^{\frac{1}{2}} \mathbf{y}(t) - (y_m(t)/\pi_m(t)) \mathbf{v}_t) = D_t^{\frac{1}{2}} \mathbf{y}(t) - \rho_t (y_m(t) + n(t) \pi_m(t)^{\frac{1}{2}}) \mathbf{v}_t,$$

$$\mathbf{b}_t = n(t)^{\frac{1}{2}} D_t^{-\frac{1}{2}} (D_t \mathbf{y}(t)/n(t) - \rho_t (y_m(t)/n(t) + \pi_m(t)^{\frac{1}{2}}) \mathbf{e}), \tag{6.11}$$

Table 6.1
Cattle virus data

$\log_{10}$ (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

where $\mathbf{y}(t)$ is the vector with components $(y_1(t), \ldots, y_{m-1}(t))$. In this case the components of $\mathbf{b}_t$ have a scale of $\sqrt{n(t)\pi_i(t)} = \sqrt{\mathscr{E}\{y_i(t)\}}$, while the untransformed quantities $\partial\mathscr{L}_t/\partial\eta_i$ have a corresponding scale of $n(t)$.

Remark 6.1. Here the sampling strategy must combine aspects of the two following possibilities, both of which can be relevant to the convergence analysis:

(1) Each $n(t)\to\infty$, $t \in T$ uniformly in the sense that $n(t_i)/n(t_j)$ is bounded for all $t_i, t_j \in T$, but $|T|$ is fixed (it needs to be sufficiently large for the problem to be identifiable for $\boldsymbol{\beta}$). Then

$$\varpi\{\mathscr{I}^{-1}(\mathscr{I} - \mathscr{J})\}\to 0 \text{ provided}$$
$$y_i(t)/n(t)\to \pi_i(t) > 0, \quad i = 1, 2, \ldots, m, \quad t \in T.$$

(2) Each $n(t)$ is limited in size by a condition of the form $n(t)/N \leqslant K/|T|$ but now $|T|\to\infty$.

An example of trinomial data ($m = 3$) is given in Table 6.1. It derives from a study of the effects of a cattle virus on chicken embryos. The model fitted to this data is

$$\begin{aligned} \pi_1 &= 1/(1+\exp(-\beta_1 - \beta_3 \log(t))), \\ 1-\pi_3 &= 1/(1+\exp(-\beta_2 - \beta_3 \log(t))), \\ \pi_2 &= 1 - \pi_1 - \pi_3. \end{aligned} \tag{6.12}$$

Starting values were provided with the data. Results obtained using the scoring algorithm implemented in the manner described in [16] are given in Table 6.2. It will be seen that the performance of the algorithm is very satisfactory. *It is stressed that the fast rate of convergence achieved provides persuasive evidence also that the modelling strategy is satisfactory*.

The direct application of scoring to multinomial data is discussed in Green [5]. The above presentation gives A and $\mathbf{b}$ explicitly in (4.19). The resulting algorithm appears to

Table 6.2
Results of computation with trinomial data

Iteration	$\mathscr{K}(\boldsymbol{\beta})$	$-\nabla\mathscr{K}\mathbf{h}$	β_1	β_2	β_3
0	49·31		−4·597	−3·145	0·7405
1	47·53	0·1353 + 2	−3·937	−2·369	0·7834
2	47·00	0·9778 + 0	−4·419	−2·584	0·8882
3	46·99	0·2145 − 1	−4·405	−2·620	0·9060
4	46·99	0·7370 − 5	−4·405	−2·619	0·9060
5	46·99	0·8861 − 8	−4·405	−2·619	0·9060

Table 6.3

Incidence of R.secalis on leaves of 10 varieties grown at 9 sites: percentage leaf area affected

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	0·05	0·00	0·00	0·10	0·25	0·05	0·50	1·30	1·50	1·50
(2)	0·00	0·05	0·05	0·30	0·75	0·30	3·00	7·50	1·00	12·70
(3)	1·25	1·25	2·50	16·60	2·50	2·50	0·00	20·00	37·50	26·25
(4)	2·50	0·50	0·01	3·00	2·50	0·01	25·00	55·00	5·00	40·00
(5)	5·50	1·00	6·00	1·10	2·50	8·00	16·50	29·50	20·00	43·50
(6)	1·00	5.00	5·00	5·00	5·00	5·00	10·00	5·00	50·00	75·00
(7)	5·00	0·10	5·00	5·00	50·00	10·00	50·00	25·00	50·00	75·00
(8)	5·00	10·00	5·00	5·00	25·00	75·00	50·00	75·00	75·00	75·00
(9)	17·50	25·00	42·50	50·00	37·50	95·00	62·50	95·00	95·00	95·00

be superior to the procedure described in [12] for use with the statistical package GLIM. A recent applications oriented account of the use of this package is given in [1].

Quasi-likelihood calculation. The quasi-likelihood calculation reported here follows the problem analysis presented in [19]. In Table 6.3 percentage of leaf blotch infection (R.secalis) is given for 10 varieties of barley grown on 9 sites. We let y_{ij} be the proportion of leaf blotch at site i for variety j. The quasi-likelihood calculation is defined by setting $\mu_{ij} = \mathscr{E}\{y_{ij}\}$, $i = 1, 2, \ldots, n_a$, $j = 1, 2, \ldots, n_b$ and positing the model

$$\operatorname{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = m + a_i + b_j \tag{6.13a}$$

$$\operatorname{var}(\mathbf{y}_{ij}) = \mu_{ij}^2(1-\mu_{ij})^2. \tag{6.13b}$$

To use the mean model (6.13a) it is necessary to add identifiability constraints as in (5.3). The reduction of the linear subproblem to (5.9), (5.10), and hence to (4.19) is straightforward in this case as the rows of the constraint matrix C are orthogonal. And a further simplification occurs because of the choice of model. Let

$$w_{ij} = m + a_i + b_j,$$

then

$$\mu_{ij} = \frac{e^{w_{ij}}}{1+e^{w_{ij}}}$$

Table 6.4

Performance of scoring: quasi-likelihood case

Iteration	DX	DF	NB	NR	XINC	RAT
1	2·93 + 0	−2·39 + 2	2	12	1·754	
2	2·13 + 0	−4·55 + 1	2	11	1·335	0·727
3	7·83 − 1	−1·37 + 1	2	11	1·369	0·367
4	5·36 − 1	−2·82 + 0	2	8	1·089	0·684
5	1·65 − 1	−3·36 − 1	1	6	0·971	0·308
6	6·83 − 2	−4·26 − 2	2	6	1·022	0·414
7	2·65 − 2	−7·32 − 3	1	5	0·940	0·388
8	1·10 − 2	−1·11 − 3	2	6	1·006	0·415
9	4·39 − 3	−1·98 − 4	1	4	0·932	0·399
10	1·83 − 3	−3·09 − 5	1	1	0·998	0·417
11	7·40 − 4	−5·60 − 6	1	2	0·931	0·404
12	3·12 − 4	−8·99 − 7	1	1	0·981	0·422
13	1·25 − 4	−1·60 − 7	1	1	0·943	0·401
14	5·32 − 5	−2·63 − 8	1	1	0·980	0·426

Table 6.5
Residuals by site and variety

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	0·000	−0·000	−0·000	0·000	0·001	−0·001	0·001	0·004	0·007	−0·002
(2)	−0·001	−0·000	−0·001	−0·001	0·002	−0·002	0·015	0·039	−0·022	0·061
(3)	−0·004	0·002	0·007	0·123	−0·037	−0·036	−0·152	−0·110	0·091	−0·194
(4)	0·012	−0·003	−0·014	−0·002	−0·022	−0·046	0·133	0·301	−0·176	0·017
(5)	0·033	−0·004	0·037	−0·043	−0·054	0·003	−0·021	−0·70	−0·137	−0·083
(6)	−0·016	0·033	0·028	−0·015	−0·044	−0·042	−0·118	−0·362	0·118	0·183
(7)	0·003	−0·029	−0·001	−0·064	0·339	−0·058	0·160	−0·314	−0·033	0·043
(8)	−0·047	0·037	−0·055	−0·169	−0·044	0·461	−0·028	0·012	0·037	−0·090
(9)	−0·123	0·040	0·110	−0·025	−2·247	0·334	−0·190	0·033	0·043	−0·004

and the ij'th row of $\tilde{A}$, where $A = \tilde{A}Q_2$, is given by

$$\tilde{\mathbf{a}}_{ij} = \frac{1}{\operatorname{var}(y_{ij})^{\frac{1}{2}}} \frac{d\mu_{ij}}{dw_{ij}} \nabla_\beta w_{ij}$$
$$= \nabla_\beta w_{ij}.$$

This corresponds to a balanced design so that $\mathbf{h}_\beta$ can be written down immediately. In the numerical results reported below, the secant algorithm was used to find an approximate zero of (4.6), and a standard linear least squares fit to the data used to set initial values. The behaviour of the iteration is detailed in Table 6.4 which gives, for each iteration, $DX = \|\mathbf{h}\|$, $DF = \mathscr{W}'(\boldsymbol{\beta}:\mathbf{h})$, the number of steps needed to bracket the minimum NB, the number of secant algorithm steps NR, and the computed step $XINC$. The convergence test for terminating the algorithm is based on the size of DX. The test on the linesearch is less extreme and takes account both of the size of DX and the convergence of $XINC$.

The secant algorithm is quite well behaved with the root being bracketed quickly, and the choice of a unit scale being clearly satisfactory. Equally clearly the calculation could be economized (taking $\lambda = 1$ in the line search step except possibly in the first iteration would almost certainly be satisfactory). *However, the overall rate of convergence is not very fast, and with the scoring algorithm this is a good indication of shortcomings in the modelling procedure.* In this case the power method gives a useful indication of the limiting value of ϖ. The successive ratios are given in the final column of Table 6.4. The residuals in the fit are displayed in Table 6.5.

Acknowledgements

I am indebted to Dorothy Anderson for helpful discussion on the application of scoring methods, and, in particular, for the data for the trinomial example. Also to Terry Speed for a careful reading of the paper and many constructive suggestions regarding presentation.

References

1. Atkin, M., Anderson, Dorothy, Francis, B. & Hinde, J. (1989). Statistical Modelling in GLIM, *Oxford Statistical Science Series,* **4,** Clarendon Press.
2. Ascher, U. & Osborne, M.R. (1988). A note on solving nonlinear equations and the natural criterion function, **J.O.T.A. 55,** 147–152.
3. Burke, J.J. & Moré, J.J. On the identification of active constraints, *Tech. Mem. no.* 82, Mathematics and Computer Science Division, Argonne National Laboratory.
4. Godambe, V.P. & Heyde, C.C. (1987). Quasilikelihood and optimal estimation, *Int. Stat. Rev.* **55,** 231–244.
5. Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, *J. R. Statist. Soc.* B **46,** 149–192.

6. Huber, P.J. The behaviour of maximum likelihood estimators under nonstandard conditions, in *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability,* **1,** University of California Press, Berkeley.
7. Jennrich, R.L. (1969). Asymptotic properties of nonlinear least squares estimators, *Ann. Math. Stat.* **40,** 633–643.
8. Jorgenson, B. (1987). Exponential dispersion models, *J. R. Statist. Soc.* B, **49,** 127–162.
9. Kass, R.E. & Smyth, G.K. (1989). The rate of convergence of the Fisher scoring: a geometric interpretation, *Technical Report,* Department of Mathematics (1989), University of Queensland.
10. Kent, J.T. (1982). Robust properties of likelihood ratio tests, *Biometrika* **69,** 19–27.
11. McCullagh, P. (1983). Quasi-likelihood functions, *Ann. Statist.* **11,** 59–67.
12. McCullagh, P. & Nelder, J.A. (1989). Generalised Linear Models, 2nd edition. Chapman & Hall.
13. McLeish, D.L. & Small, C.G. (1988). The Theory and Application of Statistical Inference Functions, *Lecture Notes in Statistics* **44,** Springer-Verlag.
14. Moran, P.A.P. (1971). The uniform consistency of maximum likelihood estimators, *Proc. Cam. Phil. Soc.* **70,** 435–439.
15. Moran, P.A.P. (1971). Maximum-likelihood estimation in non-standard conditions, *Proc. Cam. Phil. Soc.* **70,** 441–450.
16. Osborne, M.R. (1987). Estimating nonlinear models by maximum likelihood for the exponential family, *S.I.S.S.C.* **8,** 446–456.
17. Osborne, M.R. Notes on the method of scoring, *DMS report ACT* 87/20, C.S.I.R.O.
18. Pratt, J.W. (1981). Concavity of the log likelihood, *J. Amer. Statist. Assoc.* **76,** 103–106.
19. Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss–Newton method, *Biometrika* **61,** 439–447.
20. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48,** 817–838.
21. Zeger, S.L., Kung-Yee Liang & Albert, P.S. (1988). Models for longitudinal data: a generalised estimating equation approach, *Biometrics* **44,** 1049–1060.

Résumé

Nous analysons la méthode des cotes de Fisher pour maximiser les vraisemblances et pour résoudre les équations basées sur la quasivraisemblance. Nous montrons qu'un estimateur consistent du vecteur des paramètres est nécessaire pour obtenir une convergence rapide, mais si cette condition est satisfaite, cet algorithme est très efficace. Nous soulignons la connexion entre la performance de l'algorithme des cotes et la justesse de la modélisation du problème. Nous discutons l'effet de constraintes linéaires et nous donnons des exemples de calcul des vraisemblances et des quasi-vraisemblances.

[*Received October* 1990, *accepted September* 1991]