Scoring with Constraints

Michael R. Osborne

December 15, 1994

Abstract

This report considers the solution of estimation problems based on the maximum likelihood principle when a fixed number of equality constraints are imposed on the problem. Consistency and the asymptotic distribution of the parameter estimates as $n \to \infty$, where n is the number of observations, are discussed, and it is shown that a suitably scaled limiting multiplier vector is known. It is suggested that this information should permit the good properties of Fisher's method of scoring for the unconstrained case to be extended to a class of augmented Lagrangian methods for the constrained case.

1 Introduction

The method of maximum likelihood provides an important paradigm in many modelling situations which require a parametrized class of models to be fitted to observed data. Its general good behaviour, and the existence of an effective numerical algorithm in the method of scoring are now well understood [3]. Behavior of the method when the likelihood is subject to equality constraints has received much less attention, in part because in many applications knowledge of the constraints permits elimination of variables and simplifies the resulting optimisation problem. In the case of linear inequality constraints the situation is rather similar because methods are available which permit the active constraint set to be identified in a finite number of iterations [3].

Here the problem is studied in part because the influence of statistical linearity as the number of observations increases without bound on the performance of algorithms is of intrinsic interest. However, it is relevant also to the problem of developing methods for estimating systems of ordinary differential equations which have the property that they are independent of the need to use auxilliary information in order to establish a parametrised class of solutions to provide competition in the optimization problem. Methods which use such auxilliary information are discussed in the report [2]. The current report does not solve this problem in the case that the number of discretization points for integrating the differential system is allowed to tend to ∞ with the number of observations. However, the results do apply when a *fixed* number of discretization points are employed with the solution values being estimated by (say) linear interpolation between them. The main results obtained are as follows. It is shown that the Lagrange multipliers associated with the equality constraints are statistically orthogonal to the estimates of the natural problem parameters, that consistent estimates of suitably scaled Lagrange multipliers are available a priori, and that, as a consequence, augmented Lagrangian methods appear to offer attractive solution procedures which inherit the good properties of the scoring method in this case. These good properties are not inherited by the straightforward application of scoring as an obvious modification of Newton's method applied to the Lagrangian formulation of the problem. This approach does lead to an algorithm with good local convergence properties, but it lacks a suitable merit function to use in conjuction with a line search to stabilize the computation.

The basic data for the estimation problem consists of:

1. Observations $\mathbf{y}_i,\ i = 1, 2, \cdots, n$ indexed by an associated configuration descriptor $t_i,\ i = 1, 2, \cdots, n$ which could be time for example. Observations made at different times are assumed to be independent;
2. A density $f(t, \eta(t, \beta), \mathbf{y})$ giving the distribution of the observations $\mathbf{y}_i$. Here $\eta(t, \beta)$ can be considered a model for the process generating the sample values $\mathbf{y}_i$, and it is parametrised by $\beta \in R^p$;
3. It is assumed that there is a true parameter vector β^*. It will be necessary to distinguish between expectations computed for different values of β, and this will be done using the notation $\mathcal{E}_\beta$. If the true β is used then the expectation is written $\mathcal{E}_*$;
4. Constraints providing additional information about the parameter vector β. These may be either linear

$$C\beta = \mathbf{d}, \tag{1}$$

where $C \in R^p \to R^m$ is required to have full rank $m < p$, and $\mathbf{d} \in R^m$; or nonlinear

$$g_i(\beta) = 0, \quad i = 1, 2, \cdots, m \tag{2}$$

where the $\nabla_\beta g_i(\beta^*)$ are linearly independent; and

5. It is assumed that the limiting situation corresponding to increasing n without bound can be conceptualised. This requires a method for assigning the observation points t_i and this could be either random or deterministic. In either case it is assumed there exists a density $w(t)$ describing the limiting process in the sense that

$$\frac{1}{n}\sum_{i=1}^{n} q(t_i) \to \int_0^1 q(t)dw(t), \quad n \to \infty. \tag{3}$$

for all suitably smooth functions $q(t)$.

The method of maximum likelihood seeks to estimate β^* by maximizing the likelihood

$$\mathcal{L} = \prod_{i=1}^{n} f(t_i, \eta(t_i, \beta), \mathbf{y}_i).$$

Here this is to be maximised subject to the constraints (1), or (2). Let

$$L_i(t_i, \beta, \mathbf{y}_i) = \log f(t_i, \eta(t_i, \beta), \mathbf{y}_i), \tag{4}$$

then the problem is equivalent to minimizing

$$K_n(\beta) = -\sum_{i=1}^{n} L_i(t_i, \beta, \mathbf{y}_i) \tag{5}$$

subject to the same constraint set.

The plan of the report is as follows. In the next section the properties of the problem are developed in the case of linear constraints. The scoring algorithm is discussed in section 3., and the possibility of using an augmented Lagrangian formulation of the problem is explored. The final section discusses the extension to nonlinear constraints.

2 Scoring with linear constraints

The problem considered in this section is

$$\min_{\beta} K_n(\beta); \quad C\beta = \mathbf{d}. \tag{6}$$

The necessary conditions for a minimum give

$$\nabla_\beta K_n = \zeta^T C, \tag{7}$$

where ζ is the vector of Lagrange multipliers. Equation (6), together with the constraint equations, give $p+m$ equations for the $p+m$ unknowns β, ζ. The limiting form for this equation as $n \to \infty$ is obtained by considering

$$\frac{1}{n}\left\{\nabla_\beta K_n - \mathcal{E}_*\{\nabla_\beta K_n\}\right\} + \frac{1}{n}\mathcal{E}_*\{\nabla_\beta K_n\} = (\zeta/n)^T C.$$

The first term on the left hand side tends to zero as $n \to \infty$ by the law of large numbers. Making use of (3) the second has the limiting form

$$-\int_0^1 \mathcal{E}_*\left\{\nabla_\beta L(t,\beta,\mathbf{y})\right\} dw(t).$$

Thus the limiting system is

$$\begin{aligned} -\int_0^1 \mathcal{E}_*\left\{\nabla_\beta L(t,\beta,\mathbf{y})\right\} dw(t) &= \zeta^{*T} C \\ C\beta &= \mathbf{d}, \end{aligned} \tag{8}$$

where $\zeta^* = \lim_{n\to\infty} \zeta/n$. This equation has the solution

$$\beta = \beta^*, \ \zeta^* = 0$$

as a consequence of the standard identity $\mathcal{E}_\beta\{\nabla_\beta L(t,\beta,\mathbf{y})\} = 0$. It is an isolated solution provided the Jacobian of the system is nonsingular. This requires that the augmented matrix

$$A(\beta^*) = \begin{bmatrix} \mathcal{I} & C^T \\ C & 0 \end{bmatrix}$$

be nonsingular. A form of second order sufficiencey will serve. Let C have the orthogonal factorization

$$C^T = \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \begin{bmatrix} U \\ 0 \end{bmatrix} \tag{9}$$

then $A(\beta^*)$ is nonsingular provided:

1. U is nonsingular - a consequence of the linear independence assumption, and

2. $Q_2^T \mathcal{I} Q_2$ is nonsingular.

If $\hat{\beta}_n$ minimizes (6) then the argument in [3] serves to prove consistency almost surely as $n \to \infty$. Here it is convenient to apply Newton's method with $\beta = \beta^*$ as initial guess to the system comprising

$$\nabla_\beta K_n Q_2 = 0 \tag{10}$$

and the constraint equations (1). Equation (10) is obtained from(8) using the factorization (9) to eliminate the Lagrange multipliers.

Scoring can be defined for either system by applying Newton's method with the variation that ${\nabla_\beta}^2 K_n$ is replaced by its formal expectation

$$\mathcal{I}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}_\beta \left\{ \nabla_\beta L_i^T \nabla_\beta L_i \right\}. \tag{11}$$

For example, the system corresponding to (10) is

$$\begin{aligned} Q_2^T \mathcal{I}_n \delta_\beta &= -\frac{1}{n} Q_2^T \nabla_\beta K_n^T, \\ C\delta_\beta &= -\left(C\beta - \mathbf{d}\right). \end{aligned}$$

The left hand side matrix is non singular if and only if the corresponding augmented matrix is nonsingular. The rate of convergence for the resulting iteration is most readily analysed by considering it as the fixed point iteration

$$\beta \leftarrow \beta - \begin{bmatrix} Q_2^T \mathcal{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{n} Q_2^T \nabla_\beta K_n^T \\ C\beta - \mathbf{d} \end{bmatrix} \tag{12}$$

Let $\varpi\{A\}$ denote the spectral radius of the matrix A, that is the magnitude of the largest eigenvalue in modulus. Then the condition for an attractive fixed point at $\beta = \hat{\beta}_n$ is

$$\varpi \left\{ \left[I - \begin{bmatrix} Q_2^T \mathcal{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{n} Q_2^T {\nabla_\beta}^2 K_n^T \\ C \end{bmatrix} \right] \right\} =$$

$$\varpi \left\{ \begin{bmatrix} Q_2^T \mathcal{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} Q_2^T \left(\frac{1}{n} \sum_{i=1}^n {\nabla_\beta}^2 L_i + \mathcal{I}_n(\hat{\beta}_n) \right) \\ 0 \end{bmatrix} \right\} < 1.$$

This condition is satisfied for large enough n almost surely. In fact the much stronger result $\varpi \to 0$ holds. This follows by consistency (ϖ is not changed much by evaluating it at β^*) followed by an application of the law of large numbers. This shows that scoring is locally an effective algorithm for maximizing a likelihood subject to linear equality constraints, and that the rate of convergence asymptotically approaches second order as $n \to \infty$. However, an important feature of scoring for unconstrained problems is that K_n provides a natural merit function which can be used in a linesearch strategy to stabilize the computation and improve its global properties. This follows because $\nabla_\beta K_n \delta < 0$ whenever $\mathcal{I}_n$ is nonsingular. It is not clear that a similar effective merit function exists in the constrained case. The emphasis here is on 'effective'. The situation corresponds to that for Newton's method for solving $\mathbf{f} = 0$. Here $\sum f_i^2$ provides a merit function which is always available, but which suffers from being poorly scaled, and this often translates into poor performance in practice [1]. In our case different scales for the likelihood function and the constraints add further complications.

Distributional properties for $\hat{\beta}_n$ can be derived using (10). Expanding gives

$$\begin{aligned}
0 &= Q_2^T \left\{ \frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T + \frac{1}{n} \nabla_\beta{}^2 K_n(\beta^*) \sqrt{n}(\hat{\beta}_n - \beta^*) + o_P(1) \right\}, \\
0 &= C\sqrt{n}(\hat{\beta}_n - \beta^*).
\end{aligned}$$

An application of the central limit theorem shows that asymptotically

$$\frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T \sim N(0, \mathcal{I}). \tag{13}$$

Combining gives the asymptotic result

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta^*) &= -Q_2(Q_2^T \mathcal{I} Q_2)^{-1} Q_2^T \frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T \\
&\sim N(0, Q_2(Q_2^T \mathcal{I} Q_2)^{-1} Q_2^T). \qquad (14)
\end{aligned}$$

The distribution of the multiplier vector ζ can be found from (7). Expanding this in similar fashion gives

$$\begin{aligned}
\hat{\zeta}/\sqrt{n} &= (CC^T)^{-1} C^T \frac{1}{\sqrt{n}} \nabla_\beta K_n(\hat{\beta}_n)^T, \\
&= (CC^T)^{-1} C^T [\frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T \\
&\quad + \frac{1}{n} \nabla_\beta{}^2 K_n(\beta^*) \sqrt{n}(\hat{\beta}_n - \beta^*) + o_P(1)]. \qquad (15)
\end{aligned}$$

It follows that $\hat{\zeta}/\sqrt{n}$ is asymptotically normally distributed with bounded variance. A consequence is that $\hat{\zeta}/n$ is a $1/\sqrt{n}$ consistent estimate of $\zeta^* = 0$. The following calculation is of interest (terms small in probability are ignored without further comment).

$$\begin{aligned}
\mathcal{E}_*\{\hat{\zeta}_n(\hat{\beta}_n - \beta^*)^T\} &= (CC^T)^{-1}C^T\mathcal{E}_*\{\frac{1}{\sqrt{n}}\nabla_\beta K_n(\hat{\beta}_n)^T\sqrt{n}(\hat{\beta}_n - \beta^*)^T\}, \\
&= (CC^T)^{-1}C^T[\mathcal{E}_*\{-\frac{1}{n}\nabla_\beta K_n^T\nabla_\beta K_n Q_2(Q_2^T\mathcal{I}Q_2)^{-1}Q_2^T\} \\
&\quad +\mathcal{I}Q_2(Q_2^T\mathcal{I}Q_2)^{-1}Q_2^T,] \\
&= 0.
\end{aligned}$$

This shows that $\hat{\zeta}_n/\sqrt{n}$ and $\sqrt{n}(\hat{\beta}_n - \beta^*)$ are orthogonal parameter vectors.

3 Computation of $\hat{\beta}_n$

Direct application of scoring to solve the necessary conditions (7) does not appear too attractive because of the need to find a suitable monitor function to improve the global convergence properties of the method. Thus it appears to have similar utility to that of Newton's method in this application, but it does require one less order of differentiation. However, the a priori knowledge that $\zeta^* = 0$ may make attractive the Powell–Hestenes method (augmented Lagrangian method). The idea here is to proceed iteratively by two steps at each stage. In the first step, given $\sigma^{(i)}$, $\theta^{(i)}$, an estimate $\beta^{(i)}$ of $\hat{\beta}_n$ is computed by minimizing

$$H_n(\beta, \theta^{(i)}, \sigma^{(i)}) = \frac{1}{n}K_n + \sigma^{(i)}\sum_{j=1}^{m}(\mathbf{c}_j^T\beta - d_j + \theta_j^{(i)})^2. \tag{16}$$

Here $\theta^{(i)}$ is an auxilliary vector of parameters and σ is a penalty parameter which governs the rate of convergence of the two step method (which is geometric with ratio $1/\sigma$) and must be chosen large enough. In the second step either σ is increased to increase the rate of convergence or, more usually, θ is adjusted to make β satisfy the constraint equations.The appeal of the method is in the simplicity of the formula for updating θ when σ is large:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} + C\beta^{(i)} - \mathbf{d}. \tag{17}$$

A derivation of this formula is given in Appendix 1 for completeness. The necessary conditions for a minimum of (16) with respect to β give

$$\frac{1}{n}\nabla_\beta K_n + 2\sigma^{(i)}\sum_{j=1}^{m}(\mathbf{c}_j^T\beta - d_j + \theta^{(i)})\mathbf{c}_j^T = 0.$$

Thus it follows that $2\sigma\theta_j$ estimates $\zeta_j^*,\ j = 1, 2, \cdots, m$.
Newton's method applied to minimize (16) gives a correction δ_H satisfying

$$\left\{\frac{1}{n}\nabla_\beta{}^2 K_n + 2\sigma\sum_{j=1}^{m}\mathbf{c}_j\mathbf{c}_j^T\right\}\delta_H = -\frac{1}{n}\nabla_\beta K_n - 2\sigma\sum_{j=1}^{m}(\mathbf{c}_j^T\beta - d_j + \theta_j)\mathbf{c}_j. \quad (18)$$

One step consistency is an immediate consequence of the argument used in [3]. It is only necessary to start the Newton iteration with $\beta = \beta^*$ and $\sigma\theta^{(1)} = 0$, and to note that, as the constraints are satisfied initially, the right hand side of 18 is $\mathcal{O}(1/\sqrt{n})$. Now, because the initial estimates are $\mathcal{O}(1/\sqrt{n})$ accurate, and because the estimates are improved by a factor $\mathcal{O}(1/\sigma)$ at each complete iteration, it follows that $C\beta^{(1)} - \mathbf{d}$ must be $\mathcal{O}(1/(\sigma\sqrt{n}))$ if (17) is to give the required improved estimate for $\theta^{(2)}$. Thus choosing $\sigma = \mathcal{O}(\sqrt{n})$ ensures a sequence of $1/\sqrt{n}$ consistent estimators for ζ^*. In this context scoring appears an attractive method for minimizing (16) with respect to β given θ. The idea is to replace $\frac{1}{n}\nabla_\beta{}^2 K_n$ in (18) by $\mathcal{I}_n$ using (11). This gives a correction δ_S satisfying

$$\left\{\mathcal{I}_n + 2\sigma\sum_{j=1}^{m}\mathbf{c}_j\mathbf{c}_j^T\right\}\delta_S = -\nabla_\beta H_n^T. \quad (19)$$

The rate of convergence of the corresponding fixed point iteration can be calculated in the same way as before. We have

$$\beta \leftarrow \beta - \left\{\mathcal{I}_n + 2\sigma\sum_{j=1}^{m}\mathbf{c}_j\mathbf{c}_j^T\right\}^{-1}\nabla_\beta H_n^T,$$

so the condition for β_{PH}, the minimizer of (16), to be a fixed point becomes

$$\varpi\left\{(\mathcal{I}_n + 2\sigma\sum_{j=1}^{m}\mathbf{c}_j\mathbf{c}_j^T)^{-1}(\frac{1}{n}\nabla_\beta{}^2 K_n - \mathcal{I}_n)\right\} < 1. \quad (20)$$

Here the contribution from the constraints cancels in the numerator, so the left hand side $\to 0$, $n \to \infty$ provided consistency and the law of large numbers can be used as before. Thus the existence of a consistent estimate of ζ^* to ensure that $\beta^{(1)}$ is a consistent estimator of β^* is critical to the argument. This shows that scoring can be expected to have a good rate of convergence for n large enough. But another attractive feature of scoring in the unconstrained case is also available because here the Powell–Hestenes objective function can be used directly to monitor progress when a linesearch is used to improve global convergence characteristics. This follows because

$$\nabla_\beta H_n \delta_S = -\nabla_\beta H_n \delta_S (\mathcal{I}_n + 2\sigma C^T C)^{-1} \nabla_\beta H_n < 0 \tag{21}$$

provided the generic condition $\mathcal{I}_n$ positive definite is satisfied. There are some negative aspects, however. The need to choose $\sigma = \mathcal{O}(\sqrt{n})$ means that the condition number of $(\mathcal{I}_n + 2\sigma C^T C)$ is $\mathcal{O}(\sqrt{n})$, and reflects some imbalance in the scaling of the objective function H_n.

4 Extensions to nonlinear constraints

If the constraint equations are nonlinear then details of the results in the linear case go over largely unchanged, but the arguments become somewhat more complicated. The estimation problem is written

$$\min_\beta \frac{1}{n} K_n(\beta); \quad \mathbf{g}(\beta) = 0. \tag{22}$$

The necessary conditions become

$$\begin{aligned} \frac{1}{n}\nabla_\beta K_n(\beta) &= \zeta^T \nabla_\beta \mathbf{g}(\beta), \\ \mathbf{g}(\beta) &= 0. \end{aligned} \tag{23}$$

Let

$$P(\beta) = \nabla_\beta \mathbf{g}(\beta)^T (\nabla_\beta \mathbf{g} \nabla_\beta \mathbf{g}^T)^{-1} \nabla_\beta \mathbf{g}(\beta), \tag{24}$$

and define V^* by

$$\nabla_\beta \mathbf{g}(\beta^*) V^* = 0, \quad V^{*T} V^* = I. \tag{25}$$

Then $\forall \beta \in B(\rho, \beta^*)$, with ρ small enough that $\nabla_\beta \mathbf{g}$ has its full rank in B,

$$V(\beta) = (I - P(\beta)) V^* \to V^*, \quad \beta \to \beta^*. \tag{26}$$

and $V(\beta)$ inherits the smoothness of $\nabla_\beta \mathbf{g}$, $\beta \in B$. This permits a reduced system analogous to (10) to be defined in the nonlinear case. This is

$$\begin{aligned} \frac{1}{n}\nabla_\beta K_n V(\beta) &= 0, \\ \mathbf{g}(\beta) &= 0. \end{aligned} \tag{27}$$

It reduces to (10) in the linear case if the choice $V^* = Q_2$ is made. It has $\hat{\beta}_n$ as an isolated solution provided the Jacobian of the system is nonsingular. Compared to the linear case there is an extra term to be considered which comes from differentiating $V(\beta)$. However, the law of large numbers can be used to show this term is small so the result holds almost surely provided n is large enough. The arguments used to show consistency and to derive limiting distributions also follow through in a similar manner. As before the limiting equations have the solution $\beta = \beta^*$, $\zeta^* = 0$.

To develop the extension of the Powell–Hestenes algorithm the following definitions are appropriate:

$$H_n(\beta, \theta^{(i)}, \sigma^{(i)}) = \frac{1}{n}K_n + \sigma^{(i)}\sum_{j=1}^{m}(g_j(\beta) + \theta_j^{(i)})^2, \tag{28}$$

$$\nabla_\beta H_n = \frac{1}{n}\nabla_\beta K_n + 2\sigma^{(i)}\sum_{j=1}^{m}(g_j(\beta) + \theta_j^{(i)})\nabla_\beta g_j, \tag{29}$$

$$\begin{aligned} \nabla_\beta^2 H_n = {} & \frac{1}{n}\nabla_\beta^2 K_n \\ & +2\sigma^{(i)}\sum_{j=1}^{m}(\nabla_\beta g_j^T \nabla_\beta g_j + (g_j(\beta) + \theta_j^{(i)})\nabla_\beta^2 g_j). \end{aligned} \tag{30}$$

In developing the scoring algorithm it is desirable to avoid calculating second derivatives also in the constraint terms. To see that this is possible consider the scoring correction δ_S given by

$$\{\mathcal{I}_n + 2\sigma\sum_{j=1}^{m}\nabla_\beta g_j^T \nabla_\beta g_j\}\delta_S = -\nabla_\beta H_n^T. \tag{31}$$

The associated fixed point iteration is

$$\beta \leftarrow \beta - \{\mathcal{I}_n + 2\sigma\sum_{j=1}^{m}\nabla_\beta g_j^T \nabla_\beta g_j\}^{-1}\nabla_\beta H_n.$$

The condition for $\beta^{(i)}$ to be an attractive fixed point of the i'th step of the Powell–Hestenes iteration requires the variational matrix associated with the iteration to have spectral radius < 1. The variational matrix is

$$\begin{aligned}
J^{(i)} &= I - \{\mathcal{I}_n + 2\sigma \sum_{j=1}^{m} \nabla_\beta g_j^T \nabla_\beta g_j\}^{-1} \nabla_\beta^2 H_n, \\
&= -\{\mathcal{I}_n + 2\sigma \sum_{j=1}^{m} \nabla_\beta g_j^T \nabla_\beta g_j\}^{-1} \{\nabla_\beta^2 H_n - (\mathcal{I}_n + 2\sigma \sum_{j=1}^{m} \nabla_\beta g_j^T \nabla_\beta g_j)\}, \\
&= -\{\mathcal{I}_n + 2\sigma \sum_{j=1}^{m} \nabla_\beta g_j^T \nabla_\beta g_j\}^{-1} \\
&\quad \{\frac{1}{n} \nabla_\beta^2 K_n - \mathcal{I}_n - 2\sigma \sum_{j=1}^{m} (g_j + \theta_j) \nabla_\beta^2 g_j\}.
\end{aligned}$$

The term $\frac{1}{n}\nabla_\beta^2 K_n - \mathcal{I}_n$ gets small by the usual argument involving the law of large numbers and consistency of the estimates. The term involving the constraint second derivatives is of similar order provided $\sigma^{(i)}\theta^{(i)}$ is a $1/\sqrt{n}$ consistent estimator of $\zeta_j^* = 0$. Note that $\sigma g_j(\beta^{(i)})$ must be of at most the same order if (17) is to provide an improved estimate for $\theta^{(i+1)}$. Thus the consistency of the estimates at each stage of the iteration is crucial to the argument. Given this, then

$$\varpi(J^{(i)}) \to 0, \quad n \to \infty. \tag{32}$$

The significant conclusions are:

1. The Powell–Hestenes algorithm generates $1/\sqrt{n}$ consistent estimates of the solution variables at each combined step of the iteration, $i = 1, 2, \cdots$.

2. The convergence of the algorithm has a characteristic $\mathcal{O}(1/\sqrt{n})$ rate provided σ is chosen appropriately. Thus there is little point in proceeding beyond the first few steps of the process.

3. The convergence of the iterates in each of the scoring steps has the characteristic speed provided n is large enough.

5 References

References

[1] U.Ascher and M.R.Osborne. *A note on solving nonlinear equations and the natural criterion function*, J.O.T.A, 55, p.147–152, 1988.

[2] S.B.Childs and M.R.Osborne. *Fitting solutions of ordinary differential equations to observed data*, Technical Report CS94-071,TAMU, 1994.

[3] M.R.Osborne. *Fisher's method of scoring*, Int. Stat. Rev., 60, p.99–117, 1992.

Appendix 1: The Powell–Hestenes method

The necessary conditions for a minimum of H_n give

$$\frac{1}{n}\nabla_\beta K_n + 2\sigma \sum_{i=1}^{n} (\mathbf{c}_i^T \beta - d_i + \theta_i)\mathbf{c}_i^T = 0,$$

and these determine β as a function of θ. The aim is to adjust θ so that

$$C\beta(\theta) - \mathbf{d} = 0.$$

If a Newton iteration is used to solve this equation then a correction to the current θ is given by

$$C\frac{\partial \beta}{\partial \theta}\delta_\theta = -(C(\beta(\theta) - \mathbf{d}).$$

To calculate $\frac{\partial \beta}{\partial \theta}$ differentiate the necessary conditions to obtain

$$\left\{\frac{1}{n}\nabla_\beta^2 K_n + 2\sigma C^T C\right\}\frac{\partial \beta}{\partial \theta} = -2\sigma C^T.$$

Transforming this using the factorization (9) gives

$$Q^T\left\{\frac{1}{n}\nabla_\beta^2 K_n + 2\sigma C^T C\right\}QQ^T\frac{\partial \beta}{\partial \theta} = -2\sigma\begin{bmatrix} U \\ 0 \end{bmatrix}.$$

The inverse of $Q^T \nabla^2_\beta H_n Q$ when σ is large is given by

$$\left[\begin{array}{cc} \frac{1}{2\sigma} U^{-T} U^{-1} + \mathcal{O}(\frac{1}{\sigma^2}) & -\frac{1}{2\sigma} U^{-T} U^{-1} + \mathcal{O}(\frac{1}{\sigma^2}) \\ -\frac{1}{2\sigma} U^{-T} U^{-1} + \mathcal{O}(\frac{1}{\sigma^2}) & \left(\frac{1}{n} Q_2^T \nabla^2_\beta K_n Q_2 \right)^{-1} + \mathcal{O}(\frac{1}{\sigma}) \end{array} \right]$$

Thus

$$Q_1^T \frac{\partial \beta}{\partial \theta} = -U^{-T} + \mathcal{O}(\frac{1}{\sigma}),$$

so that

$$C \frac{\partial \beta}{\partial \theta} = -I + \mathcal{O}(\frac{1}{\sigma}).$$

Substituting in the Newton step gives

$$\delta_\theta = C\beta - \mathbf{d} + \mathcal{O}(\frac{1}{\sigma}).$$