ANZIAM J. 42(2000), 9–25

SCORING WITH CONSTRAINTS

MICHAEL R. OSBORNE[1]

(Received 16 March 1998)

Abstract

This paper considers the solution of estimation problems based on the maximum likelihood principle when a fixed number of equality constraints are imposed on the parameters of the problem. Consistency and the asymptotic distribution of the parameter estimates are discussed as $n \to \infty$, where n is the number of independent observations, and it is shown that a suitably scaled limiting multiplier vector is known. It is also shown that when this information is available then the good properties of Fisher's method of scoring for the unconstrained case extend to a class of augmented Lagrangian methods for the constrained case. This point is illustrated by means of an example involving the estimation of a mixture density.

1. Introduction

The method of maximum likelihood provides an important paradigm in many modelling situations which require a parameterised 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 [6]. Simple constraints on the likelihood are familiar and include examples such as:

(1) Constraints defining discrete pdf's:

$$\pi_i \geq 0, \quad i = 1, 2, \dots, m; \quad \sum_1^m \pi_i = 1;$$

(2) Constraints imposed to ensure identifiability:

$$g(\mu_{ij}) = \mu + a_i + b_j, \quad i = 1, 2, \dots, n_a, \quad j = 1, 2, \dots, n_b,$$

[1] Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200 Australia.

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

$\Big($typical examples of such adjoined constraints include

$$\sum_1^{n_a} a_i = 0, \quad \sum_1^{n_b} b_j = 0\Big);$$

(3) Constraints implied by additional information such as constraints of physical origin (for example bound constraints)

$$x_i \geq 0, \quad i = 1, 2, \dots, p; \quad \text{and}$$

(4) Constraints imposed by exploratory techniques in which components of the parameter vector are controlled or set in defining a sequence of test models.

Typically these can be taken into account by simple devices such as the elimination of variables and the scoring algorithm can then be applied with little change. However, there are cases where it can be convenient to treat the constraints explicitly. Examples include:

(1) Likelihoods based on the Kent density [4]:

$$f(\mathbf{y}_t \mid \mathbf{a}, A) = \frac{1}{C(\mathbf{a}, A)} \exp\left(-\mathbf{a}^T \mathbf{y}_t - \mathbf{y}_t^T A \mathbf{y}_t\right),$$

where the parameters $\mathbf{a}$, A must satisfy the constraints

$$\mathbf{a}^T A = 0,$$
$$\operatorname{trace}(A) = 0;$$

(2) Estimating $p \times p$ systems of ordinary differential equations from observations made on a solution trajectory in the presence of noise. Here p additional pieces of information are needed to parameterise the set of possible solutions. This can be done by adjoining suitable auxiliary conditions [3], but also by imposing the suitably discretized differential equation [1] as a system of constraints. This latter approach has been explored further in [7]. It would appear to have significant advantages; and

(3) Testing complex hypotheses in contingency tables [10]. Here the y_i/n give consistent estimates of the frequencies π_i. This information can be used to obtain consistent estimates of the multipliers.

This paper is concerned with the application of scoring when there are explicit constraints on the likelihood. The scope of the approach used extends to problems where an *a priori* consistent estimate of the Lagrange multipliers associated with the constraints is known. However, consideration here is restricted to the case where the consistent estimate is provided by zero. In other cases it is necessary to modify the algorithm to take explicit account of constraint second derivatives. The basic data for the class of estimation problem considered consists of:

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

(1) Observations $\mathbf{y}_i$, $i = 1, 2, \ldots, n$, indexed by an associated configuration descriptor t_i, $i = 1, 2, \ldots, n$, 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 parameterised 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 $\mathscr{E}_\beta$. If the true β is used then the expectation is written $\mathscr{E}_*$;

(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,

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

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

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

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

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

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

subject to the same constraint set.

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. In particular, the augmented Lagrangian can be used as a merit function in the line search in order to stabilize the computation. These good properties are not inherited by the straightforward application of scoring as an obvious modification of Newton's method applied to the usual 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 the line search. Thus scoring here loses one of the most attractive features it possesses in the unconstrained case, and, in this respect, will be seen to compare unfavourably with the approach based on the augmented Lagrangian. This approach is illustrated by applying the sequential Powell-Hestenes technique to estimate the components of a mixture in the case that the proportions are prescribed functions of the individual population means. The Powell-Hestenes technique involves a sequence of minimizations of the objective function with multiplier estimates being updated after each minimization [5]. Scoring is applied to minimize the augmented Lagrangian in each of the sequence of steps, and its characteristic fast rate of convergence provided n is large enough is demonstrated.

The plan of the paper is as follows. In the next section the properties of the problem are developed in the case of linear constraints and the naive use of a scoring algorithm discussed. Then the possibility of using an augmented Lagrangian formulation of the problem is explored first for linear constraints and then for nonlinear constraints where additional problems resulting from constraint curvature are considered. The final section describes the application to estimating a mixture of densities. A derivation of the main formulae of the Powell-Hestenes method is summarised in an appendix.

2. Scoring with linear constraints

The problem considered in this section is

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

It is considered first because the vanishing of the constraint curvature makes this somewhat simpler to treat than the general case. In practice it may well be that the linear constraints are used to solve for a subset of the variables. The resulting unconstrained system can then be minimised by scoring in the usual way. Here the

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

necessary conditions for a minimum give

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

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

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

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

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

Thus the limiting system is

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

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

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

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

$$\operatorname{Aug}(\beta^*) = \begin{bmatrix} \mathscr{I} & -C^T \\ -C & 0 \end{bmatrix}$$

be nonsingular where the information matrix $\mathscr{I}$ is given by

$$\mathscr{I} = -\int_0^1 \mathscr{E}_*\left\{\nabla_\beta{}^2 L(t, \beta, \mathbf{y})\right\} dw(t) = \int_0^1 \mathscr{E}_*\left\{\nabla_\beta L^T \nabla_\beta L\right\} dw(t). \tag{9}$$

A form of second-order sufficiency will serve. Let C have the orthogonal factorization

$$C^T = [Q_1 \mid Q_2]\begin{bmatrix} U \\ 0 \end{bmatrix}, \tag{10}$$

then $\operatorname{Aug}(\beta^*)$ is nonsingular provided:

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

(1) U is nonsingular—a consequence of the linear independence assumption; and
(2) $Q_2^T \mathscr{I} Q_2$ is nonsingular.

If $\hat{\beta}_n$ minimizes (6) then the argument in [6] 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

$$\nabla_\beta K_n Q_2 = 0 \tag{11}$$

and the constraint equations (1). Equation (11) is obtained from (8) using the factorization (10) 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

$$\mathscr{I}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{E}_\beta \left\{ \nabla_\beta L_i^T \nabla_\beta L_i \right\}. \tag{12}$$

For example, the system corresponding to (11) is

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

The left-hand side matrix is nonsingular 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

$$\beta \leftarrow \beta - \begin{bmatrix} Q_2^T \mathscr{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{n} Q_2^T \nabla_\beta K_n^T \\ C\beta - \mathbf{d} \end{bmatrix}. \tag{13}$$

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

$$\begin{aligned} &\varpi \left\{ \left[I - \begin{bmatrix} Q_2^T \mathscr{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{n} Q_2^T {\nabla_\beta}^2 K_n^T \\ C \end{bmatrix} \right] \right\} \\ &\quad = \varpi \left\{ \begin{bmatrix} Q_2^T \mathscr{I}_n \\ C \end{bmatrix}^{-1} \begin{bmatrix} Q_2^T \left(\frac{1}{n} \sum_{i=1}^{n} {\nabla_\beta}^2 L_i + \mathscr{I}_n(\hat{\beta}_n) \right) \\ 0 \end{bmatrix} \right\} < 1 \end{aligned}$$

as (11) holds at $\hat{\beta}_n$. 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

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

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 $\mathscr{I}_n$ is nonsingular and $\beta \neq \hat{\beta}_n$. 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 [2]. In our case, different scales for the likelihood function and the constraints have the potential to add further complications.

Distributional properties for $\hat{\beta}_n$ can be derived using (11). The argument is sketched below. For additional details see [9] (for example). Expanding (11) 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

$$\frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T \sim N(0, \mathscr{I}). \tag{14}$$

Using (3), (9) and noting that $\hat{\beta}_n - \beta^*$ is in the span of the columns of Q_2, gives the asymptotic result

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \beta^*) &= -Q_2(Q_2^T \mathscr{I} Q_2)^{-1} Q_2^T \frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T \\ &\sim N(0,\, Q_2(Q_2^T \mathscr{I} Q_2)^{-1} Q_2^T). \end{aligned} \tag{15}$$

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

$$\begin{aligned} \frac{\hat{\zeta}}{\sqrt{n}} &= (CC^T)^{-1} C \frac{1}{\sqrt{n}} \nabla_\beta K_n(\hat{\beta}_n)^T \qquad (16) \\ &= (CC^T)^{-1} C \left[\frac{1}{\sqrt{n}} \nabla_\beta K_n(\beta^*)^T + \frac{1}{n} \nabla_\beta{}^2 K_n(\beta^*) \sqrt{n}(\hat{\beta}_n - \beta^*) + o_P(1) \right]. \end{aligned}$$

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 $\zeta^* = 0$.

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

REMARK 2.1. The following calculation is of interest (terms small in probability are ignored without further comment):

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

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 (also known as the augmented Lagrangian method) [5]. The idea here is to proceed iteratively by two steps at each stage. In the first step, given $\omega^{(i)}$ and $\theta^{(i)}$, an estimate $\beta^{(i)}$ of $\hat{\beta}_n$ is computed by minimizing

$$H_n(\beta,\theta^{(i)},\omega^{(i)}) = \frac{1}{n}K_n+\omega^{(i)}\sum_{j=1}^{m}\left(\mathbf{c}_j^T\beta-d_j+\theta_j^{(i)}\right)^2. \tag{17}$$

Here $\theta^{(i)}$ is an auxiliary vector of parameters and ω is a penalty parameter which governs the rate of convergence of the two-step method (which is geometric with ratio $1/\omega$) and must be chosen large enough. In the second step, either ω 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 ω is large:

$$\theta^{(i+1)}\leftarrow\theta^{(i)}+C\beta^{(i)}-\mathbf{d}. \tag{18}$$

A derivation of this formula is given in Appendix 1 for completeness. The necessary conditions for a minimization of (17) with respect to β give

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

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

This must be equivalent to the Lagrange multiplier equations (7) when $\beta = \hat{\beta}_n$ because the constraints are satisfied at $\hat{\beta}_n$. It follows that $2\omega\theta_j$ estimates $\zeta_j^*, j = 1, 2, \ldots, m$. The importance of a good estimate follows from the result sketched in Appendix 1 that if $2\omega\theta_j^{(i)} - (\hat{\zeta}_n)_j = \epsilon_j$ then

$$\left\| \beta^{(i)} - \hat{\beta}_n \right\| = \mathcal{O}(\|\epsilon\|/\omega).$$

Newton's method applied to minimize (17) gives a correction δ_H satisfying

$$\left\{ \frac{1}{n} \nabla_\beta{}^2 K_n + 2\omega \sum_{j=1}^{m} \mathbf{c}_j \mathbf{c}_j^T \right\} \delta_H = -\frac{1}{n} \nabla_\beta K_n - 2\omega \sum_{j=1}^{m} (\mathbf{c}_j^T \beta - d_j + \theta_j) \mathbf{c}_j . \tag{19}$$

One-step consistency is an immediate consequence of the argument used in [6]. It is only necessary to start the Newton iteration with $\beta = \beta^*$ and $\omega\theta^{(1)} = 0$, and to note that, as the constraints are satisfied initially, the right-hand side of (19) is $\mathcal{O}(1/\sqrt{n})$ in probability as a consequence of (15). Now, because the initial estimates are $\mathcal{O}(1/\sqrt{n})$ accurate, and because the estimates are improved by a factor $\mathcal{O}(1/\omega)$ at each complete iteration provided the θ update is acceptable, it follows that $C\beta^{(1)} - \mathbf{d}$ must be $\mathcal{O}(1/(\omega\sqrt{n}))$ if (18) is to give the required improved estimate for $\theta^{(2)}$. Choosing $\omega = \mathcal{O}(\sqrt{n})$ appears a convenient choice which should ensure both rapid convergence of the $\theta^{(j)}$ and $1/\sqrt{n}$ consistency of the estimator of ζ^*. In this context, scoring appears an attractive method for minimizing (17) with respect to β given θ. The idea is to use (12) to replace $\frac{1}{n}\nabla_\beta{}^2 K_n$ in (19) by $\mathcal{I}_n$. This gives a correction δ_S satisfying

$$\left\{ \mathcal{I}_n + 2\omega \sum_{j=1}^{m} \mathbf{c}_j \mathbf{c}_j^T \right\} \delta_S = -\nabla_\beta H_n^T . \tag{20}$$

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

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

at the minimum $\nabla_\beta H_n^T = 0$, so the condition for β_{PH}, the minimizer of (17), to be a fixed point becomes

$$\varpi \left\{ \left(\mathcal{I}_n + 2\omega \sum_{j=1}^{m} \mathbf{c}_j \mathbf{c}_j^T \right)^{-1} \left(\frac{1}{n} \nabla_\beta{}^2 K_n - \mathcal{I}_n \right) \right\} < 1. \tag{21}$$

Here the contribution from the constraints has cancelled in the numerator, so the left-hand side → 0 as $n \to \infty$, assuming as before that the estimates are consistent and

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

that the law of large numbers applies. 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

$$\nabla_\beta H_n \delta_S = -\nabla_\beta H_n (\mathcal{I}_n + 2\omega C^T C)^{-1} \nabla_\beta H_n < 0, \tag{22}$$

provided the generic condition of $\mathcal{I}_n$ being positive definite is satisfied. There are some negative aspects, however. Choosing $\omega = \mathcal{O}(\sqrt{n})$ means that the condition number of $(\mathcal{I}_n + 2\omega C^T C)$ is $\mathcal{O}(\sqrt{n})$ provided $\operatorname{Aug}(\beta)$ is nonsingular, 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

$$\min_{\beta} \frac{1}{n} K_n(\beta); \quad \mathbf{g}(\beta) = 0. \tag{23}$$

The necessary conditions become

$$\begin{aligned} \frac{1}{n} \nabla_\beta K_n(\beta) &= \zeta^T \nabla_\beta \mathbf{g}(\beta), \\ \mathbf{g}(\beta) &= 0. \end{aligned} \tag{24}$$

Let

$$P(\beta) = \nabla_\beta \mathbf{g}(\beta)^T (\nabla_\beta \mathbf{g} \nabla_\beta \mathbf{g}^T)^{-1} \nabla_\beta \mathbf{g}(\beta) \tag{25}$$

and define $V^* : R^p \to R^{p-m}$ by

$$\nabla_\beta \mathbf{g}(\beta^*) V^* = 0, \qquad V^{*T} V^* = I. \tag{26}$$

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

$$V(\beta) = (I - P(\beta)) V^* \to V^*, \quad \beta \to \beta^*, \tag{27}$$

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

and $V(\beta)$ inherits the smoothness of $\nabla_\beta \mathbf{g}$, $\beta \in B$. This permits a reduced system analogous to (11) to be defined in the nonlinear case. This is

$$\begin{aligned} \frac{1}{n}\nabla_\beta K_n V(\beta) &= 0, \\ \mathbf{g}(\beta) &= 0. \end{aligned} \tag{28}$$

It reduces to (11) in the linear case if the choice $V^* = Q_2$ is made. It has $\hat{\beta}_n$ as an isolated solution provided the Jacobian is nonsingular. It is the right system to use in order to derive the properties of the parameter estimates; but compared to the linear case there is an extra term to be considered. This comes from differentiating $V(\beta)$. However, the law of large numbers can be used to show this is small so the corresponding results hold 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 $\beta = \beta^*$ and $\zeta^* = 0$.

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

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

$$\nabla_\beta H_n = \frac{1}{n}\nabla_\beta K_n + 2\omega^{(i)} \sum_{j=1}^{m} \left(g_j(\beta) + \theta_j^{(i)}\right) \nabla_\beta g_j, \tag{30}$$

$$\nabla_\beta^2 H_n = \frac{1}{n}\nabla_\beta^2 K_n + 2\omega^{(i)} \sum_{j=1}^{m} \left(\nabla_\beta g_j^T \nabla_\beta g_j + \left(g_j(\beta) + \theta_j^{(i)}\right) \nabla_\beta^2 g_j\right). \tag{31}$$

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

$$\left\{ \mathcal{I}_n + 2\omega \sum_{j=1}^{m} \nabla_\beta g_j^T \nabla_\beta g_j \right\} \delta_S = -\nabla_\beta H_n^T. \tag{32}$$

The associated fixed-point iteration is

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

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

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

spectral radius < 1. The variational matrix is

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

In this expression the term $\frac{1}{n}\nabla_\beta^2 K_n - \mathscr{I}_n$ gets small by the usual argument involving the law of large numbers provided that $\beta^{(i)}$ is $1/\sqrt{n}$ consistent for each i, and that the extra term in the nonlinear case involving the constraint second derivatives is of similar order. This requires that $\omega^{(i)}\theta^{(i)}$ is $O(1/\sqrt{n})$ at most for each i. This follows because $\hat{\zeta}_n/n$ is a consistent estimator of $\zeta^* = 0$ so that the error in the initial estimate of $\hat{\zeta}_n/n$ by $\omega\theta^{(1)}$ obtained by setting $\theta^{(1)} = 0$ is of the right order. Also the rate of convergence result shows that this estimate improves by a factor of $O(1/\omega)$ in each outer iteration. Thus $\mathrm{g}(\beta^{(1)})$ can be at most $(1/\omega\sqrt{n})$ and $\mathrm{g}(\beta^{(i)}) = o(1/\omega\sqrt{n})$, $i > 1$, in agreement with the estimates of $\|\beta^{(i)} - \hat{\beta}\|$ given in the appendix. Given this, then

$$
\varpi(J^{(i)}) \to 0, \quad n \to \infty, \quad i = 1, 2, \dots . \tag{33}
$$

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, $i = 1, 2, \dots$;

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

(3) The convergence of the iterates in each of the scoring steps has the characteristic speed associated with the method, provided n is large enough.

5. An example

Consider the mixture density

$$
\begin{aligned}
f_R(y|\mu_1, \mu_2, \sigma_1, \sigma_2) = {}& \frac{1}{\sqrt{2\pi}\sigma_1} \frac{\mu_1}{\mu_1 + \mu_2} \exp - \frac{(y - \mu_1)^2}{2\sigma_1^2} \\
&+ \frac{1}{\sqrt{2\pi}\sigma_2} \frac{\mu_2}{\mu_1 + \mu_2} \exp - \frac{(y - \mu_2)^2}{2\sigma_2^2}.
\end{aligned}
$$

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

Random numbers generated according to a realisation of this density can be considered also to be generated according to the density

$$f\,(y\,|\mathbf{a}) = \frac{1}{\sqrt{2\pi}\sigma_1}\alpha_1 \exp - \frac{(y-\mu_1)^2}{2\sigma_1^2} + \frac{1}{\sqrt{2\pi}\sigma_2}\alpha_2 \exp - \frac{(y-\mu_2)^2}{2\sigma_2^2},$$

where

$$\mathbf{a}^T = [\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2],$$

subject to the constraints

$$g_1(\mathbf{a}) = \alpha_1 - \frac{\mu_1}{\mu_1+\mu_2} = 0, \qquad g_2(\mathbf{a}) = \alpha_2 - \frac{\mu_2}{\mu_1+\mu_2} = 0.$$

Thus it should be possible to recover μ_1, μ_2, σ_1 and σ_2 from data generated according to f_R by considering the likelihood defined by f subject to the above constraints.

Let

$$e_i(y) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \frac{(y-\mu_i)^2}{2\sigma_i^2}, \qquad i = 1, 2.$$

Then

$$f\,(y\,|\mathbf{a}) = \alpha_1 e_1(y) + \alpha_2 e_2(y).$$

We have:

$$K_n(\mathbf{a}) = -\frac{1}{n}\sum_{i=1}^{n} \log f\,(y_i|\mathbf{a}),$$

$$\nabla_\alpha K_n(\mathbf{a}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{f\,(y_i|\mathbf{a})}\mathbf{v}_i^T, \tag{34}$$

where

$$\mathbf{v}_i^T = \left[\left(1, \alpha_1\frac{y_i-\mu_1}{\sigma_1^2}, \alpha_1\left(\frac{(y_i-\mu_1)^2}{\sigma_1^3} - \frac{1}{\sigma_1}\right)\right)e_1, \right.$$
$$\left.\left(1, \alpha_2\frac{y_i-\mu_2}{\sigma_2^2}, \alpha_2\left(\frac{(y_i-\mu_2)^2}{\sigma_2^3} - \frac{1}{\sigma_2}\right)\right)e_2\right].$$

To avoid computing the expectation of the Hessian note that the operations of taking expectations and summation can be interchanged so that

$$\begin{aligned}\mathscr{E}\left\{\nabla_\alpha^2 K_n\right\} &= \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}\left\{\nabla_\alpha L_i^T \nabla_\alpha L_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} - \left\{\nabla_\alpha L_i^T \nabla_\alpha L_i - \mathscr{E}\left\{\nabla_\alpha L_i^T \nabla_\alpha L_i\right\}\right\} + \frac{1}{n}\sum_{i=1}^{n}\nabla_\alpha L_i^T \nabla_\alpha L_i \\ &\sim \frac{1}{n}\sum_{i=1}^{n}\nabla_\alpha L_i^T \nabla_\alpha L_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{f\,(y_i|\mathbf{a})^2}\mathbf{v}_i\mathbf{v}_i^T.\end{aligned} \tag{35}$$

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

This shows that the expected Hessian K_n can be computed to within working accuracy (that is, to within the errors committed in using the law of large numbers) by a sum of quantities computed using the observed data.

In this case the augmented Lagrangian is

$$H_n = K_n + \omega \sum_{i=1}^{2} (g_i(\mathbf{a}) + \theta_i)^2.$$

It follows from (34), (35) and (32) that the scoring method using the above estimate for the expected Hessian gives a set of equations for the correction δ_S which can be written as the linear least squares problem:

$$\min_{\delta} \mathbf{r}^T\mathbf{r};$$

$$\mathbf{r} = \left[\begin{array}{c} \sqrt{2\omega}\nabla_\alpha g_1 \\ \sqrt{2\omega}\nabla_\alpha g_2 \\ \hline \mathbf{v}_1^T/(\sqrt{n}f\,(y_1|\mathbf{a})) \\ \cdots \\ \mathbf{v}_i^T/(\sqrt{n}f\,(y_i|\mathbf{a})) \\ \cdots \\ \mathbf{v}_n^T/(\sqrt{n}f\,(y_n|\mathbf{a})) \end{array}\right] \delta + \left[\begin{array}{c} \sqrt{2\omega}(g_1+\theta_1) \\ \sqrt{2\omega}(g_2+\theta_2) \\ \hline -\mathbf{e}/\sqrt{n} \end{array}\right]. \qquad (36)$$

Note that the constraint contributions appear first. This is because they have a larger scale than the likelihood contributions, and this ordering is advisable for numerical stability when a QR factorization is used to solve the least squares problem [8].

Numerical results are presented for computations carried out using $\mu_1 = 1.0$, $\mu_2 = 2.0$ for two cases:

- $\sigma_1 = \sigma_2 = 0.5$ and
- $\sigma_1 = \sigma_2 = 0.7$.

A random number generator was used to produce random numbers to provide data on the mixture density for $n = 100$, 1000 and 10000. Results for two different seeds for the uniform generator `drand48` are displayed in the tables given below. In both the estimated values and a summary of the iteration progress are given. The latter is given in the column headed 'P-H steps' which summarises the number of scoring iterations in each Powell-Hestenes step. In each case the exact values were taken as starting values and appear to provide a fair test. The estimate computed by the 'inner' scoring algorithm is accepted when

$$1 - \frac{\|\mathbf{r}\|}{\|\mathbf{f}\|} < \text{tol},$$

where $\mathbf{f}$ is the data vector in (36) and tol $= 10^{-4}$. The 'outer' iteration is terminated when $\|\mathbf{g}\| <$ tol. As expected, the performance of the algorithm improves significantly

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

TABLE 1. Results for first seed.

n	μ_1	μ_2	σ_1	σ_2	P-H steps
100	0.9079	2.0137	0.4209	0.4579	(3,2,2,1)
1000	1.0289	2.0104	0.4761	0.4948	(3,2,1)
10000	1.0106	2.0048	0.4870	0.4952	(2,2)
100	0.5018	1.9324	0.7718	0.6688	17 P-H steps
1000	0.9555	2.0023	0.6583	0.6564	(3,2,1)
10000	1.0211	2.0106	0.7040	0.7013	(2,2)

TABLE 2. Results for second seed.

n	μ_1	μ_2	σ_1	σ_2	P-H steps
100	1.0090	2.0933	0.4233	0.4489	(5,2,2,1)
1000	0.9749	2.0067	0.4734	0.5079	(2,2,1)
10000	0.9652	2.0020	0.4749	0.4934	(2,2)
100	1.1709	2.0311	0.6723	0.6234	(5,2,2,1,1,1,1)
1000	1.1462	1.9789	0.7495	0.7192	(3,2,1)
10000	1.0372	1.9894	0.7094	0.6941	(2,2)

as n is increased. Note that the sample replacement (35) for the expected Hessian does not appear to have caused any deleterious effects in the scoring iteration.

Acknowledgements

This work was carried out in part while the author was a visiting Professor in the Computer Science and Mathematics Departments of Texas A&M University. Andy Wood provided valuable assistance, especially in suggesting applications. Mike Powell suggested a number of improvements in presentation. John Reid made the interesting observation that recent work on the connection between Householder reduction and modified Gram-Schmidt orthogonalisation actually showed why row pivoting is not required in the latter procedure.

References

[1] U. Ascher, R. Mattheij, and R. D. Russell, *The Numerical Solution of Boundary Value Problems* (Birkhauser, 1988.)

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

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

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

[4] N. I. Fisher, T. Lewis, and B. J. J. Embleton, *Statistical Analysis of Spherical Data* (Chapman and Hall, 1992).

[5] R. Fletcher, *Practical Methods of Optimization, Vol.2: Constrained Optimization* (Wiley, 1981).

[6] M. R. Osborne, "Fisher's method of scoring", *Int. Stat. Rev.* **60** (1992) 99–117.

[7] M. R. Osborne, "Cyclic reduction, dichotomy, and the estimation of differential equations", *J. Comp. and Appl. Math.* **86** (1997) 271–286.

[8] M. J. D. Powell and J. K. Reid, "On applying Householder's method to linear least squares problems", *Proc. IFIP Congress* (1968) 122–126.

[9] P. K. Sen and J. M. Singer, *Large Sample Methods in Statistics* (Chapman and Hall, 1993).

[10] J. J. Shuster and D. J. Downing, "Two-way contingency tables for complex sampling schemes", *Biometrica* **63** (1976) 271–276.

Appendix 1: The Powell-Hestenes method

The computation of the Powell-Hestenes correction is carried out for the linear constraint case for simplicity. The necessary conditions for a minimum of H_n give

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

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 $\partial\beta/\partial\theta$, differentiate the necessary conditions to obtain the equation

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

The special form of the right-hand side should be noted. Transforming this equation using the factorization (10) gives

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

The inverse of $Q^T\nabla_\beta^2 H_n Q$ when ω is large is given by

$$\begin{bmatrix}\frac{1}{2\omega}U^{-T}U^{-1} + \mathcal{O}\left(\frac{1}{\omega^2}\right) & -\frac{1}{2\omega}U^{-T}U^{-1} + \mathcal{O}\left(\frac{1}{\omega^2}\right)\\ -\frac{1}{2\omega}U^{-T}U^{-1} + \mathcal{O}\left(\frac{1}{\omega^2}\right) & \left(\frac{1}{n}Q_2^T\nabla_\beta^2 K_n Q_2\right)^{-1} + \mathcal{O}\left(\frac{1}{\omega}\right)\end{bmatrix}.$$

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press

Thus

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

so that

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

Substituting in the Newton step gives

$$\delta_\theta = [I + O(1/\omega)]\,[C\beta - \mathbf{d}]\,.$$

The advantage of a good estimate for the Lagrange multipliers can be seen by arguing in a similar fashion. Let $2\omega\theta_i = \lambda_i + \epsilon_i$, where λ_i is the exact multiplier and $\hat{\beta}$ the solution of the constrained problem. Then

$$\frac{1}{n}\nabla_\beta K_n(\beta) + 2\omega \sum (C\beta - \mathbf{d} + \theta)^T C = 0,$$

$$\frac{1}{n}\nabla_\beta K_n(\hat{\beta}) + \sum \lambda_i (\mathbf{e}_i^T C) = 0.$$

Subtracting, and arguing as above, gives

$$\|\beta - \hat{\beta}\| = \mathcal{O}(\|\epsilon\|/\omega).$$

https://doi.org/10.1017/S1446181100011561 Published online by Cambridge University Press