Statistical Science
1989, Vol. 4, No. 4, 394–408

The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms

Guido del Pino

Abstract. This expository paper deals with the role of iterative generalized least squares as an algorithm for the computation of statistical estimators. Relationships between various algorithms, such as Newton–Raphson, Gauss–Newton, and scoring, are studied. A parallel is made between statistical properties of the model and the structure of the numerical algorithm employed to find parameter estimates. In particular a general linearizability property that extends the concept of link function in generalized linear models is considered and its computational meaning is discussed. Maximum quasilikelihood estimators are reinterpreted so that they may exist even when there is no quasilikelihood function.

Key words and phrases: Iterative generalized least squares, maximum likelihood estimation, scoring algorithm, quasilikelihood, generalized linear models.

1. INTRODUCTION

The present paper attempts to provide a unified view of the role of iteratively reweighted least squares (IRLS) in statistical estimation problems. IRLS algorithms have proven to be a powerful computing tool in a variety of estimation problems, notably the computation of M estimators in location and regression problems (Andrews, 1974; Huber, 1981) and remarkably so in the fitting of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983). For this reason, it is of interest to characterize the kind of models for which parameter estimation may be done through IRLS. In linear regression, weighted least squares is used when the observations are independent with different variances, while generalized least squares is used when they are dependent. In a more general situation, IRLS is primarily appropriate when the observations are statistically independent. An analogy with linear regression then suggests iterative generalized least squares (IGLS) as a natural extension of IRLS. From a mathematical view point, however, it is IGLS which plays the dominant role, IRLS being treated as an important particular case. A short review of generalized least squares is given in Section 2.

Many estimation problems, like the two mentioned in the previous paragraph, reduce to minimizing a smooth function subject to certain constraints. In Section 3 we show that, for a rather broad class of iterative minimization algorithms, each iteration consists of solving a GLS problem, and in this sense these algorithms are examples of IGLS. A general purpose minimization algorithm is that of Newton–Raphson (NR). Although this algorithm does not fall in the IGLS class, a modification of it does. This modified Newton–Raphson (MNR) algorithm reduces to NR when constraints are linear.

One minimization problem of special interest in statistics is nonlinear least squares, Gauss–Newton (GN) being perhaps one of the best known algorithms for its solution. In Section 4 we show that MNR may be considered as a generalization of GN. In this way, the extensive experience existing about GN may be brought to bear on the behavior of MNR.

In generalized linear models, the mean parameters are transformed through the so called link function in such a way that the new parameter space becomes a linear manifold. In Section 5, we extend the concept of link function to that of a general linearizing transformation. This includes as a special case the composite links first suggested in Thompson and Baker (1981) and used in the analysis of some genetic models. The significance of this transformation in relation to the structure of the equations defining each IGLS iteration is discussed.

Section 6 studies the computation of maximum likelihood estimators (MLE) with emphasis on the method of scoring (SCO), also an important example of IGLS. It is examined there how statistical

Guido del Pino is Professor, Department of Probability and Statistics, Pontificia Universidad Católica de Chile, Casilla 6177, Santiago 22, Chile.

properties of the fitted model are related to the behavior of the coefficients defining each IGLS iteration. For instance, the way in which some of these coefficients depend on the data or on nuisance parameters is related to the type of distribution (e.g., exponential family) followed by the observations. This examination shows that generalized linear models may be in some sense characterized by the particularly simple form taken by the IRLS algorithm. At the same time, it suggests how to adapt the algorithm to cover more general classes of statistical models.

In Sections 3 through 6, generalized least squares estimators play a key role. The famous Gauss–Markov theorem states that these estimators are, in fact, best linear unbiased when the covariance structure of the observations is known. By allowing it to depend on the mean vector one is led to the framework of quasilikelihood. The concepts of quasilikelihood function and maximum quasilikelihood estimators were introduced by Wedderburn (1974) in the independent case and further extended by McCullagh (1983) to the dependent case. In Section 7 we give a simple argument that leads to the algorithm (of IGLS type) proposed in McCullagh (1983). It is claimed that the same procedure makes sense even if a quasilikelihood function does not exist, and so we propose that the algorithm just mentioned as well as the resulting estimators be called generalized Gauss–Markov (GGM).

2. GENERALIZED LEAST SQUARES

In its most general form, we have a vector space E with inner product $\langle\ ,\ \rangle$ and corresponding norm $\|\ \|$ and a linear manifold $M = a + L$, where L is subspace. Then the minimization problem

$$\min_{\mu \in M} \| Y - \mu \|^2 \tag{2.1}$$

may be called a *generalized least squares* (GLS) problem, whose solution is

$$\hat{\mu} = a + P_L(Y - a), \tag{2.2}$$

where P_L denotes the orthogonal projection, with respect to the given inner product, onto the subspace L.

More concretely, let X be a $n \times k$ matrix whose columns span the subspace L and let A be a $n \times n$ positive definite matrix. Let $\langle\ ,\ \rangle_A$ and $\|\ \|_A$ be defined by

$$\langle u, v \rangle_A = u'Av,$$
$$\| u \|_A = \langle u, u \rangle_A^{1/2}.$$

Any inner product and norm in $\mathscr{R}^n$ has this form. Then (2.1) and (2.2) become

$$\min_{\mu \in \{a + X\beta \mid \beta \in \mathscr{R}^k\}} (Y - \mu)'A(Y - \mu) \tag{2.3}$$

and

$$\hat{\mu} = a + X\hat{\beta} \tag{2.4}$$

with

$$\hat{\beta} = (X'AX)^{-1}X'A(Y - a). \tag{2.5}$$

A central problem in the statistical theory of linear models is the search for best linear unbiased estimators (BLUE). The Gauss–Markov theorem states the mathematical equivalence between this problem and the GLS one. Some details about this important theorem are included below since they will be useful in Section 7 in relation to quasilikelihood concepts.

Let Y be a $n \times 1$ random vector with mean μ, V be a positive definite matrix, and α be a positive scalar such that

$$\mu \in a + L, \quad \mathrm{Var}(Y) = \alpha V. \tag{2.6}$$

The class of linear unbiased estimators of μ is formed by those linear transformations of Y with expected value μ. The BLUE $\hat{\mu}$ of μ is the element of this class with smallest covariance matrix (with respect to the partial order associated with positive definite matrices (see Rao, 1973)). The Gauss–Markov theorem states that $\hat{\mu}$ is given by (2.2) where the orthogonal projection is taken with respect to the inner product $\langle\ ,\ \rangle_A$ or, in matrix form, by (2.4), (2.5), with $A = V^{-1}$. Another problem which has the same solution is that of finding the maximum likelihood estimator of μ under (2.6), when Y has a multivariate normal distribution.

In the remainder of this paper we will see that many iterative statistical estimation algorithms may be interpreted as the solution of GLS problems corresponding to local linearization of the constraints and quadratic approximation of the function being minimized. Numerical and computational aspects of the solution to (2.3) are discussed for instance in Lawson and Hanson (1974).

3. NONLINEAR MINIMIZATION AND ITERATIVE GENERALIZED LEAST SQUARES

The GLS problem (2.1) will now be extended by replacing the quadratic function $\| Y - \mu \|^2$ and/or the linear manifold M by a smooth function g and a nonlinear manifold respectively. It will be seen that the steps of many numerical algorithms may be interpreted as solutions to GLS problems, in which the norm $\|\ \|$ and the linear manifold change at each iteration. We first discuss a general class of descent algorithms and then we analyze the important Newton–Raphson (NR) method. The reader is referred to Dennis and Schnabel (1983), hereafter DS, for a general discussion. Other relevant references are Avriel (1976) and Fletcher (1980, 1981).

We start by establishing the notation and definitions. Let f be a function defined on an open set in $\mathscr{R}^m$ with values in $\mathscr{R}^p$ and let $f(x) = (f_1(x), \cdots, f_p(x))$. The function f is said to be of class $\mathscr{C}^k$ if all partial derivatives of order k of the component functions $f_1, \cdots, f_p$ are continuous. The Jacobian matrix of f at the point x is the $m \times p$ matrix with entries $\partial f_r/\partial x_i$. When f is real valued the transpose of the Jacobian matrix (in this case a vector) is the gradient of f, its value at x being denoted by $T_f(x)$ or $T(x)$. Let g be a real valued function of class $\mathscr{C}^1$. We say that $p(x)$ is a descent direction for g at x if $g(x + \lambda p(x)) < g(x)$ for λ sufficiently small. Any descent direction for f may be written as $-B(x)T(x)$, where $B(x)$ is a positive definite matrix. A large class of iterative algorithms for minimizing f has the form

$$x^{q+1} = x^q - \lambda^q B(x^q)T(x^q)$$

where λ^q is a positive number. Consider now the problem

$$\min_{\theta \in M = a+L} g(\theta)$$

where θ is the vector of parameters and L is a k-dimensional subspace of $\mathscr{R}^n$, which may be written as $\{X\beta, \beta \in \mathscr{R}^k\}$, for an $n \times k$ matrix X of rank k. The gradient of $G(\beta) = g(a + X\beta)$ is $X'T_g(a + X\beta)$, and a descent direction for G is $-B_X(\theta)X'T_g(\alpha + X\beta)$. In terms of the function g, $-XB_X(\theta)X'T_g(a + X\beta)$ is a descent direction contained in L. It seems natural to impose the condition that this last vector depends only on g and L and not on the particular X chosen. This leads to the choice $B_X(\theta) = (X'A(\theta)X)^{-1}$. A geometric interpretation is as follows: The direction in L given by the indicated choice for $B_X(\theta)$ is the orthogonal projection of $A(\theta)^{-1}T_g(\theta)$ on the subspace L, with respect to the inner product $\langle\ ,\ \rangle_{A(\theta)}$, and this projection does not depend on the choice of the matrix X whose columns span L. The suggested algorithm is then

$$\beta^{q+1} = \beta^q + \lambda_q \delta^g,$$

with

$$\delta^q = (X'A^qX)^{-1}X'A^qY^q \tag{3.1a}$$

where

$$\theta^q = a + X\beta^q, \quad A^q = A(\theta^q), \qquad Y^q = -(A^q)^{-1}T_g(\theta^q). \tag{3.1b}$$

The basic algorithm corresponds to the choice $\lambda^q = 1$. From (2.3) and (2.5), (3.1a) is also the solution to the GLS problem

$$\min_{\delta} (Y^q - X\delta)'A^q(Y^q - X\delta). \tag{3.2}$$

So far $A(\theta)$ has been an arbitrary positive definite matrix. If we assume further that the function g is of class $\mathscr{C}^2$, then g may be approximated in the neighborhood of θ^q by the quadratic function

$$\bar{g}(\theta^q + \delta) = \tfrac{1}{2}\| Y^q - \delta \|^2_{H(\theta^q)} + g(\theta^q) - \tfrac{1}{2}Y^{q\prime}H(\theta^q)Y^q \tag{3.3}$$

where $H(\theta)$ is the Hessian of g at θ. If g is exactly quadratic, using (3.1) with $A(\theta) = H(\theta)$ solves the minimizing problem in one step. More relevant perhaps is that this choice for $A(\theta)$ will provide fast convergence when the initial point is close to the solution. It turns out that this algorithm is just a special case of NR applied to $G(\beta) = g(a + X\beta)$. It is well known that NR is quadratically convergent (DS, page 90).

We next lift the linearity condition on M by letting

$$M = \{h(\beta), \beta \in B\} \tag{3.4}$$

where B is open in $\mathscr{R}^k$ and h is a function of class $\mathscr{C}^1$ with full rank Jacobian matrix $X(\beta)$ for all β in B. A natural modification of algorithm 3.1 is to minimize $g(\theta)$ subject to $\theta \in M$ by replacing X in (3.1) by $X^q = X(\beta^q)$. The iteration then becomes $\beta^{q+1} = \beta^q + \lambda^q\delta^q$ with

$$\delta^q = (X(\beta^q)'A(\theta^q)X(\beta^q))^{-1}X(\beta^q)'A(\theta^q)Y(\theta^q) \tag{3.5a}$$

where

$$Y(\theta) = -A(\theta)^{-1}T_g(\theta). \tag{3.5b}$$

An equivalent formula is

$$\beta^{q+1} = (X(\beta^q)'A(\theta^q)X(\beta^q))^{-1} \cdot X(\beta^q)'A(\theta^q)[X(\beta^q)\beta^q + Y(\theta^q)]. \tag{3.6}$$

We call (3.5) or (3.6) the IGLS algorithm. Unlike the linear manifold case, the IGLS algorithm with $A(\theta) = H(\theta)$ does not reduce to NR. In fact, NR corresponds to $\lambda^q = 1$ and

$$\delta^q = ((X(\beta^q)'H(\beta^q)X(\beta^q) + E(\beta^q))^{-1}X(\beta^q)'H(\theta^q)Y(\theta^q) \tag{3.7}$$

with

$$E_{ij}(\beta) = \sum_{r=1}^{n} T_r(h(\beta))\frac{\partial^2 h_r}{\partial\beta_i\partial\beta_j}, \quad i,j = 1, \cdots, k. \tag{3.8}$$

Note that $E(\beta) = 0$ for a linear manifold. Thus IGLS with $A(\theta) = H(\theta)$ differs from NR by a term which is null for a linear manifold. For this reason we will refer to it as modified Newton–Raphson (MNR).

In many statistical problems, the function g to be minimized has separable structure, in the sense that

$$g(\theta) = \sum_{1}^{n} g_r(\theta_r), \tag{3.9}$$

where each g_r is a smooth function of one real variable. This type of structure is often associated with

independence properties in the statistical model. When (3.9) holds, it is natural to choose $A(\theta)$ in (3.5) as a diagonal matrix, whose rth element is $a_r = u_r(\theta_r)$ for some functions u_r, $r = 1, \cdots, n$. It is easily seen that MNR satisfies this condition. With this choice of A, each iteration in (3.5) reduces to the solution of a weighted least squares problem and so IGLS becomes IRLS.

Next we point out how the ideas considered so far may be applied in two statistical problems. The first problem is the computation of M estimators in robust regression. The mathematical problem is the minimization of

$$g(\theta) = \sum_1^n \rho(U_r - \theta_r), \tag{3.10}$$

subject to $\theta = X\beta$, $\beta \in \mathscr{R}^k$, where $U_1, \cdots, U_n$ are the observed responses (assumed to be independent). The function ρ is assumed to be an even function, nondecreasing for positive values of the argument, and null when this argument is 0. If f is of class $\mathscr{C}^2$, a natural algorithm is MNR (which coincides with NR), and it will be of IRLS type. Note, however, that the dependent variable does not coincide with the response. Andrews (1974) and Beaton and Tukey (1974) suggest replacing the second derivative $\delta''(v)$ by the secant approximation

$$\frac{\rho'(v) - \rho'(0)}{v - 0} = \frac{\rho'(v)}{v}$$

so that

$$a_r = \frac{\rho'(u_r - \theta_r)}{u_r - \theta_r},$$

$$Y_r = \frac{\rho'(u_r - \theta_r)}{a_r} = U_r - \theta_r,$$

and the rth component of the last term in (3.4) becomes U_r.

Huber (1974, 1977) proposes the choice $a_r = 1$, i.e., $A(\theta) = I$, the identity matrix. For further discussion of these algorithms, we refer the reader to Birch (1980), Byrd and Pyne (1979), Holland and Welsch (1977), Huber (1981) and Peters, Klema and Holland (1978).

The second problem we discuss is the fitting of linear regression with censored data. We follow Aitkin's (1981) discussion of Schmee and Hahn (1979). The essential point is that the negative log-likelihood of normal data under censoring has the form (3.9). If furthermore A is chosen as the identity matrix, (3.6) becomes

$$\beta^{q+1} = (X'X)^{-1}X'W^q$$

with

$$W_r^q = \theta_r^q - g_r'(\theta_r^q), \quad r = 1, \cdots, n.$$

In the special application considered, $g_r(\theta_r) = v_r(u_r - \theta_r)$, where the form of function v_r depends on whether the rth observation is censored or not.

4. NONLINEAR LEAST SQUARES

The statistical origin of the nonlinear least-squares problem is as follows: Let $U_1, \cdots U_n$ be independent with constant variance and $EU_r = h_r(\beta)$, $r = 1, \cdots, n$, where h_r are the components of the function h in Section 3. The method of least squares to estimate β consists of minimizing

$$\sum_1^n (U_r - h_r(\beta))^2. \tag{4.1}$$

If U_r is also normally distributed, $r = 1, \cdots, n$, this method coincides with that of maximum likelihood. It is clear that (4.1) is a particular case of the problem of Section 3 with

$$g(\theta) = \| U - \theta \|_I^2, \tag{4.2}$$

where I is the identity matrix. The MNR algorithm applied to (4.2) coincides with the popular method of Gauss–Newton (GN) (see DS, Chapter 10). A natural extension is to replace (4.2) by

$$g(\theta) = \| U - \theta \|_A^2, \tag{4.3}$$

and the corresponding MNR algorithm may still be referred to as GN. We might consider MNR as a generalized Gauss–Newton (GGN) algorithm, in the sense that the quadratic function (4.2) is replaced at each iteration by a local quadratic approximation of $g(\theta)$. It is also possible to apply the general IGLS algorithm of Section 3. We remark, however, that the popular Marquardt algorithm, which consists of replacing $X'H(\theta^q)X$ by

$$X'H(\theta^q)X + \eta^q I,$$

is not a particular case of IGLS since it is not invariant with respect to the choice of X. If we allow A^q to depend on $X(\beta^q)$ then the choice $A^q = H(\theta^q) + \eta^q(X(\beta^q)'X(\beta^q))^{-2}$ leads to

$$\delta^q = (X(\beta^q)'X(\beta^q) + \eta^q I)^{-1}X(\beta^q)'U,$$

which correspond to Marquardt's algorithm. As a final remark, we point out that the problem will have separable structure if the matrix A defining the norm is diagonal.

5. LINEARIZABLE MANIFOLDS

In a typical statistical problem the numbers n and k will correspond to the number of observations and the number of independent parameters in the model, respectively. Often n may be very large so that the cost of storing and computing the matrices involved in each step of algorithm (3.5) would consequently be

very high. We concentrate here on the MNR algorithm since it is directly related to iterative generalized least squares. To simplify the notation, we omit the superscript q in (3.5) and related equations.

We point out first that the matrix A and the vector Y in (3.5) depend explicitly on the value of θ and not the value of β yielding $\theta = h(\beta)$. Matrix X, on the other hand, will generally depend on β. In the separable case (3.9), matrix A is diagonal and so only n elements need to be computed and stored. In this situation, the most expensive computation would be that associated with the $n \times k$ matrix X. The most favorable case arises when X is a constant matrix so that it does not need to be recomputed at each step. This condition actually means that M is a linear manifold so that MNR will coincide with NR and enjoy the rapid convergence properties of this algorithm near the optimum. Assume now M is not linear. We pose the following question: May (3.5) be replaced by an equivalent system of equations

$$Z'_{k\times n}B_{n\times n}Z_{n\times k}\delta_{k\times 1} = Z'_{k\times n}B_{n\times n}W_{n\times 1} \tag{5.1}$$

so that Z is constant and B, W are functions of θ? Note that if there is an invertible matrix E, depending on θ, such that $X = EZ$, with Z being a constant matrix, then (5.1) holds with

$$B = E'AE, \quad W = E^{-1}Y. \tag{5.2}$$

Since X is the Jacobian matrix of h at β, a sufficient condition for (5.1) to hold is then

$$\frac{\partial h_r}{\partial \beta_j} = \sum_{s=1}^{n} e_{rs}(\theta)Z_{sj}, \qquad r = 1, \cdots, n; \quad j = 1, \cdots, k, \tag{5.3}$$

where e_{rs} are functions of θ and Z_{sj} are constants. To get further insight on the meaning of (5.3), it is convenient to make the following definition.

DEFINITION 5.1. A k-dimensional manifold M of class $\mathscr{C}^1$ is said to be linearizable if there exists an invertible mapping F of class $\mathscr{C}^1$, with F^{-1} also of class $\mathscr{C}^1$, such that the following condition holds: F maps an open set in $\mathscr{R}^n$ containing M onto an open set and $F(M)$ is contained in a k-dimensional linear manifold.

Definition 5.1 means that there is a change of variable $\eta = F(\theta)$ such that $\theta \in M$ is equivalent to $\eta \in F(M)$ and $F(M)$ is contained in a linear manifold N of the same dimension as M. Any element of N may then be written as

$$\eta = v + Z\beta, \quad \beta \in \mathscr{R}^k$$

with Z of order $n \times k$ and rank k. Hence any point θ in M may be written as

$$\theta = F^{-1}(v + Z\beta), \quad \beta \in B \text{ open in } \mathscr{R}^k. \tag{5.4}$$

In what follows we will not deal explicitly with the restriction $\beta \in B$, and so for our purposes we will essentially assume $B = \mathscr{R}^k$. By the chain rule (5.4) implies (5.3), the $n \times n$ matrix $E(\theta)$ with components $e_{rs}(\theta)$ being the Jacobian matrix of F^{-1}. In turn this implies that (5.1) holds with B and W given by (5.2), where E is the matrix $E(\theta^q)$.

In the separable case (3.9), it is natural to impose the condition that B, and hence E, be diagonal. Then (5.3) becomes

$$\frac{\partial h_r}{\partial \beta_r} = t_r(h_r(\beta))Z_{rj}, \qquad r = 1, \cdots, n; j = 1, \cdots, k. \tag{5.5}$$

The equations $m'_r = 1/t_r$, $r = 1, \cdots, n$, define m_r uniquely up to additive constants. The transformation F, given by

$$(F(\theta))_r = m_r(\theta_r), \tag{5.6}$$

linearizes the manifold and the arbitrary constants in m_r, $r = 1, \cdots, n$, may be absorbed into the vector v in (5.4). The special case (5.5, 5.6) arises in the theory of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983) in relation to link functions, although only the case $m_1 = \cdots = m_n$ and $v = 0$ is considered there. The case of general v is allowed, however, in the GLIM package (Baker and Nelder, 1978) through the OFFSET option. The more general choice of F is relevant for the extension to composite links (Thompson and Baker, 1981). It is somewhat surprising how purely computational considerations may suggest meaningful statistical ideas. We will have the opportunity to do a similar analysis in the next section. Given a linearizable manifold, a natural procedure is to make the parameter transformation $\eta = F(\theta)$ and minimize $G(\eta) = g(F^{-1}(\eta))$ with η restricted to a linear manifold. NR and MNR are identical when applied to G and this common algorithm will in general be different from NR or MNR applied to the original function g.

Although it seems natural to make the change of parameters, it is possible that a computer program has been written for minimizing only special cases of the function g. This, in fact, is the case for generalized linear models and the GLIM package. To illustrate the two alternatives consider the following simple example:

EXAMPLE 5.1. Let $g(\theta) = \frac{1}{2}\sum_{r=1}^{n}(u_r - \theta_r)^2$ and let $\theta_r = \exp(\beta_1 + \beta_2 r)$. Then $Z_{1r} = 1$, $Z_{2r} = r$. At each step of MNR applied to the function g, the following system of equations is solved (we omit the superscript q and the evaluation at θ^q for simplicity):

$$\sum_{j=1}^{k}\sum_{r=1}^{n} Z_{ri}Z_{rj}\theta_r^2\delta_j = \sum_{r=1}^{n} Z_{ri}\theta_r(u_r - \theta_r) \tag{5.7}$$

whereas MNR applied after the transformation $\eta = \log \theta$ yields

$$(5.8) \quad \sum_{j=1}^{k} \sum_{r=1}^{n} Z_{ri} Z_{rj} (\theta_r^2 + \theta_r(\theta_r - u_r)) = \sum_{r=1}^{n} Z_{ri} \theta_r (u_r - \theta_r).$$

We see that the right-hand sides of (5.7) and (5.8) are identical, a property easily seen to hold in general. The only difference lies in the coefficient multiplying $Z_{ri}Z_{rj}$, the coefficient in (5.8) having an extra term $\theta_r(\theta_r - U_r)$. In (5.7) the coefficient is always positive, while in (5.7) it is possible that it has a negative value. This suggests that (5.6) might be safer for practical use. If $u_1, \cdots, u_n$ are the observed values of random variables $U_1, \cdots, U_n$ such that $EU_r = \theta_r$, $r = 1, \cdots, n$, the expected value of the extra term $2\theta_r(\theta_r - U_r)$, *evaluated assuming the current value of* θ_r *in the iterative procedure is the true one*, is null. This in turn suggests that the MNR applied to the original function g will be close to the NR applied to the transformed problem if the initial estimates are close to the true values and the observations have high precision.

Example 5.1 suggests that there is not a clear cut choice between an algorithm based on (5.1) and the MNR algorithm applied after a parameter transformation that linearizes the manifold. Similar discussions may also be made in connection with the scoring algorithm of the next section.

6. MAXIMUM LIKELIHOOD ESTIMATION

In this section, we attempt to relate the material already discussed to the special case where the function g to be minimized is the negative of a log-likelihood. We also want to consider a new element coming into the picture, namely the data. The data may be represented by a vector u, and u may be considered as the realization of a random vector U.

In order that the n-dimensional vector θ be a convenient description of the statistical model, we impose the condition that the log-likelihood $L(\cdot, u)$ be defined in an open set Θ. The corresponding model will be called the full model and the model to be fitted corresponds to the restricted model $\theta \in M$, where M is a k-dimensional manifold. It will be assumed that the full model satisfies all the usual regularity conditions. In particular it is assumed that the Fisher information matrix $I(\theta)$ exists, is positive definite, and coincides with the expected value of the Hessian of $-L(\cdot, U)$.

Consider first the NR algorithm. The expected value of $E(\beta^q)$ in (3.7) under the assumption that β^q is the true value, is null. This suggests simplifying NR by dropping the E term, thereby arriving at the NMR algorithm. A further step consists of replacing the Hessian by its expected value. This leads to an IGLS algorithm called scoring (SCO). This was originally proposed by Fisher (1925) in the unidimensional case and a convenient general reference is Rao (1973). Consider again Example 5.1 and assume $U_1, \cdots, U_n$ are iid and $U_r \sim N(\theta_r, 1)$, $r = 1, \cdots, n$. Then $-L(\theta, u)$ differs from $g(\theta)$ only by an additive constant. Since the Hessian of $-L(\cdot, u)$ is constant, it follows that (5.6) coincides with the scoring algorithm applied to the original parametrization. The same algorithm applied to the new parametrization $\eta_r = \log \theta_r$, $r = 1, \cdots, n$ is obtained by replacing the coefficient of $Z_{ri}Z_{rj}$ in the left-hand side by its expected value θ_r^2. It is seen that the scoring algorithm is invariant under changes in the θ parameter. This property holds in general and may be considered as an advantage of SCO over MNR and NR. What is claimed is that the system of equations associated with two different parametrizations will be equivalent. However, the actual sets of equations and the corresponding numerical properties may be quite different. In the linearizable case, the convenient form (5.1) is directly obtained by applying SCO to the transformed parameter $\eta = F(\theta)$ that linearizes the manifold M. Using the results of Section 3, it is clear that SCO is a special case of IGLS with $A(\theta) = I(\theta)$. From Section 4 it follows that SCO may also be considered as a modification of GGN, in which the Hessian is replaced by its expected value. The connection between MLE and nonlinear least squares has been much discussed in the literature (see e.g., Bradley, 1973; Wedderburn, 1974; Jennrich and Moore, 1975; and Green, 1984). If U_r, $r = 1, \cdots, n$, are independent and the distribution of U_r depends only on θ_r, the log-likelihood has the separable form (3.9). SCO shares with MNR the property that it reduces to a special case of IRLS.

Next we examine three issues related to the special way in which the observations, nuisance parameters and linearizable manifolds affect the solution to the likelihood equations, as well as the steps in the MNR and scoring algorithms.

Hessian of log-likelihood. When MNR is used, it may be convenient that the Hessian of the log-likelihood does not depend on the observations. This is equivalent to the condition that the Hessian coincides with $-I(\theta)$ and it is well known that this will hold if and only if

$$(6.1) \quad L(\theta, U) = \theta' T(U) + S(U) + C(\theta)$$

for suitable functions T, S, and C, i.e., if the model corresponds to an exponential family with the natural parameter space.

Nuisance parameters. Let θ be the vector parameter of interest and let $\phi \in \Phi$ be a nuisance parameter. Let $\hat{\theta}(\phi, M)$ be the MLE of θ in the model $\theta \in M$. A sufficient condition for

$$(6.2) \quad \hat{\theta}(\phi_1, M) = \hat{\theta}(\phi_2 M)$$

for all $\phi_1, \phi_2 \in \Phi$ and all manifolds M is

(6.3) $L(\theta, \phi, u) = f(g(\theta, u), \phi, u)$ for all θ, ϕ, u,

for some functions f and g of class $\mathscr{C}^1$, such that the partial derivative of f with respect to the first argument is non-null. A special case of (6.2) is

(6.4) $L(\theta, \phi, u) = \alpha(\phi, u)g(\theta, u) + \gamma(u, \theta)$

for suitable functions g, α, γ, with $\alpha > 0$. Functions α, g and γ are not unique, but it is possible to impose the condition that $g(\theta, u)$ be equal to $L(\theta, \phi_0, u)$ for some ϕ_0, in which case we have $\alpha(\phi_0, u) = 1, \gamma(\phi_0, u) = 0$.

When $\alpha(\phi, u)$ does not depend on u, the Hessian of L at ϕ is the Hessian of L at $\phi = \phi_0$ multiplied by $\alpha(\phi_0)$ and the same happens with the information matrix. We may write the log-likelihood as

(6.5) $L(\theta, \phi, u) = \alpha(\phi)g(\theta, u) + \gamma(\phi, u)$,

which corresponds to the extended class of generalized linear models proposed by Jorgensen (1983). When U corresponds to a vector $(U_1, \cdots, U_n)'$ of independent observations, with the distribution of U_r depending only on θ_r,

(6.6) $$L(\theta, \phi, u) = \sum_1^n l_r(\theta_r, \phi, u_r)$$

and the form (6.4) may be actually derived from (6.3).

Under (6.4) the steps of MNR will not depend on ϕ while the steps of SCO will depend on this nuisance parameter, unless we have a natural exponential family for fixed ϕ, i.e.,

(6.7) $L(\theta, \phi, u) = \alpha(\phi)(\theta' T(u) - b(\theta)) + \gamma(\phi, u)$

in which case MNR and SCO coincide.

Linearizable manifolds. We have already mentioned the invariance of SCO under changes in parametrization. The convenient form (5.1) is obtained applying it to the transformed parameter $\eta = F(\theta)$.

The popular class of generalized linear models corresponds to the following set of assumptions:

(1) (6.4) holds.
(2) $U_1, \cdots, U_n$ are independent and U_r has distribution $Q_{\theta_r, \phi}$, $r = 1, \cdots, n$.
(3) For a fixed ϕ, the family $Q_{\theta, \phi}$ is an exponential family.
(4) Manifold M is linearizable.

The useful fact, pointed out by Nelder and Wedderburn, that MLE for these models may be conveniently computed using IRLS, follows from the general framework discussed in this paper. More precisely, NR applied to the natural parametrization of the exponential family coincides with SCO applied to the same parametrization. By the invariance of SCO, it will also coincide with SCO applied to the transformed parameter η linearizing the manifold and will have the convenient form (5.1). It should be recalled that NR applied to the parameter η leads to a different algorithm. More discussion along these lines may be found in del Pino (1984). Perhaps a more valuable property of the approach presented is that it makes clear that similar methods may be valuable for estimating parameters in more general classes of models. Some relevant references are Jorgensen (1983, 1987) and Green (1984).

7. GENERALIZED GAUSS–MARKOV ESTIMATION

Consider the problem of estimating $\theta = EY$ given

(7.1) $$\mathrm{Var}(Y) = \alpha V(\theta)$$

and

(7.2) $$\theta \in M,$$

where as usual M is a manifold of class $\mathscr{C}^1$. When M is linear and $V(\theta) = V$ constant, the BLUE $\hat{\theta}$ of θ is characterized by $\hat{\theta} \in M$ and

$$\langle Y - \hat{\theta}, x \rangle = 0, \quad \forall x \in L,$$

where L is the translate of M to the origin and $\langle a, b \rangle = a' V^{-1} b$. This suggests that a reasonable estimator $\hat{\theta}$ of θ in the general case, having some "local optimality properties" is given by

(7.3) $\hat{\theta} \in M$ and $\langle Y - \hat{\theta}, x \rangle = 0, \quad \forall x \in L,$

where L is the tangent subspace to M at $\hat{\theta}$ and $\langle a, b \rangle = a' V(\hat{\theta})^{-1} b$.

We call $\hat{\theta}$ satisfying (7.3) a generalized Gauss–Markov estimator (GGME) of θ. An iterative solution to (7.3) may proceed as follows.

Let $\theta^q = h(\beta^q)$ be the current estimate of θ after q steps, let $\langle a, b \rangle_q = a' V(\theta^q)^{-1} b$, and let L^q be the tangent subspace to M at θ^q. By analogy with (7.3) consider the problem: Find $\delta^q \in \mathscr{R}^k$ such that $\langle Y - \theta^q - X(\beta^q)\delta^q, x \rangle = 0$, for all $x \in L^q$. More explicitly, since an arbitrary element of L^q can be expressed as $X(\beta^q)\delta$, where δ is an arbitrary element of $\mathscr{R}^k$, it follows that the orthogonal projection of $Y - \theta^q$ onto L^q, with respect to the inner product $\langle , \rangle_q$, is $X(\beta^q)\delta^q$, and

(7.4) $$\delta^q = (X(\beta^q)' V(\theta^q)^{-1} X(\beta^q))^{-1} \cdot X(\beta^q)' V(\theta^q)^{-1} (Y - \theta^q).$$

The $(q + 1)$th iteration of the algorithm is then $\beta^{q+1} = \beta^q + \delta^q$, $\theta^{q+1} = h(\beta^{q+1})$. We will refer to this as the GGM algorithm. The iteration (7.4) becomes identical to (3.5) for some function g if $A(\theta) = V(\theta)^{-1}$ and

if the gradient $T_g(\theta)$ satisfies

$$T_g(\theta) = -V(\theta)^{-1}(Y - \theta). \tag{7.5}$$

A necessary and sufficient condition for (7.5) to hold is as follows.

LEMMA 7.1. *Let Θ be a convex open set in $\mathscr{R}^n$ and let the matrix $V^{-1}(\theta)$ be continuous on θ. A necessary and sufficient condition for the existence of a function g of class $\mathscr{C}^2$ such that*

$$T_g(\theta) = V^{-1}(\theta)(Y - \theta)$$

is that $V^{-1}(\theta)$ be the Hessian of some function G, evaluated at θ.

COROLLARY 7.1. *If $V(\theta)$ is diagonal and $V_{rr}(\theta) = a_r(\theta_r)$, for some continuous functions $a_1, \cdots, a_n$ then (7.5) holds. Furthermore, if $1/a_r = A_r''$ then*

$$g(\theta) = \sum_{r=1}^{n} A_r(\theta_r) + \sum_{r=1}^{n} A_r'(\theta_r)(Y_r - \theta_r)$$

up to an additive function of Y.

As an example, if Y_r are independent Poisson random variables with $E(Y_r) = \theta_r$, $r = 1, \cdots, n$, then $a_r(\theta_r) = \theta_r$, $A_r(\theta_r) = \theta_r \log \theta_r - \theta_r$, $A_r'(\theta_r) = \log \theta_r$, and $g(\theta) = \sum_1^n (\theta_r \log \theta_r - \theta_r) + \sum_1^n \log \theta_r (Y_r - \theta_r) = \sum_1^n Y_r \log \theta_r - \sum \theta_r$, up to a function of Y.

If Y has an exponential family distribution with $EY = \theta$ and $\mathrm{Var}(Y) = V(\theta)$, for θ in an open convex set in $\mathscr{R}^n$, then the negative of the log-likelihood function expressed in terms of θ, may be used as the function g in Lemma 7.1 and $V^{-1}(\theta)$ coincides with the Fisher information matrix. In this special case the GGM algorithm coincides with SCO and the GGM estimator coincides with MLE.

A function of θ whose gradient $T(\theta)$ satisfies (7.5) is called a quasilikelihood (Wedderburn 1974; McCullagh 1983). The GGM estimator then maximizes the quasilikelihood and it may be called a maximum quasilikelihood estimator (MQLE). The properties of MQLE discussed in McCullagh (1983) appear to hold for GGM even if a quasilikelihood does not exist. When the manifold M is linearizable (as in generalized linear models), GGM may be implemented in the form (5.1). As shown in McCullagh and Nelder (1983), this algorithm may be heuristically obtained by using a "working dependent variable" $F(\theta^q) + E(\theta^q)^{-1}(Y - \theta^q) - v$, whose expected value and covariance matrix are approximately given by $Z\beta$ and $\alpha E(\theta^q)^{-1} V(\theta^q) E(\theta^q)^{-1\prime}$. A related approach to estimate the mean θ is to minimize the statistic

$$S(\theta) = (Y - \theta)' V(\theta)^{-1}(Y - \theta) \tag{7.6}$$

for $\theta \in M$. When Y has a multinomial distribution, (7.6) reduces the Pearson chi-squared statistic (actually $B(\theta)$ is singular and a generalized inverse must be used). This is probably the reason why the method of estimation based on minimizing (7.6) is called minimum chi-squared (MCS). There has been a long discussion in the literature about the relative merits of MLE and MCS. We refer the reader to Berkson (1980) and the discussion therein. We point out that for a multivariate normal $-2L(\theta, Y) = S(\theta) + \log\det V(\theta)$, so that MLE and MCS do not agree. Although (7.6) may be interpreted as a Mahalanobis squared-distance between Y and θ, the dependence of the covariance matrix on θ, is a source of theoretical as well as practical difficulties. In fact, MCS are not usually consistent and the minimization of (7.6) may involve the analytic computation of complicated derivatives.

Another use of (7.6) is in testing $\theta = \theta_0$ vs. $\theta \neq \theta_0$. In fact, for an exponential family, $S(\theta_0)$ is the score test statistic. Pregibon (1982) shows that the GLIM package provides the value of $S(\theta^{q-1})$ and this may be used for the computation of score tests. The MCS may be then considered as the least rejected value of θ, when using the score test. The fact that score tests may be good against local alternatives does not imply that MCS is a good estimator.

When comparing SCO, GGM and MCS, a relevant point is that the last two use only information about the covariance matrix $V(\theta)$. This suggests that SCO would generally lead to a more efficient estimation. As a counterpart, GGM and MCS may be more robust in the sense that they do not depend on a full distributional assumption. The following example illustrates several of the estimation algorithms discussed.

EXAMPLE 7.1. Let $U_1, \cdots, U_n$ be independent random variables and let U_r be normal with mean θ_r and variance θ_r^α, where $\theta_r > 0$ are unknown parameters and $\alpha > 0$ is a known constant. Assume that θ_r satisfies the parametric model $\theta_r = e^{\beta_1} Z_r^{\beta_2}$, $r = 1, \cdots, n$, where the Z_r are known constants.

We discuss below the GGM, MLE and MCS estimators of $\beta = (\beta_1, \beta_2)'$. More precisely, we consider the following five algorithms: 1. GGM; 2. MNR applied to MCS; 3. MNR with expected second order derivatives applied to MCS; 4. MNR applied to MLE; and 5. SCO. These five algorithms have the IGLS structure (3.5) so that it is only necessary to specify the elements a_r of the diagonal matrix $A(\theta)$ and the elements t_r of the vector $T(\theta)$. The matrix X (of dimension $n \times 2$) is the same in all cases and its elements are $X_{r1} = e^{\beta_1} Z_r^{\beta_2}$, $X_{r2} = (\log Z_r) X_{r1}$, $r = 1, \cdots, n$. The superscripts of a_r and t_r below indicate the number of the corresponding method of estimation. To make the coefficients easier to compare, we will

minimize half of the expression (7.5) when computing MCS.

(1) $t_r^{(1)} = -(U_r - \theta_r)/\theta_r^{\alpha}, a_r^{(1)} = \theta_r^{-\alpha},$

(2) $t_r^{(2)} = t_r^{(1)} - \alpha(U_r - \theta_r)^2/(2\theta_r^{\alpha+1}),$
$a_r^{(2)} = a_r^{(1)}(1 + 2\alpha w_r + \alpha(\alpha+1)w_r^2/2)$
with $w_r = (U_r - \theta_r)/\theta_r,$

(3) $t_r^{(3)} = t_r^{(2)}, a_r^{(3)} = a_r^{(1)}(1 + \alpha(\alpha+1)\theta_r^{\alpha-2}/2),$

(4) $t_r^{(4)} = t_r^{(2)} + \alpha/2\theta_r, a_r^{(4)} = a_r^{(2)} - \alpha/2\theta_r^2,$

(5) $t_r^{(5)} = t_r^{(4)}, a_r^{(5)} = a_r^{(3)} - \alpha/2\theta_r^2.$

As expected, MNR, applied to either MCS or MLE, generates a data dependent term a_r, which may be negative, possibly implying lack of convergence. This problem is avoided by using expected second derivatives, although this is likely to decrease the speed of convergence in well-behaved cases. The computational difficulty of MCS and MLE is about the same, so that the last should be preferred. The simplest computations are those corresponding to GGM. For $\alpha = 1(\alpha = 2)$ the computations may be performed through GLIM, using ERROR POISSON (ERROR GAMMA). Although the GGM is less efficient than MLE, it will be safer to use if the assumption of normality is suspect.

In this example M is linearizable since $\log \theta_r = \beta_1 + \beta_2 \log Z_r$. Any of the five algorithms may be then better implemented replacing (X_{r1}, X_{r2}) by $(1, \log Z_r)$, a_r by $\theta_r^2 a_r$, and t_r by $\theta_r t_r$.

ACKNOWLEDGMENTS

The author wants to thank the referees for their useful detailed comments, which led to a substantial rewriting of the manuscript. This work was partially supported by Dirección de Investigación Universidad Católica de Chile, under Grant 40-84. The revision of the manuscript was done while the author was visiting Carnegie Mellon University.

REFERENCES

AITKIN, M. (1981). A note on the regression analysis of censored data. *Technometrics* **23** 161–163.

ANDREWS, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16** 523–531.

AVRIEL, M. (1976). *Nonlinear Programming: Analysis and Methods.* Prentice Hall, Englewood Cliffs, N.J.

BAKER, R. J. and NELDER, J. A. (1978). *Generalized Linear Interactive Modelling* (Release 3). Numerical Algorithms Group, Oxford.

BATES, D. M. and WATTS, D. G. (1980). Relative curvature measures of nonlinearity (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 1–25.

BEATON, A. E. and TUKEY, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopical data. *Technometrics* **16** 147–185.

BERKSON, J. (1980). Minimum chi-square, not maximum likelihood! (with discussion). *Ann. Statist.* **8** 457–487.

BIRCH, J. B. (1980). Some convergence properties of iterated least squares in the location model. *Comm. Statist. B—Simulation Comput.* **9** 359–369.

BRADLEY, E. L. (1973). The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J. Amer. Statist. Assoc.* **68** 199–200.

BYRD, R. H. and PYNE, D. A. (1979). Some results on the convergence of the iteratively reweighted least squares algorithm for robust regression. *ASA Proc. Statist. Comput. Sec.* 87–90.

DEL PINO, G. E. (1984). Generalized linear models and iteratively weighted least squares. *Revista de la Sociedad Chilena de Estadistica* **1** 9–18. (In Spanish.)

DENNIS, J. E. and SCHNABEL, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, Englewood Cliffs, N.J.

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

FLETCHER, R. (1980). *Practical Methods of Optimization* **1**: *Unconstrained Optimization.* Wiley, New York.

FLETCHER, R. (1981). *Practical Methods of Optimization* **2**: *Constrained Optimization.* Wiley, New York.

GOLUB, G. H. (1969). Matrix decompositions and statistical computations. In *Statistical Computation* (R. C. Milton and J. A. Nelder, eds.). Academic, New York.

GREEN, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. Roy. Statist. Soc. Ser. B* **46** 149–192.

HOLLAND, P. W. and WELSCH, R. E. (1977). Robust regression using iteratively reweighted least squares. *Comm. Statist. A—Theory Methods* **6** 813–827.

HUBER, P. J. (1974). Numerical solution of robust regression problems. In COMPSTAT 1974 *Proc. Symposium on Computational Statistics* (G. Bruckmann, ed.). Physica, Vienna.

HUBER, P. J. (1977). *Robust Statistical Procedures.* SIAM, Philadelphia.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

JENNRICH, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40** 633–643.

JENNRICH, R. I. and MOORE, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *ASA Proc. Statist. Comput. Sec.* 57–65.

JORGENSEN, B. (1983). Maximum likelihood estimation and large sample inference for generalized linear and nonlinear regression models. *Biometrika* **70** 19–28.

JORGENSEN, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 127–162.

KALE, B. K. (1962). On the solution to likelihood equations by iteration processes: Multiparametric case. *Biometrika* **49** 479–486.

KASS, R. E. (1983). The rate of convergence of the Fisher scoring and Gauss-Newton algorithms. Technical Report 284, Dept. Statistics, Carnegie Mellon Univ.

LAWSON, C. L. and HANSON, R. J. (1974). *Solving Least Squares Problems.* Prentice Hall, Englewood Cliffs, N.J.

MCCULLAGH, P. (1983). Quasilikelihood functions. *Ann. Statist.* **11** 59–67.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MOORE, R. H. and ZEIGLER, R. (1967). The use of nonlinear regression methods for analyzing sensitivity and quantal response data. *Biometrics* **23** 563–566.

NELDER, J. A. and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* **74** 221–232.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

PETERS, S. C., KLEMA, V. C. and HOLLAND, P. (1978). Software for iteratively reweighted least squares computations. In *Proc. Computer Science and Statistics: Eleventh Annual Symposium on the Interface* (A. R. Gallant and T. M. Gerig, eds.) 380–384.

PREGIBON, D. (1982). Score tests in GLIM, with applications. *GLIM 82: Proc. International Conference on Generalised Linear Models. Lecture Notes in Statist.* **14**. Springer, New York.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York.

SCHMEE, J. and HAHN, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics* **21** 417–432.

THOMPSON, R. and BAKER, R. J. (1981). Composite link functions in generalized linear models. *Appl. Statist.* **30** 125–131.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447.

WU, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513.

Comment

Bent Jørgensen

del Pino is to be congratulated for his extensive survey of iterative least squares methods. In particular, I welcome the emphasis on the parallel between statistical properties of the model and the structure of the algorithm. This is where statistical computing distinguishes itself from the general area of optimization. An example of this is the role of orthogonality of parameters (cf. Cox and Reid, 1987), which implies an exact or approximate block diagonal structure of the Hessian of the log-likelihood function, with consequent simplification of the calculations. Another example is the discussion in Jørgensen (1984) of marginal and conditional maximum likelihood calculations.

Actually, I think that this marriage between algorithms and statistical theory will be taken much further in the future, and while, at the moment, iterative weighted least squares algorithms are probably the best general class of statistical algorithms available, I predict that the use of iterative least-squares methods will soon be changing. One of the driving forces in this development is the theory related to Barndorff-Nielsen's formula (cf. Barndorff-Nielsen, 1988; Reid, 1988 and references therein) and associated methods, such as saddlepoint approximations, modified profile likelihoods and so on. It is possible that these developments, in particular their geometric aspects, will lead to new and improved statistical algorithms.

To illustrate the potential influence of statistical theory on computing habits, consider the fact that the iterative weighted least-squares algorithm effectively ignores the second derivative of the model function h, denoted $E(\beta)$ by del Pino. On the other hand the theory associated with Barndorff-Nielsen's formula is effectively the systematic exploration of high-order derivatives of the likelihood, which certainly involves quantities such as $E(\beta)$. Hence, the advantage of iterative weighted least-squares methods, that $E(\beta)$ need not be calculated, will soon become unimportant, because $E(\beta)$ is needed for other purposes. In conclusion, statistical calculations involve much more than just the maximization of the likelihood or of some other objective function, and future statistical computer systems will to a larger extent than is the case today, involve a complete system of procedures for answering various types of inferential problems concerning the data. No doubt, automatic execution of symbolic mathematical calculations will play a crucial role in these developments.

In the meantime, I would like to mention some aspects of iterative weighted least-squares methods considered in Jørgensen (1984). There, I considered what I call the delta algorithm, which is nothing more than the iterative weighted least-squares algorithm with a general A-matrix, concentrating mainly on the case of a separable structure for the likelihood, and the possibility for implementing the algorithm in GLIM. The paper discussed the relation with various other algorithms and mentioned the algorithm for robust estimation considered by del Pino in connection with (3.10), which I referred to as the case of "score weights." In fact, this algorithm may be used in connection with any objective function and is not specific to robust estimation. Among other choices for A considered in Jørgensen (1984) was the case (referred to as "deviance weights"), which, in the language of generalized linear models, corresponds to a data-dependent link function, such that the objective function g becomes exactly quadratic. In other words, all the nonlinearity of the model is "thrown" into the link function. The point here is that there exists a

Bent Jørgensen is Visiting Professor, Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, 22460 Rio de Janeiro RJ, Brazil.