range of possible choices for A, which allows the construction of good algorithms for many different types of statistical models.

This brings me to the question of choice of algorithm, or, in the case of iterative weighted least squares, the choice of the matrix A. The choice of algorithm depends very much on the expected use of the algorithm, and there is a world of difference between an all-purpose algorithm and an algorithm tailored for a specific application. For example, the Fisher scoring algorithm may be considered a good general algorithm. However, in many specific applications, it is easy to find better algorithms, for example the algorithms based on the deviance weights or score weights mentioned above. Another example is the case of a convex objective function, for which the Newton–Raphson algorithm is the natural choice for a general algorithm. However, if the objective function is close to being nonconvex, as is the case for example for the hyperbolic distribution mentioned in Jørgensen (1984), the Newton–Raphson algorithm may become unstable, and, again, one of the two algorithms mentioned above may offer a more stable performance. An extreme case of this is L_1-estimation, where the Newton–Raphson algorithm fails, whereas the algorithm with score weights may be used.

Finally, I want to point out that our understanding of the relative performance of algorithms is still, at best, incomplete. I believe that the study of convergence, as practiced in the mathematics of optimization, is a fairly crude and incomplete tool for the understanding of the performance of algorithms, at least for statistical algorithms. For example, I have, until now, never seen a satisfactory explanation of the fact that Fisher's scoring algorithm works extremely well in the case of generalized linear models, as exemplified by GLIM. I have rarely seen an example of a generalized linear model where the algorithm diverges, in spite of the fact that no steplength calculation is performed (in GLIM), and the number of iterations to convergence is, in the majority of cases, around three to five. This is in contrast to the case of more general, non-exponential, models where the Fisher scoring algorithm may become excruciatingly slow, even when a steplength calculation is included. To draw a parallel, the simplex algorithm for linear programming is known to perform much better in praxis than expected on the basis of a worst-case analysis. Not surprisingly, at least to a statistical audience, a more complete understanding of the effectiveness of the simplex algorithm was obtained only after a probabilistic analysis of the algorithm was performed (cf. Borgwardt, 1987 and references therein). Similarly, I suspect that our understanding of the performance of iterative weighted least-squares algorithms will remain incomplete until a probabilistic analysis of the algorithm has been undertaken.

ADDITIONAL REFERENCES

BARNDORFF-NIELSEN, O. E. (1988). *Parametric Statistical Models and Likelihood. Lecture Notes in Statist.* **50**. Springer, New York.

BORGWARDT, K. H. (1987). *The Simplex Method. A Probabilistic Analysis. Algorithms and Combinatorics* **1**. Springer, New York.

COX, D. R. and REID, N. (1987). Parameter orthogonality and conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

JØRGENSEN, B. (1984). The delta algorithm and GLIM. *Internat. Statist. Rev.* **52** 283–300.

REID, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.* **3** 213–238.

Comment

Peter McCullagh

TERMINOLOGY

del Pino draws a distinction between *iteratively weighted least squares* (IWLS), in which the response vector $\mathbf{Y}$ is assumed to have a diagonal covariance matrix $\mathbf{V}$, and *iterative generalized least squares* (IGLS), in which $\mathbf{V}$ is an arbitrary covariance matrix. For purposes of exposition this distinction seems rather inconsequential, and, to my mind, insufficient to justify the usage of two four-letter acronyms. For numerical purposes, however, the savings in computational effort and organizational overhead resulting from the assumption of independence are very substantial. Thus, as the title suggests, the most useful distinction relates to computational organization rather than to conceptual issues.

Peter McCullagh is Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

ESTIMATING EQUATIONS VERSUS MINIMIZATION CRITERIA

del Pino is correct in his claim that the generalization of Gauss–Markov estimation is most naturally

expressed in geometrical terms as in (7.3) rather than in algebraic terms as the solution to a minimization problem. In essence (7.3) asserts that the residual vector must be orthogonal to the tangent space of M at the maximum likelihood point. It is not immediately obvious that any minimization is involved. What follows is an attempt to reconcile these two points of view.

The generalized Gauss–Markov estimator given by del Pino in equation (7.3) may be written in the form $\mathbf{U}(\hat{\boldsymbol{\beta}}; \mathbf{y}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \tag{1}$$

In this notation, points in the manifold M are expressed in the form $\boldsymbol{\mu}(\boldsymbol{\beta}) = E(\mathbf{Y}; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a p-dimensional vector. The tangent space of M at $\boldsymbol{\mu}(\boldsymbol{\beta})$ is spanned by the columns of $\mathbf{D}$, where $D_{ir} = \partial\mu_i/\partial\beta_r$. Thus generalized Gauss–Markov estimates or quasi-likelihood estimates are obtained as the root of what Godambe (1976) and Godambe and Thompson (1989) call an optimal linear estimating equation. The estimation equation, but not the estimate itself, is linear in $\mathbf{y}$.

If the components of $\mathbf{U}$ are denoted by U_r, it is ordinarily the case that

$$\frac{\partial U_r}{\partial \beta_s} \neq \frac{\partial U_s}{\partial \beta_r},$$

even though their expectations are equal. Thus, unless some conditions are imposed on the functional form of the matrix $\mathbf{V}$, $\mathbf{U}(\boldsymbol{\beta}; \mathbf{y})$ cannot be the gradient vector with respect to $\boldsymbol{\beta}$ of any scalar function. It is neither necessary nor sufficient that $\mathbf{V}$ be diagonal. It is unclear to me what statistical implications the asymmetry in the derivative matrix might have.

Although $\mathbf{U}(\boldsymbol{\beta}; \mathbf{y})$ cannot, in general, be the gradient vector of a scalar function, it is nevertheless possible to express the root of (1) as the solution to a minimization problem. We simply define $Q(\boldsymbol{\beta}; \mathbf{y})$ to be the squared length of the projection of $(\mathbf{y} - \boldsymbol{\mu})$ on to the tangent space of M at $\boldsymbol{\mu}(\boldsymbol{\beta})$. The appropriate projection matrix is

$$\mathbf{P}_V = \mathbf{D}(\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{V}^{-1}.$$

The weighted squared length of $\mathbf{P}_V(\mathbf{y} - \boldsymbol{\mu})$ is

$$\begin{aligned} Q(\boldsymbol{\beta}; \mathbf{y}) &= (\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}\mathbf{P}_V(\mathbf{y} - \boldsymbol{\mu}) \\ &= \mathbf{U}^T(\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{U}. \end{aligned} \tag{2}$$

This function is strictly positive except at the roots of (1). By contrast, as del Pino points out, minimization of Pearson's statistic

$$X^2(\boldsymbol{\mu}; \mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{3}$$

with respect to $\boldsymbol{\beta}$ does not ordinarily lead to consistent estimates, claims to the contrary by Berkson (1980) notwithstanding.

Note that (3) depends only on the point $\boldsymbol{\mu} \in R^n$, but not otherwise on the manifold M. By contrast, (2) depends not just on $\boldsymbol{\mu}$, but on the tangent space of M at $\boldsymbol{\mu}$.

For the purpose of setting approximate confidence limits or approximate Bayesian intervals for $\boldsymbol{\beta}$, or components thereof, it is tempting to treat $-\frac{1}{2}Q(\boldsymbol{\beta}; \mathbf{y})$ as *if* it were a log-likelihood function. Thus, if there are no nuisance parameters, the approximate confidence set for $\boldsymbol{\beta}$ is

$$\{\boldsymbol{\beta}: Q(\boldsymbol{\beta}; \mathbf{y}) < \chi^2_{p,\alpha}\}$$

at level $1 - \alpha$. In many cases this procedure gives sensible results. For example, the statistic has desirable invariance properties, and the χ^2_p approximation for $Q(\boldsymbol{\beta}; \mathbf{Y})$ is often quite accurate. There is, however, at least one undesirable property of the method that may be important in small samples. If the likelihood function is such that it has a minimum as well as a maximum, and if (1) is in fact the log-likelihood derivative, then (1) will have two roots, $\check{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$. By construction $Q(\boldsymbol{\beta}; \mathbf{y})$ is zero at both roots even though the likelihood at $\check{\boldsymbol{\beta}}$ may be negligible compared with the likelihood at $\hat{\boldsymbol{\beta}}$. Thus regions based on small values of (2) may include the least plausible parameter points as well as the most plausible points. In large samples these regions should be well separated, and no confusion should arise. In small samples the two regions may overlap, making it more difficult to present unambiguous results.

ADDITIONAL REFERENCES

GODAMBE, V. P. (1976). Conditional likelihood and optimum estimating equations. *Biometrika* **63** 277–284.

GODAMBE, V. P. and THOMPSON, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference* **22** 137–172.