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.*

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