# Rejoinder

**Guido del Pino**

My primary objective in writing this paper was just to illustrate the wide applicability of iterative generalized least squares and its special case IRLS as algorithms for parameter estimation. A secondary objective was to show the simplicity of the geometric concepts underlying the algorithm. I am indebted to the discussants for their contributions, which greatly enhance the value of the paper. Their comments point to interesting extensions and open new problems.

I must apologize to Professor Jørgensen for overlooking his important 1984 paper. The delta method essentially coincides with the IGLS algorithm (3.1), although he stressed the application to likelihood maximization. The score and deviance weights are, in fact, based on a quadratic approximation to the function $g$ being minimized (in this case, the negative of the log-likelihood). If $g$ and its gradient are computed for $n + 1$ points, and the quadratic approximation holds globally in a region containing this point, there are several ways of approximating the Hessian of $g$ at one of these points. In the separable case, the Hessian is diagonal and it may be estimated using just two points. The score and deviance weights are based on the current point and the point that minimizes $g$ without restrictions. These considerations, and the examples in Jørgensen (1984) show that one must keep in mind the possibility of employing alternative weights in IWLS.

I agree that the gain in simplicity obtained by neglecting the curvature of the manifold $M$ will become in time less important than the corresponding loss of efficiency. Even in that case, IGLS will provide good starting points for the more refined algorithms that take into account higher order properties of the likelihood.

Professor McCullagh rightly points out that the quasilikelihood estimator may be obtained as a solution to a minimization problem. In fact, any vector equation $G(\beta, y) = 0$ may be translated into the minimization of the function

$$G'(\beta, y)A(\beta)G(\beta, y),$$

where $A(\beta)$ is a positive definite matrix with the smallest eigenvalue bounded away from zero. Equation (2) of McCullagh is the special case

$$A(\beta) = V^{-1}(\mu(\beta)),\ \dot{G}(\beta, y) = U(\beta, y).$$

It seems worthwhile to explore the connections between estimating equations and GGM in further detail and I now address this point. Assume an iterative estimation procedure for estimating the unrestricted parameter $\theta$ is given and that we want to estimate $\theta$ under the restriction $\theta = h(\beta)$. For this purpose write the iterations in the unrestricted case as

$$\theta^{q+1} = \theta^q + S^q(Y) \tag{1}$$

and assume the following approximate equality holds for $\theta^q$ close to $\theta$:

$$ES^q(Y) \doteq \theta - \theta^q. \tag{2}$$

Regarding $\theta^q$ as fixed and writing

$$\mathrm{Var}(S^q(Y)) = V^q(\theta^q + \theta - \theta^q),$$

we are led to a GGM setup for the "working parameter" $\theta - \theta^q$ and the "working dependent variable" $S^q(Y)$. This suggests using (7.4) with $V^q(\theta^q)$ and $S^q(Y)$ substituted for $V(\theta^q)$ and $Y - \theta^q$, respectively.

In the original GGM formulation, the natural estimator of the mean vector $\theta$ is the data vector $Y$. This corresponds to $S^q(Y) = Y - \theta^q$, so that (2) holds exactly and $V^q(\theta) = \mathrm{Var}(Y)$.

An important special case of (1) is $S^q(Y) = S(\theta^q, Y)$, with $S$ continuous in its first argument. Any limit point of (1) is then a solution of the estimating equation

$$S(\theta, y) = 0 \tag{3}$$

and any limit point of the corresponding sequence $\beta^q$ is a limit point of

$$S(\theta(\beta), y) = 0.$$

The scoring method corresponds to the choice

$$S(\theta, Y) = I(\theta)^{-1}T(\theta, Y).$$

Under some regularity conditions, condition (1) will be satisfied for standardized unbiased estimating equations (Godambe, 1976).

Let me now comment on the meaning of $Q(\beta; y)$. For notational simplicity I will keep the dependence on $y$ implicit. Let $T(\theta)$ be the gradient of the log-likelihood $L(\theta)$ and let $U(\theta) = I(\theta)^{-1}T(\theta)$. In del Pino (1987), I considered the orthogonal projection, $U(\theta, M)$, of $U(\theta)$ onto the tangent subspace to $M$ at $\theta$. The square norm of $U(\theta, M)$, evaluated at $\theta = \theta_0$, is then the score test statistic for $\theta = \theta_0$ versus $\theta \neq \theta_0$, subject to $\theta \in M$. Orthogonality and norm must be understood with respect to the inner product $\langle\ ,\ \rangle_{I(\theta)}$ For an exponential family, parametrized in terms of the mean parameter $\mu$, $I(\mu) = V^{-1}(\mu)$, and $U(\mu, M)$ coincides with $P_V(y - \mu)$ (McCullagh's notation).

Hence $Q(\beta_0; y)$ is the score test statistic for $\beta = \beta_0$ versus $\beta \neq \beta_0$ (where $\theta_0 = \mu(\beta_0)$). From

$$\text{Var}(U(\theta)) = I(\theta)^{-1}, \quad \text{Var}((y - \mu)) = V(\mu),$$

it is clear that substituting $y - \mu$ and $\langle\ ,\ \rangle_{V^{-1}(\mu)}$ for $V(\theta)$ and $I(\theta)$, respectively, keeps the geometric structure intact. Extension to composite hypothesis must be done in terms of $C_\alpha$-tests (Neyman, 1959; Breusch and Pagan, 1980; Engle, 1981), since maximum likelihood estimators are not available from second order information alone.

Dr. Hill's comments deal with related topics that fall quite a bit outside the scope of this paper. I am afraid I cannot do justice to his interesting questions and will limit myself to a few remarks. From the point of view of quasilikelihood estimation, the mixture model is only used to provide the variance function of the marginal distribution. The key issue is then the dependence of this variance function on unknown parameters (like $\theta$ in the example given by him). Although its properties are not yet completely understood, the extended quasilikelihood method of estimation is, to my knowledge, the only general method that is based on just the variance function. Empirical comparisons of this method with alternatives like MLE would improve our understanding of its behavior. That kind of comparions has been done by Hill and Tsai (1988), although the emphasis there is on estimation of the regression parameters. From a theoretical point of view, the efficiency loss due to the restriction to first- and second-order information must be related to the curvature of the log-likelihood. As Hill says himself, this is also related to the problem of properly evaluating the precision of statistical estimators.

In conclusion, I would once again like to thank the discussants for their comments. I would particularly like to thank Professor Morris DeGroot for taking an active interest in the manuscript and for his encouragement through various revisions of this paper.

## ADDITIONAL REFERENCES

BREUSCH, T. S. and PAGAN, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Rev. Econom. Studies* **47** 239–253.

DEL PINO, G. (1987). A coordinate free approach to score tests. Presented at IMS meeting, San Francisco.

ENGLE, R. F. (1981). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics* (Z. Griliches and M. Intriligator, eds.). North-Holland, Amsterdam.

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

HILL, J. R. and TSAI, C. L. (1988). Calculating the efficiency of maximum quasilikelihood estimation. *Appl. Statist.* **37** 219–230.

JØRGENSON, B. The delta algorithm and GLIM. *Internat. Statist. Rev.* **52** 283–300.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics: The Harald Cramér Volume* (U. Grenander, ed.) 213–234. Wiley, New York.