On the Method of Scoring and Use of the Observed Information in Obtaining Maximum Likelihood Estimates

by

David Hinkley and George Runger

School of Statistics, University of Minnesota

Technical Report No. 323

June 1978

*This research was supported in part by National Science Foundation Grant MCS-77-00959.

On the Method of Scoring and Use of the Observed Information in Obtaining Maximum Likelihood Estimates

David Hinkley and George Runger
School of Statistics, University of Minnesota

Abstract

Statistical properties of the method of scoring and the Newton-Raphson method are contrasted and illustrated for estimation of the location of a Cauchy distribution. The use of observed information as weight when pooling estimates is discussed.

Key Words: Maximum likelihood estimate; Newton-Raphson; scoring; observed information; curvature; Cauchy distribution; asymptotics; Fisher.

1. Introduction.

The method of scoring was suggested by R. A. Fisher (1925) as a simple means of approximating a solution to the likelihood equation. Starting with a root-n consistent estimator, one application of the method of scoring gives a (first-order) efficient estimator. Fisher claimed that a second iteration would produce a second-order efficient statistic. Curiously, Fisher did not discuss the Newton-Raphson method of solving the likelihood equation, which in effect uses the observed information instead of the expected information in the method of scoring. One possible explanation is the relative numerical simplicity of the scoring method in 1925.

There are other methods of solving the likelihood equation and of maximizing the likelihood function. Some of these are described by Barnett (1966) and in elementary numerical analysis texts. In this note we are concerned solely with the method of scoring and the Newton-Raphson method. The object is to document basic statistical properties of these two methods (§2) and to give illustrative numerical results for estimation of the center of a Cauchy distribution (§3). A secondary aspect of the work is an empirical demonstration of the importance of combining estimates using observed, not expected, informations as weights, as described by Fisher (1925).

2. Outline of Theory.

2.1 The Scoring Methods

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed according to probability density $f_\theta(x)$, so that the log likelihood

is

$$\ell_\theta = \sum_{j=1}^{n} \log_e f_\theta(x_j) \quad .$$

We assume $f_\theta(x)$ to be regular, possessing continuous third derivatives with respect to θ. We denote first and second derivatives of ℓ_θ with respect to θ by $\dot{\ell}_\theta$ and $\ddot{\ell}_\theta$. The maximum likelihood estimate $\hat{\theta}$ of θ is that solution to $\dot{\ell}_\theta = 0$ which maximizes ℓ_θ. Two important quantities of interest are the observed and expected informations,

$$I = -\ddot{\ell}_{\hat{\theta}} \quad \text{and} \quad \mathcal{I}_\theta = E(-\ddot{\ell}_\theta) \doteq E(I) \quad . \tag{1}$$

Fisher (1925, §5) indicated the following method for approximating $\hat{\theta}$. Let T_0 be an initial (consistent) estimate of θ. Then

$$T_1 = T_0 + \dot{\ell}_{T_0} / \mathcal{I}_{T_0} \tag{2}$$

is efficient, in the sense that $\mathrm{Var}(T_1)/\mathrm{Var}(\hat{\theta}) \to 1$ as $n \to \infty$. This iterative method is known now as the "method of scoring" (Rao, 1974, p. 366). It is based on the truncated expansion

$$\dot{\ell}_{\hat{\theta}} \doteq \dot{\ell}_{T_0} + (\hat{\theta} - T_0)\ddot{\ell}_{T_0} \quad , \tag{3}$$

with substitution of the average $\mathcal{I}_{T_0}$ for $-\ddot{\ell}_{T_0}$. The obvious alternative to T_1 is the Newton-Raphson approximation for $\hat{\theta}$

$$T_1' = T_0 + \dot{\ell}_{T_0} / (-\ddot{\ell}_{T_0}) \quad . \tag{4}$$

Both procedures (2) and (4) can be repeated, thus generating the iterative schemes

$$T_{j+1} = T_j + \dot{\ell}_{T_j} / \mathcal{I}_{T_j} \qquad (j = 0,1,\ldots) \tag{5}$$

$$T'_{j+1} = T'_j + \dot{\ell}_{T'_j} / (- \ddot{\ell}_{T'_j}) \qquad (j = 0,1,\ldots) \tag{6}$$

Fisher (1925, §13) claims that T_2 is second-order efficient, which would imply that as $n \to \infty$

$$n^2\{\mathrm{Var}(T_2) - \mathrm{Var}(\hat{\theta})\} \to 0$$

in contrast with the first-order efficiency of T_1 which implies only

$$n\{\mathrm{Var}(T_1) - \mathrm{Var}(\hat{\theta})\} \to 0 \ .$$

2.2 Convergence Rates

It is well known in numerical analysis that $\{T_j\}$ "converges linearly" and that $\{T'_j\}$ "converges quadratically" to a root $\hat{\theta}$ of $\dot{\ell}_\theta = 0$. To obtain detailed results about the stochastic nature of $\{T_j\}$ and $\{T'_j\}$ we carry out Taylor series expansions for the derivatives in (5) and (6) to obtain

$$T_{j+1} - \hat{\theta} \doteq (T_j - \hat{\theta})\{(1 - \frac{I}{\dot{\ell}_{\hat{\theta}}}) + \frac{1}{2}(T_j - \hat{\theta}) \frac{\dddot{\ell}_{\hat{\theta}}}{\dot{\ell}_{\hat{\theta}}}\} \tag{7}$$

$$T'_{j+1} - \hat{\theta} \doteq - \frac{1}{2}(T'_j - \hat{\theta})^2 \frac{\dddot{\ell}_{\hat{\theta}}}{I} \ , \tag{8}$$

where $T'_0 = T_0$ is chosen to differ from $\hat{\theta}$ by an amount of order $n^{-1/2}$. Clearly, the convergence rates depend on whether or not $E(\dddot{\ell}_\theta) = 0$. A more detailed analysis requires the following results.

<u>Lemma.</u> Let $\mu_a = E\left\{\frac{\partial^a \log f_\theta(X)}{\partial \theta^a}\right\}$ $(a = 2,3)$ and define

$$\sigma_{ijk} = E\left[\left\{\frac{\partial \log f_\theta(X)}{\partial \theta}\right\}^i \left\{\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2} - \mu_2\right\}^j \left\{\frac{\partial^3 \log f_\theta(X)}{\partial \theta^3} - \mu_3\right\}^k\right] ,$$

$$\gamma^2 = (\sigma_{020} - \sigma_{110}^2/\sigma_{200})/\sigma_{200}^2$$

$$\delta^2 = (\sigma_{002} - \sigma_{101}^2/\sigma_{200})/\sigma_{200}^2 \quad .$$

Then as $n \to \infty$

$$1\text{-}I/\vartheta_{\hat{\theta}} \sim N(0, n^{-1}\gamma^2) \tag{9}$$

and

$$\dddot{\ell}_{\hat{\theta}}/I - n\mu_3/\vartheta_\theta \sim N(0, n^{-1}\delta^2) \quad . \tag{10}$$

Proof. Result (9) is proved by Efron and Hinkley (1978), and (10) may be proved similarly.

If $\mu_3 = 0$ we deduce from (7) through (10) that

$$\frac{T_{j+1} - \hat{\theta}}{T_j - \hat{\theta}} \sim N(0, n^{-1}\gamma^2) \tag{11}$$

and

$$\frac{T'_{j+1} - \hat{\theta}}{(T'_j - \hat{\theta})^2} \sim N(0, \tfrac{1}{4}n^{-1}\delta^2) \quad . \tag{12}$$

Notice that $T'_1 - \hat{\theta}$ and $T_2 - \hat{\theta}$ are of the same order, $n^{-3/2}$. When $\mu_3 \neq 0$, $T'_1 - \hat{\theta}$ is of order n^{-1}. There is, then, an advantage to the parametrization for which $\mu_3 = 0$. Since γ is invariant under reparametrization, the order of the convergence rate of scoring cannot be affected.

The results (11) and (12) give some quantitative indication of convergence rates, but we find in our numerical example that convergence is somewhat faster than the variances in (11) and (12) would predict.

Monte Carlo results for scoring and Newton-Raphson in estimating the location parameter of a Cauchy distribution are described in Section 3.

2.3 Combining Independent Estimates

A somewhat related point discussed by Fisher (1925) has to do with pooling estimates. Suppose that we have m samples of the type previously discussed, the kth sample size being $n_k \equiv n$. Denote the corresponding values of $\hat{\theta}$, I and $\mathcal{I}_\theta$ by $\hat{\theta}_k$, I_k and $\mathcal{I}_k \equiv \mathcal{I}$. Fisher suggests that, if the overall maximum likelihood estimator is $\hat{\theta}_0$ (from the pooled likelihood), then

$$T(I) = \Sigma\, \hat{\theta}_k I_k / \Sigma\, I_k \tag{13}$$

is equivalent to $\hat{\theta}_0$ to second order, whereas

$$T(\mathcal{I}) = \Sigma\, \hat{\theta}_k \mathcal{I}_k / \Sigma \mathcal{I}_k = m^{-1} \Sigma \hat{\theta}_k \tag{14}$$

is equivalent to $\hat{\theta}_0$ only to first order. That is,

$$\frac{\mathrm{Var}\{T(I)\} - \mathrm{Var}(\hat{\theta}_0)}{\mathrm{Var}\{T(\mathcal{I})\} - \mathrm{Var}(\hat{\theta}_0)} \text{ is of order } n^{-1} ;$$

the appropriateness of $T(I)$, in particular the use of I^{-1} as a variance approximation for $\hat{\theta}$, is also discussed by Efron and Hinkley (1978). For the case where $E(\hat{\theta}) = \theta$, Efron (1975, equ. (10.1)) gives

$$\mathrm{Var}(\hat{\theta}_k) = \frac{1}{\mathcal{I}_k}\{1 + \gamma^2/n_k + O(n_k^{-2})\} \quad (k = 0,1,\ldots,m)$$

with $n_0 = mn$, $\mathcal{I}_0 = m\mathcal{I}$ and γ^2 as in the lemma. Therefore,

$$\mathrm{Var}\{T(\mathcal{I})\} - \mathrm{Var}(\hat{\theta}_0) = \frac{1}{\mathcal{I}_0} \cdot \frac{\gamma^2}{n}\left(1 - \frac{1}{m}\right) + O\{(mn)^{-2}\} . \tag{15}$$

Some Monte Carlo results comparing $T(I)$ with $T(\mathcal{I})$ in the Cauchy location case are given in Section 3.

3. Monte Carlo Results for Cauchy Location Family

We consider now the case of the Cauchy location family

$$f_\theta(x) = \pi^{-1}\{1 + (x-\theta)^2\} ,$$

for which $\mathcal{I}_\theta \equiv \frac{1}{2}n$ and $\gamma_\theta^2 = \frac{5}{2}$. The Monte Carlo results summarized by Efron (1975) suggest that second-order asymptotics are good approximations for $n \geq 15$.

Monte Carlo Cauchy samples were generated on the CYBER 74 computer using ratios of $N(0,1)$ variables constructed by the Marsaglia-Bray method. In computing estimates of variance for the various statistics the "location swindle" was used (Simon, 1976); the "scale swindle" cannot be used when dealing with likelihood derivatives. We used 5,000 samples of size $n = 15$ to obtain the tables given below.

It is important to note that we take $\hat{\theta}$ to be the value of T_4', i.e. the result of four Newton-Raphson iterations, starting with T_0 = sample median.

Table 1 gives estimates of $\mathrm{Var}(T_j)$ and $\mathrm{Var}(T_j')$, with standard errors of these estimates in brackets. These numbers may be compared to

$$\frac{1}{\text{Fisher information in } T_0} = 0.1838 ,$$

given by Fisher (1925),

$$\frac{1}{\text{total Fisher information } \mathcal{I}} = 0.1333 ,$$

$$\frac{1}{\text{Fisher information in } \hat{\theta}} = \frac{1}{\mathcal{I}(1+n^{-1}\gamma^2)} = 0.1556$$

given by Efron (1975). Table 2 gives corresponding estimates of $\mathrm{Var}(T_j-\hat{\theta})$ and $\mathrm{Var}(T_j'-\hat{\theta})$, indicating the superiority of T_j'.

Table 1. Estimates of Variance for Iterative Approximations to M.L.E. of Cauchy Location, n = 15 .

j	T_j	T_j'
0	0.1920 (.0055)	0.1920 (.0055)
1	0.1629 (.0047)	0.1622 (.0054)
2	0.1599 (.0046)	0.1584 (.0048)
3	0.1592 (.0046)	0.1583 (.0048)

Table 2. Estimates of Variance for Errors of Iterative Approximations to M.L.E. of Cauchy Location, n = 15 .

j	$T_j-\hat{\theta}$	$T_j'-\hat{\theta}$
0	0.0369 (18×10^{-4})	0.0369 (18×10^{-4})
1	0.0064 (64×10^{-5})	0.0024 (13×10^{-4})
2	0.0024 (29×10^{-5})	66×10^{-6} (26×10^{-6})
3	0.0013 (20×10^{-5})	1×10^{-9} (5×10^{-10})

The third table relates to the relative convergence rates of equations (7), (8), (11) and (12). We have denoted $(T_{j+1}-\hat{\theta})/(T_j-\hat{\theta})$ by R_{j+1} and $(T'_{j+1}-\hat{\theta})/(T'_j-\hat{\theta})^2$ by R'_{j+1} . By (11) and (12) the theoretical values of $\mathrm{Var}(R_j)$ and $\mathrm{Var}(R'_j)$ are respectively 0.167 and 0.067, clearly overestimating the observed variances.

Table 3. Estimates of Mean and Variance for Relative Errors R_j, R_j' of Iterative Approximations to M.L.E. of Cauchy Location, n = 15.

	R_j		R_j'	
j	mean	variance	mean	variance
1	-6.1×10^{-2}	8.2×10^{-2}	2.6×10^{-2}	2.7×10^{-2}
2	5.3×10^{-2}	1.8×10^{-2}	3.4×10^{-2}	2.1×10^{-2}

Finally, Table 4 concerns the pooling of estimates described in Section 2.3. Here $m = 4$ samples of size $n = 15$ are pooled. The table compares $T(I)$ and $T(\mathcal{I})$, as defined in (13) and (14), with the overall m.l.e. $\hat{\theta}_0$ (obtained by Newton-Raphson iteration from $T(\mathcal{I})$ using the combined sample log likelihood). The first term on the right of (15) is 0.0042.

Table 4. Comparison of Pooled Estimates and M.L.E. from m = 4 Cauchy Samples of Size n = 15.

	$\mathrm{Var}\{T(\mathcal{I})\}$	$\mathrm{Var}\{T(I)\}$	$\mathrm{Var}(\hat{\theta}_0)$	$\mathrm{Var}\{T(\mathcal{I})-\hat{\theta}_0\}$	$\mathrm{Var}\{T(I)-\hat{\theta}_0\}$
estimate	0.0392	0.0348	0.0339	0.0059	0.0012
st. err.	0.0017	0.0015	0.0014	0.0006	0.0002

These numerical results confirm the theoretical support for the use of observed information I both in iterative estimation of θ and in pooling estimates of θ.

References

Barnett, V. D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. Biometrika, 53, 151-165.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Statist., 3, 1189-1242.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimate: observed versus expected information. Biometrika, 65, (to appear).

Fisher, R. A. (1925). Theory of statistical estimation. Proc. Camb. Phil. Soc., 22, 700-725.

Rao, C. R. (1974). Linear Statistical Inference and Its Applications (2nd Edition), New York: Wiley.

Simon, G. (1976). Computer simulation swindles with applications to estimates of location and dispersion. App. Statist., 25, 266-274.