On iterative procedures of asymptotic inference

by K. O. DZHAPARIDZE*

Abstract An informal discussion is given on performing an unconstrained maximization or solving non-linear equations of statistics by iterative methods with the quadratic termination property. It is shown that if a miximized function, e.g. likelihood, is asymptotically quadratic, then for asymptotically efficient inference finitely many iterations are needed.

Key Words: methods of NEWTON-RAPHSON, scoring, quasi-NEWTON, DAVIDON-FLETCHER-POWELL, conjugate gradient; quadratic termination; asymptotically differentiable; asymptotically quadratic.

In this paper we briefly discuss some applications of modern iteration methods of numerical analysis to the problems of mathematical statistics.

Certain applications of the most basic iteration method of NEWTON-RAPHSON (or its stochastic modification - *the scoring method*) have been well-known since FISHER, and are included in many statistical textbooks (see, e.g., KENDALL and STUART (1961), Section 18.31; RAO (1965), Section 5g; ZACKS (1971), Section 5.2).

Although *the* NEWTON-RAPHSON *method* is theoretically very attractive, it may turn out to be highly unsuitable in practice, especially when the number of unknown parameters, involved in the statistical model under study, is large.

In order to mitigate some of the computational difficulties, unavoidable in *the* NEWTON-RAPHSON *method,* various developments of this method are intensively discussed in the literature on numerical analysis. The most important are the so-called *quasi*-NEWTON methods, and their alternatives, the *conjugate gradient methods.*

We intend to demonstrate here that the application of certain stochastic modifications of this kind of methods will, in general, lead to a statistical inference which is at least as efficient as that of the NEWTON method. It should be noted, however, that the considerations presented below are highly informal, as they are in fact aimed at showing why the above statement should be true, rather then at proving strict mathematical results (to be found, in principle, in the enclosed references).

Returning to FISHER's ideas let us recall that he has applied the NEWTON-RAPHSON method to the classical problem of estimating the unknown parameter θ involved in the distribution F_θ, when a sample

$$X_1, \dots, X_n \tag{1}$$

is drawn from a population specified by this distribution function F_θ.

Assuming that the population is of the continuous type and f_θ is the density of F_θ, FISHER (1925) used the NEWTON-RAPHSON method for maximizing the loglikelihood function

$$L_n(\theta) = L_n(X_1, \dots, X_n; \theta) = \sum_{i=1}^{n} \log f_\theta(X_i). \tag{2}$$

[*] Department of Mathematical Statistics, The Mathematical Centre, Amsterdam, on grant B62-165 of the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Attractiveness and universality of the maximum likelihood method is justifiable by the existence, under fairly wide conditions, of a value of θ that renders the loglikelihood (2) as large as possible, at least when the sample size n is sufficiently large. Conditions under which the maximizing value of θ – the maximum likelihood estimator $\hat{\theta}_n$ – is $\sqrt{n}$-consistent are also fairly broad.

By $\sqrt{n}$-consistency of $\hat{\theta}_n$ we mean that the sequence of the distributions

$$\mathscr{L}\{\sqrt{n}(\hat{\theta}_n-\theta)\}, \qquad n=1,2,\ldots \tag{3}$$

converges to a non-degenerate distribution.

Moreover, additional conditions guarantee that the limit of (3) is Gaussian with mean zero and variance, given by the reciprocal of FISHER's information amount I_θ per single observation, that is

$$\mathscr{L}\{\sqrt{n}(\hat{\theta}_n-\theta)\}\Rightarrow N(0,I_\theta^{-1}). \tag{4}$$

After FISHER, we can therefore call $\hat{\theta}_n$ asymptotically efficient. These and some further theoretical properties provide the basis for *"a quasi-hypnotic attraction the m.l. estimates seem to exert"* (LECAM (1960), p. 94).

However in practice complications may arise when one starts to maximize the loglikelihood (2) by, for instance, looking for roots of the corresponding likelihood equation

$$(\partial/\partial\theta)L_n(X_1,\ldots,X_n;\theta)=0 \tag{5}$$

(if there are any for fixed n), especially if this equation turns out to be highly nonlinear (as frequently happens).

The additional task of choosing an appropriate root among several of them is also difficult.

Aware of these problems, FISHER (1925) suggested looking for iterative solutions of the equation (5) (or the corresponding maximization problem) by the NEWTON method defined as iterations

$$\theta_n^{i+1}=\theta_n^i-\left(\frac{\partial^2L_n}{\partial\theta^2}\right)^{-1}_{\theta_n^i}\left(\frac{\partial L_n}{\partial\theta}\right)_{\theta_n^i}, \qquad i=0,1,\ldots, \tag{6}$$

Alternatively, observing that

$$-\frac{1}{n}\frac{\partial^2L_n}{\partial\theta^2}\to I_\theta \tag{7}$$

in probability (under θ), he suggested also the asymptotically equivalent procedure of scoring

$$\theta_n^{i+1}=\theta_n^i+\frac{1}{n}(I_\theta^{-1})_{\theta_n^i}\left(\frac{\partial L_n}{\partial\theta}\right)_{\theta_n^i}, \qquad i=0,1,\ldots \tag{8}$$

Besides, he pointed out that if the starting value θ^0 is any $\sqrt{n}$-consistent estimator for θ (for instance, constructed, when it is possible, by using the method of moments), then the result of the very first iteration, θ_n^1, is an estimator for θ as efficient asymptotically as the maximum likelihood estimator $\hat{\theta}_n$.

Indeed, FISHER did not worry about the mathematical accuracy of his statements. The first careful treatment of the subject containing a further study of asymptotic properties of the estimator θ_n^1, is due to LECAM (1956).

Later, LECAM (1960) extended his studies to a considerably more general class* of experiments than those generated by independent identically distributed (i.i.d.) observations:
the function $L_n(X_1, ..., X_n; \theta)$ was treated as a general loglikelihood function and not necessarily that of the i.i.d. observations (as in (2)). He observed that for sufficiently large n the Taylor expansion of L_n involves significant terms which are related to the first and second order derivatives of L_n only, as all other terms become asymptotically negligible when $n \to \infty$. That is,

$$L_n(\theta + h/\sqrt{n}) - L_n(\theta) = h\Delta_n(\theta) - \tfrac{1}{2}h^2 I_\theta + o_p(1) \tag{9}$$

where

$$\Delta_n(\theta) = \frac{1}{\sqrt{n}} \frac{\partial L_n}{\partial \theta} + o_p(1) \tag{10}$$

and I_θ is the stochastic limit of the second derivative of $(1/n)L_n$ with opposite sign (recall (7)), while $o_p(1)$ in (9) and (10) (or anywhere below) stands for those terms which are asymptotically negligible in the sense that they tend to 0 stochastically as $n \to \infty$. Thus Δ_n can be sought** as a *principal part* of $n^{-1/2}(\partial/\partial\theta)L_n$.

Further, under fairly wide conditions the random variable $\Delta_n(\theta)$ is asymptotically normal:

$$\mathscr{L}(\Delta_n(\theta)) \Rightarrow N(0, I_\theta) \tag{11}$$

and asymptotically differentiable in the sense, that if θ_n^0 is any $\sqrt{n}$-consistent estimator for θ, then

$$\Delta_n(\theta_n^0) - \Delta_n(\theta) = I_\theta \sqrt{n}(\theta_n^0 - \theta) + o_p(1). \tag{12}$$

Note that in the case of i.i.d. observations (11) is a simple consequence of the central limit theorem and the well-known fact that

$$\frac{1}{n} E\left(\frac{\partial L_n}{\partial \theta}\right)^2 = I_\theta.$$

[*] Deviating from the i.i.d. case, one often encounters situations in which the formulae (9)-(11) below hold with some differential $\delta_n > 0$ such that $\delta_n \to 0$, different from $1/\sqrt{n}$, and this is taken into account in the later works of LECAM (1969), (1974).
It should be noted also, that in the case of a vector-valued parameter θ the normalization of each component by $\sqrt{n}$ (to be discussed below) often fails: these components even may have different rates of convergence, and then the normalization by some positive definite matrix with a vanishing (as $n \to \infty$) norm is needed (see IBRAGIMOV and HAS'MINSKII (1981) where an excellent treatment of estimating problems can be found, in the spirit of LECAM).

[**] In the theory of i.i.d. observations Δ_n is usually taken to be equal to the first term on the right-hand side of (10). In more complicated situations, however, it often happens that $n^{-1/2}(\partial/\partial\theta)L_n$ cannot be obtained explicitly, or it is too complicated to be used in practice, while its *principal part*, Δ_n, can be chosen among asymptotically equivalent candidates as simple and smooth as possible to ensure, in particular, asymptotic relations of type (12).

(In general this last equation holds only asymptotically when $n \to \infty$).

Equation (12) also has a natural interpretation in terms of the derivatives of L_n.

The equations (11) and (12) have a very important consequence.

Proposition 1
If (11) *and* (12) *hold, the estimator*

$$\theta_n^1 = \theta_n^0 + \frac{1}{\sqrt{n}} I_{\theta_n^0}^{-1} \Delta_n(\theta_n^0) \tag{13}$$

is asymptotically normal:

$$\mathscr{L}(\sqrt{n}(\theta_n^1 - \theta)) \Rightarrow N(0, I_\theta^{-1}). \tag{14}$$

Note the similarity of (13) and (8) with $i = 0$, and also the coincidence of the right-hand sides of (4) and (14).

The proof is very simple: By (12) and (13),

$$\sqrt{n}(\theta_n^1 - \theta) = \sqrt{n}(\theta_n^0 - \theta) + I_{\theta_n^0}^{-1}[\Delta_n(\theta) - I_\theta \sqrt{n}(\theta_n^0 - \theta)] + o_p(1).$$

If we now replace θ_n^0 in $I_{\theta_n^0}^{-1}$ by θ (this is justifiable if I_θ is continuous in θ), then

$$\sqrt{n}(\theta_n^1 - \theta) = I_\theta^{-1} \Delta_n(\theta) + o_p(1).$$

Hence (14) is an immediate consequence of (11).

According to this proposition the estimator θ_n^1 has the same asymptotic properties as the maximum likelihood estimator. In other words, instead of looking for the maximum likelihood estimators $\hat{\theta}_n$ one can use without loss of efficiency (at least for samples of large size n) the following two-step* procedure:

(i) construct a preliminary estimator θ_n^0 of θ satisfying $\sqrt{n}$-consistency, and then
(ii) defining for the particular problem under study $\Delta_n(\theta)$ and I_θ from a corresponding likelihood function $L_n(\theta)$, construct θ_n^1 as indicated in (13).

It should be noted that, in principle, this alternative procedure *"applies also to cases that certain authors may deem pathological - cases in which m.l. estimates do not behave or do not exist. This is somewhat irrelevant. What is relevant is that statistical life is plagued with situations involving dependent variables or other more or less complicated situations in which it seems to be a waste of time to try to prove that m.l. estimates do behave. Even in cases in which the m.l. estimates are asymptotically well behaved it may be preferable not to use them"* (LeCam (1960), Appendix II).

That seems to be why the just cited *"author is firmly convinced that a recourse to maximum likelihood is justifiable only when one is dealing with families of distributions that are extremely regular. The cases in which m.l. estimates are easily obtainable and have been proved to have good properties are extremely restricted. One of the purposes of*

* Obviously this procedure can be used iteratively by continuing as in (8). That is why Fisher called the method of estimation by formula (8) the *scoring method* (the word *scoring* is used here to stress that the procedure scores iteratively the corrections).

this paper (LeCam (1960)) *is precisely to deëmphasize the role of m.l. estimates".*

"The drawback in having a liberal amount of flexibility in the choice of the estimates is that one is likely to have to choose between radically different formulas which all lead to the same asymptotic properties. From a practical point of view, it should be emphasized that a purely asymptotic theory does not say anything about a particular problem. The standard practice of letting a parameter tend to infinity is a mathematical device which leads to fairly simple theorems..."

The reason for such an extensive quotation should become clear below, for we shall now follow *"the standard practice of letting the sample size n tend to infinity"*, and define alternative procedures of estimation which lead to the same asymptotic properties as that of m.l. of Newton-Raphson (scoring). Also, the procedures defined below can in fact be utilized under the same circumstances as the two-step procedure mentioned above, so that the reasonings of LeCam concerning the latter procedure could in principle be applied to the case we shall discuss.

Observe meanwhile that the considerations which are followed above can be easily extended to the s vector-valued parameter case when $\Delta_n(\theta)$ is an s vector-valued random variable and I_θ is a positive definite $(s \times s)$-matrix.

In this case the application of formula (13) (or the iterative procedures of type (6) and (8)) requires the inversion of $(s \times s)$-matrices. This may be difficult, when the number of unknown parameters, s, is large.

It is natural to try also other methods of unconstrained maximization (or solving essentially nonlinear system of corresponding equations) provided by modern numerical analysis.

The justification of nearly all such methods is based on the presumption that the maximized quantity, in a neighborhood of a maximum point, can be well approximated by a quadratic function. Thus a number of methods are advanced in numerical analysis which efficiently maximize quadratic functions, in the hope that they do perform well on more general functions at least in a neighborhood of a maximum point. This motivation leads in the first place to the derivation of Newton's method which gives the maximum of a quadratic function*, $c + b^T x - \frac{1}{2} x^T A x$ say, after the very first iteration, $x^1 = A^{-1} b$, for any initial value x^0.

Also, the extension of the classical Newton method mentioned in the beginning of this paper, such as the quasi-Newton methods and conjugate gradient methods, possess a special property with respect to quadratic functions: the maximum is found in at most s iterations where s is the number of unknowns. Therefore, it is often said that these methods possess the property of *quadratic termination.*

On the orher hand, in view of the asymptotic relation (9) the function L_n can be regarded as *"asymptotically quadratic"*.

* As for the maximization of a general function, the nice feature of Newton's method consists in the fact that when the iterations do converge, the rate of convergence is quadratic. However, Newton's iterations often fail to converge - when the results are far from a maximum point difficulties may arise. Nevertheless, the attraction of the quadratic convergence, in a neighborhood of a maximum, keeps all methods as close to Newton's iterations as possible, only introducing modifications to gain more reliability.

Basically, this determines the fine asymptotic properties of the first iteration in (6) (or (8)) as an estimator of θ, specifically, the property determined by (14). Realizing these facts one should come to the conjecture that the *quadratic termination* property of a utilized method ought to guarantee the same asymptotic properties for the result of at most s iterations treated as the estimator for θ.

An attempt in this direction is made in BEINICKE and DZHAPARIDZE (1982), where our conjecture is confirmed for the special method of DAVIDON-FLETCHER-POWELL (DFP), which is one of a family of quasi-NEWTON methods.

The concept of a quasi-NEWTON method for the solution of the system (5), with $(\partial/\partial\theta)$ to be understood now as the gradient vector (or for the maximization of $L_n(\theta)$), consists of an algorithm which proceeds as follows. Choosing the *initial value* (any $\sqrt{n}$-consistent estimator for θ) θ_n^0 beforehand, along with a symmetric positive definite matrix H_n^0 (for instance, H_n^0 can be chosen as the $s\times s$ unit matrix), at iteration j, define

$$\theta_n^{j+1}=\theta_n^j+\frac{1}{\sqrt{n}}a_n^jH_n^j\Delta_n(\theta_n^j) \tag{15}$$

where a_n^j is determined by an *exact line search,* that is, it is a scalar determined as the value a that maximizes the function

$$L_n\left(\theta_n^j+\frac{1}{\sqrt{n}}aH_n^j\Delta_n(\theta_n^j)\right).$$

Neglecting again the omitted terms in (9) and replacing $I_{\theta_n^j}$ by a consistent estimator I_n^* for I_θ (by $I_{\theta_n^0}$, say), we get

$$a_n^j=\frac{\Delta_n^T(\theta_n^j)H_n^j\Delta_n(\theta_n^i)}{\Delta_n^T(\theta_n^j)H_n^jI_n^*H_n^j\Delta_n(\theta_n^j)} \tag{16}$$

(Δ_n^T denotes the transpose of Δ_n).

As for the matrices $H_n^j, j=1,2,\ldots$ in (15) and (16), they must possess the property

$$H_n^{j+1}q_n^j=r_n^j \tag{17}$$

where $r_n^j=\sqrt{n}(\theta_n^{j+1}-\theta_n^j)$, $q_n^j=-[\Delta_n(\theta_n^{j+1})-\Delta_n(\theta_n^j)]$.

The following specific choice of the matrices $H_n^j, j=1,2,\ldots$, satisfying (17) determines the DFP method (see, e.g. ORTEGA and RHEINBOLDT (1979));

$$H_n^{j+1}=H_n^j+\frac{r_n^j(r_n^j)^T}{(r_n^j)^Tq_n^j}-\frac{H_n^jq_n^j(H_n^jq_n^j)^T}{(q_n^j)^TH_n^jq_n^j}. \tag{18}$$

Theorem 1 now shows the ability of a stochastic modification of the DFP method to produce asymptotically efficient estimators:

Theorem 1

If (11) and (12) are satisfied, then the estimator θ_n^s defined by (15), (16) and (18) (s being the number of unknowns) is asymptotically normal; more specifically,

$$\mathscr{L}(\sqrt{n}(\theta_n^s-\theta))\Rightarrow N(0,I_\theta^{-1}). \tag{19}$$

In addition, H_n^s is a consistent estimator for the inverse of FISHER'S *information matrix I_θ per single observation.*

The proof of this result can be found in BEINICKE and DZHAPARIDZE (1982). Note that the considerations of this paper are based on the definition (18) of the matrices H_j, $j=1,2,...,$ while, in general, results of DIXON (1972) allow extensions to the full BROYDEN *family* (see, e.g. BRODLIE (1977)).

Following considerations similar to those of BEINICKE and DZHAPARIDZE (1982), the former author has shown in his Ph.D. thesis at Tbilisi State University (1979) that the *conjugate gradient* method, appropriately modified, leads to an analoguous result. Specifically, the following theorem holds.

Theorem 2

Define the stochastic modification of the conjugate gradient iterations:

$$\theta_n^{i+1} = \theta_n^i - \frac{1}{\sqrt{n}} \alpha_n^i p_n^i$$

where

$$\alpha_n^i = \frac{(\Delta_n(\theta_n^i))^T p_n^i}{(p_n^i)^T I_n^* p_n^i},$$

$$p_n^0 = \Delta_n(\theta_n^0), p_n^{i+1} = \Delta_n(\theta_n^{i+1}) - \beta_n^i p_n^i, \beta_n^i = \frac{\Delta_n(\theta_n^{i+1})^T I_n^* p_n^i}{(p_n^i)^T I_n^* p_n^i}.$$

Then under the conditions of Theorem 1 the estimator θ_n^s has property (19).

In conclusion, we briefly remark on further statistical applications. The first remark is concerned with certain situations in which the distribution of observations (1) is not (or rather cannot be) fully defined, none the less some of its characteristics are known to involve parameters θ about which certain inference has to be drawn. [To exemplify such situations we mention two problems of inference on a parameter of

(i) a (deterministic) signal masked with (random) noise of an unspecified distribution,
(ii) a spectrum of a (wide sense) stationary time series].

Aiming at solving, specifically, estimation problems, one cannot now base one's inference on the corresponding loglikelihood $L_n(\theta)$, and so extend directly the above reasoning to these circumstances. In many applications, however, other criterion functions can be sought which are, essentially, free from any kind of nuisance quantities and thus depend only on θ (and on observations). Of course, this function, say $U_n(\theta) = U_n(X_1, ..., X_n; \theta)$, has to be chosen so as to guarantee the sensibility of the estimator for θ defined as the value of θ that maximizes (or minimizes) $U_n(\theta)$. (As an illustrative example of this kind of practice, the utilization in various settings of the *least squares* method should be mentioned; see, e.g., JENNRICH (1969) on non-linear regression, or KOHN (1978), DZHAPARIDZE and YAGLOM (1982) on time series analysis).

The demands on $U_n(\theta)$ made above are usually met by requiring of its *asymptotical* *differentiability* in the sense that for the difference $U_n(\theta + h/\sqrt{n}) - U_n(\theta)$ there exists a (multivariate) relation analoguous to (9) with some (s vector-valued) random variable $\Delta_n(\theta)$ and positive definite matrix I_θ. Moreover, these quantities are usually related as in (12). Often the asymptotic normality of $\Delta_n(\theta)$ can also be shown, although the covariance matrix W_θ appearing in the limiting distribution may in general differ from I_θ.

It might be clear now that under these circumstances the considerations followed above remain valid for $U_n(\theta)$ in place of the likelihood function $L_n(\theta)$, although in the conclusions (namely in (14) and (19)) I_θ^{-1} has to be replaced by $I_\theta^{-1} W_\theta I_\theta^{-1T}$ (Beinicke and Dzhaparidze (1982)).

Observe, finally, that the result $H_n^s \to I_\theta^{-1}$ (stochastically), stated in Theorem 1, can be used in constructing test statistics for certain tests-of-fit based on χ^2-distributions. Indeed structurally these kinds of test statistics may be described as quadratic forms in random variables, generated by the inverses of their limiting covariance matrices. Under condition (11), for instance, the statistics

$$\Delta_n^T(\theta_0) I_{\theta_0}^{-1} \Delta_n(\theta_0) \approx \Delta_n^T(\theta_0) H_n^s \Delta_n(\theta_0)$$

can be used for testing the hypothesis: $\theta = \theta_0$.

Acknowledgement

I would like to thank Prof. dr. M. Kean and the referee for helpful suggestions.

References

Beinicke, G. and K. O. Dzhaparidze (1982), On parameter estimation by the Davidon-Fletcher-Powell method, *Teor. Veroyatnost. i Primenen. 27,* 374-380.

Brodlie, K. W. (1977), Unconstrained minimization, in: D. Jacobs (ed.), *The State of the Art in Numerical Analysis,* Academic Press, London, pp. 229-269.

Dzhaparidze, K. O. and A. M. Yaglom (1982), Spectrum Parameter Estimation in Time Series Analysis, in: P. R. Krishnaiah, (ed.), *Developments in Statistics,* Academic Press, New York, Vol. 4, Ch. 1, pp. 1-96.

Dixon, L. C. W. (1972), Quasi-Newton algorithms generate identical points, *Math. Prog. 2,* 383-387.

Fisher, R. A. (1925), Theory of Statistical Estimation, *Proc. Cambridge Philos. Soc. 22,* 700-725.

Ibragimov, I. A. and R. Z. Has'minskii (1981), *Statistical Estimation, Asymptotic Theory,* Springer, New York.

Jennrich, R. I. (1969), Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist. 40,* pp. 633-643.

Kendall, M. G. and A. Stuart (1968), *The Advanced Theory of Statistics,* Vol. 2, Griffin, London.

Kohn, R. (1978), Asymptotic properties of time domain Gaussian estimators, *Adv. Appl. Prob. 10,* 339-359.

LeCam, L. (1956), On the asymptotic theory of estimation and testing hypotheses, in: J. Neyman, (ed.), *Proceedings in the Third Berkeley Symposium on Mathematical Statist. and Probab.,* Vol. I, pp. 129-156. California Univ. Press, Berkeley.

LeCam, L. (1960), Locally Asymptotically Normal Families of Distributions, *Univ. California Publ. in Statist. Vol. 3, No. 2,* pp. 37-98, California Univ. Press. Berkeley.

LeCam, L. (1969), *Théorie asymptotique de la décision statistique,* Presses Univ. Montréal, Montréal.

LECAM, L. (1974), Notes on Asymptotic Methods in Statistical Decision Theory, *Centre Rech. Math., Univ. Montréal,* Montréal.

ORTEGA, J. M. and W. C. RHEINBOLDT (1970), *Iterative Solutions of Non-linear Equations in Several Variables,* Academic Press, New York.

RAO, C. R. (1965), *Linear Statistical Inference and its Applications,* Wiley, New York.

ZACKS, S. (1971), *The Theory of Statistical Inference,* Wiley, New York.

Received April 1983, Revised July 1983.