PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 11, Fasc. 2 (1991), pp. 149–156

THE SELECTION FUNCTIONAL

BY

BRENTON R. CLARKE (MURDOCH)

Abstract. The paper illustrates a solution of the problem of choosing a root from estimating equations which have multiple roots. This solution is applicable also to multivariate parameter estimation. In the univariate parameter case, the consistency of the M-estimator is illustrated in a way which shows how other estimation methods can easily invoke the same technique. Multivariate parameter extensions are then indicated.

1. Introduction. The idea of a selection functional arises out of a need to bring together notions of choosing a root from estimating equations which have multiple roots. Important examples of this occur in different classical estimation methods for parameters in mixtures of two normal distributions and again in the maximum likelihood estimation of a location parameter in a Cauchy parametric family with scale known. Numerous particular examples of the former type are illustrated in [23], while a solution of the latter problem is actually given in [7].

Several authors, including Cramér [9], Huzurbazaar [18], Tarone and Gruenage [22] and Foutz [11], all examine consistency arguments for local solutions θ_n^* of maximum likelihood equations. They consider both existence and uniqueness of local solutions of the equations which pertain to the efficient solution. Here it is assumed that a parametric family $\mathscr{F} = \{F_\tau\colon \tau \in \Theta\}$, where $\Theta \subset E^r$, Euclidean r-space, describes the random sequence of independent, identically distributed random variables $X_1, \ldots, X_n$ at a particular parameter $\theta \in \Theta$. Consistency implies $\theta_n \to \theta$ in probability or almost surely. There of course may exist other sequences of solutions $\{\tilde{\theta}_n\}$ which are not consistent. On the other hand, global arguments for consistency of a properly defined maximum likelihood estimator are given in [24] and [17]. Similar arguments to these are also used in minimum distance estimation in [25]. Here the argument is the same. It is the extremum value which is the estimator. The extremum is assumed to be both unique and to exist, whence the estimator is well defined. Yet, asymptotic optimality properties of the estimator are more frequently defined through the solution of the equations, as for example in the

description of von Mises expansions defining asymptotic normality [6]. The connection between an extremum value and a solution of the equations is therefore illuminating, when it is known to exist. In this paper the argument associating the extremum with "consistent" root, that is, the root that is known to offer asymptotic optimality properties, is illustrated and further generalized to overcome the pathological cases that appear in some likelihood and robustness theory [5, 26].

There may also be real numerical advantages in highlighting the link between the two consistency arguments through the selection functional. A simple but illustrative example is in the solution of the Cauchy location estimating equation, where the selection functional is equivalent to the distance from the median. That is, the root closest to the median is defined to be the estimator. Multivariate parameter extensions of this innovation are indicated in the summary of Section 5.

Finally, the approach given here of showing the consistency of an estimator can prove both simple and illustrative for students of asymptotic theory.

2. Definition of a selection functional. The idea of a selection functional is used initially in Clarke [7] to retain properties of weak continuity and strict Fréchet differentiability of M-estimators given as solutions of equations

$$K_{F_n}(\tau) = \int \psi(x, \tau) dF_n(x) = 0, \tag{2.1}$$

where F_n is the empirical distribution function that attributes atomic weight n^{-1} to observations $X_1, \ldots, X_n$, assumed to be independent, identically distributed with a common distribution $F_\theta \in \mathscr{F}$.

Already then the range of application is broad since M-estimators include maximum likelihood estimators, some minimum distance estimators (cf., e.g., [19]) and a whole host of robust proposals for M-estimators since the initial foray into the area by Huber [16], and Andrews et al. [1]. The selection functional is designed to select for all sufficiently large n the root of the equations that is consistent from among all the roots of equations (2.1). The functional ϱ: $\mathscr{G} \times \Theta \to E$, where $\mathscr{G}$ is the space of distribution functions, is defined to have the property

$$\forall \text{ neighbourhood } N \text{ of } \theta \quad \inf_{\tau \notin N} \varrho(F_\theta, \tau) - \varrho(F_\theta, \theta) > 0. \tag{2.2}$$

It is assumed that $\varrho(F_\theta, \tau)$ is continuous in $\tau \in \Theta$. If $I(\psi, \tau)$ is the set of all solutions of (2.1), the functional ϱ satisfying (2.2) when used to define the estimator $\hat{\theta}_n$ via

$$\inf_{\tau \in I(\psi, \tau)} \varrho(F_n, \tau) = \varrho(F_n, \hat{\theta}_n) \tag{2.3}$$

is known as a *selection functional.*

3. Interesting examples of selection functionals. For those estimators minimizing a distance the selection functional corresponds to the distance. Examples are found in [8] and [4] for estimating the mixture parameters, for example. A typical formulation of such a distance is

$$\varrho(F_n, \tau) = \int (F_n(x) - F_\tau(x))^2 dK(x) \tag{3.1}$$

for suitable weight functions K. Minimizing (3.1) can be shown to give equations (2.1) (cf. [2, 19]). Here assumption (2.2) is equivalent to

$$\inf_{\tau \notin N} \int (F_\theta(x) - F_\tau(x))^2 dK(x) > 0, \tag{3.2}$$

which for usual choices of K, including Lebesgue measure and exponential weight functions, can be shown to be a result of identifiability of the parametric family $F_{\theta_1} = F_{\theta_2} \Leftrightarrow \theta_1 = \theta_2$. Assumption (3.2) is found in the minimum distance theory of Pollard [20] and Wolfowitz [25] to name but two examples.

On the other hand, the maximum likelihood estimator is adopted into the framework of the selection functional by setting

$$\varrho(F_n, \tau) = -\int \log f_\tau(x) dF_n(x),$$

where f_τ is the density associated with distribution F_τ. Denoting E as expectation, assumption (2.2) is

$$\inf_{\tau \notin N} \mathrm{E}_{F_\theta}[-\log f_\tau(X)] - \mathrm{E}_{F_\theta}[-\log f_\theta(X)] > 0$$

or

$$\sup_{\tau \notin N} \mathrm{E}_{F_\theta}[\log f_\tau(X)] < \mathrm{E}_{F_\theta}[\log f_\theta(X)].$$

In comparison to Wald's [24] global consistency argument this assumption is close to Lemma 1 of that paper which shows under suitable conditions on $\{f_\tau\}$ for $\tau \neq 0$ that

$$\mathrm{E}_{F_\theta}[\log f_\tau(X)] < \mathrm{E}_{F_\theta}[\log f_\theta(X)].$$

When $\varrho(F_\theta, \tau)$ is continuous in τ the two statements are equivalent.

The important innovation in the argument for a selection functional, which delineates it from the typical loss functional, is that equations (2.1) may be defined separately from the selection functional (2.2). The selection functional can thus be used as a tool for solving pathological problems of classical statistical estimation theory and also for examining more recent robustness theory techniques. Since the latter are derivatives of the former, it should not be surprising that the selection functional should be applicable to both areas.

The best example already in application is that illustrated in the theory of redescending ψ-functions of M-estimators, for example, as illustrated in [5] and [13]. The selection functional is

$$\varrho(F_n, \tau) = |F_n^{-1}(\tfrac{1}{2}) - \tau|. \tag{3.3}$$

Though multiple roots of the equations may exist a single root is chosen to be the estimator via a functional which is unrelated to the estimating equations.

4. The consistency argument. Arguments for the existence of a consistent local root of the estimating equations based on the Fisher consistency requirement

$$\mathrm{E}_{F_\theta}[\psi(X, \theta)] = 0 \qquad \forall \theta \in \Theta$$

can be found in varying forms. Cramér [9], Huzurbazaar [18] and Chanda [3] give variations of a proof of consistency of a root of maximum likelihood equations (2.1) with

$$\psi(x, \theta) = f_\theta^{-1}(x) \frac{\partial}{\partial \theta} f_\theta(x).$$

Multivariate parameter extensions are proved in [11] and [22] for example. The univariate parameter generalization of these results to M-estimators is covered by the following proposition:

PROPOSITION 1. *There exists a $\varkappa > 0$ such that a root θ_n^* of equations* (1.1) *exists and is unique in $(\theta-\varkappa, \theta+\varkappa)$ for all sufficiently large n (f.a.s.l. n). For arbitrary $0 < \varkappa^* < \varkappa$, $\theta_n^* \in (\theta-\varkappa^*, \theta+\varkappa^*)$ f.a.s.l. n.*

Typical proofs of Proposition 1 using the theory of uniform convergence can be found in [5, 7, 11]. It proves convenient to adopt the probability framework of these papers to describe almost sure convergence, namely the event E_n is said to converge for all sufficiently large n whenever

$$\mathrm{P}\Big(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} E_n\Big) = 1.$$

Using the framework of uniform convergence, global consistency of the estimator $\hat{\theta}_n$ defined by (2.3) can be shown using the assumption

$$\forall \varepsilon > 0 \qquad \sup_{\tau \in \Theta} |\varrho(F_n, \tau) - \varrho(F_\theta, \tau)| < \varepsilon \qquad \text{f.a.s.l. } n. \tag{4.1}$$

This assumption will follow directly from assumption (4.2) in [7] when neighbourhoods are generated by either Kolmogorov or Lévy metrics, though (4.1) is not even as strong as this. No connection to equations (2.1) is needed in making this assumption.

THEOREM 1. *Assume θ_n^* satisfies Proposition* 1. *Let ϱ satisfy* (2.2) *and* (4.1). *Then $\hat{\theta}_n$ is the unique root in $(\theta-\varkappa, \theta+\varkappa)$ f.a.s.l. n and $\hat{\theta}_n$ converges almost surely to θ.*

A proof of Theorem 1 is accompanied by Figure 1. A fictitious selection functional is used to illustrate the argument of the proof.

Let $\varkappa$ be defined by Proposition 1 and let

$$\delta(\varkappa) = \inf_{\tau \notin (\theta-\varkappa, \theta+\varkappa)} \varrho(F_\theta, \tau) - \varrho(F_\theta, \theta). \tag{4.2}$$

By continuity choose $0 < \varkappa^* < \varkappa$ so that

$$\sup_{\tau \in (\theta - \varkappa^*, \theta + \varkappa^*)} \varrho(F_\theta, \tau) - \varrho(F_\theta, \theta) < \delta(\varkappa)/4. \tag{4.3}$$

Note from Proposition 1 any other root $\tilde{\theta}_n$ of equations (2.1) lies outside of $(\theta-\varkappa, \theta+\varkappa)$ f.a.s.l. n, while there exists a unique root $\theta_n^* \in (\theta-\varkappa^*, \theta+\varkappa^*)$ f.a.s.l. n. By setting $\varepsilon = \delta(\varkappa)/4$ in (4.1) we obtain

$$\begin{aligned} \varrho(F_n, \tilde{\theta}_n) &> \varrho(F_\theta, \tilde{\theta}_n) - \delta(\varkappa)/4 \quad \text{f.a.s.l. } n \text{ by (4.1)} \\ &> \varrho(F_\theta, \theta) + \tfrac{3}{4}\delta(\varkappa) \quad \text{by (4.2)} \\ &> \sup_{\tau \in (\theta - \varkappa^*, \theta + \varkappa^*)} \varrho(F_\theta, \tau) + \tfrac{1}{2}\delta(\varkappa) \quad \text{by (4.3)} \\ &> \varrho(F_n, \theta_n^*) \quad \text{f.a.s.l. } n. \end{aligned}$$

Consequently, by definition (2.3), $\hat{\theta}_n = \theta_n^*$ f.a.s.l. n. That is, θ_n^* minimizes $\varrho(F_n, \tau)$ among all the roots of (2.1) f.a.s.l. n, whence θ_n^* is almost surely

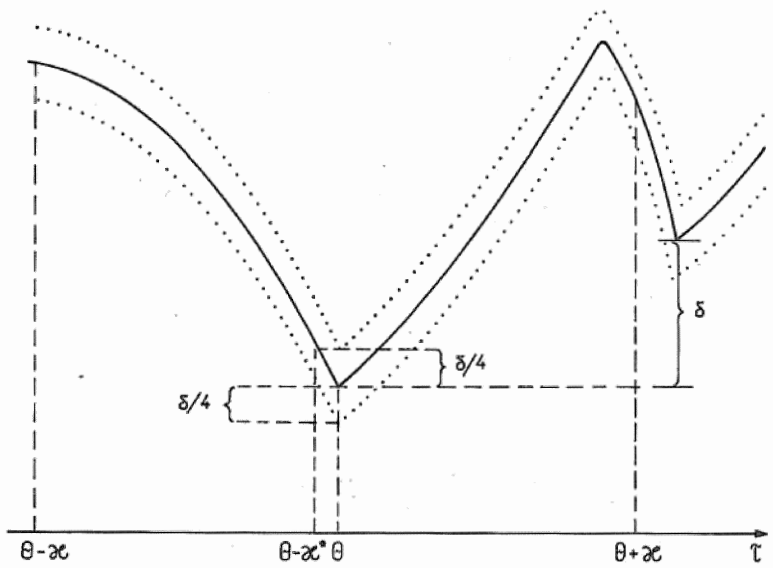

Fig. 1. ——— $\varrho(F_\theta, \tau)$, · · · · · $\varrho(F_\theta, \tau) \pm \delta/4$

equivalent to $\hat{\theta}_n$ and the consistency result of Proposition 1 carries over globally to $\hat{\theta}_n$.

The multivariate parameter extension of Theorem 1 is made by replacing the univariate interval $(\theta-\varkappa, \theta+\varkappa)$ by the open ball $B_\theta(\varkappa)$ defined on Euclidean r-space.

5. Some further examples. Many examples of M-estimators are constructed on the basis of an optimality criterion (cf. e.g. [14]). Such criteria are often derived from a von Mises expansion of equations (2.1). Consequently, the selection functional can play a useful role whenever the optimal ψ-function is selected in order to complete the global consistency argument and also to offer guidance in numerical solution of the equations (cf. [5]). A tentative but attractive proposal for possible robust selection of functionals is to choose

$$\varrho(F_n, \tau) = \left| \int \left\{ \frac{\partial}{\partial \tau} \psi(x, \tau) \right\} dF_\tau(x) - \int \left\{ \frac{\partial}{\partial \tau} \psi(x, \tau) \right\} dF_n(x) \right|. \tag{5.1}$$

This is based on the following geometric argument, which assumes uniform convergence of the curves $K_{F_n}(\tau)$ and $\partial K_{F_n}(\tau)/\partial\tau$ to their respective asymptotic curves $K_{F_\theta}(\tau)$ and $\partial K_{F_\theta}(\tau)/\partial\tau$. Since for a consistent root θ_n^*, $\varrho(F_n, \theta_n^*) \xrightarrow{P} 0$, it is expected for a realization of $X_1, \ldots, X_n$ that $\partial K_{F_n}(\tau)/\partial\tau$ evaluated at θ_n^* should at least have the same sign as $\partial K_{F_\theta}(\tau)/\partial\tau$ evaluated at θ_n^*, even though θ is unknown. Roots of equations (2.1) adjacent to θ_n^* will naturally have opposite signs for $\partial K_{F_n}(\tau)/\partial\tau$ simply by the geometry. Also, since in most examples the shape of the curve $K_{F_{\theta_1}}(\tau)$ is not too dissimilar to that of $K_{F_\theta}(\tau)$ for $\theta_1 \neq \theta$, the main difference, if any, being a translation of $|\theta_1 - \theta|$ coordinates (cf. symmetric location case), the sign of $\partial K_{F_{\theta_1}}(\tau)/\partial\tau|_{\tau=\theta_1}$ is the same as that of $\partial K_{F_\theta}(\tau)/\partial\tau|_{\tau=\theta}$. Consequently, for $\tilde{\theta}_n$ adjacent to θ_n^* the signs of

$$\int \left\{ \frac{\partial}{\partial \tau} \psi(x, \tau) \right\} dF_\tau(x) \quad \text{and} \quad \int \left\{ \frac{\partial}{\partial \tau} \psi(x, \tau) \right\} dF_n(x)$$

are opposite to each other, increasing the value of $\varrho(F_n, \tilde{\theta}_n)$. This approach was adopted in identifying roots in the simulation studies of Clarke [5] for example, where the Newton Raphson equation was the focus of study. A multivariate parameter analogue is when, after solving minimizing equations (2.1), the Jacobian is checked to identify the solution as a minima as opposed to a maxima. It is the author's suggestion here that using (5.1) is in many instances a more powerful tool than this latter approach.

In classical estimation theory also, there exist examples where solutions of equations are sometimes not easily identified as being consistent. Numerous difficulties in solving estimating equations for mixtures of two normal distributions are recorded in [23] for example. It is interesting that this problem has already been tackled in practice, where several authors, including Everitt and Hand [10], Fukunaga and Flick [12], consider the relative goodness of fit of more than one solution to the moment estimating equations, while using auxiliary criteria, including a χ^2 goodness-of-fit statistic. Hawkins [15] reports a case where different solutions to the equations do not fit equally well.

Acknowledgement. I wish to thank Professor Andrzej Kozek for the opportunity to visit the Institute of Computer Sciences, Wrocław University, which helped to complete the final version of this manuscript.

REFERENCES

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton 1972.

[2] D. D. Boos, *Minimum distance estimators for location and goodness of fit*, J. Amer. Statist. Assoc. 76 (1981), pp. 663–670.

[3] K. C. Chanda, *A note on the consistency and maxima of the roots of likelihood equations*, Biometrika 41 (1954), pp. 56–61.

[4] B. R. Clarke, *An unbiased minimum distance estimator of proportion in a mixture of two normal distributions*, Statist. Probab. Lett. 7 (1989), pp. 275–281.

[5] — *Asymptotic theory for description of regions in which Newton Raphson iterations converge to location M-estimators*, J. Statist. Plann. Inference 15 (1986), pp. 71–85.

[6] — *Robust estimation, limit theorems and their applications*, Ph. D. Thesis, Australian National University, 1980.

[7] — *Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations*, Ann. Statist. 11 (1983), pp. 1196–1205.

[8] — and C. R. Heathcote, *Comment on "Estimating mixtures of normal distributions and switching regressions"*, J. Amer. Statist. Assoc. 73 (1978), pp. 730–752.

[9] H. Cramér, *Mathematical Methods in Statistics*, Princeton University Press, Princeton 1946.

[10] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*, Chapman and Hall, London 1981.

[11] R. V. Foutz, *On the unique consistent solution to the likelihood equations*, J. Amer. Statist. Assoc. 72 (1977), pp. 147–148.

[12] K. Fukunaga and T. E. Flick, *Estimation of the parameters of a Gaussian mixture using the method of moments*, IEEE Trans. Patt. Anal. Match. Intell., PAMI-5 (1983), pp. 410–416.

[13] F. R. Hampel, *The influence curve and its role in robust estimation*, J. Amer. Statist. Assoc. 69 (1974), pp. 383–393.

[14] — E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: the Approach Based on Influence Functions*, Wiley, New York 1986.

[15] R. H. Hawkins, *A note on multiple solutions to the mixed distribution problem*, Technometrics 14 (1972), pp. 973–976.

[16] P. J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. 35 (1964), pp. 73–101.

[17] — *The behaviour of maximum likelihood estimates under nonstandard conditions*, Proc. V Berkeley Symposium, Math. Statist. Prob. 1 (1967), pp. 221–233.

[18] V. S. Huzurbazaar, *The likelihood equation, consistency and the maxima of the likelihood function*, Ann. Eug. 14 (1948), pp. 185–200.

[19] L. F. Knüsel, *Über minimum-distance-schätzungen*, Ph. D. Thesis, Swiss Federal Institute of Technology, Zürich 1969.

[20] D. Pollard, *The minimum distance method of testing*, Metrika 27 (1980), pp. 43–70.

[21] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, New York 1980.

[22] R. E. Tarone and G. Gruenage, *A note on the uniqueness of roots of the likelihood equations for vector-valued parameters*, J. Amer. Statist. Assoc. 70 (1975), pp. 903–904.

[23] D. M. Titterington, A. F. M. Smith and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York 1986.

[24] A. Wald, *A note on the consistency of maximum likelihood estimation*, Ann. Math. Statist. 20 (1949), pp. 595–601.

[25] J. Wolfowitz, *The minimum distance method*, ibidem 28 (1957), pp. 75–88.

[26] C. F. J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist. 11 (1983), pp. 95–103.

Murdoch University
Murdoch W. A. 6150
Australia

Received on 12. 4. 1987;
new version on 10. 2. 1989