The Annals of Statistics
1981, Vol. 9, No. 4, 758–767

ON THE EXISTENCE AND UNIQUENESS OF THE MAXIMUM LIKELIHOOD ESTIMATE OF A VECTOR-VALUED PARAMETER IN FIXED-SIZE SAMPLES[1]

By Timo Mäkeläinen, Klaus Schmidt and George P. H. Styan

University of Helsinki, University of Warwick, and McGill University

The maximum likelihood estimate is shown to exist and to be unique if a twice continuously differentiable likelihood function is constant on the boundary of the parameter space and if the Hessian matrix is negative definite whenever the gradient vector vanishes. The condition of constancy on the boundary cannot be completely removed, cf. Tarone and Gruenhage (1975). The theory is illustrated with several examples.

1. Introduction. In maximum likelihood estimation the solution of the likelihood equations frequently presents difficulties. In general, one is faced with the problem of sorting out those stationary values which maximize the likelihood, cf. e.g., Barnett (1966). If it is not clear that the likelihood function has just a single local maximum some check on the global form of the likelihood function is desirable, as Cox and Hinkley (1975, pages 308–309) point out. For example the occurrence of several maxima of about the same magnitude would mean that the likelihood-based confidence regions are formed from disjoint regions and summarization of data by means of a maximum likelihood estimate and its asymptotic variance could be very misleading. On the other hand, asymptotic theory ensures, for a sufficiently regular family of distributions, that a weakly consistent sequence of solutions to the likelihood equations will be unique from some sample size onwards. Thus it is of interest to find out, as a partial check on the applicability of asymptotic maximum likelihood theory or, more generally, as a step in inspecting the likelihood function, whether the likelihood equations admit a unique solution and whether such a solution actually maximizes the likelihood.

Huzurbazar (1949), Styan (1969, pages 80–81), Turnbull (1974), and Copas (1975) have suggested looking at the Hessian matrix of second derivatives of the likelihood function when the parameters satisfy the likelihood equations. They claimed that if the Hessian matrix is negative definite at these points then the likelihood equations admit a unique solution. While this conclusion is correct when there is only one parameter, Tarone and Gruenhage (1975) noted that with more than one parameter the assertion may not be true, for the function

$$g(x, y) = -e^{-2y} - e^{-y}\sin x,$$

which is defined on the plane, has a countably infinite number of maxima and no other critical points. We note, however, that the behaviour of g as $x, y \to \pm\infty$ differs from the behaviour frequently (though not always) expected of a likelihood function in that there

Received September, 1979; revised November, 1980.

[1] Part of this research was completed while the second and third authors were visiting the Indian Statistical Institute in 1976. They are very grateful to Dr. C. R. Rao for most generous provision of facilities at the Delhi Campus. This research was supported in part by the Canada Council, the Natural Sciences and Engineering Research Council Canada, and by the Gouvernement du Québec, Programme de formation de chercheurs et d'action concertée.

AMS 1980 *subject classifications.* Primary 62F10; secondary 62H12.

Key words and phrases. Hessian matrix, likelihood equations, likelihood surface, multivariate analysis, unimodality, constancy on the boundary.

is no common limiting value of g for all kinds of approaches to infinity of the arguments. In contrast, the likelihood functions considered by Styan (1973), Turnbull (1974), and Copas (1975) approach zero whenever one or more parameters approach either infinity or zero, the latter limit being appropriate for parameters which are intrinsically positive (see Section 4 below). Such a function is said to vanish or, more generally, to be *constant on the boundary*.

It seems rather rare for the likelihood function of an identifiable parameter to be periodic, so g might not be considered to be a convincing counter-example; also, g cannot be related in any obvious way to an actual likelihood function. For a further counter-example which looks more like a likelihood function see McMillan (1978, pages B-2, 3).

As an example of a likelihood function which is constant on the boundary but has many stationary points, consider the multivariate t-distribution with known degrees of freedom k and covariance matrix $k\mathbf{I}_p/(k-2)$ but with unknown location vector $\boldsymbol{\theta} \in \mathscr{R}^p$; then the likelihood function of a random sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is proportional to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} (1 + \|\mathbf{x}_i - \boldsymbol{\theta}\|^2/k)^{-(p+k)/2},$$

where $k > 0$ and $\|\mathbf{y}\|^2 = \sum_1^p y_j^2$. Then

$$\nabla \log L(\boldsymbol{\theta}) = \frac{p+k}{k} \sum_{i=1}^{n} \frac{\mathbf{x}_i - \boldsymbol{\theta}}{1 + \|\mathbf{x}_i - \boldsymbol{\theta}\|^2/k}.$$

If $\|\mathbf{x}_j - \boldsymbol{\theta}\|$ is large the jth term will be small. Consequently if the other observations lie sufficiently far away from $\mathbf{x}_i$ the behaviour of $\nabla \log L$ in a neighbourhood of $\mathbf{x}_i$ will be determined by the ith term alone. If all observations lie sufficiently far apart it can be shown, extending the argument of Barnett (1966), that one can have as many as n local maxima, each in a neighbourhood of its own observation. The larger the number k of degrees of freedom, the farther apart must the observations lie, and in the limiting normal case we will have unimodality.

Theorem 2.1 of this paper establishes existence and uniqueness of a local (and hence global) maximum for a twice continuously differentiable likelihood function defined on an open connected subset of Euclidean space, subject to constancy of the function on the boundary and negative definiteness of the Hessian at the points where the gradient vector vanishes. Conversely, we note that the nonexistence of maximum likelihood estimates is sometimes a consequence of the behaviour of the likelihood function on the boundary of the parameter space.

Theorem 2.6 deals with a different approach to the uniqueness of local and global maxima, one utilizing (generalized) concavity. In order to establish the existence of a maximum the latter approach must, however, be supplemented with additional considerations.

Aspects of both approaches are illustrated with the two-parameter Cauchy distribution (Section 4.1), a survival model involving double-censoring (Section 4.2), and the estimation of variances in a multivariate normal distribution with known correlation matrix (Section 4.3).

2. The main results. Let Θ be an open subset of $\mathscr{R}^p$. A sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots$, in Θ is said to converge to the boundary $\partial\Theta$ of Θ if for every compact set $K \subset \Theta$ there exists an integer $k_0 \geq 1$ such that $\boldsymbol{\theta}^{(k)} \notin K$ for every $k \geq k_0$. When $\Theta = \mathscr{R}^p$ this condition is equivalent to $\lim_{k\to\infty} \|\boldsymbol{\theta}^{(k)}\| = \infty$.

A real-valued function f defined on Θ is said to be *constant on the boundary* $\partial\Theta$ when there exists an extended real number c such that for every neighbourhood $\mathscr{N}$ of c there exists a compact subset $K \subset \Theta$ with $\{\boldsymbol{\theta}: f(\boldsymbol{\theta}) \notin \mathscr{N}\} \subset K$. This condition is equivalent to $\lim_{k\to\infty} f(\boldsymbol{\theta}^{(k)}) = c$ for every sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots$ in Θ converging to $\partial\Theta$. We shall, therefore, use the notation $\lim_{\boldsymbol{\theta}\to\partial\Theta} f(\boldsymbol{\theta}) = c$ to express that f is constant (in fact, equal to the constant c) on the boundary.

THEOREM 2.1. *Let $L(\boldsymbol{\theta})$ be a twice continuously differentiable likelihood function with $\boldsymbol{\theta}$ varying in a connected open subset $\Theta \subset \mathscr{R}^p$. Suppose that*

(i) $$\lim_{\theta \to \partial\Theta} L(\boldsymbol{\theta}) = 0,$$

and that

(ii) *the Hessian matrix*

$$\mathbf{H}(\boldsymbol{\theta}) = \left\{ \frac{\partial^2 L}{\partial\theta_i \partial\theta_j} (\boldsymbol{\theta}) \right\}$$

of second partial derivatives is negative definite at every point $\boldsymbol{\theta} \in \Theta$ for which the gradient vector

$$\nabla L = \{\partial L / \partial \theta_i\}$$

vanishes. Then

(1) *there is a unique maximum likelihood estimate $\hat{\boldsymbol{\theta}} \in \Theta$, and*

(2) *the likelihood function attains*

  (a) *no other maxima in Θ,*

  (b) *no minima or other stationary points in Θ,*

  (c) *its infimum value 0 on the boundary $\partial\Theta$ and nowhere else.*

Theorem 2.1 is an immediate consequence of the following slightly more general result.

THEOREM 2.2. *Let Θ be a connected open subset of $\mathscr{R}^p$, $p \geq 1$, and let f be a twice continuously differentiable real-valued function on Θ with $\lim_{x \to \partial\Theta} f(\mathbf{x}) = 0$. Suppose that the Hessian matrix $\mathbf{H} = \{\partial^2 f / \partial x_i \partial x_j\}$ of second partial derivatives is negative definite at every point $\mathbf{x} \in \Theta$ for which the gradient vector $\nabla f = \{\partial f / \partial x_i\}$ vanishes. Then f has a unique local (and hence global) maximum and no other critical points. Furthermore $f(\mathbf{x}) > 0$ for every $\mathbf{x} \in \Theta$.*

Theorem 2.2 can be shown to be a simple consequence of Morse theory, cf. e.g., Eells (1967). Since some readers may not be experts in differential topology we shall give a brief outline of an elementary proof of Theorem 2.2. For more details see Mäkeläinen, Schmidt and Styan (1979) or Barndorff-Nielsen and Blæsild (1980). We will assume f and Θ to be as in the statement of Theorem 2.2.

LEMMA 2.3. *Let $\mathbf{x}_0 \in \Theta$ be a critical point of f, i.e., $\nabla f(\mathbf{x}_0) = \mathbf{0}$. Then $\mathbf{x}_0$ is a strict local maximum. Furthermore there exists an open neighbourhood $\mathscr{N}(\mathbf{x}_0)$ of $\mathbf{x}_0$ in Θ such that*

(1) *the Hessian matrix $\mathbf{H}(\mathbf{x})$ is negative definite for every $\mathbf{x} \in \mathscr{N}(\mathbf{x}_0)$,*

(2) *the gradient $\nabla f(\mathbf{x}) \neq \mathbf{0}$ for every $\mathbf{x} \in \mathscr{N}(\mathbf{x}_0)$ with $\mathbf{x} \neq \mathbf{x}_0$.*

Lemma 2.3 is an elementary consequence of Taylor's formula and the continuity of $\mathbf{H}(\mathbf{x})$. The constancy of f on the boundary of Θ implies that we can extend f to a continuous function f on the one-point compactification $\tilde{\Theta}$ of Θ. Using this fact one obtains the following lemma.

LEMMA 2.4. *The function f has a global maximum in Θ. Furthermore, $f(\mathbf{x}) > 0$ for every $\mathbf{x} \in \Theta$.*

PROOF OF THEOREM 2.2. We choose and fix a point $\mathbf{x}_0 \in \Theta$ at which f attains its global maximum M and set, for every real number a with $0 \leq a < M$, $S(a) = \{\mathbf{x} \in \Theta\colon a < f(\mathbf{x}) \leq M\}$. In order to prove Theorem 2.2 by contradiction we assume the existence of a point

$\mathbf{x}_1 \neq \mathbf{x}_0$ in Θ at which f attains a local maximum $M_1 \leq M$. For every a with $0 \leq a < M_1$, and for $i = 1, 2$, let $S_i(a)$ be the connected component of $S(a)$ containing $\mathbf{x}_i$. Since we can connect $\mathbf{x}_0$ and $\mathbf{x}_1$ by a compact path $\mathscr{C}$ in Θ, and since $\min_{x \in \mathscr{C}} f(\mathbf{x}) > 0$ by Lemma 2.4, $S_1(a) = S_2(a)$ for all positive a sufficiently close to zero. Lemma 2.3, on the other hand, implies that $S_1(a) \cap S_2(a) \neq \varnothing$ for a sufficiently close to, but less than, M_1. We define $a_0 = \inf\{a: 0 \leq a < M_1 \text{ and } S_1(a) \cap S_2(a) = \varnothing\}$. Using either Theorem 3.1 in Milnor (1963) or an ad hoc argument involving the gradient flow of f we conclude the existence of a critical point y of f in $\overline{\cup_{a>a_0} S_1(a)} \cap \{\mathbf{x}: f(\mathbf{x}) = a_0\}$, where the bar denotes closure. It is obvious that y cannot be a strict local maximum of f, so that we have arrived at a contradiction to the assertion of Lemma 2.3. The theorem is proved. □

An essentially nonmathematical proof of Theorem 2.2 may be given as follows: Consider a function $f: \Theta \to \mathscr{R}$ satisfying all the conditions stated in Theorem 2.2, and apply Lemma 2.4 to see that $f(\mathbf{x}) > 0$ for every $\mathbf{x} \in \Theta$. To help our intuition, we replace f by $-\log f = \psi$ and look for a unique minimum of ψ instead. It may be helpful to the reader to imagine a country completely surrounded by infinitely high mountains where ψ describes the altitude above sea level at any given point on the map. Suppose there are at least two valleys in the country which are separated by a ridge of mountains (i.e., there are two distinct local minima). Suppose, furthermore, that there is a very serious flood. As the water rises in each of the valleys, several lakes will form, whose size will increase all the time. Eventually at least two of these lakes will merge and this is the crucial moment: the point at which they merge must be a pass, i.e., a lowest point in the mountain ridge separating two valleys. But a pass (in a continuously differentiable landscape) is a point at which the gradient of ψ vanishes. Our assumptions on ψ imply that every such point is, in fact, the bottom of a valley, and this is where we arrive at our contradiction. The conclusion is that there exists only one valley, i.e., only one local (and hence global) minimum.

The following corollary to Theorem 2.2 is useful when considering the log-likelihood function rather than the likelihood function itself.

COROLLARY 2.5. *Let Θ be a connected open subset of $\mathscr{R}^p$, $p \geq 1$, and let f be a twice continuously differentiable real-valued function on Θ with $\lim_{x \to \partial\Theta} f(\mathbf{x}) = c$, where c is either a real number or $-\infty$. Suppose that the Hessian matrix $\mathbf{H} = \{\partial^2 f/\partial x_i \partial x_j\}$ of second partial derivatives is negative definite at every point $\mathbf{x} \in \Theta$ for which the gradient vector $\nabla f = \{\partial f/\partial x_i\}$ vanishes. Then f has a unique local (and hence global) maximum and no other critical points. Furthermore, $f(\mathbf{x}) > c$ for every $\mathbf{x} \in \Theta$.*

PROOF. If $c \in \mathscr{R}$, replace f by $f - c$, and apply Theorem 2.2. If $c = -\infty$, f must be bounded above. Hence there exists a number $M \in \mathscr{R}$ with $f(\mathbf{x}) \leq M$ for every $\mathbf{x} \in \Theta$. Theorem 2.2 can then be applied to the function $-1/(f(\mathbf{x}) - M - 1)$ to give the desired result. □

If the Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$ in (ii) in Theorem 2.1 is negative definite at every point $\boldsymbol{\theta} \in \Theta$ then the condition (i) may be weakened. We note that Theorem 2.1 remains valid if L is replaced by $l = \log L$ and (i) by $\lim_{\boldsymbol{\theta} \to \partial\Theta} l(\boldsymbol{\theta}) = -\infty$, cf. Corollary 2.5. In Theorem 2.6 we use l since it is often concave even though L typically is not.

THEOREM 2.6. *Let $l(\boldsymbol{\theta})$ be the logarithm of a twice continuously differentiable likelihood function, with $\boldsymbol{\theta}$ varying in a connected open subset $\Theta \subset \mathscr{R}^p$. Suppose that*

(i) *the gradient vector ∇l vanishes in at least one point $\boldsymbol{\theta} \in \Theta$ and that*

(ii) *the Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$ of $l(\boldsymbol{\theta})$ is negative definite at every point $\boldsymbol{\theta} \in \Theta$.*

Then

(a) *$l(\boldsymbol{\theta})$ is a strictly concave function of $\boldsymbol{\theta}$,*

(b) *there is a unique maximum likelihood estimate* $\hat{\boldsymbol{\theta}} \in \Theta$, *and*
(c) $l(\boldsymbol{\theta})$ *has no other maxima or minima or other stationary points in* Θ.

In view of the comments just before Theorem 2.6, we note that its condition (ii) implies condition (ii) of Theorem 2.1. Theorem 2.6 has frequently been used in the statistical literature [on occasion without condition (i) in which case it is wrong!] but we have not been able to find it stated explicitly. Its proof is straightforward and will be omitted.

3. Discussion. Condition (i) in Theorem 2.1 seems to be the only general and yet reasonable simple way of ensuring that the likelihood has at least one extremal value inside the parameter space, cf. e.g., Anderson (1958, page 47), and condition (ii) will, of course, guarantee that any such extremal value is, in fact, a local maximum. The advantage of Theorem 2.1 is that it not only yields a unique global maximum and no other stationary points *inside* the parameter space, but that it also excludes any local maxima *on the boundary*. In fact the boundary of the parameter space must necessarily be the region of "minimal likelihood," which is another useful feature of our approach, cf. (2.c) in Theorem 2.1. The practical consequence of all this is that the likelihood-based confidence regions will always be connected, and will stay away from the boundary.

Even though the conditions of Theorem 2.1 seem quite natural and are certainly satisfied by wide classes of likelihood functions, it is important to bear in mind that a likelihood function L may have all the desirable properties (2.a)–(2.c) of Theorem 2.1 without being constant on the boundary of the parameter space. However, we do not know of any other result or method in the literature which proves the properties (2.a)–(2.c) for any equally general class of likelihood functions, and which at the same time stands up to mathematical scrutiny.

We now turn to some of the most important alternative methods available for proving uniqueness of the maximum likelihood estimate. In some cases the likelihood equations "separate" in such a way that one variable may be explicitly expressed in terms of the other variables. In the case of two parameters the problem can then be solved by considering the single-parameter likelihood "ridge" obtained by maximization with respect to the other parameter. This approach has been taken, for example, with the two-parameter Weibull distribution when either the location or the shape parameter is known, cf. Pike (1966), Rockette, Antle and Klimko (1974). Here the problem can be solved by inspection of the appropriate partial derivatives. This method of "separating the variables" can be extended to higher dimensions in a straightforward manner. Cox (1976) describes a method of pivotal reduction whenever the likelihood equations can be solved sequentially for the kth variable in terms of $\theta_{k+1}, \theta_{k+2}, \cdots, \theta_p$, $1 \le k \le p$, $p \ge 2$, and uses it to determine the nature of the various critical points of the likelihood function. This sequential method can sometimes offer great computational advantages and can give detailed information about the shape of the likelihood surface. It is interesting to note that the uniqueness of the maximum likelihood estimate for the two-parameter Weibull distribution may also be established using Theorem 2.1, with the added conclusion that the boundary of the parameter space is a region of "minimal likelihood." Furthermore, this provides another example of the situation where the Hessian is *not* negative definite everywhere.

Another group of conditions frequently imposed on likelihood functions centers on various notions of concavity. Between the log-concavity of Theorem 2.2 and condition (ii) of Theorem 2.1 there are various weaker forms of concavity, any one of which could have been taken as a basic assumption. Thus it is possible to obtain the conclusions of Theorem 2.2 by relaxing the concavity assumption (ii) to the assumption of quasi-concavity, i.e., to the assumption that all the level sets $\{\boldsymbol{\theta} : l(\boldsymbol{\theta}) \ge a\}$ be convex, and by assuming that the Hessian is negative definite at every critical point. In all these cases, however, one has to impose the condition that the gradient vanish in at least one point of Θ, a condition which often is extremely difficult to check in the absence of constancy on the boundary.

In some applications we may be able to reparameterize so that the likelihood function becomes strictly concave throughout the new parameter space. We may then apply Theorem 2.6. As an example, consider the family of linear models as given by Burridge (1981). Let

$$y_i = \mu + \gamma'\mathbf{x}_i + \sigma\varepsilon_i; \qquad i = 1, \cdots, n, \tag{3.1}$$

where the error terms ε_i are assumed to be independent and to have a common strictly log-concave density function. Examples include the standard normal, logistic and extreme value distributions. Burridge (1981) then makes the reparameterization:

$$\alpha = \mu/\sigma, \qquad \boldsymbol{\beta} = \gamma/\sigma, \qquad \phi = 1/\sigma,$$

and shows that the logarithm of the joint likelihood of (3.1) is strictly concave throughout the $(\alpha, \boldsymbol{\beta}, \phi)$-parameter space. This result is also seen to hold when some of the y_i are not known exactly but are known only to lie between given constants a_i and b_i.

Even though the notion of constancy on the boundary (combined with a suitable condition on the Hessian or any more global concavity assumption) appears to have a wide range of applications and allows one to replace many ad hoc arguments by a single method, there seem to be few references to this phenomenon in the literature. Wedderburn (1976), however, formulates this notion explicitly for certain generalized linear models. From Barndorff-Nielsen (1978), one may conclude that, under suitable conditions, the likelihood function of a regular exponential family vanishes on the boundary [cf. his equation (5), page 105, Theorem 7.1, page 103, and Corollary 5.1, page 78; see also Huzurbazar (1949)].

4. Examples. In this section we present some examples to give an idea of the techniques and difficulties involved in proving that a likelihood function is constant on the boundary, and to show how our method can be applied. Further examples are given by Pukelsheim and Styan (1979), Barndorff-Nielsen and Blæsild (1980), and by Burridge (1980).

4.1. *The two-parameter Cauchy distribution.* Given the sample point $\mathbf{X} = (x_1, x_2, \cdots, x_n)$ of independent observations from the Cauchy distribution with location-scale parameters (α, β) we consider the version

$$L(\alpha, \beta) = \frac{1}{\beta^n} \prod_{i=1}^{n} \left[1 + \left(\frac{x_i - \alpha}{\beta}\right)^2\right]^{-1}, \qquad \alpha \text{ real}, \beta > 0, \tag{4.1}$$

of the likelihood function. This likelihood function has been studied by Copas (1975) with different methods.

We shall prove that, for a typical sample point, L vanishes on the boundary of the region $\mathscr{G} = \mathscr{R} \times \mathscr{R}^+$. We fix $\varepsilon > 0$ arbitrarily, and exhibit a rectangle $\mathscr{K} \subset \mathscr{G}$ outside of which we have $L < \varepsilon$.

It is clear that

$$\lim_{\beta\to\infty} \sup_{\alpha\in\mathscr{R}} L(\alpha, \beta) \le \lim_{\beta\to\infty} \frac{1}{\beta^n} = 0.$$

Now let

$$d = \tfrac{1}{2} \min\{|x_i - x_j| : 1 \le i < j \le n,\ x_i \ne x_j\},$$

$$k_i = \#\{j : x_j = x_i\}, \qquad \text{and}$$

$$m = n - \max\{k_i : 1 \le i \le n\}.$$

Then

$$\lim_{\beta\to 0} \sup_{\alpha\in\mathscr{A}} L(\alpha, \beta) \le \lim_{\beta\to 0} \min_{1\le j\le n} \frac{1}{\beta^n} \textstyle\prod_{i=1}^{n-k_j} \left(1 + \dfrac{d^2}{\beta^2}\right)^{-1}$$

$$= \lim_{\beta\to 0} \frac{1}{\beta^n} \left(1 + \frac{d^2}{\beta^n}\right)^{-m} \le \lim_{\beta\to 0} \beta^{2m-n}/d^{2m}.$$

This shows that $\lim_{\beta\to 0} \sup_{\alpha\in\mathscr{A}} L(\alpha, \beta) = 0$ under the assumption $2m - n > 0$, i.e., that fewer than $(\frac{1}{2})n$ observations coincide. Under this assumption we conclude that, for every $\varepsilon > 0$, there exists a constant $b > 0$ with

$$\{(\alpha, \beta)\colon L(\alpha, \beta) \ge \varepsilon\} \subset \{(\alpha, \beta)\colon |\log \beta| \le b\}.$$

Finally, we note that with $\mathscr{B} = \{\beta\colon |\log \beta| \le b\}$

$$\lim_{|\alpha|\to\infty} \sup_{\beta\in\mathscr{B}} L(\alpha, \beta) \le \lim_{|\alpha|\to\infty} e^{bn} \textstyle\prod_{i=1}^{n} [1 + (\alpha - x_i)^2 e^{-b}]^{-1} = 0.$$

In other words, we can find a constant $a > 0$ such that $L(\alpha, \beta) < \varepsilon$ whenever $|\alpha| > a$ and $|\log \beta| \le b$. Combining these arguments we have shown that, for every $\varepsilon > 0$, there exist positive constants a, b with $\{(\alpha, \beta)\colon L(\alpha, \beta) \ge \varepsilon\} \subset \{(\alpha, \beta)\colon |\alpha| \le a \text{ and } e^{-b} \le \beta \le e^{b}\}$. This proves that $\lim_{(\alpha,\beta)\to\partial\mathscr{G}} L(\alpha, \beta) = 0$ whenever $2m - n > 0$.

Let us now look at the case when at least half of the observations are equal, say $x_1 = \cdots = x_k$, with $k \ge (\frac{1}{2})n$. Then, setting $D = \max|x_i - x_j|$, we get, for sufficiently small β,

$$\begin{aligned} L(x_1, \beta) &= \beta^{-n} \textstyle\prod_{i=k+1}^{n} [1 + (x_i - x_1)^2/\beta^2]^{-1} \\ &\ge \beta^{-n}(1 + D^2/\beta^2)^{-(n-k)} \ge \beta^{-n}(2D^2/\beta^2)^{-(n-k)} \\ &= \beta^{n-2k}(2D^2)^{n-k}. \end{aligned}$$

Thus $L(x_1, \beta) \not\to 0$ as $\beta \downarrow 0$.

In conclusion, the Cauchy likelihood vanishes on the boundary if and only if every subset of equal observations contains fewer than half of all the observations. If exactly half of the observations are equal then L is bounded, but not constant on the boundary; in fact there is a global maximum on the boundary. If more than half of the observations are equal, to x_1 say, one even gets $\lim_{\beta\to 0} L(x_1, \beta) = \infty$, i.e., L tends to an absolute supremum on the boundary and has no global maximum.

Copas (1975) proved that the matrix of second derivatives of L is negative definite whenever the likelihood equations hold. It then follows from Theorem 2.1 that if fewer than half of the observations are coincident then L is unimodal and the two-parameter Cauchy distribution has a unique maximum likelihood estimate. See also McMillan (1978, Appendix B). By contrast the concavity approach of Theorem 2.2 does not work since for β fixed and small enough, L is multimodal (cf. Barnett, 1966) contradicting (quasi-)concavity.

It is interesting to note that the Cauchy distribution is a member of the Pearson VII family, whose likelihood functions may be written as

$$L(\alpha, \beta) = \frac{1}{\beta^n} \textstyle\prod_{i=1}^{n} \left[1 + c\left(\dfrac{x_i - \alpha}{\beta}\right)^2\right]^{-(1+1/c)/2}, \qquad \alpha \text{ real}, \beta > 0, \tag{4.2}$$

where c is a known positive constant. For $c = 0$,

$$L(\alpha, \beta) = \frac{1}{\beta^n} \textstyle\prod_{i=1}^{n} \left[\exp\left(-\dfrac{1}{2}\left(\dfrac{x_i - \alpha}{\beta}\right)^2\right)\right],$$

and the distribution is normal. For $c = 1/r$, the distribution is Student's t with r degrees of freedom and location-scale parameters (α, β). For $c = 1$ (4.2) becomes (4.1).

For every $c \ge 0$, one can again easily verify negative definiteness of the Hessian matrix whenever the likelihood equations hold. Furthermore, with $c < n - 1$ it can be shown that the likelihood function is constant on the boundary of the parameter space $\{(\alpha, \beta): \alpha \in \mathscr{R}, \beta > 0\}$, provided that fewer than $n/(1 + c)$ observations coincide; under this condition Theorem 2.1 implies once again the existence of a unique maximum likelihood estimate.

4.2. *A survivorship function with double-censoring.* Turnbull (1974) considered a random sample of the lifetimes $T_1, T_2, \cdots, T_n$, of n items, where not all the T_i are observed exactly but some are censored on the right and some on the left. From each item $i = 1, 2, \cdots, n$, let $L_i < U_i$ be lower and upper limits of observation so that $(L_i, U_i]$ is a "window" of observation and the recorded information is

$$X_i = \max(\min\{T_i, U_i\}, L_i); \qquad i = 1, 2, \cdots, n.$$

It is known whether $X_i = U_i$ (i.e., $T_i > U_i$ and the item is right-censored or a "loss"), or $X_i = L_i$ (i.e., $T_i \le L_i$ and the item is left-censored or a "late entry"), or $X_i = T_i$ (i.e., $L_i < T_i \le U_i$ and the item is uncensored or a "death").

The lifetimes are recorded only as belonging to one of the m time intervals $(t_{j-1}, t_j]$; $j = 1, 2, \cdots, m$, and $0 \equiv t_0 < t_1 < \cdots < t_m$. Let

$$\lambda_j = \text{number of losses at age } t_j;$$

$$\mu_j = \text{number of late entires at age } t_j;$$

$$\delta_j = \text{number of deaths in } (t_{j-1}, t_j];$$

$$P_j = G(t_j),$$

where $j = 1, 2, \cdots, m$ and $G(t) = P(T > t)$ is the survivorship function. Note that $1 - G(t) = P(T \le t)$ is the lifetime distribution function.

The problem is to estimate the P_j given the λ_j, μ_j and δ_j; $j = 1, 2, \cdots, m$. The likelihood function is

$$L = c \prod_{j=1}^{m} P_j^{\lambda_j}(1 - P_j)^{\mu_j}(P_{j-1} - P_j)^{\delta_j},$$

where c is a known positive constant and the observations satisfy

$$(4.3) \qquad \lambda_m > 0; \quad \lambda_j \ge 0, \quad \mu_j \ge 0, \quad \delta_j > 0, \qquad j = 1, 2, \cdots, m.$$

The parameter space is the open convex region $\mathscr{G}$ in $\mathscr{R}_+^m$ given by:

$$0 < P_m < \cdots < P_1 < P_0 \equiv 1.$$

It is clear that L vanishes on the boundary of $\mathscr{G}$. Moreover, if $l = \log L$ then the Hessian matrix

$$\mathbf{H} = \{\partial^2 l/\partial P_i \partial P_j\} = -\mathbf{A} - \mathbf{U}\mathbf{B}\mathbf{U}^T,$$

where the $m \times m$ diagonal matrices

$$\mathbf{A} = \text{dg}\left\{\frac{\lambda_j}{P_j^2} + \frac{\mu_j}{(1 - P_j)^2}\right\},$$

and

$$\mathbf{B} = \text{dg}\{\delta_j/(P_{j-1} - P_j)^2\},$$

while the $m \times m$ upper bidiagonal matrix

$$\mathbf{U} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Under the assumptions (4.3) it follows that $\mathbf{H}$ is negative definite throughout $\mathscr{G}$. We may, therefore, apply Theorem 2.1 to conclude that L is unimodal and that there is a unique maximum likelihood estimate for the $m \times 1$ vector $\{P_j\}$ in the region $\mathscr{G}$. The assumptions (4.3) may be relaxed to include the possibility that

$$\delta_1 = 0 \qquad \text{and} \qquad \mu_1 > 0.$$

In other words, we may assume that

$$\text{(4.4)} \qquad \begin{cases} \lambda_m > 0; \quad \lambda_1 \geq 0, \quad \mu_1 \geq 0, \quad \delta_1 \geq 0; \quad \mu_1 + \delta_1 > 0; & \\ \lambda_j \geq 0, \quad \mu_j \geq 0, \quad \delta_j > 0, & j = 2, 3, \cdots, m. \end{cases}$$

When (4.4) holds L still vanishes on the boundary of $\mathscr{G}$; to see that $\mathbf{H}$ is still negative definite we note that

$$-\mathbf{H} \geq \mathbf{K} = \begin{pmatrix} a_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \cdot & \cdot \\ \cdot & \mathbf{U}_1\mathbf{B}_1\mathbf{U}_1^T \end{pmatrix},$$

where a_1 is the leading element of $\mathbf{A}$, while $\mathbf{U}_1$ and $\mathbf{B}_1$ are the lower right $(m - 1) \times (m - 1)$ submatrices of $\mathbf{U}$ and $\mathbf{B}$, respectively. It follows that $\mathbf{U}_1\mathbf{B}_1\mathbf{U}_1^T > 0$ and $|\mathbf{K}| = a_1 |\mathbf{U}_1\mathbf{B}_1\mathbf{U}_1^T| > 0$, the second matrix in $\mathbf{K}$ having rank $m - 1$. We note that our method allowed us to avoid the rather difficult question of whether the likelihood equations have a solution in the parameter space $\mathscr{G}$. Furthermore the methods mentioned in Section 3 for establishing uniqueness in exponential families do not appear to be readily applicable here since L is not regular, even with the constraints $\lambda_j + \mu_j + \delta_j = n$; $j = 1, 2, \cdots, m$.

4.3. *The multivariate normal distribution with known correlation matrix.* Consider the $p \times 1$ random vector $\mathbf{X}$ which follows the multivariate normal distribution $\mathscr{N}(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \mathbf{\Delta}\mathbf{R}\mathbf{\Delta};$$

here $\mathbf{\Delta}$ is the positive definite diagonal matrix of unknown standard deviations σ_i and $\mathbf{R}$ is the known positive definite correlation matrix. The parameter space is, therefore, $\mathscr{R}_+^p = \{\sigma_i \colon \sigma_i > 0\}$.

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ denote a random sample from this distribution and let

$$\mathbf{S} = \textstyle\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^T/n.$$

If $n \geq p$ then $\mathbf{S}$ is positive definite with probability 1. Let L denote the likelihood; then

$$\begin{aligned} -l &= -(2/n)\log L - p \log 2\pi \\ &= \operatorname{tr} \mathbf{\Sigma}^{-1}\mathbf{S} + \log|\mathbf{\Sigma}| \\ &= \boldsymbol{\theta}^T(\mathbf{R}^{-1}*\mathbf{S})\boldsymbol{\theta} - \textstyle\sum_{i=1}^p \log \theta_i^2 + \log|\mathbf{R}|, \end{aligned}$$

where $\boldsymbol{\theta} = \{1/\sigma_i\}$, cf. Styan (1973, page 233). The operation $*$ denotes the elementwise Hadamard (or Schur) product: $\mathbf{A}*\mathbf{B} = \{a_{ij}b_{ij}\}$. As $\sigma_i \to 0$ or ∞ so does θ_i and $l \to -\infty$ for

$$-l - \log|\mathbf{R}| \geq \textstyle\sum_{i=1}^p \log(e^{a\theta_i^2}/\theta_i^2),$$

where a is the smallest characteristic root of the positive definite (with probability 1) matrix $\mathbf{R}^{-1}*\mathbf{S}$ (the theorem that the Hadamard product of two positive definite matrices is positive definite is due to Schur, cf. Styan, 1973, Theorem 3.1). Hence we can apply Theorem 2.1 to see that

$$\partial l/\partial\boldsymbol{\theta} = -2[(\mathbf{R}^{-1}*\mathbf{S})\boldsymbol{\theta} - \boldsymbol{\sigma}]$$

vanishes at least once (with probability 1) in $\mathscr{R}_+^p$. Here $\boldsymbol{\sigma} = \{\sigma_i\}$. Furthermore,

$$\partial^2 l/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T = -2(\mathbf{R}^{-1}*\mathbf{S} + \mathbf{\Delta}^2)$$

is negative definite throughout $\mathscr{R}_+^n$, since $\mathbf{R}^{-1} * \mathbf{S}$ is nonnegative definite and $\mathbf{\Delta}^2$ is positive definite.

It then follows from Theorem 2.1 and the remarks following it that the likelihood equation

$$(\mathbf{R}^{-1} * \mathbf{S})\boldsymbol{\theta} = \boldsymbol{\sigma}$$

for a multivariate normal distribution with known correlation matrix admits a unique solution in $\mathscr{R}_+^n$, and that the associated likelihood is strictly concave (with probability 1). Notice that here the likelihood belongs to the exponential family but is not regular.

Acknowledgment. The authors are grateful to Alex McMillan, as well as to Simon Dahan, Jeok-Eng Goh and Friedrich Pukelsheim, for helpful discussions.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester.

BARNDORFF-NIELSEN, O. and BLÆSILD, P. (1980). Global maxima, and likelihood in linear models. Research Report No. 57, Dept. Theoretical Statist., Inst. Statist., Univ. Aarhus, Denmark.

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

BURRIDGE, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B* **43** 41-45.

COPAS, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62** 701-704.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

COX, N. R. (1976). A note on the determination of the nature of turning points of likelihoods. *Biometrika* **63** 199-201.

EELLS, J. (1967). *Singularities of Smooth Maps.* Gordon and Breach, London.

HUZURBAZAR, V. S. (1949). On a property of distributions admitting sufficient statistics. *Biometrika* **36** 71-74.

MÄKELÄINEN, T., SCHMIDT, K. and STYAN, G. P. H. (1979). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. Technical Report Series A, No. 24, Dept. Math., Univ. Helsinki, Finland.

MCMILLAN, A. (1978). Finite-sample properties of maximum-likelihood estimators. M.Sc. thesis, Dept. Math., McGill Univ., Montréal.

MILNOR, J. (1963). *Morse Theory.* Ann. Math. Studies No. 51. Princeton Univ. Press.

PIKE, M. C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrika* **22** 142-161.

PUKELSHEIM, F. and STYAN, G. P. H. (1979). Nonnegative definiteness of the estimated dispersion matrix in a multivariate linear model. *Bull. Acad. Polon. Sci. Sér. Sci. Math.* **27** 327-330.

ROCKETTE, H., ANTLE, C. and KLIMKO, L. A. (1974). Maximum likelihood estimation with the Weibull model. *J. Amer. Statist. Assoc.* **69** 246-249.

STYAN, G. P. H. (1969). Multivariate normal inference with correlation structure. Ph.D. dissertation, Dept. Math. Statist., Columbia Univ., New York.

STYAN, G. P. H. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra Appl.* **6** 217-240.

TARONE, R. E. and GRUENHAGE, G. (1975). A note on the uniqueness of roots of the likelihood equations for vector-valued parameters. *J. Amer. Statist. Assoc.* **70** 903-904.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69** 169-173.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** 27-32.

TIMO MÄKELÄINEN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HELSINKI
HALLITUSKATU 15
SF-00100 HELSINKI 10
FINLAND

GEORGE P. H. STYAN
DEPARTMENT OF MATHEMATICS
MCGILL UNIVERSITY
805 OUEST, RUE SHERBROOKE
MONTRÉAL, QUÉBEC
CANADA H3A 2K6

KLAUS SCHMIDT
MATHEMATICS INSTITUTE
UNIVERSITY OF WARWICK
COVENTRY
ENGLAND CV4 7AL