Scand J Statist 1: 49–58, 1974

Maximum Likelihood Theory for Incomplete Data from an Exponential Family

ROLF SUNDBERG

University of Stockholm

Received February 1973

ABSTRACT. In this paper we study such classes of distributions which are generated from exponential families by loss of information due to the fact that only some function of the exponential family variable is observable. Examples of such classes are mixtures and convolutions of exponential type distributions as well as grouped, censored and folded distributions. Their common structure is analysed. The existence is demonstrated of a $n^{1/2}$-consistent, asymptotically normally distributed and asymptotically efficient root of the likelihood equation which asymptotically maximizes the likelihood in every compact subset of the parameter space, imposing only the natural requirement that the information matrix is positive definite. It is further shown that even the weaker requirement of local parameter identifiability, which admits of application to non-regular cases, is sufficient for the existence of consistent maximum likelihood estimates. Finally the subject of large sample tests based on maximum likelihood estimates is touched upon.

Key words: Incomplete data, grouping, mixtures, exponential families, maximum likelihood.

1. Introduction

This paper is concerned with classes of probability distributions generated from distributions of exponential type by loss of information. In accordance with Lehmann (1959), we say that a distribution is of *exponential type* (or belongs to an *exponential family*) if it possesses a probability density

$$p(x;\alpha) = C(\alpha)^{-1}e^{\alpha\cdot t(x)} \tag{1.1}$$

with respect to a σ-finite measure μ over a Borel set X in a Euclidean space, and where $\alpha\cdot t$ is the scalar product of an r-dimensional parameter α (in "natural" parametrization) and a statistic t of the same dimension. $C(\alpha)$ is a norming constant,

$$C(\alpha) = \int e^{\alpha\cdot t(x)}d\mu(x). \tag{1.2}$$

The importance of the exponential families as statistical models is well-known.

Since the statistic t is sufficient for the parameter α, it contains all the relevant information contained in x for inference about the parameter. The situation is sometimes such, however, that some relevant information contained in x is unobtainable, disregarded or lost, with the consequence that not having x we have to be satisfied with the value of a function $y = y(x)$, from which t can only be partially or approximately determined. The resulting distribution of y is in general not of exponential type. The term *incomplete data* from an exponential family will be used for this kind of situation.

Incomplete data from an exponential family appear not infrequently, as should be clear from the examples in section 2, and many papers have been devoted to the study of special cases. A general view seems first to have been adopted by Martin-Löf (1966), who recognized the common structure of the examples below and discovered the form (4.10) of the maximum likelihood equations and the form (4.11) of the information matrix. A further analysis appeared in Sundberg (1972), from which the present paper is essentially an excerpt.

For simplicity the asymptotic theory in this paper is restricted to random samples, that is samples of independent observations from one and the same distribution. Note that a random sample $(x_1, ..., x_n)$, from a distribution with the density (1.1), has a distribution of exponential type with the same parameter α and the sufficient statistic $\Sigma_1^n t(x_i)$.

2. Examples

(The numbering of the examples will be adhered to throughout the paper.)

Example 1. Grouping and censoring. Grouping means that we are given some partitioning of the sample space and observe $y(x)$ = 'the class into which x falls' instead of x itself. As a rule, grouping implies loss of information. A more general case is partial grouping, cf. for instance Kulldorff (1961). A special case of that is type I censoring, in which

the observations x falling outside some fixed interval are grouped. In type II censoring, a fixed proportion of the sample is censored and the resulting observations are no longer independent, so that case will not be covered by our asymptotic theory.

In this context the essentially different nature of truncation should be observed. We may say that censoring acts upon the sample, whereas truncation acts upon the underlying measure μ. As a consequence, truncation does not bring us outside the exponential families.

Example 2. Finite mixtures. By a mixture of k distributions of exponential type with densities $p_j(x;\alpha_j)$, $j=1, \ldots, k$, with respect to some common measure μ, we mean a distribution with the density

$$\pi(y;\alpha) = \sum_{j=1}^{k} \theta_j p_j(y;\alpha_j), \tag{2.1}$$

where $\theta_j > 0$ are known or unknown mixing proportion parameters with $\Sigma\theta_j = 1$.

That this is a situation with incomplete date from an exponential family can be seen in the following way, following Martin-Löf (1966). If the underlying k populations were distinguishable, so that we could observe not only y but also a label indicating from which population y was sampled, we would have observations from a distribution of exponential type. For let

$$e_j = \begin{cases} 1 \text{ if } y \text{ is sampled from } p_j(y;\alpha_j) \\ 0 \text{ otherwise} \end{cases}$$

for $j=1, \ldots, k$. Then the density of $x=(y, e_1, \ldots, e_k)$ is

$$\prod_{j=1}^{k} (\theta_j p_j(y;\alpha_j))^{e_j}$$

$$= \exp\left[\sum_{j=1}^{k}\left(\alpha_j \cdot e_j t_j(y) + \log\frac{\theta_j}{C_j(\alpha_j)} \cdot e_j\right)\right], \tag{2.2}$$

and as a consequence it is of exponential type. Since the e_j's are linearly related by $\Sigma e_j = 1$, corresponding to $\Sigma\theta_j = 1$, the parameter dimension is one less than it appears from (2.2). Furthermore, the parameter coordinates can be permitted to be linearly related, in which case the actual dimension is still less.

A hybrid example has been used by Mendenhall & Hader (1958) as a model for a life-testing experiment. They consider maximum likelihood estimation when observing from a mixture of two exponential life-time distributions censored at a fixed point of time, where only the censored observations are indistinguishable.

When we are concerned with applications to practical cases, mixtures of two normal distributions with or without common standard deviation seem to be of the greatest importance, and parameter estimation for such mixtures have been studied ever since the days of Karl Pearson.

Example 3. Convolutions. Another example arises if we can only observe the sum $y = \Sigma_1^k x_j$ of k independent variables $x_1, \ldots, x_k$, each with a distribution of exponential type. A more specified model, which has been applied in practical situations, is the following: We desire observations on a real variable x with a distribution of exponential type, the parameters of which are to be estimated. But we cannot avoid a normally distributed additive error component, with the result that the observable statistic is $y = x + z$, where z has a normal distribution with more or less unknown parameters.

Example 4. Folded distributions. The folded normal distribution appears when only the numerical value $y = |x|$ of a normally distributed variable x is observable. Leone et al. (1961) give examples of situations arising in industrial practice in which this model may provide an appropriate description of the data.

What might analogously be called a folded binomial distribution has been considered by Urbakh (1967) as giving an appropriate statistical description of an experiment to study differences in the rates of DNA synthesis for two morphologically indistinguishable types of pair-wise occurring chromosomes.

Example 5. The wrapped normal distribution. The wrapped normal distribution is a distribution on a circle, generated by wrapping an ordinary normal distribution around the circle. Let the circumference of the circle be of unit length. Then the wrapping is equivalent to addition modulo one of the ordinary normal distribution. Instead of observing x we can only observe

$$y(x) = x - [x], \tag{2.3}$$

where $[x]$ is the greatest integer less than or equal to x.

Example 6. Incomplete data in multivariate analysis. Much attention has been paid to the problem of missing values in multivariate analysis. Assume given incomplete observation vectors from a multivariate normal distribution, where the incompleteness is due to pure chance or to design. Then we have an example of the general structure which is the subject of our analysis. For recent discussions on maximum likelihood estimation in such models and literature reviews, see Hartley & Hocking (1971) and Woodbury & Orchard (1972).

Example 7. The negative binomial distribution. The negative binomial distribution, with density

$$\pi(y;\varkappa,\theta) = \binom{\varkappa+y-1}{y} \theta^y(1-\theta)^\varkappa, \quad y=0, 1, 2, \ldots, \tag{2.4}$$

is an example of incomplete data from an exponential family, and it can be used to illustrate two kinds of generation of such examples.

If y has a Poisson distribution with parameter λ which is itself a random variable with a two-dimensional gamma distribution, the simultaneous distribution of $x=(y, \lambda)$ is of exponential type, since the gamma distribution itself is of exponential type. But it is common knowledge that the distribution of y alone is the negative binomial. In this way the negative binomial distribution shows that some infinite mixtures appear as examples of incomplete data from exponential families.

If z is Poisson distributed and y is the sum of z independent variables following a common logarithmic distribution, the simultaneous distribution of $x=(y, z)$ is of exponential type and the distribution of y alone is well known to be the negative binomial. In this way the negative binomial distribution shows that some compound (Poisson) distributions appear as examples of incomplete data from exponential families.

One interpretation of a model as a kind of incomplete data may contribute to our understanding of the data, but any interpretation of that kind works as a technical aid in permitting us to use the general theory in the following sections.

Example 8. The Cauchy distribution. The Cauchy distribution family, given by the density

$$\pi(y; \lambda, \mu) = \frac{1}{\pi} \frac{\lambda}{\lambda^2+(y-\mu)^2}, \tag{2.5}$$

can also be looked upon as the distribution of incomplete data from an exponential family. One possibility is to consider it as a mixture, like the negative binomial, namely by giving the inverse variance in a normal distribution with mean μ the distribution of the square of a normal variable with mean zero and variance $1/\lambda^2$. Another possibility is to consider the quotient of two random variables having a bivariate normal distribution with means zero and common variance. We omit the details, which are not new.

3. Basic properties of exponential families

This section consists of some definitions and properties of exponential families, taken from Martin-Löf (1970). Most of the material can also be found in Lehmann (1959).

$p(x; \alpha)$ as introduced in (1.1) is a probability density provided that the integral $C(\alpha)$ given by (1.2) is convergent. The interior of the set of points α for which $C(\alpha)$ is finite is called the *natural parameter space* and is denoted by A. Most results in the sequel are not valid on the boundary of A, for which reason A has been made open by definition. We assume A to be non-empty. A is an open and convex subset of R^r.

The function $C(\alpha)$ can also be expressed as the integral

$$C(\alpha) = \int e^{\alpha \cdot t} d\lambda(t) \tag{3.1}$$

with respect to the measure λ, which is induced from μ by the transformation $x \to t(x)$ from X into R^r and given by

$$\lambda(E) = \int_{t^{-1}(E)} d\mu(x). \tag{3.2}$$

It is easily verified that $\lambda(E)$ is finite for any compact set E in R^r. Following Martin-Löf we call λ the *structure measure*. $C(\alpha)$ is the Laplace transform of the structure measure.

Let ζ be a complex vector with $\alpha = \mathrm{Re}\, \zeta$. The function $C(\zeta)$ is analytic in the open set $\alpha \in A$. The differentiation of $C(\zeta)$ may be performed under the integral sign, from which operation we obtain the formula (with m as a multiindex)

$$D^m C(\zeta) = \int t^m e^{\zeta \cdot t} d\lambda(t). \tag{3.3}$$

Choosing ζ real, $\zeta = \alpha$, it appears that all moments $E_\alpha(t^m)$ of t exist and can be written

$$E_\alpha(t^m) = D^m C(\alpha)/C(\alpha). \tag{3.4}$$

Since $C(\alpha) > 0$ for all $\alpha \in A$, $\log C(\zeta)$ is well-defined and analytic in some open set containing the real set A. Differentiating $\log C$ we get the semi-invariants (cumulants) $\varkappa^{(m)}$,

$$D^m \log C(\alpha) = \varkappa_\alpha^{(m)}(t). \tag{3.5}$$

In particular this implies that the gradient vector $D \log C(\alpha)$ equals $E_\alpha(t)$ and that the second and third order derivatives equal the corresponding second and third order central moments of t. Denoting the variance–covariance matrix by $\mathrm{Var}_\alpha(t)$ and the matrix of second order derivatives by D^2, we thus have the relation

$$\begin{aligned} \mathrm{Var}_\alpha(t) &= D^2 \log C(\alpha) = -D^2 \log p(x; \alpha) \\ &= E_\alpha(-D^2 \log p(x; \alpha)). \end{aligned} \tag{3.6}$$

From this we can conclude that $\mathrm{Var}_\alpha(t)$ is Fisher's information matrix, which is known to play a fundamental role in the asymptotic maximum likelihood theory.

If the structure measure λ were concentrated to a hyperplane of R^r, we should have replaced t by a statistic of lower dimension. Without restrictions we can therefore assume that λ is not concentrated to any such hyperplane, from which follows that $\mathrm{Var}_\alpha(t)$ is strictly positive definite.

From the relation

$$D \log p(x;\alpha) = t - D \log C(\alpha) = t - E_\alpha(t) \tag{3.7}$$

we find the form

$$\frac{1}{n}\sum_{i=1}^{n} t(x_i) = E_\alpha(t) \tag{3.8}$$

for the likelihood equation when a random sample $(x_1, \ldots, x_n)$ is observed. The positive definiteness of $\mathrm{Var}_\alpha(t)$ implies strict convexity of $\log C(\alpha)$ and one-to-one correspondence between α and $E_\alpha(t)$. From the latter fact it follows that the likelihood equation has at most one root $\hat{\alpha} \in A$. It is easily verified that this $\hat{\alpha}$ maximizes the likelihood function in A whenever it exists, and thus is the maximum likelihood estimate in the strict sense.

During the recent years several authors independently have studied asymptotic properties of the maximum likelihood estimate for exponential families (Andersen, 1969; Martin-Löf, 1970; and others). Martin-Löf has shown the following uniformity version:

Let $\alpha_0 \in A$ be the true parameter vector. $\hat{\alpha}$ is defined with a probability tending to one as $n \to \infty$, $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is asymptotically normally distributed with mean 0 and variance $\mathrm{Var}_{\alpha_0}(t)^{-1}$ (in particular, $\hat{\alpha}$ is consistent and asymptotically efficient), and the convergences are uniform in α_0 on each compact subset of A.

4. Basic results on distributions generated when observing $y(x)$

Our starting-point is the following pair of assumptions:

(i) x has a distribution of exponential type, as defined in (1.1);

(ii) $y = y(x)$ is observed, where $y(x)$ is a measurable function of x with values in a Euclidean space Y.

We first introduce a well-behaving probability density for y. It is no restriction to assume that μ is a probability distribution. As shown for instance in Lehmann (1959), there exists for each y a conditional probability distribution μ_y for x, generated by μ, and denoting the marginal distribution of y by ν we have for each Borel set B in Y (with characteristic function I_B) and each $\alpha \in A$

$$P\{y(x) \in B;\alpha\} = \int I_B(y)\frac{1}{C(\alpha)} e^{\alpha \cdot t(x)} d\mu(x) = \int I_B(y)\frac{1}{C(\alpha)}\left(\int e^{\alpha \cdot t(x)} d\mu_y(x)\right) d\nu(y). \tag{4.1}$$

The integral in parentheses differs from $C(\alpha)$ only by the subscript y on μ, so it comes natural to introduce the notation

$$C_y(\alpha) = \int e^{\alpha \cdot t(x)} d\mu_y(x). \tag{4.2}$$

Relation (4.1) states that for each $\alpha \in A$,

$$\pi(y;\alpha) = C_y(\alpha)/C(\alpha) \tag{4.3}$$

is a probability density for y (w.r.t. ν). Analogously

$$\frac{1}{C_y(\alpha)} e^{\alpha \cdot t(x)} = p(x;\alpha)/\pi(y;\alpha) \tag{4.4}$$

is a version of the conditional density of x with respect to the measure μ_y for fixed $y = y(x)$.

It should be noted that since $\exp(\alpha \cdot t)$ is everywhere positive, the sets of probability zero in Y as well as in X are parameter independent.

Being proportional to a probability density, $C_y(\alpha)$ is an almost surely finite function of y for each fixed $\alpha \in A$. What is more, we now prove that $C_y(\alpha)$ has the desirable property of being almost surely finite on all of A simultaneously. For each fixed y, $C_y(\alpha)$ as a function of α obviously has the same kind of properties as $C(\alpha)$. In particular, the set on which $C_y(\alpha)$ is finite is a convex subset of R^r. Consider now a dense, countable set P of points in A. P being countable, $C_y(\alpha)$ is almost surely finite on all of P simultaneously. But then $C_y(\alpha)$ is almost surely finite on all of the convex hull of P, which is identical with A.

$C_y(\alpha)$ has the same kind of properties as $C(\alpha)$, and $C_y(\alpha)$ plays the same role in the conditional distribution of x for given y as $C(\alpha)$ plays in the original distribution of x. Therefore we can immediately state the following properties of $C_y(\alpha)$ analogous to the properties of $C(\alpha)$ and valid for almost all y.

Properties of the function $C_y(\alpha)$. $C_y(\zeta)$ is analytic in the set $\mathrm{Re}\,\zeta \in A$, and the differentiation may be performed under the integral sign;

$$D^m C_y(\alpha) = C_y(\alpha) E_\alpha(t^m | y); \tag{4.5}$$

$$D^m \log C_y(\alpha) = \varkappa_\alpha^{(m)}(t|y), \text{ and in particular} \tag{4.6}$$

$$D \log C_y(\alpha) = E_\alpha(t|y) \text{ and} \tag{4.7}$$

$$D^2 \log C_y(\alpha) = \mathrm{Var}_\alpha(t|y). \tag{4.8}$$

Directing our interest to the density $\pi(y;\alpha) = C_y(\alpha)/C(\alpha)$ of y, the properties of $C_y(\alpha)$ and $C(\alpha)$ immediately provide us with expressions for the likelihood equation and for Fisher's information matrix. We first remind of the general identity

$$\mathrm{Var}\,(t) = E(\mathrm{Var}\,(t|y)) + \mathrm{Var}\,(E(t|y)). \tag{4.9}$$

It now follows that the likelihood equation for a sample $(y_1, \ldots, y_n)$ is

$$\frac{1}{n}\sum_{i=1}^{n} E_\alpha(t|y_i) = E_\alpha(t), \tag{4.10}$$

and that Fisher's information matrix can be written

$$\mathrm{Var}_\alpha(E_\alpha(t|y)) = \mathrm{Var}_\alpha(t) - E_\alpha(\mathrm{Var}_\alpha(t|y)). \tag{4.11}$$

The expression (4.10) for the likelihood equation was noted by Kale (1962) in the special case of observations from the convolution of a general distribution of exponential type and a completely specified normal distribution. But its general form seems first to have been found by Martin-Löf (1966). The likelihood equation (4.10) for incomplete data is obtained by taking the conditional expectation of the likelihood equation (3.8) for complete data. In fact, this is true for general parametric families, as indicated by Woodbury & Orchard (1972), under suitable regularity conditions. Analogously the expression (4.11) for the information matrix is extended to general families by substituting $D \log p$ for t.

The conditional expectation not being parameter independent makes equation (4.10) more complicated than equation (3.8). The likelihood equation (3.8) for complete data has at most one root in A, and that root maximizes the likelihood. The likelihood equation (4.10), for incomplete data can have several solutions in A, and it may happen that none of them corresponds to a global maximum in A of the likelihood. As an illustration we consider an example.

Example 2. Mixtures of two normal distributions. Let the two normal distributions have means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, and let the mixing proportions be θ and $1-\theta$ respectively. The five-dimensional natural parameter space A is specified by the inequalities $0<\theta<1$, $\sigma_1>0$, $\sigma_2>0$. The likelihood for a sample $(y_1, \ldots, y_n)$ of observations from the mixture is

$$L(\theta, \mu_1, \sigma_1, \mu_2, \sigma_2) = \prod_{i=1}^{n}\left(\frac{\theta}{\sigma_1}\varphi\left(\frac{y_i-\mu_1}{\sigma_1}\right) + \frac{1-\theta}{\sigma_2}\varphi\left(\frac{y_i-\mu_2}{\sigma_2}\right)\right), \tag{4.12}$$

where φ denotes the standardized normal frequency function. Put $\mu_1 = y_j$ for some j, $1 \leqslant j \leqslant n$, and let $\sigma_1 \to 0$. For fixed $0<\theta<1$, μ_2, $\sigma_2>0$ the likelihood $L(\theta, y_j, \sigma_1, \mu_2, \sigma_2)$ tends to infinity as σ_1 tends to zero. Consequently the likelihood function has no global maximum in A.

This was pointed out by Day (1969), who also stated that any pair, triplet, etc. of distinct observations sufficiently close together will generate a local maximum of the likelihood function. The latter fact shows that the likelihood equation can have more than one root. On the basis of these observations, Day stated that maximum likelihood estimation clearly breaks down in this case. The results in the following sections will show that Day's opinion was too pessimistic.

The likelihood equation for mixtures of normal distributions has several roots for an entirely different reason. It is a consequence of the arbitrariness of the numbering of the components of the mixture that whenever $(\theta, \mu_1, \sigma_1, \mu_2, \sigma_2)$ is a root, $(1-\theta, \mu_2, \sigma_2, \mu_1, \sigma_1)$ is also a root. This non-unicity reflects the lack of a unique identifiability of the parameter. However, unless $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, the parameter may be called locally identifiable. (*End of ex.*)

We conclude this section with a lemma, to be applied in the next section.

Lemma. *Each $\alpha_0 \in A$ possesses a neighbourhood in which any conditional moment $E_\alpha(|t^m|\,|y)$ can be uniformly dominated by a function which is integrable with respect to the density $\pi(y;\alpha_0)$.*

Proof. For $|\alpha - \alpha_0| \leqslant \delta$ we have

$$\begin{aligned} E_\alpha(|t^m|\,|y) &= \int |t^m| e^{\alpha\cdot t} d\mu_y(x) \Big/ \int e^{\alpha\cdot t} d\mu_y(x) \\ &\leqslant \int |t^m| e^{\alpha_0\cdot t + \delta|t|} d\mu_y(x) \Big/ \int e^{\alpha_0\cdot t - \delta|t|} d\mu_y(x) \\ &= E_{\alpha_0}(|t^m| e^{\delta|t|}\,|y) \Big/ E_{\alpha_0}(e^{-\delta|t|}|y) \\ &\leqslant E_{\alpha_0}(|t^m| e^{\delta|t|}\,|y)\cdot E_{\alpha_0}(e^{\delta|t|}\,|y). \end{aligned} \tag{4.13}$$

The last inequality is an application of Jensen's inequality to the convex function

$$f(x) = 1/x, \quad x>0. \tag{4.14}$$

The dominating function above is independent of α, and by repeated application of the Schwarz inequality it is shown to be integrable:

$$E_{\alpha_0}(E_{\alpha_0}(\,|t^m|\,e^{\delta|t|}\,|y)\cdot E_{\alpha_0}(e^{\delta|t|}\,|y))$$
$$\leqslant\{E_{\alpha_0}(E_{\alpha_0}(|t^m|\,e^{\delta|t|}\,|y)^2)\cdot E_{\alpha_0}(E_{\alpha_0}(e^{\delta|t|}\,|y)^2)\}^{\frac{1}{2}}$$
$$\leqslant\{E_{\alpha_0}(|t^{2m}|\,e^{2\delta|t|})\cdot E_{\alpha_0}(e^{2\delta|t|})\}^{\frac{1}{2}}<\infty. \quad (4.15)$$

for $\delta>0$ sufficiently small. (*End of proof.*)

In particular the lemma ensures that

$$E_{\alpha_0}(E_\alpha(\,|t^m|\,|y))<\infty \quad (4.16)$$

for some $\delta>0$ and $|\alpha-\alpha_0|<\delta$. It should be noted that δ only depends on the distance to the boundary of A and thus may be chosen to be independent of α_0 when α_0 varies within a compact.

An indirect implication of the lemma is that expectations such as $E_\alpha(E_\alpha(t^m|y)^k)$ are continuous functions of α. Of particular interest is the continuity of the information matrix $\mathrm{Var}_\alpha(E_\alpha(t|y))$. For $E_\alpha(E_\alpha(t^m|y)^k)$ to be continuous at $\alpha=\alpha_0$, it is sufficient that the integrand

$$\pi(y;\alpha)\,E_\alpha(t^m|y)^k=C_y(\alpha)\,C(\alpha)^{-1}E_\alpha(t^m|y)^k \quad (4.17)$$

is almost surely continuous at α_0 and is dominated by an integrable (α_0) function independent of α for α in some neighbourhood of α_0, see e.g. Cramér (1946) § 7.3. In the same way as in the proof of the lemma it is seen that

$$C_y(\alpha)\leqslant C_y(\alpha_0)\,E_{\alpha_0}(e^{\delta|t|}\,|y) \quad (4.18)$$

for $|\alpha-\alpha_0|\leqslant\delta$. For $E_\alpha(t^m|y)$ the lemma provides a suitable dominating function. The integrability of the resulting dominating function for the product (4.17) for δ small enough is now a consequence of the Hölder inequality, as in the proof of the lemma.

Differentiability can be shown in complete analogy to the continuity proof above.

5. Local $n^{1/2}$-consistent maximum likelihood estimates

In the last section we made the discouraging observation that the likelihood equation for incomplete data can have several essentially different solutions, none of which necessarily corresponds to a global maximum of the likelihood function. The situation for large sample sizes is far from hopeless, however. We first prove a consistency theorem of Cramér type, for random samples of incomplete data from an exponential family. All our theorems are formulated in terms of the natural parameter α. They can easily be converted to asymptotic statements concerning any regular function of α, but that contains nothing specific for the present situation.

Let $\alpha_0\in A$ be the true parameter. All we need impose in order to obtain the following local theorem is one natural condition on the information matrix for y, ensuring that the observable statistic y provides information on all of the parameter α_0. This condition is not necessary for consistency, cf. Theorem 7.2, but when the condition is not satisfied we cannot expect a maximum likelihood estimator $\hat{\alpha}$ to be $n^{1/2}$-consistent, that is $n^{1/2}(\hat{\alpha}-\alpha_0)$ to be bounded in probability as $n\to\infty$.

$n^{1/2}$-consistency condition. $\mathrm{Var}_{\alpha_0}(E_{\alpha_0}(t|y))$ *is a strictly positive definite matrix.*

This condition refers to the local behaviour of the likelihood function in the vicinity of α_0. A closer examination follows in section 6.

Theorem. *Let the $n^{1/2}$-consistency condition be satisfied. With a probability tending to one as the sample size n tends to infinity, the likelihood equation then has a solution $\hat{\alpha}$ being a consistent estimator of α_0. There exists a neighbourhood of α_0 in which this root is unique with a probability tending to one. The estimator $\hat{\alpha}$ is asymptotically efficient and asymptotically normally distributed with mean α_0 and variance*

$$\frac{1}{n}\mathrm{Var}_{\alpha_0}(E_{\alpha_0}(t\,|\,y))^{-1} \quad (5.1)$$

as n tends to infinity.

Remark. From the identity (4.11) it follows that the relative loss of efficiency of estimation, or the relative loss of information, when y is observed instead of x, is expressed by the matrix

$$\mathrm{Var}_{\alpha_0}(t)^{-1}E_{\alpha_0}(\mathrm{Var}_{\alpha_0}(t|y)). \quad (5.2)$$

Proof of the theorem. First below we state general conditions on distribution families sufficient for the statements in the theorem to hold. We next demonstrate that these conditions are satisfied by the distribution families under consideration. The general conditions are almost straightforward extensions to arbitrary parameter dimension of Cramér's well-known consistency conditions in the one-dimensional case, cf. Cramér (1946), p. 500. The sufficiency of the multi-parametric conditions for the proclaimed results has been demonstrated by Aitchison & Silvey (1958), to which we refer for a a strict proof. They did not consider the uniqueness property, however, so finally we show that the conditions are sufficient also for the uniqueness property to hold.

Sufficient conditions on a general density $\pi(y;\alpha)$ for the theorem to hold are:

(i) $\log \pi(y; \alpha)$ is almost surely three times differentiable with respect to α in some neighbourhood of α_0.

(ii) $E_{\alpha_0}(D \log \pi(y; \alpha_0)) = 0$

(iii) $E_{\alpha_0}(-D^2 \log \pi(y; \alpha_0)) = \mathrm{Var}_{\alpha_0}(D \log \pi(y; \alpha_0))$, and this information matrix is strictly positive definite.

(iv) Uniformly for α in some neighbourhood of α_0 the third order partial derivatives of $\log \pi(y; \alpha)$ are dominated by some function which is integrable with respect to the density $\pi(y; \alpha_0)$.

For the distribution families under consideration

$$\log \pi(y; \alpha) = \log C_y(\alpha) - \log C(\alpha), \tag{5.3}$$

and the information matrix for y is precisely

$$\mathrm{Var}_{\alpha_0}(E_{\alpha_0}(t|y)). \tag{5.4}$$

From the properties of $C(\alpha)$ (section 3) and $C_y(\alpha)$ (section 4) it immediately follows that conditions (i)–(iii) are satisfied whenever the $n^{1/2}$-consistency condition holds true. It remains to be shown that condition (iv) is satisfied.

Only $\log C_y(\alpha)$ depends on y in the decomposition (5.3) of $\log \pi(y; \alpha)$. By the property (4.6) of C_y the third order partial derivatives of $\log C_y(\alpha)$ equal the corresponding third order central conditional moments of t. The central moments are dominated by non-central moments, as may be seen from

$$\begin{aligned} E_\alpha(|(t - E_\alpha(t|y))^m||y) &\leq \sum_{j=1}^{r} E_\alpha((|t_j| + E_\alpha(|t_j||y))^{|m|}|y) \\ &= \sum_{j=1}^{r} \sum_{k=0}^{|m|} \binom{|m|}{k} E_\alpha(|t_j|^k|y)(E_\alpha(|t_j||y))^{|m|-k} \\ &\leq 2^{|m|} \sum_{j=1}^{r} E_\alpha(|t_j|^{|m|}|y), \end{aligned} \tag{5.5}$$

the last inequality being an application of Hölder's inequality to each factor. The lemma in section 4 now implies that condition (iv) is satisfied.

Only the local uniqueness property remains to be shown. By Taylor expansion of the second order partial derivatives of $\log \pi(y; \alpha)$ about α_0 it is seen from the conditions (iii) and (iv) that there exists a neighbourhood of α_0 in which the log-likelihood function is strictly concave with a probability tending to one as the sample size tends to infinity. In that neighbourhood the likelihood function consequently has at most one extremal point with a probability tending to one, by which the uniqueness is settled. (*End of proof.*)

As an illustration we next apply the result to one of the examples.

Example 1. Grouped normal samples. Even in situations as simple as observing from a grouped normal distribution with either the mean μ or the variance σ^2 regarded as known constants, our approach to the problems reveals some of its power and possibilities. As far as it is known to the author, the best asymptotic results in the literature on maximum likelihood estimates in these situations state consistency only under restrictions imposing certain kinds of upper bounds on the number of intervals in the grouping (Kulldorff, 1961). Now, to apply our consistency theorem we need only investigate when the information matrix is positive. This is elementary and gives the following results:

There exists a consistent estimator for (μ, σ) as soon as there are at least three intervals in the grouping.

When σ is known there exists a consistent estimator for μ even for two intervals.

When μ is known there exists a consistent estimator for σ even for two intervals, unless they are $(-\infty, \mu)$ and (μ, ∞). (*End of ex.*)

In the same way any other grouped exponential family could be analyzed.

6. The $n^{1/2}$-consistency condition and identifiability of parameters

Let us write $\alpha \sim \beta$ whenever $\pi(y; \alpha) = \pi(y; \beta)$ a.s. In this way we introduce an equivalence relation in A, separating A in equivalence classes. Obviously the theorem in section 5 remains valid if the true α_0 is replaced by any $\alpha \sim \alpha_0$ in the formulations of the theorem and its assumption. From observations on y the true parameter is only identifiable with respect to equivalence class. We first give some examples of such equivalence classes.

Example 1. For the normal distribution grouped in the two classes $(-\infty, c)$ and (c, ∞) the equivalence classes form the curves

$$\frac{c - \mu}{\sigma} = \text{constant} \tag{6.1}$$

in the parameter space.

Example 2. For a mixture of two normal distributions,

$$(\theta, \mu_1, \sigma_1, \mu_2, \sigma_2) \sim (1 - \theta, \mu_2, \sigma_2, \mu_1, \sigma_1), \tag{6.2}$$

and these pairs of points form the equivalence classes except when $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. In the latter case the equivalence classes are the curves obtained as θ runs through (0, 1). Note that identifiability of the true parameter in this context requires more than the usual concept, the identifiability of the mixing meas-

ure. In the present example the mixing measures correspond to the equivalence classes.

Example 4. For the folded normal distribution, when $y = |x|$,

$$(\mu, \sigma) \sim (-\mu, \sigma). \tag{6.3}$$

Example 5. For the wrapped normal distribution, when $y = x - [x]$,

$$(\mu, \sigma) \sim (\mu + k, \sigma) \tag{6.4}$$

for each integer k. (*End of ex.*)

We might say that α is *locally indentifiable* at α_0 if α_0 is an isolated point in its equivalence class. We will now show that this is the case if $\mathrm{Var}_{\alpha_0}(E_{\alpha_0}(t|y))$ is strictly positive. In fact we prove a slightly stronger assertion.

Theorem. *If the information matrix is strictly positive definite at $\alpha_0 \in A$, there exists some neighbourhood of α_0 that contains at most one point from each equivalence class.*

Proof. Assume the contrary, that is the existence of a sequence $\{(\alpha_k, \alpha_k')\}$ of pairs of distinct equivalent points such that $\alpha_k \to \alpha_0$, $\alpha_k' \to \alpha_0$ as $k \to \infty$. Consider the Taylor expansion of $\log \pi(y; \alpha_k')$ about α_k up to a third order remainder term of third order unconditional and conditional moments, which are majorized by means of relation (5.5) and the lemma in section 4. Taking expectations, the linear term vanishes and we obtain

$$\begin{aligned} 0 &= E_{\alpha_k}(\log \pi(y; \alpha_k') - \log \pi(y; \alpha_k)) \\ &= -\tfrac{1}{2}(\alpha_k' - \alpha_k)' \,\mathrm{Var}_{\alpha_k}(E_{\alpha_k}(t|y))\,(\alpha_k' - \alpha_k) \\ &\quad + O(|\alpha_k' - \alpha_k|^3). \end{aligned} \tag{6.5}$$

For α_k and α_k' sufficiently close to α_0, the remainder term is uniformly bounded by some constant times $|\alpha_k' - \alpha_k|^3$. This is a consequence of the fact that the δ of the lemma can be chosen independently of the α_k and that the majorization (4.15) is bounded on compact subsets of A. On the other hand, the quadratic form is bounded from below by some positive constant times $|\alpha_k' - \alpha_k|^2$ for α_k sufficiently close to α_0, since the information matrix is continuous and strictly positive at α_0. As a consequence, the positive quadratic form dominates as $k \to \infty$ and (6.5) cannot be zero for all k, in contradiction to the assumption. (*End of proof.*)

If the information matrix is strictly positive for all α equivalent to α_0, the theorem implies that there are at most a finite number of points α equivalent to α_0 in every compact subset of A. A further discussion of local identifiability and its relations to the behaviour of the information matrix may be found in Sundberg (1972).

7. Asymptotic likelihood maximization on compacts

This section is independent of the $n^{1/2}$-consistency condition. Let the true value α_0 be fixed. Our starting-point is the information inequality

$$E_{\alpha_0}(\log \pi(y; \alpha)) \leqslant E_{\alpha_0}(\log \pi(y; \alpha_0)), \tag{7.1}$$

with equality if and only if $\alpha \sim \alpha_0$. Since our distributions have a common support it is no restriction to assume that the right hand side is finite and that the left hand side exists for all α, finite or $-\infty$. Under strong assumptions on the distribution family, Wald (1949) showed that (7.1) implies strong consistency of the 'global' maximum likelihood estimate. In the present situation Wald's conditions are much too strong and his result does not necessarily hold. But the information inequality (7.1) makes it plausible that some results similar to that of Wald should hold if we restrict the parameter to compact subsets of A. More precisely, we have the following two results. The proof reminds of Wald's proof inasmuch as it contains a majorization of the likelihood within neighbourhoods, a finite number of which cover the compact.

Let $\tilde{\alpha}_0$ denote the set of points equivalent to α_0.

Theorem 7.1. *Let K be an arbitrary compact subset of A, containing some element of $\tilde{\alpha}_0$. The maximum point in K of the likelihood function is strongly consistent w.r.t. $\tilde{\alpha}_0$, in the sense that its distance to the set $\tilde{\alpha}_0$ converges almost surely to 0 as $n \to \infty$.*

Theorem 7.2. *Suppose α_0 is isolated from equivalent points in A. Almost surely there exists for n large enough at least one root $\hat{\alpha}$ of the likelihood equation such that $\hat{\alpha} \to \alpha_0$ and such that $\hat{\alpha}$ asymptotically maximizes the likelihood function in every compact subset K of A, in the sense that the difference between $(1/n)\,\Sigma_i \log \pi(y_i; \hat{\alpha})$ and $\sup_{\alpha \in K} (1/n)\,\Sigma_i \log \pi(y_i; \alpha)$ converges to 0 as $n \to \infty$.*

The theorems are immediate consequences of the following pair of lemmas.

Lemma 7.1. *To each compact subset K of A which has no point in common with $\tilde{\alpha}_0$ we can find an $\varepsilon > 0$ such that*

$$\sup_{\alpha \in K} \frac{1}{n} \sum_i \log \pi(y_i; \alpha) < \frac{1}{n} \sum_i \log \pi(y_i; \alpha_0) - \varepsilon \tag{7.2}$$

for n large enough with probability one.

Lemma 7.2. *For each compact subset K of A and each $\varepsilon > 0$*

$$\sup_{\alpha \in K} \frac{1}{n} \sum_i \log \pi(y_i; \alpha) < \frac{1}{n} \sum_i \log \pi(y_i; \alpha_0) + \varepsilon \tag{7.3}$$

for n large enough with probability one.

Proof of Lemma 7.1. We first choose a compact $K' \subset A$, which also contains no point in the equivalence class of α_0, but such that the interior of K' contains K. We next choose $\varepsilon > 0$ such that

$$\sup_{\alpha \in K'} E_{\alpha_0}(\log \pi(y; \alpha)) \leq E_{\alpha_0}(\log \pi(y; \alpha_0)) - 3\varepsilon. \tag{7.4}$$

This is possible by the inequality (7.1) and the continuity of $E_{\alpha_0}(\log \pi(y; \alpha))$. This continuity follows from the expression

$$\log \pi(y; \alpha) = \log C_y(\alpha) - \log C(\alpha), \tag{7.5}$$

where the convexity of $\log C_y(\alpha)$ is preserved when taking expectations. For each fixed $\alpha \in K'$ the strong law of large numbers now ensures that

$$\frac{1}{n} \sum_i \log \pi(y_i; \alpha) < \frac{1}{n} \sum_i \log \pi(y_i; \alpha_0) - 2\varepsilon \tag{7.6}$$

for n large enough with probability one.

Let us now reconsider the expression (7.5). Since $\log C_y(\alpha)$ is a convex function of α, the supremum of $\log C_y(\alpha)$ in a closed cube is attained at some of the 2^r corners. To each point in K we can find a cubic neighbourhood contained in K' such that the continuous function $\log C(\alpha)$ varies with less than ε in that neighbourhood. Since K is compact a finite number of these neighbourhoods are sufficient to cover K. Let the finite set of their corners be denoted by $F \subset K'$. It follows that

$$\sup_{\alpha \in K} \frac{1}{n} \sum_i \log \pi(y_i; \alpha) \leq \sup_{\alpha \in F} \frac{1}{n} \sum_i \log \pi(y_i; \alpha) + \varepsilon, \tag{7.7}$$

and since F is finite we can finally conclude from (7.6) that

$$\sup_{\alpha \in K} \frac{1}{n} \sum_i \log \pi(y_i; \alpha) < \frac{1}{n} \sum_i \log \pi(y_i; \alpha_0) - \varepsilon \tag{7.8}$$

for n large enough with probability one. (*End of proof.*)

Proof of Lemma 7.2. To obtain a proof of Lemma 7.2, we need only simplify the last proof by excluding the common restriction on the compacts K and K' that they should not contain any point $\alpha \sim \alpha_0$ and excluding the critical choice of ε. (*End of proof.*)

Theorem 7.1 may be regarded as a large sample justification of the method of maximum likelihood for the distribution families under consideration. The condition of Theorem 7.2 is certainly satisfied when the information matrix is strictly positive, by the theorem of section 6, but the interesting applications of Theorem 7.2 are to non-regular cases. As a single example we take folded distributions.

Example 4. Folded distributions. For the folded normal and binomial distributions the folding in the original sample space X has a direct correspondence in the parameter space A, where the equivalence classes are obtained by folding the parameter space at the line $\mu = 0$ and the point $\theta = 1/2$ respectively. The $n^{1/2}$-consistency condition holds except at the folds $\mu = 0$ and $\theta = 1/2$. Theorem 7.2 ensures the existence of consistent roots even when the true parameters happen to fall at these folds. For the folded normal distribution, an exhaustive investigation is given by Sundberg (1974).

8. Large sample tests

The theory above can be used to derive or to justify the usual large sample tests of parametric hypotheses based on maximum likelihood estimates. For simplicity, write $\alpha = (\beta, \gamma)$, the dimensions of β and γ being p and $q = r - p > 0$, respectively, and let the hypothesis to be tested be $\gamma = 0$. It is assumed that A contains points satisfying the hypothesis. We will only consider the local behaviour of the tests; more specifically we apply the usual technique to consider a sequence of alternatives $\alpha^{(n)}$ approaching some $\alpha_0 = (\beta_0, 0)$ as $n \to \infty$, at such a rate that

$$|\alpha^{(n)} - \alpha_0| = O(n^{-1/2}). \tag{8.1}$$

We impose the $n^{1/2}$-consistency condition (section 5) at α_0. As one consequence this will save us from all troubles of local non-identifiability. Furthermore this condition ensures the existence of $n^{1/2}$-consistent roots $\hat{\alpha}_0 = (\hat{\beta}_0, 0)$ and $\hat{\alpha} = (\hat{\beta}, \hat{\gamma})$ of the likelihood equations when the hypothesis is assumed true and when it is not, respectively, where $n^{1/2}$-consistency means convergence to α_0 at a rate of order $n^{-1/2}$ as $n \to \infty$ and $\{\alpha^{(n)}\}$ satisfies (8.1) For details and proof we refer to Sundberg (1972).

In order to simplify the expressions below we introduce a projection matrix Q, which associates with each r-dimensional vector the vector of the q last coordinates. Then we can for instance write $\gamma = Q\alpha$. Furthermore let the component of t corresponding to γ be v, that is $v = Qt$.

Theorem. *Let the $n^{1/2}$-consistency condition be satisfied at α_0. Then the following test statistics (i)–(iv) are locally asymptotically equivalent in the sense that the differences tend to zero in probability as $n \to \infty$*

and the parameter $\alpha^{(n)}$ *runs through a sequence satisfying* (8.1). *Their common asymptotic distribution is the (non-central)* χ^2 *distribution with* q *degrees of freedom and non-centrality parameter*

$$n(\gamma^{(n)})'(Q(\mathrm{Var}_{\alpha_0} E_{\alpha_0}(t\,|\,y))^{-1}Q')^{-1}\gamma^{(n)}.$$

(i) $$-2\log\Lambda = -2\sum_i \log\frac{\pi(y_i;\,\hat{\alpha}_0)}{\pi(y_i;\,\hat{\alpha})}$$

(the local likelihood ratio test statistic).

(ii) $$\frac{1}{n}\sum_i (E_{\hat{\alpha}_0}(t\,|\,y_i) - E_{\hat{\alpha}_0}(t))'(\mathrm{Var}_{\hat{\alpha}_0} E_{\hat{\alpha}_0}(t\,|\,y))^{-1}$$

$$\sum_i (E_{\hat{\alpha}_0}(t\,|\,y_i) - E_{\hat{\alpha}_0}(t)) = \frac{1}{n}\sum_i (E_{\hat{\alpha}_0}(v\,|\,y_i) - E_{\hat{\alpha}_0}(v))'$$

$$(Q(\mathrm{Var}_{\hat{\alpha}_0} E_{\hat{\alpha}_0}(t\,|\,y))^{-1}Q')\sum_i (E_{\hat{\alpha}_0}(v\,|\,y_i) - E_{\hat{\alpha}_0}(v)).$$

A general version of this test statistic was first proposed by Rao (1948).

(iii) $n\hat{\gamma}'(Q(\mathrm{Var}_{\hat{\alpha}} E_{\hat{\alpha}}(t\,|\,y))^{-1}Q')^{-1}\hat{\gamma}$.

A general version of this test statistic was introduced by Wald (1943).

(iv) $n(\hat{\alpha} - \hat{\alpha}_0)'\,\mathrm{Var}_{\hat{\alpha}_0} E_{\hat{\alpha}_0}(t|y)(\hat{\alpha} - \hat{\alpha}_0)$ or

$n(\hat{\alpha} - \hat{\alpha}_0)'\,\mathrm{Var}_{\hat{\alpha}} E_{\hat{\alpha}}(t|y)(\hat{\alpha} - \hat{\alpha}_0)$.

The theorem can be proved by Taylor expansions of the likelihood function and its derivatives. The remainder terms can be handled by means of the lemma in section 4. However, the proof is long and trite, and for the details we refer to Sundberg (1972).

An example of what can happen when the $n^{1/2}$-consistency condition is not satisfied can be found in Sundberg (1974), where large sample tests of $\mu = 0$ for the folded normal distribution (example 4) are studied.

Acknowledgement

I am greatly indebted to Docent Per Martin-Löf for his proposal of the treated topic and for many very helpful and encouraging discussions.

References

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–828.

Andersen, A. H. (1969). Asymptotic results for exponential families. *Bull. Int. Statist. Inst.* **43**:2, 241–242.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton Univ. Press, Princeton.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.

Hartley, H. O. & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics* **27**, 783–823.

Kale, B. K. (1962). A note on a problem in estimation. *Biometrika* **49**, 553–557.

Kulldorff, G. (1961). *Contributions to the theory of estimation from grouped and partially grouped samples.* Almqvist & Wiksell, Uppsala.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* Wiley, New York.

Leone, F. C., Nelson, L. & Nottingham, R. (1961). The folded normal distribution. *Technometrics* **3**, 543–550.

Martin-Löf, P. (1966). Statistics from the point of view of statistical mechanics. Lecture notes, Math. Inst., Aarhus Univ.

Martin-Löf, P. (1970). Statistiska modeller. Lecture notes (in Swedish), Inst. Math. Stat., Stockholm Univ.

Mendenhall, W. & Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* **45**, 504–520.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cam. Phil. Soc.* **44**, 50–57.

Sundberg, R. (1972). Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable. Doctoral Thesis, Inst. Math. Stat., Stockholm Univ.

Sundberg, R. (1974). On estimation and testing for the folded normal distribution. To appear in *Comm. Statist.* **3**.

Urbakh, V. Yu. (1967). Statistical testing of differences in causal behaviour of two morphologically indistinguishable objects. *Biometrics* **23**, 137–143.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54**, 426–482.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595–601.

Woodbury, M. A. & Orchard, T. (1972). A missing information principle: theory and applications. *6th Berkeley Symp. Math. Statist. Prob.* **1**, 697–715.

Dr Rolf Sundberg
Department of Mathematics
The Royal Institute of Technology
S-100 44 Stockholm 70
Sweden