COMMUN. STATIST.-SIMULA. COMPUTA., B5(1), 55-64 (1976)

AN ITERATIVE METHOD FOR SOLUTION OF THE LIKELIHOOD EQUATIONS FOR INCOMPLETE DATA FROM EXPONENTIAL FAMILIES

Rolf Sundberg

Royal Institute of Technology, S-100 44 Stockholm 70

Key Words & Phrases: maximum likelihood estimation; grouped data; censored data; mixtures.

ABSTRACT

The paper deals with the numerical solution of the likelihood equations for incomplete data from exponential families, that is for data being a function of exponential family data. Illustrative examples especially studied in this paper concern grouped and censored normal samples and normal mixtures. A simple iterative method of solution is proposed and studied. It is shown that the sequence of iterates converges to a relative maximum of the likelihood function, and that the convergence is geometric with a factor of convergence which for large samples equals the maximal relative loss of Fisher information due to the incompleteness of data. This large sample factor of convergence is illustrated diagrammatically for the examples mentioned above. Experiences of practical application are mentioned.

1. INTRODUCTION AND THEORETICAL BACKGROUND

The statistical distributions to which the iterative method of the present paper could be applied are the distributions of incom-

plete data from exponential families. A distribution is said to be of exponential type (or belong to an exponential family) if it has a probability density

$$p_\alpha(x) = C(\alpha)^{-1} e^{\alpha \cdot t(x)} \qquad (1)$$

with respect to some σ-finite measure $d\mu(x)$ over a Borel set in a Euclidean space. Here $\alpha \cdot t$ denotes the usual scalar product of a parameter vector α (in "natural" parametrization) and a minimal sufficient statistic t, and $C(\alpha)$ is a norming constant. The parameter α belongs to the natural parameter space A, which will here mean the interior of the maximal set of possible α values. The distribution of a sample $(x_1, \ldots, x_n)$ from (1) is of exponential type with $t = \sum_i t(x_i)$. The importance of the exponential families as statistical models is well-known, some simple examples being the binomial, the Poisson, the normal and the exponential distribution families.

The term incomplete data will be used to mean that some relevant information contained in x is unobtainable, disregarded or lost, such that we observe 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. This situation appears frequently.

Examples. Grouped or censored data from a distribution of exponential type. Missing values in multivariate analysis. Data from a mixed normal distribution. Data from an exponential family distribution observed with normal additive error (i.e. observations from a convolution). For details and further examples, see Sundberg (1974).

The following basic properties of exponential families and their incomplete data distribution families will be used in section 2. For proofs and other details the reader is referred to Sundberg (1974).

The likelihood equation system for parameter estimation in the exponential family (1) may be written

$$E_\alpha[t]=t \quad \text{or} \quad m_t(\alpha)=t \quad , \tag{2}$$

where $E_\alpha[t]$ and $m_t(\alpha)$ both are used to denote the expectation vector of t. The Fisher information matrix I_x is given by the variance-covariance matrix $V_\alpha[t]$ of t,

$$I_x=V_\alpha[t] \quad . \tag{3}$$

This covariance matrix also represents the Jacobian matrix of $m_t(\alpha)$. Without restriction we may assume that $V_\alpha[t]$ is strictly positive definite for all $\alpha \in A$. This ensures that $m_t(\alpha)$ is a one-to-one function of α.

Let $y=y(x)$ be a measurable function of x. The likelihood equations when y is observed are obtained by taking the conditional expectation of (2),

$$E_\alpha[t]=E_\alpha[t|y] \quad \text{or} \quad m_t(\alpha)=m_{t|y}(\alpha) \quad . \tag{4}$$

The Jacobian matrix of $m_{t|y}(\alpha)$ is $V_\alpha[t|y]$, the conditional covariance matrix of t, and the Fisher information matrix I_y is the expectation of $V_\alpha[t]-V_\alpha[t|y]$,

$$I_y=V_\alpha[t]-E_\alpha[V_\alpha[t|y]] \quad . \tag{5}$$

Comparison with (3) shows that the second term in (5) measures the loss of information due to observing $y(x)$ instead of x.

Frequently $m_t(\alpha)$ is explicitly invertible, and the unique solution of the complete data equation system is readily obtained as

$$\hat{\alpha}=m_t^{-1}(t) \quad . \tag{6}$$

In equation (4) we read $m_{t|y}(\alpha)$ instead of the t of equation (2), and moreover this function of α is usually quite complicated. The need for an iterative method of solution of the likelihood equation system (4) is apparent.

2. THE ITERATION METHOD, THEORY

The two standard methods of iteration for solving likelihood equations are the Newton-Raphson and scoring methods. They require repeated matrix inversion. The matrices to be inverted are

$V_\alpha[t] - V_\alpha[t|y]$ and its expectation I_y given by (5). If these matrices are laborious to calculate or to compute, or if the inversion of the matrices is problematic, or if the method is sensitive to the choice of trial value, then a simpler and more robust but less rapid method may be a good alternative. In many particular models, various such alternative methods have been proposed (cf. below).

The iteration method proposed here is usually simple, its convergence is geometric, and it is insensitive to the choice of trial value because it always finds a relative maximum of the likelihood function. These statements will be made more precise and be demonstrated below. The iteration method runs as follows. Choose a trial value α_0 of α and compute successive iterates by the iteration step

$$\alpha_{k+1} = f(\alpha_k) = m_t^{-1}(m_{t|y}(\alpha_k)) \quad , \tag{7}$$

cf. the formula (6) for complete data. For this method to be of any practical value, $m_t(\alpha)$ must be explicitly invertible. As mentioned above, in most cases of importance this is so.

For the most common examples of incomplete data from specific classes of distributions, such as grouped and censored data, multivariate analysis with missing values, and data from mixtures, several special iteration methods have been proposed in the literature. The method (7) was suggested by A. Martin-Löf (1967, personal communication). For the special case of grouped or censored samples from an exponential family, the method was proposed by Blight (1970), although somewhat obscured. However, he failed to prove its convergence. Also worth mentioning when discussing the method (7) is the method by Hasselblad (1969) for finite mixtures of distributions of the same exponential type. Hasselblad's method is similar but not identical to the method (7). Neither did he manage to prove the convergence of his method.

To prove the convergence of the method (7) we will make use of the following criterion (see Ostrowski (1960), ch. 18) for a root

$\hat{\alpha} \in A$ to be a point of attraction for the iterative process $\alpha_{k+1} = f(\alpha_k)$. Let $|\lambda|_{max}$ be the value of the numerically largest eigenvalue of the Jacobian matrix of $f(\alpha)$ at $\alpha = \hat{\alpha}$.

Criterion. For a root $\hat{\alpha} \in A$ to be a point of attraction it is necessary that $|\lambda|_{max} \leq 1$ and sufficient that $|\lambda|_{max} < 1$.

The quantity $|\lambda|_{max}$ is the factor of convergence. Asymptotically as $k \to \infty$ the error $|\alpha_k - \hat{\alpha}|$ decreases by this factor at each iteration step, provided $|\lambda|_{max} < 1$.

Lemma. The eigenvalues corresponding to the iteration process (7) are all real and non-negative. At local maximum points of the likelihood function they satisfy $\lambda_{max} \leq 1$ and at other extremal points they cannot satisfy $\lambda_{max} < 1$.

Outline of proof. The Jacobian of the composite function in (7) at $\alpha = \hat{\alpha}$ is

$$V_{\hat{\alpha}}[t]^{-1} V_{\hat{\alpha}}[t|y] \tag{8}$$

The properties of its eigenvalues, as stated in the lemma, follow from the positive (semi-)definiteness of covariance matrices in combination with the local extremality of $\hat{\alpha}$. For details, see Sundberg (1972).

Remark. The lemma indicates that the set of local maximum points in A and the set of attraction points in A are identical. A strict proof in the one-dimensional case of this nice property is given in Sundberg (1972). Hence, in contrast to what is the case with the Newton-Raphson and similar methods, we need not fear to arrive at a local minimum or any other non-maximum.

We now approach a large sample result being an application of the lemma and its proof. Assume that the (imagined) complete data constitute a sample $(x_1, \ldots, x_n)$ from a distribution of exponential type. The distribution of the sample is of exponential type with minimal sufficient statistic $\Sigma t(x_i)$. Let us further assume that we observe $(y_1, \ldots, y_n)$, where $y_i = y(x_i)$ for a given measurable function $y(x)$. In Sundberg (1974) the following sufficient condition and result can be found.

$n^{1/2}$-consistency condition. The Fisher information matrix I_y,

see (5), is strictly positive definite at the true α.

$n^{1/2}$-consistency theorem. Provided that the condition above is satisfied, with probability tending to one as $n \to \infty$ there exists a (unique) consistent root $\hat{\alpha}$ of the likelihood equations, and asymptotically $\hat{\alpha} \sim N(\alpha, (nI_y)^{-1})$.

Combining this theorem and the lemma above the following result may be proved.

Theorem. Provided that the $n^{1/2}$-consistency condition is satisfied, with a probability tending to one as $n \to \infty$ the consistent root $\hat{\alpha}$ is a point of attraction with a factor of convergence asymptotically equal to the maximal eigenvalue of

$$V_\alpha[t]^{-1} E_\alpha[V_\alpha[t|y]] \tag{9}$$

Outline of proof. In this case of a sample $(y_1, \ldots, y_n)$ and at the point $\alpha = \hat{\alpha}$, the matrix (8) reads

$$V_{\hat{\alpha}}[t]^{-1} \frac{1}{n} \sum_{i=1}^{n} V_{\hat{\alpha}}[t|y_i] \tag{10}$$

As $n \to \infty$ this matrix converges to (9) and the eigenvalues of (10) converge to the eigenvalues of (9). A strict proof of this is given by Sundberg (1972). Finally, the $n^{1/2}$-consistency condition ensures that the eigenvalues of (9) are strictly less than 1.

Remark. The matrix (9) describes the relative loss of Fisher information, as is seen from (3) and (5). Hence the factor of convergence of the iteration method (7) for large samples equals the maximal relative loss of information due to observing $y(x)$ instead of x (maximal with respect to the parameter components).

3. THE ITERATION METHOD, NUMERICAL RESULTS.

In this section we illustrate by some examples how the large sample factor of convergence of the iteration method (= the maximal eigenvalue λ_{max} of the matrix (10)) depends on the parameter values.

Example 1. Grouped normal sample.

Figure 1 shows how the large sample factor of convergence depends on the class-width h in relation to the standard

deviation σ of the two-parametric $N(\mu,\sigma)$, when the sample is grouped according to the class limits

$$\mu+kh \quad , \quad k=0, \pm1, \pm2, \ldots$$

The factor of convergence is then independent of μ .

It is seen that convergence is rapid (or equivalently that the information loss due to grouping is small) even for a moderate class-width.

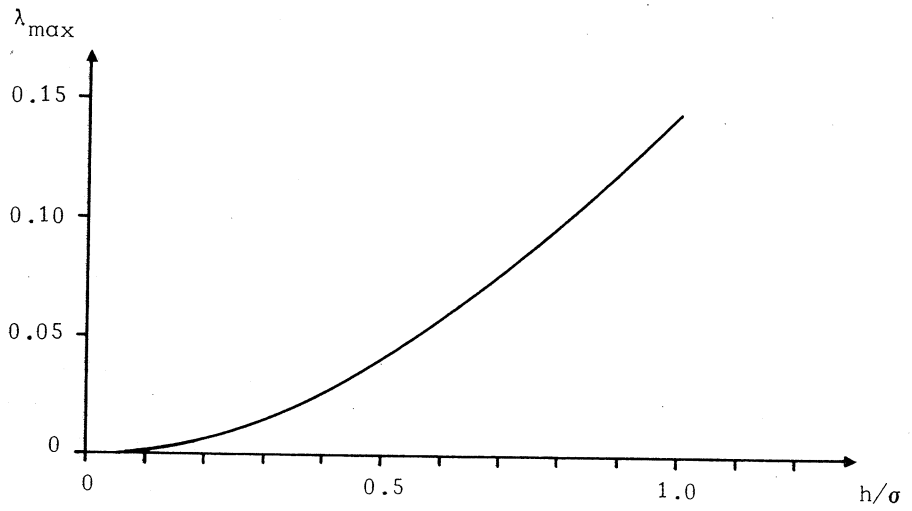

FIG. 1

Large sample factor of convergence λ_{max} for grouped normal distribution, $N(\mu,\sigma)$, class-width h .

Example 2. Censored normal sample (type I, double).

Figure 2 shows how the large sample factor of convergence depends on the points of censoring in relation to the standard deviation σ of the two-parametric $N(\mu,\sigma)$. It is assumed that the fixed points of censoring happen to fall equidistant from μ , at $\mu \pm h$, and that the observations in the intervals $(-\infty, \mu-h)$ and $(\mu+h, \infty)$ are separately counted.

It is seen that convergence is rapid as long as a small part of the data are censored.

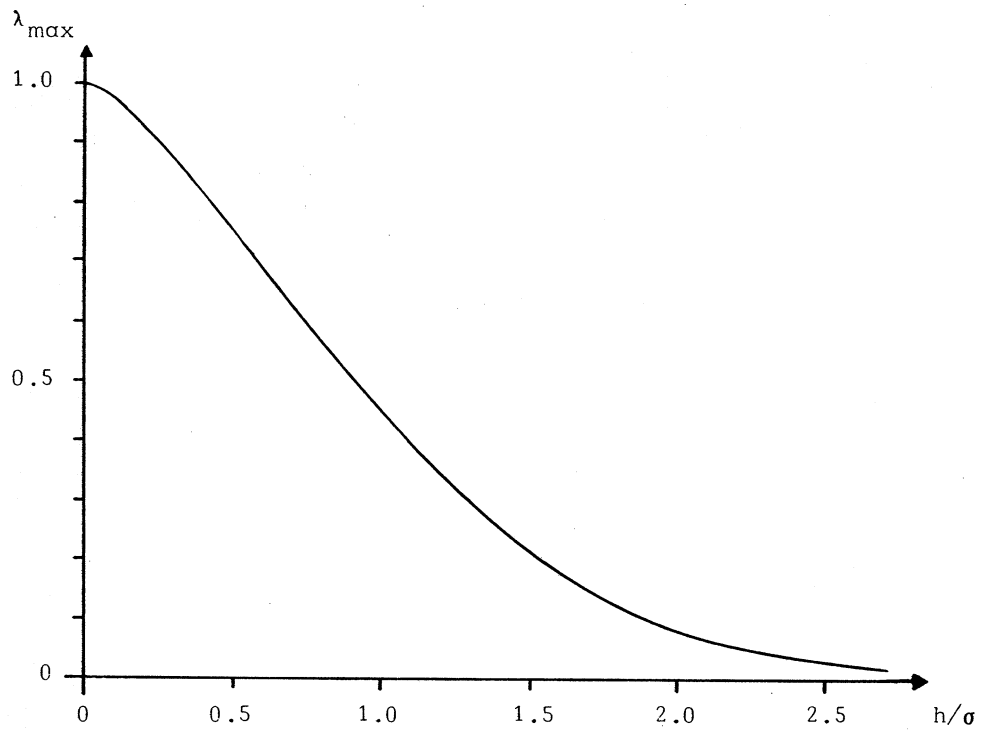

FIG. 2

Large sample factor of convergence λ_{max} for censored normal distribution, $N(\mu,\sigma)$, points of censoring $\mu \pm h$.

Example 3. Mixture of two normal samples.

Figure 3 illustrates how the large sample factor of convergence depends on $(\sigma_1+\sigma_2)/|\mu_1-\mu_2|$ for a 5-parametric mixture of two normal distributions when $\sigma_1=\sigma_2$ happens to hold, and for a 4-parametric mixture when $\sigma_1=\sigma_2=\sigma$ by assumption. The factor of convergence also depends on the mixing proportion $\theta : 1-\theta$, more specifically it increases with $|\theta-0.5|$. Curves are given for $\theta = 0.5$ and $\theta = 0.1$.

For these two mixture models the iteration method has also been applied in practice, inter alia on various samples of wing and bill lengths of birds (mixture of males and females of equal

appearance). The sample sizes ranged from slightly more than 100 to several thousands. The characteristic $2\hat{\sigma}/|\hat{\mu}_1-\hat{\mu}_2|$ ranged from .5 to 1. For all these samples the observed rates of convergence agreed well with the large sample factor of convergence as shown in Figure 3. In no case did the choice of trial value seem to be critical for the convergence.

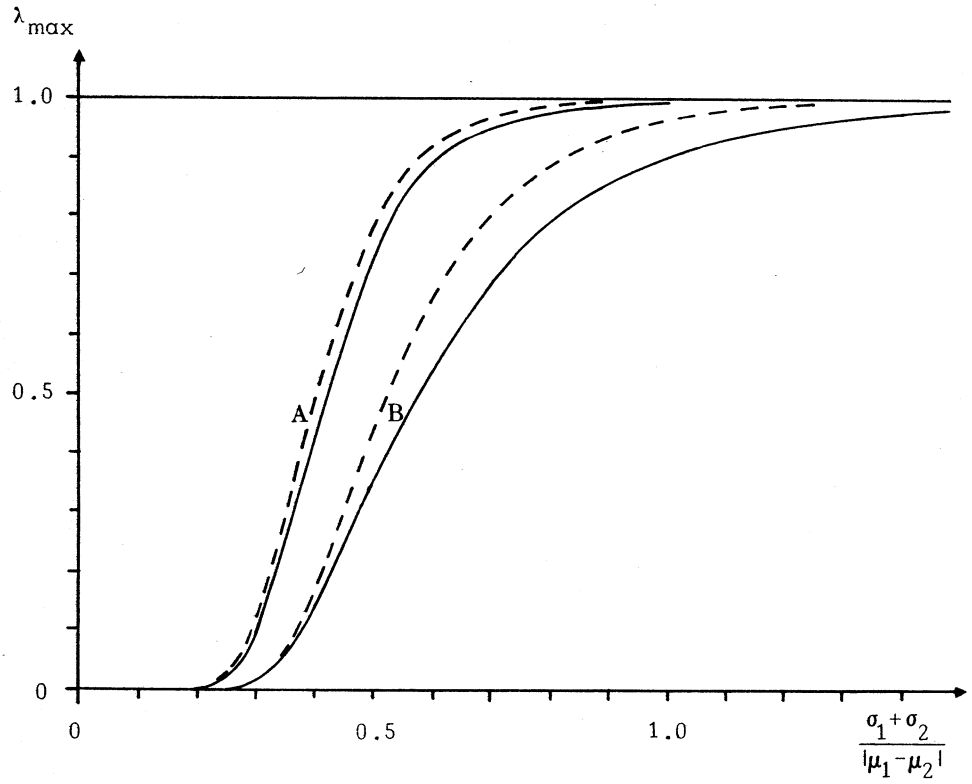

FIG. 3

Large sample factor of convergence λ_{max} for mixture of two normal distributions, $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$, mixing proportions $\theta:(1-\theta)$.

A. Five-parametric model, $\sigma_1=\sigma_2$ by chance.

B. Four-parametric model, $\sigma_1=\sigma_2=\sigma$ by assumption.

Continuous curves: $\theta=0.5$

Dashed curves: $\theta=0.1$ or 0.9

Example 4. Convolution of exponential and normal distributions.

The iteration method was also applied to a few censored samples of life-time measurements of short-lived atomic kernel states. The model assumed that exponentially distributed life-times were measured with an additive normally distributed measurement error of the same magnitude. In some of these cases extremely slow convergence was observed, and the Newton-Raphson method applied to a simplified equation system was a better alternative.

ACKNOWLEDGEMENT

This paper is based on part of the author's doctoral thesis at Stockholm University, see Bibliography.

BIBLIOGRAPHY

Blight, B.J.N. (1970). Estimation from a censored sample for the exponential family. Biometrika 57, 389-395.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. J. Amer. Statist. Ass. 64, 1459-1471.

Ostrowski, A.M. (1960). Solution of equations and systems of equations. New York: Academic Press.

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

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scand. J. Statist. 1, 49-58.

Associate Editor: *Michael Fryer*

Referee: *Michael Fryer*