Søren Tolver Jensen
Søren Johansen
Steffen L. Lauritzen

An Algorithm for Maximizing a Likelihood Function

Preprint
April
1988

3

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

Søren Tolver Jensen, Søren Johansen
and Steffen L. Lauritzen*

AN ALGORITHM FOR MAXIMIZING A LIKELIHOOD FUNCTION

Preprint 1988 No. 3

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

April 1988

Abstract

In the present note we show convergence, under very general assumptions, of iterative maximization procedures with cyclic fixing of groups of parameters, maximizing over the remaining. Further we show that a slightly modified Newton-Raphson procedure applied to the reciprocal likelihood function in a one dimensional exponential family, is convergent.

By combining these two results we obtain a general convergent iterative procedure for maximizing a likelihood function. It applies, for example, to a any full, regular k-dimensional exponential family.

Key words : Iterative proportional scaling, Newton-Raphson algorithm, exponential families.

1. INTRODUCTION

A crucial point in any statistical analysis based on the method of maximum likelihood is concerned with the actual maximization of the likelihood function. In many cases an explicit method is not available and the maximization has to be performed using numerical methods, often of iterative character. A common way is to use techniques of *Newton-Raphson* type although one knows very little about its convergence properties. A related method is the *method of scoring* due to Fisher (1958). These methods get in general impractical in problems involving several parameters, an exception being in the generalised linear models of Nelder & Wedderburn (1972), where the method of scoring can be performed as an iterative weighted least squares procedure.

Alternative methods have been developed for particular cases with many parameters, often of the type of *iterative scaling*, see for example, Andersen (1974), Darroch & Ratcliff (1972), and Speed & Kiiveri (1986). The convergence of the algorithms has been established in each of these cases (loc.cit.). All these algorithms have the flavour of using the tempting idea of cyclically keeping some parameters fixed while maximizing over the remaining.

In the present note we show that this idea often works, i.e. it gives algorithms that are convergent under quite general assumptions. The most critical of these assumptions is probably that *the existence and uniqueness of the maximum* must have been established by other means. In full, regular exponential families this has been done by Barndorff-Nielsen (1978) and we therefore first consider such cases.

2. EXPONENTIAL FAMILIES

In this section we shall give a simple algorithm which yields the maximum likelihood estimate in any k-dimensional regular exponential family. The following lemma will be useful.

LEMMA 1. *Let f have a continuous and positive derivative on an interval of the real line and let $f(x^*) = 0$. If*

$$f \text{ is convex for } x \geq x^*,$$

then the Newton-Raphson algorithm, modified to stay within the interval, converges to x^* for any starting value in the interval.

Proof. Let first f be defined on R, and let

$$g(x) = x - f(x)/f'(x).$$

The Newton-Raphson algorithm takes the form

$$x_{n+1} = g(x_n), \ n = 0,1,\ldots$$

for some starting value x_0. It is easily seen from the convexity of f that if ever $f(x_n) > 0$, then x_{n+k} is decreasing to x^*. If, on the other hand, $f(x_n) < 0$ for all n, then x_n is increasing to x^*.

If f is only defined on the interval from a to b, then we modify the algorithm as follows:

$$x_{n+1} = \begin{cases} g(x_n) & \text{if } g(x\ ) < b \\ (x_n + b)/2 & \text{if } g(x_n) \geq b \end{cases}$$

and the same proof holds. The proof is illustrated in Figure 1.

[FIGURE 1]

Let now μ be a non-negative measure on R^k, which is not concentrated on any affine hyperplane. We define the exponential family generated by μ by the densities

$$f_\vartheta(x) = exp(x\vartheta)/\varphi(\vartheta) \ , \ \vartheta \in D,$$

where

$$D = \{ \ \vartheta \ | \ \varphi(\vartheta) = \int \exp(x\vartheta) \ \mu(dx) < \infty \ \}.$$

The family is called full since the parameter set is the largest possible, and regular when D is open. It is one of the fundamental results of the theory of regular exponential families, see Barndorff-Nielsen (1978), that when the observation x is contained in the interior of the convex support of μ, then the maximum likelihood estimate exists and is uniquely defined as the solution to the equation

$$E_\vartheta(X) = x.$$

This equation can often be solved explicitly, but sometimes one has to solve it numerically and the purpose of this paper is to find an algorithm for doing so.

Consider first the case $k = 1$.

THEOREM 1. *For a regular one dimensional exponential family it holds, that if the observation is contained in the interior of the convex support of μ, then the modified Newton-Raphson algorithm applied to the reciprocal likelihood function converges to the maximum likelihood estimate for any starting value in the interval D.*

Proof. The reciprocal likelihood function can be written as

$$K(\vartheta) = \int \exp(\vartheta(y-x))\mu(dy).$$

It is well known that K is convex, and that $k(\vartheta) = K'(\vartheta)$ is increasing since

$$k'(\vartheta) = \int (y-x)^2 \exp(\vartheta(y-x))\mu(dy) > 0.$$

Let $\hat{\vartheta}$ be the unique zero of k. We also find that

$$k'''(\vartheta) = \int (y-x)^4 \exp(\vartheta(y-x))\mu(dy) > 0$$

which shows that k'' is strictly increasing. There is no loss of generality in assuming that $k''(\hat{\vartheta}) \geq 0$, since otherwise one could just consider the parameter $-\vartheta$. Hence k'' is positive for $\vartheta \geq \hat{\vartheta}$ which shows that k is convex for $\vartheta \geq \hat{\vartheta}$. Now Lemma 1 implies that the algorithm converges.

Remark. If we let τ and υ denote the mean and variance

$$\tau(\vartheta) = E_{\vartheta}(X), \quad \upsilon(\vartheta) = \mathrm{var}_{\vartheta}(X)$$

we get that each iterative step looks like

$$\vartheta_{n+1} = \vartheta_n + (x-\tau(\vartheta_n))/\{\upsilon(\vartheta_n) + (x-\tau(\vartheta_n))^2\}.$$

Comparing with the more usual procedure of performing the iteration on the logarithm of the likelihood function which gives

$$\vartheta_{n+1} = \vartheta_n + (x-\tau(\vartheta_n))/\upsilon(\vartheta_n),$$

we see that the difference comes from the term $(x-\tau(\vartheta_n))^2$ which gives extra stability when ϑ_n is far from $\hat{\vartheta}$, and which has no influence when ϑ_n comes close to $\hat{\vartheta}$.

Example 1. As an illustration of this result, consider the case where Y is gamma distributed with known scale-parameter λ_0 and unknown shape-parameter β i.e. it has density

$$\Gamma(\beta)^{-1}\lambda_0^{\beta}\mathrm{y}^{\beta-1}\exp(-\lambda_0 y)$$

Here x = log$x$ is the observation of the canonical statistic and the iterative procedure above is convergent if the observation y is positive and is given by

$$\beta_{n+1} = \beta_n + \frac{x+\log\lambda_0 - \psi(\beta_n)}{\psi'(\beta_n)+ (x+\log\lambda_0 - \psi(\beta_n))^2},$$

where $\psi(\beta) = \Gamma'(\beta)/\Gamma(\beta)$ is the digammafunction, see Abramowitz & Stegun (1965).

Note that we are not directly concerned with computational feasibility of the various procedures, merely with showing their convergence. In the multidimensional case, when $k \geq 2$, we rely on the following:

THEOREM 2. *If the observation from a regular k-dimensional exponential family is contained in the interior of the convex support of* μ, *then the maximum likelihood estimate can be calculated by successively maximizing over each parameter keeping the others fixed.*

Proof. The theorem follows from the main result in the next section about iterative partial maximization. Note that when keeping all parameter values fixed except one, we get a one dimensional exponential family, where the modified Newton-Raphson algorithm can be applied to the reciprocal likelihood function.

In the case of iterative proportional scaling (IPS) as used in log-linear models for contingency tables, see e.g. Andersen (1974), and its generalisations to covariance selection models, Speed & Kiiveri (1986), and other models, Darroch & Ratcliff (1972), each partial maximization step can typically be performed explicitly. Theorem 2 establishes the (wellknown) convergence of the quoted algorithms but, in general, each partial maximization might involve iteration.

Example 2 Continuing Example 1 we shall here assume that we have observed m independent gamma distributed random variables Y_1 ,...,Y_m having common but unknown, scale parameter λ and different unknown shape-parameters $\beta^1,\dots \beta^m$. This is an exponential family with m+1 parameters and for any fixed m of them, it is a one-dimensional exponential family. Let $Y_0 = \Sigma_i Y_i$ and let y_0 be the observation of Y_0. From Theorem 2 we find that we can obtain the maximum likelihood estimate by

1. Choose a starting value λ_0
2. Find estimates $\hat{\beta}^1,\dots,\hat{\beta}^m$ iteratively as described in Example 1
3. Calculate an updated value on the scale-parameter λ as

$$\hat{\lambda} = \Sigma_i\ \hat{\beta}^i/y_0$$

4. Repeat from point 2, now with $\hat{\lambda}$ as the "known" scale-parameter.

3. ITERATIVE PARTIAL MAXIMIZATION

We now consider the general case of a problem with parameter space Θ being a topological Hausdorff space. Let L denote a continuous function. We are going to construct an iterative procedure for maximizing L under the assumptions below, where $\vartheta_0 \in \Theta$ is a starting value, such that

$$D_0 = \{\ \vartheta \in \Theta |\ L(\vartheta) \geq L(\vartheta_0)\ \} \text{ is compact.} \tag{A1}$$

As the second assumption we shall take

$$L \text{ is uniquely maximized over } D_0 \text{ for } \vartheta = \hat{\vartheta}. \tag{A2}$$

Suppose that we have given parameter functions $\psi_i : D_0 \to \Theta_i$, $i = 1,\dots,k$ and let $M_i(\vartheta)$, $\vartheta \in D_0$ be the corresponding sections:

$$M_i(\vartheta) = \{\ \eta \in D_0 |\ \psi_i(\eta) = \psi_i(\vartheta)\},\ i = 1,\dots,k.$$

Assume further that for $i = 1,\dots,k$ and $\vartheta \in D_0$

L is maximized uniquely by $T_i(\vartheta)$ on the section $M_i(\vartheta)$ and

that $T_i(\vartheta)$ is continuous on D_0. (A3)

In some specific cases one can choose ψ_i such that each partial maximization can be performed explicitly.

Finally assume that we have enough sections, or more precisely that

$$\sup_{\eta \in M_i(\vartheta)} L(\eta) = L(\vartheta),\quad i = 1,\dots,k \quad \text{implies } \vartheta = \hat{\vartheta}, \tag{A4}$$

or equivalently

$$T_i(\vartheta) = \vartheta\ ,\ i = 1,\dots,k \text{ implies } \vartheta = \hat{\vartheta}.$$

In other words that the point of global maximum $\hat{\vartheta}$ is uniquely determined by the condition that it is the partial maximum along each section $M_i(\hat{\vartheta})$.

Before we give the proof of the main result we shall show what the conditions mean in the case where L is the likelihood function in a regular k-dimensional exponential family.

In this case it is known, see Barndorff-Nielsen (1978) that if the minimal sufficient statistic is in the interior of the convex support then the maximum likelihood estimate is uniquely defined (A2) and, moreover $\{\vartheta | L(\vartheta) \geq L(\vartheta_0)\}$ is compact (A1) for any $\vartheta_0 \in D$.

For a subset of R^k it is often convenient to use the mappings

$$\psi_i(\vartheta_1,\ldots,\vartheta_k) = (\vartheta_1,\ldots,\vartheta_{i-1},\vartheta_{i+1},\ldots,\vartheta_k)$$

to define the sections, and the condition (A3) states that for fixed values of all variables but one, the maximum likelihood estimator should exist and depend continuously on ϑ.

Finally the assumption in condition (A4) implies in this case that the derivative with respect to ϑ_i is zero at the point ϑ, but the condition is then that the maximum likelihood estimate is uniquely defined by the condition that the derivative vanishes.

These conditions, (A3) and (A4), are also satisfied for a k-dimensional exponential family, which shows that Theorem 2 is proved.

If the parameter set Θ is any closed subset of the parameter set, D_0 will also be compact and (A1) is automatically fulfilled. But in this case (A2),(A3) and (A4) need to be established separately.

We shall then prove the main result:

THEOREM 3. *Under the assumptions above the algorithm*

$$\vartheta_{n+1} = T_1\circ\ldots\circ T_k\,(\vartheta_n) \tag{2.1}$$

converges to $\hat{\vartheta}$.

Proof. Since D_0 is assumed compact (A1), ϑ_n has a convergent subsequence (ϑ_{n_k}) with limit ϑ^*, say. We need to show that $\vartheta^* = \hat{\vartheta}$.

Define

$$S = T_1\circ\ldots\circ T_k$$

Since L and S are assumed continuous (A3) and $L(\vartheta_n)$ is increasing we have

$$L(S(\vartheta^*)) = \lim_{k\to\infty} L(S(\vartheta_{n_k})) \leq \lim_{k\to\infty} L(\vartheta_{n_{k+1}}) = L(\vartheta^*).$$

But we also have from (A3) that

$$L(S(\vartheta^*)) \geq L(T_2\circ\ldots\circ T_k(\vartheta^*)) \geq\ldots\geq L(T_k(\vartheta^*)) \geq L(\vartheta^*),$$

and we must everywhere have equality. Reading the chain of equalities from right to left and recalling the uniqueness of the partially maximizing values $T_i(\vartheta^*)$ we get

$$\vartheta^* = T_k(\vartheta^*) = T_{k-1}(\vartheta^*) =\ldots= T_1(\vartheta^*),$$

whereby (A4) implies $\vartheta^* = \hat{\vartheta}$.

We shall call the algorithm (2.1) the IPM algorithm, for *Iterative Partial Maximization*.

In log-linear models it is occasionally an advantage to allow zero-probabilities by considering the standard models extended by weak limits. In such extended families the parameter space is not topologically identical to an open subset of R^k. Still a minor modification - just avoiding division by zero - of the usual IPS-algorithm is convergent, as was shown by Jensen (1978), see also Lauritzen (1982). This follows from the above general result and the proof above is indeed a small modification of the proof given in the latter of these references.

For the case involving curved exponential families we have as our final example the following:

Example 3 Consider $X_1,\ldots,X_n$ independent k-dimensional Gaussian variables with mean zero and

$$\mathrm{var}(X_i) = \lambda_i\Omega,\ i = 1,\ldots,n.$$

Assume for identification that $\prod_{i=1}^{n} \lambda_i = 1$.

It was shown by Eriksen (1987) and Jensen & Johansen (1987) that the maximum likelihood estimate is uniquely defined in this curved exponential family.

For fixed λ we get a regular exponential family with explicit maximum likelihood estimation

$$\Omega(\lambda) = n^{-1}\sum_{i=1}^{n} X_i X_i^{*}/\lambda_i \qquad (3.1)$$

and for fixed Ω, we get a curved exponential family with explicit estimation

$$\lambda_i = X_i^{*}\Omega^{-1}X_i/c \qquad (3.2)$$

where $c^n = \prod_{i=1}^{n} X_i^{*}\Omega^{-1}X_i$.

The convergence of the algorithm defined by iteration of (3.1) and (3.2) is now established by Theorem 3.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I.A. (1965). *Handbook of Mathematical Functions*. New York: Dover.

ANDERSEN, A.H. (1974). Multidimensional contingency tables. *Scand. J. Statist.* 1,115-127.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential families in Statistical Theory*. New York: Wiley.

DARROCH, J.N. & RATCLIFF, D. (1972). Generalised iterative scaling for log-linear models. *Ann. Math. Statist.* **43** , 1470-1480.

ERIKSEN, P.S. (1987). Proportionality of covariance matrices. *Ann. Statist.* **15**, 732-748.

FISHER, R.A. (1958). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

JENSEN, S.T. (1978). *Flersidede kontingenstabeller*. Lecture notes. University of Copenhagen (in Danish).

JENSEN, S.T. & JOHANSEN, S. (1987). Estimation of proportional covariances. *Statist. Prob. Letters* **6**, 83-85.

LAURITZEN S.L. (1982). *Lectures on contingency tables*. 2 nd ed. Aalborg University Press.

NELDER, J.A. & WEDDERBURN, R.W.M. (1972). Generalised linear models. *J. R. Statist. Soc.* **A**. **135**, 370-380.

SPEED, T.P. & KIIVERI, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138-150.

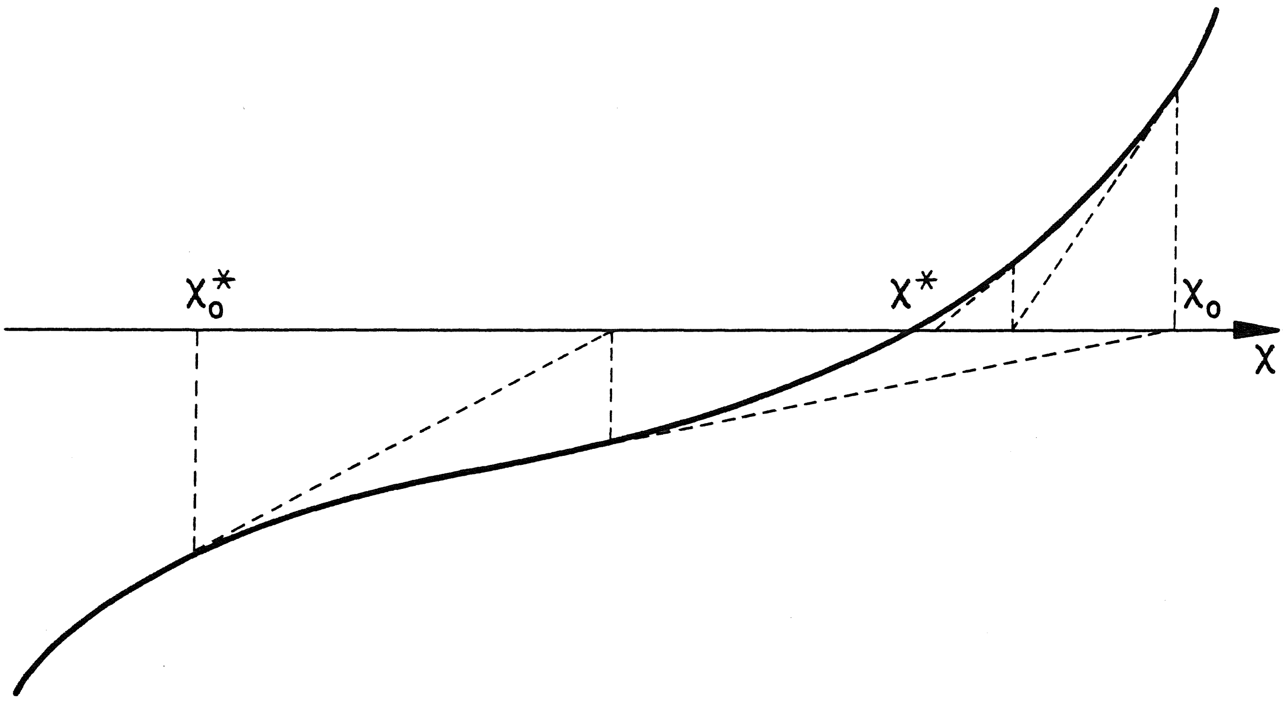

Fig 1. Convergence of the Newton-Raphson algorithm. If the algorithm starts at x_0, then x_n is decreasing. If it starts at x_0^* then it jumps to the other side of x^* and then decreases to x^*.

PREPRINTS 1987

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33.

No. 1 Jensen, Søren Tolver and Johansen, Søren: Estimation of Proportional Covariances.

No. 2 Rootzén, Holger: Extremes, Loads, and Strengths.

No. 3 Bertelsen, Aksel: On the Problem of Testing Reality of a Complex Multivariate Normal Distribution.

No. 4 Gill, Richard D. and Johansen, Søren: Product-Integrals and Counting Processes.

No. 5 Leadbetter, M.R. and Rootzén, Holger: Extremal Theory for Stochastic Processes.

No. 6 Tjur, Tue: Block Designs and Electrical Networks.

No. 7 Johansen, Søren: Statistical Analysis of Cointegration Vectors.

No. 8 Bertelsen, Aksel: On the Problem of Testing Reality of a Complex Multivariate Normal Distribution, II.

No. 9 Andersson, S.A. and Perlman, M.D.: Group-Invariant Analogues of Hadamard's Inequality.

No. 10 Hald, Anders: Two Generalizations of the Problem of Points by Bernoulli, de Moivre and Montmort.

No. 11 Andersson, Steen Arne: The Lattice Structure of Orthogonal Linear Models and Orthogonal Variance Component Models.

PREPRINTS 1988

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33.

No. 1 Jacobsen, Martin: Discrete Exponential Families: Deciding when the Maximum Likelihood Estimator Exists and is Unique.

No. 2 Johansen, Søren and Juselius, Katarina: Hypothesis Testing for Cointegration Vectors - with an Application to the Demand for Money in Denmark and Finland.

No. 3 Jensen, Søren Tolver, Johansen, Søren and Lauritzen, Steffen L.: An Algorithm for Maximizing a Likelihood Function.