On the convergence of the iterative solution of the likelihood equations

R. Moddemeijer

University of Groningen, Department of Computing Science, P.O. Box 800, NL-9700 AV Groningen, The Netherlands, e-mail: rudy@cs.rug.nl

Abstract

To determine the maximum likelihood estimate in case of the iterative solution of the likelihood equations we need a convergence criterion. Usually numerical convergence criteria are used; we propose a statistical convergence criterion. Our criterion reduces the number of iterations significantly. The iterative algorithm to solve the equations is known to be unstable in the neighborhood of the exact solution. Our method avoids this numerically unstable region. The method has been validated by simulations.

Key words: likelihood equations, convergence, maximum likelihood estimation, simulation, stability

1 Introduction.

The maximum likelihood (ML-) method to estimate a priori unknown parameters is widely used [8]. The solution of the likelihood equations to estimate these parameters is a well-known problem. We distinguish three different cases:

(1) The likelihood equations are a set of linear equations. Using linear algebra a solution of the equations can be found. This is for example the case in finding the parameters of an autoregressive (AR-) process.
(2) The likelihood equations are a set of non-linear equations which can be solved analytically; for example the estimation of the parameters of a bivariate normal distribution.
(3) The likelihood equations form a set of non-linear equations and no analytical solution is known. An iterative algorithm to solve the equations is used. Typical examples are: the estimation of the parameters of a moving average (MA-) model or of a mixed autoregressive and moving average (ARMA-) model [1,6,18], and recursive identification [16].

In relation to the last case there are many computational aspects; Gupta and Mehra state *it is not uncommon to find situations where the likelihood surface has multiple maxima, saddle-points, discontinuities, and singular Hessian matrix in the parameter space* [7, p. 774]. Gupta and Mehra also mention convergence as an important aspect.

In the same paper Gupta and Mehra conclude: *Several computational aspects of the maximum likelihood estimation have been discussed with the hope of stimulating further research in this area. From the practical standpoint, this is one of the most important problems in the applicability of system identification techniques to large scale systems. It is clear from our discussion that an extensive amount of work on related problems has been done by numerical analysts. An assimilation of this work and further extensions would lead us towards a routine use of maximum likelihood methods for estimation of parameters in dynamic systems* [7, p. 782]. After the papers of Bard [2] and of Gupta and Mehra hardly any publications about this convergence problem, which is characterized by an estimated function to be optimized, have appeared. Several recent authors mention convergence as one of the aspects of the method they used [17,19]. There exists a rather extensive literature about numerical optimization methods; see for example the references in the thesis of MacMillan [9].

To determine the position of the maximum we can use the standard methods of differential calculus, i.e. find the first derivatives of the (logarithmic) likelihood function with respect to the parameters and set these derivatives equal to zero. This results in a set of likelihood equations to be solved [5, p. 498]. Frequently these likelihood equations cannot be solved analytically.

There exist various iterative gradient methods to obtain a numerical approximation of the maximum of the (logarithmic) likelihood function [3, p. 147] [2]. Finding a good convergence criterion is a problem. Bard [2] used a numerical convergence criterion, as suggested by Marquardt [10]. The actual numerical convergence criterion is rather irrelevant; the fact that it is a *numerical* convergence criterion is essential. Why should we try to obtain numerical convergence if the parameter estimates are disturbed by an estimation error which is often several orders larger than the numerical error? A *statistical* convergence criterion would probably perform better!

2 Interpretation of the ML-method

Assume N independent and equally distributed observations $\mathbf{x}_n$, where $1 \leq n \leq N$, of the stochastic vector $\underline{\mathbf{x}}$ distributed according to the a priori unknown joint probability density function (pdf) $g(\mathbf{x})$. We are going to estimate the

parameter vector $\mathbf{p}$ such that the *correct* pdf $g(\mathbf{x})$ will optimally be modeled by the *model* pdf $f(\mathbf{x};\mathbf{p})$ which depends on the parameter vector.

To select the optimal parameter vector $\mathbf{p}$ we use the log-likelihood of the model pdf $f(\mathbf{x};\mathbf{p})$ given N observations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ as a criterion:

$$\text{log-likelihood} = \prod_{n=1}^{N} f(\mathbf{x}_n;\mathbf{p}) \tag{1}$$

We want to use the log-likelihood as an N independent criterion to validate the model pdf. This log-likelihood tends to decrease (or increase) with N. As a result this log-likelihood has no limit for $N \to \infty$. To overcome this problem we normalize the log-likelihood and take the limit for $N \to \infty$:

$$\text{MLL}\,(f;\mathbf{p}) = \lim_{N\to\infty} \tfrac{1}{N} \sum_{n=1}^{N} \log f(\mathbf{x}_n;\mathbf{p}) = \text{E}\,\{\log f(\underline{\mathbf{x}};\mathbf{p})\} \tag{2}$$

A unique criterion, the *Mean Log-Likelihood* (MLL), is assigned to every value of $\mathbf{p}$. The MLL measures the acceptability of $f(\mathbf{x};\mathbf{p})$ as an approximation of $g(\mathbf{x})$. The vector for which the criterion reaches a maximum is considered to be the optimal choice $\mathbf{p}^*$ to model $g(\mathbf{x}_n)$ by $f(\mathbf{x}_n;\mathbf{p}^*)$. The MLL is due to the *information inequality* [1] [4, p. 27] bounded by the neg(ative)-entropy:

$$\text{MLL}\,(f;\mathbf{p}) \le \text{E}\left\{\log g(\underline{\mathbf{x}})\right\} = -\text{H}\,\{\underline{x}\} \tag{3}$$

The equality holds for $\mathbf{p} = \mathbf{p}^*$ if and only if the correct pdf can exactly be modeled $f(\mathbf{x};\mathbf{p}^*) = g(\mathbf{x})$.

We introduce a new concepts: the *stochastic single observation logarithmic likelihood*:

$$\underline{l}(\mathbf{p}) = \log f(\underline{\mathbf{x}};\mathbf{p}) \tag{4}$$

which is a stochastic function. The stochastic single observation logarithmic likelihood of the parameter vector $\mathbf{p}$ given the observed random vector $\underline{\mathbf{x}}$ is merely the value of $\log f(\mathbf{x};\mathbf{p})$ selected by the stochastic vector $\underline{\mathbf{x}}$; we will use the term *stochastic log-likelihood* instead. This stochastic log-likelihood is a random function with a mean (function), the MLL-function, a variance, a pdf

[1] The information inequality [4, p. 27] in the information theory literature should not be confused with the information, Cramèr-Rao or Fréchet inequality [3, pp. 138-145] in the statistical literature

and so on. All techniques to determine the properties of stochastic variables can be applied to these stochastic log-likelihoods.

Unfortunately the MLL-function is a priori unknown, so it has to be estimated. We define a statistic to estimate the MLL-function; the *Average Log Likelihood* (ALL) function

$$\mathrm{ALL}\left(f;\mathbf{p}\right)=\widehat{\mathrm{E}}\left\{\underline{l}(\mathbf{p})\right\}=\tfrac{1}{N}\sum_{n=1}^{N}\log f(\mathbf{x}_n;\mathbf{p}) \tag{5}$$

where $\widehat{\mathrm{E}}\left\{\ldots\right\}$ denotes an average; a statistic to estimate the mean. The ALL-function is merely a normalized log-likelihood function (divided by N) with an essentially different interpretation: the ALL as a statistic to estimate the MLL. Analogous to the average which is an unbiased statistic to estimate the mean, the ALL is an unbiased statistic to estimate the MLL. Conform th ML-method the value of the parameter vector $\widehat{\mathbf{p}}$ for which the ALL-function reaches a maximum is the ML-estimate of $\mathbf{p}^*$.

Its now a rather trivial step to consider the variance of the statistic ALL-function. Due to the fact that this statistic is constructed by averaging N independent observations of the stochastic log-likelihood, we compute the variance as usual, in case of averaging observations:

$$\mathrm{VAR}\left\{\widehat{\underline{\mathrm{E}}}\left\{\underline{l}(\mathbf{p})\right\}\right\}=\tfrac{1}{N}\mathrm{VAR}\left\{\underline{l}(\mathbf{p})\right\} \tag{6}$$

This variance can easily be estimated:

$$\begin{aligned}\widehat{\mathrm{VAR}}\left\{\underline{l}(\mathbf{p})\right\}&=\tfrac{N}{N-1}\left(\tfrac{1}{N}\sum_{n=1}^{N}\left(l_n(\widehat{\mathbf{p}})\right)^2-\left(\tfrac{1}{N}\sum_{n=1}^{N}l_n(\widehat{\mathbf{p}})\right)^2\right)\\&=\tfrac{N}{N-1}\left(\tfrac{1}{N}\sum_{n=1}^{N}\left(\log f(\mathbf{x}_n;\widehat{\mathbf{p}})\right)^2-\left(\tfrac{1}{N}\sum_{n=1}^{N}\log f(\mathbf{x}_n;\widehat{\mathbf{p}})\right)^2\right)\end{aligned} \tag{7}$$

where $\widehat{\mathrm{VAR}}\left\{\ldots\right\}$ is a statistic to estimate the variance.

3 The convergence criterion

In the previous section we have shown that in the ML-method the MLL-function is used as a criterion; this MLL-function can be estimated by the ALL-function. This ALL-function has a variance which can easily be estimated.

To search for the maximum of the ALL-function, we solve the likelihood equations.

$$\frac{\partial}{\partial \mathbf{p}} \widehat{\mathrm{E}} \left\{ \underline{l}(\mathbf{p}) \right\} = \mathbf{0} \tag{8}$$

If these equations have a unique solution for $\mathbf{p} = \widehat{\mathbf{p}}$ and this solution corresponds with the absolute maximum of the ALL-function, the ML-estimate $\widehat{\mathbf{p}}$ is found.

These equations can have an analytical solution, but often we have the unfortunate situation that the solution can only be approximated. To approximate this solution the following recursive algorithm can be used:

$$\widehat{\mathbf{p}}_k = \widehat{\mathbf{p}}_{k-1} \; - \; \lambda \, \widehat{\mathrm{E}} \left\{ \ddot{\underline{l}}(\widehat{\mathbf{p}}_{k-1}) \right\}^{-1} \widehat{\mathrm{E}} \left\{ \dot{\underline{l}}(\widehat{\mathbf{p}}_{k-1}) \right\} \tag{9}$$

where $\widehat{\mathbf{p}}_k$ is the estimate of $\mathbf{p}^*$ in the k^{th} iteration and where λ is a conveniently chosen scalar. This recursive algorithm is a gradient method to solve the likelihood equations. Various gradient methods differ from each other in the choice of λ [2, p. 159]. We denote the vector of first order derivatives (gradient vector) and matrix of second order derivatives (Hessian matrix) by:

$$\dot{\underline{l}}(\widehat{\mathbf{p}}_k) = \frac{\partial}{\partial \mathbf{p}} \underline{l}(\mathbf{p}) \bigg|_{\mathbf{p}=\widehat{\mathbf{p}}_k} \qquad \ddot{\underline{l}}(\widehat{\mathbf{p}}_k) = \frac{\partial}{\partial \mathbf{p}} \frac{\partial}{\partial \mathbf{p}}^{\mathsf{T}} \underline{l}(\mathbf{p}) \bigg|_{\mathbf{p}=\widehat{\mathbf{p}}_k} \tag{10}$$

The computation of the Hessian matrix is computationally expensive with respect to the computation of the gradient vector. To assure convergence for a λ where $0 < \lambda \leq 1$ the inverse matrix in (9) should be negative definite (minus a positive definite matrix). The inverse of a negative definite matrix is also negative definite. To reduce the computational effort and to assure the matrix is negative definite, the approximation

$$\widehat{\mathrm{E}} \left\{ \ddot{\underline{l}}(\widehat{\mathbf{p}}) \right\} \approx -\widehat{\mathrm{E}} \left\{ \dot{\underline{l}}(\widehat{\mathbf{p}}) \cdot \dot{\underline{l}}(\widehat{\mathbf{p}})^{\mathsf{T}} \right\} \tag{11}$$

is used instead. This approximation is based on two equivalent forms of Fisher's information matrix:

$$\mathrm{E} \left\{ \ddot{\underline{l}}(\mathbf{p}^*) \right\} = -\mathrm{E} \left\{ \dot{\underline{l}}(\mathbf{p}^*) \cdot \dot{\underline{l}}(\mathbf{p}^*)^{\mathsf{T}} \right\} \tag{12}$$

This equality holds under the same conditions as the information inequality [3, p. 141].

The negative definite Hessian matrix in (9) assures that roughly a step into the direction of the gradient vector will be made. The size of this step depends on λ. If the ALL-function round the maximum has an exact parabolic shape, then $\lambda = 1$ is optimal and convergence is reached in one step. Therefore usually $\lambda = 1$ is used, but this choice may lead to divergence $\widehat{\mathrm{E}}\{\underline{l}(\widehat{\mathbf{p}}_k)\} < \widehat{\mathrm{E}}\{\underline{l}(\widehat{\mathbf{p}}_{k-1})\}$. For the k^{th} iteration there exists a λ where $0 < \lambda \leq 1$ such that the result is an improvement with respect to the $(k-1)^{th}$ iterations: $\widehat{\mathrm{E}}\{\underline{l}(\widehat{\mathbf{p}}_k)\} > \widehat{\mathrm{E}}\{\underline{l}(\widehat{\mathbf{p}}_{k-1})\}$. So the algorithm converges towards a (local) maximum. In the literature there exist several sophisticated methods to find the optimal λ for the k^{th} iteration to speed up the convergence. We use $\lambda = 2^{-m}$ where m is the least integer number such that the resulting $\widehat{\mathbf{p}}_k$ is an improvement over $\widehat{\mathbf{p}}_{k-1}$. If another value than the initial guess $\lambda = 1$ is used, we call the algorithm unstable.

Now remains the problem to decide after how many iterations an adequate estimate is reached?

- *Numerical convergence criterion*; the iterations proceed until the relative change in the estimated parameter vector is negligible. What negligible is depends on: the numerical precision of the computer, the numerical stability of the algorithm and the preferred accuracy of the parameter estimates. For convenience we use the upper limit to the relative change of all parameter estimates during the last iteration as a criterion.
- *Statistical convergence criterion*; the iterations are proceeded until the change of the estimated parameter vector is negligible with respect to the accuracy of the estimates.

Usually researchers apply numerical convergence criteria; they are easy to implement. Statistical convergence criteria reduce the number of iterations to a minimum, so they are to be preferred. The computation of error-bounds for all elements of the estimated parameter vector during the convergence is still an unsolved problem. After convergence, for example, the Cramér-Rao bound, which is actually a lower-bound on the error, can be used to estimate these error-bounds.

Why should we try to estimate error-bounds on all estimated elements of the parameter vector? Two parameter vectors are indistinguishable if their corresponding ALL's are indistinguishable. So we can also construct criteria for statistical convergence which are fully based on the convergence of the ALL! Therefore we propose the convergence criterion:

$$\widehat{\mathrm{E}}\{\underline{l}(\mathbf{p}_k)\} - \widehat{\mathrm{E}}\{\underline{l}(\mathbf{p}_{k-1})\} < \eta\sqrt{\tfrac{1}{N}\widehat{\mathrm{VAR}}\{\underline{l}(\widehat{\mathbf{p}}_{k-1})\}} \tag{13}$$

The constant η where $0 < \eta \leq 1$ is some arbitrary value to express what we mean by negligible; we use $\eta = 0.1$. The two main advantages of this criterion are:

(1) it is easy to implement and hardly requires computational effort and
(2) it is a statistical convergence criterion.

The computational effort to compute the criterion is, besides some overhead, N multiplications and N additions to estimate the standard error of the right-hand side of (13). It can be computed in parallel with the computation of $\widehat{\mathrm{E}}\{\underline{l}(\mathbf{p}_k)\}$.

4 Simulations

To validate the method we set up some simulations. We searched for a simple problem with a well-known outcome which can be solved both analytically and numerically. Although an important field of applications lies in AR-, MA- and ARMA- modeling, we have chosen for a simple statistical estimation problem to observe the convergence in a pure context: the estimation of the parameters μ_x, μ_y, σ_x, σ_y and ρ of the bivariate normal distribution

$$f(x,y;\mathbf{p}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp -\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right\} \quad (14)$$

where $\mathbf{p} = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$. In the simulation N pairs of observations (x_n, y_n) distributed according to $f(x,y;\mathbf{p})$ were generated, where $1 \leq n \leq N$, $\mu_x = 2$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 1$ and $\rho = 0.5$. From these observations the parameter vector $\mathbf{p}$ was estimated .

The set of likelihood equations has a known analytical solution [3, p. 152]. The numerical approximation of the solution is rather troublesome, because we are dealing with a highly non-linear set of equations; therefore this is an excellent example to verify our theory.

The initial guess of the parameter vector $\widehat{\mathbf{p}}_1$ was randomly generated. For each element we have chosen for a normally distributed guess, where σ is 10% of the mean.

The numerical convergence criterion is an upper limit to the relative change of all parameters. E.g. by $1:1000$ we mean that none of the parameters may change more than 0.1% during the last iteration.

N=100	1	2	3	4	5	6	7	8	9	10
1:100	0.1	8.0	47.9	84.2	95.2	97.9	98.8	99.2	99.4	99.4
1:1000	0.0	0.0	1.0	10.3	31.5	55.0	67.5	76.2	82.0	86.7
1:10000	0.0	0.0	0.0	0.0	2.1	11.2	26.1	42.1	53.8	62.6
stat	5.9	34.5	83.1	97.9	99.4	99.8	99.9	99.9	100.0	100.0
	11	12	13	14	15	16	17	18	19	20
1:100	99.7	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8
1:1000	89.6	91.0	92.2	93.4	94.2	95.5	96.3	97.0	97.5	97.5
1:10000	69.2	75.4	80.2	84.3	86.4	88.5	90.3	92.3	93.1	93.1
stat	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

N=1000	1	2	3	4	5	6	7	8	9	10
1:100	0.1	19.9	72.8	96.5	99.7	100.0	100.0	100.0	100.0	100.0
1:1000	0.0	0.0	11.2	56.9	93.0	99.2	99.9	100.0	100.0	100.0
1:10000	0.0	0.0	0.4	18.1	63.0	92.2	98.8	99.8	100.0	100.0
stat	0.7	17.8	71.0	96.4	99.6	100.0	100.0	100.0	100.0	100.0

Table 1
The estimated cumulative probability of convergence, expressed as a percentage, as a function of the number of iterations for the different convergence criteria.

5 Discussion

Table 1 shows the estimated cumulative probability of convergence, expressed as a percentage, as a function of the number of iterations for the different convergence criteria. In this table we observe that the number of iterations clearly depends on the acquired numerical accuracy. The statistical convergence criterion performs better than the numerical criteria used. The accuracy of the estimates using the statistical convergence criterion is equivalent to the accuracy using the most accurate numerical convergence criterion (see table 3). For $N = 100$ and a numerical accuracy of $1 : 100$ we observe in table 3 an enlarged standard deviation; consequently the parameter estimates are in this case less accurate. Obviously for $N = 100$ a relative change of $1 : 100$ is as a convergence criterion not stringent enough.

The iterative algorithm is for $N = 100$ in 29.7% and for $N = 1000$ in 7.2% of the observed cases unstable during the initial search (see table 2). After this initial instability the algorithm converges towards the optimum. This initial instability can be avoided by using better initial guesses. Any attempt to achieve an accurate numerical convergence leads to instability of the algorithm

N=100	1	2	3	4	5	6	7	8	9	10
1:100	29.3	29.5	30.2	30.7	31.6	31.7	31.8	31.8	31.8	31.8
1:1000	29.3	29.5	30.2	30.9	32.5	33.0	34.4	34.4	34.6	34.7
1:10000	29.3	29.5	30.2	30.9	33.6	37.4	45.1	51.4	57.8	62.0
stat	29.3	29.5	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7
	11	12	13	14	15	16	17	18	19	20
1:100	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8
1:1000	34.8	35.0	35.0	35.1	35.1	35.2	35.3	35.3	35.3	35.3
1:10000	65.4	68.4	70.3	72.6	74.2	75.4	76.7	77.4	77.8	77.8
stat	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7
N=1000	1	2	3	4	5	6	7	8	9	10
1:100	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2
1:1000	7.2	7.2	7.6	15.5	23.5	24.8	25.0	25.0	25.0	25.0
1:10000	7.2	7.2	7.6	19.5	44.3	59.1	62.4	63.1	63.2	63.2
stat	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2

Table 2
The estimated cumulative probability of occurrence of the first sign of instability, expressed as a percentage, as a function of the number of iterations for the different convergence criteria.

N=100	μ_x	μ_y	σ_x	σ_y	ρ
1:100	1.998 ± 0.133	1.998 ± 0.124	0.936 ± 0.076	0.934 ± 0.080	0.507 ± 0.078
1:1000	2.001 ± 0.100	1.997 ± 0.100	0.990 ± 0.067	0.987 ± 0.068	0.496 ± 0.074
1:10000	2.001 ± 0.101	1.997 ± 0.101	0.992 ± 0.068	0.990 ± 0.068	0.496 ± 0.075
stat	2.002 ± 0.106	1.999 ± 0.103	0.984 ± 0.071	0.981 ± 0.072	0.497 ± 0.075
N=1000	μ_x	μ_y	σ_x	σ_y	ρ
1:100	2.001 ± 0.031	2.001 ± 0.032	0.999 ± 0.022	0.998 ± 0.023	0.500 ± 0.023
1:1000	2.001 ± 0.031	2.000 ± 0.032	0.999 ± 0.022	0.999 ± 0.022	0.500 ± 0.023
1:10000	2.001 ± 0.031	2.000 ± 0.032	0.999 ± 0.022	0.999 ± 0.022	0.500 ± 0.023
stat	2.001 ± 0.034	2.001 ± 0.033	0.998 ± 0.025	0.997 ± 0.025	0.500 ± 0.023

Table 3
The average and standard deviation of the estimated parameters. The exact values are $\mu_x = 2$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 1$ and $\rho = 0.5$.

and to excessive computational effort. This effort is not rewarded by better estimates (see table 3).

It is doubtful whether sophisticated methods to search for numerical exact solutions of the likelihood equations as evaluated by Bard [2] makes any sense. The improvements with respect to the numerical accuracy are negligible with respect to the statistical accuracy. It seems that any attempt to achieve better estimates than allowed by the statistical accuracy is punished by numerical instability and excessive computational effort.

The concept of the stochastic log-likelihood opens a new field of research. The stochastic log-likelihood can be treated as any stochastic signal. A wide range of new methods in the fields of signal processing, statistics and information theory can be developed [11,12,13,14,15].

6 Conclusion

A statistical convergence criterion to test convergence in the iterative solution of the likelihood equations performs significantly better than a numerical criterion. The iterative method is unstable during the preliminary search and during the accurate search for the optimal solution. Using the statistical criterion the iteration will often be ended before the instability during the accurate search occurs. The resulting maximum likelihood estimates are statistically indistinguishable from the estimates obtained by the best numerical convergence criterion. Using the statistical criterion an increase of the number of iteration steps with respect to numerical convergence has never been observed. In our simulations the reduction of the number of iteration steps ranges from no reduction to a reduction by a factor of 3.

The statistical convergence criterion performs in all simulations equally well or better than numerical convergence. The statistical convergence criterion is easy to implement and costs hardly any computational effort during the iteration process.

References

[1] K. J. Åström. Maximum likelihood and prediction error methods. *Automatica*, 16:551–574, 1980.

[2] Y. Bard. Comparison of gradient methods for the solution of the nonlinear parameter estimation problems. *SIAM Journal on Numerical Analysis*, 7(1):157–186, 1970.

[3] S. Brandt. *Statistical and Computational Methods in Data Analysis*. North-Holland Publ. Comp., Amsterdam (NL), 2^{nd} edition, 1976.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[5] H. Cramér. *Mathematical methods of statistics*. Princeton, Princeton Univ. Press, 1945.

[6] B. Friedlander. A recursive maximum likelihood algorithm for ARMA spectral estimation. *IEEE Trans. on Information Theory*, 28(4):6539–646, 1982.

[7] N. K. Gupta and R. K. Mehra. Computational aspects of maximum likelihood estimation and reduction in sensitivity fuction calculations. *IEEE Trans. on Automatic Control*, 19(6):774–783, 1974.

[8] F. Gustafsson and H. Hjalmarsson. Twenty-one ML estimators for model selection. *Automatica*, 31(10):1377–1392, 1995.

[9] D. MacMillan. *Relaxing convergence conditions to improve the convergence rate*. PhD thesis, Univ. of Colorado, Denver, 1999.

[10] D. W. Marquardt. An algorithm for least squares estimation of nonlinear parameters. *SIAM Journal on Numerical Analysis*, 11:431–441, 1963.

[11] R. Moddemeijer. *Delay-Estimation with Application to Electroencephalograms in Epilepsy*. PhD thesis, University of Twente, Enschede (NL), 1989.

[12] R. Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16(3):233–246, 1989.

[13] R. Moddemeijer. Testing composite hypotheses applied to AR order estimation; the Akaike-criterion revised. In *Signal Processing Symposium (SPS '98)*, pages 135–138, Leuven (B), March 26-27 1998. IEEE Benelux Signal Processing Chapter.

[14] R. Moddemeijer. A statistic to estimate the variance of the histogram based mutual information estimator based on dependent pairs of observations. *Signal Processing*, 75(1):51–63, 1999.

[15] R. Moddemeijer. Testing composite hypotheses applied to AR-order estimation; the Akaike-criterion revised. In C. Beccari et. al., editor, *European Conference on Circuit Theory and Design (ECCTD '99)*, pages 723–726, Stresa (I), August 29-September 2 1999.

[16] T. Söderström, L. Ljung, and I. Gustavsson. A theoretical analysis of recursive identification methods. *Automatica*, 14:231–244, 1978.

[17] Fang-Kuo Sun. A maximum likelihood algorithm for mean and covariance of nonidentically distributed observations. *IEEE Trans. on Automatic Control*, 27(1):245–247, 1982.

[18] T. Westerlund. A digital quality control system for an industrial dry process rotary cement kiln. *IEEE Trans. on Automatic Control*, 26(4):885–890, 1981.

[19] L. B. White. An iterative method for exact maximum likelihood estimation of the parameters of a harmonic series. *IEEE Trans. on Automatic Control*, 38(2):367–370, 1993.