T. Bohlin

Comparison of Two Methods of Modeling Stationary EEG Signals

Abstract: This paper compares the performances—when applied to stationary EEG signals—of two methods of modeling stochastic time-series; viz, a maximum-likelihood search method for a mixed autoregressive and moving-average time-series, and the well-known method of least-squares identification of a prefiltered autoregressive series. Computing effort per step is derived for different-order search strategies, expressed in the number of a set of basic operations and also in equivalent number of likelihood function evaluations. Power spectra and total computing effort are evaluated and compared for four representative EEG samples. It appears that the least-squares method is generally superior because iteration is not needed, and in spite of the fact that higher orders are needed instead.

1. Introduction

The spontaneous electrical activity of the brain is normally measured by means of electrodes placed on the patient's scalp. The recordings of such measurements resemble random signals, occasionally containing a dominating frequency, called *rhythm*, and/or pulse-like activity, called *spikes*. These recordings, the electroencephalograms or EEG signals, are generally believed to hold information about the brain and are used by the physician as a means for diagnosis.

This paper originates in a joint study of EEG signals, involving scientists from the Karolinska Hospital, the Royal Institute of Technology and the IBM Nordic Laboratory in Sweden. The aim of the study is to find methods to supplement the diagnostician's visual inspection of EEG recordings with mathematical analysis, and in general to provide means for accessing the information in the EEG [23].

A fully automatic diagnosis would consist of two phases:

- Analysis of an EEG sample to obtain the particular characteristics of the EEG that distinguish it from other EEG's.
- Medical classification of the sample by means of its characteristics. One class would be that of "uncertain cases."

The problems associated with the second phase are not treated in this paper. The first phase involves three problems:

- To find a mathematical description of EEG's; i.e., a parametric class $\mathcal{M}$ of models that is wide enough to cover essential properties of a substantial class of physical EEG signals.
- To find numerical identification methods; i.e., to associate a particular element in this predefined class with a given EEG sample by assigning values to the parameters. Thus, the parameter values characterize the EEG and carry the information extracted from the EEG.
- To find a test criterion for the case in which a given EEG sample cannot be described by a model in the class $\mathcal{M}$.

At least the first two problems have been treated extensively in the literature [8,16,18]. Essentially two kinds of description have been tried:

1. Stationary stochastic processes. The information obtained comprises various time-invariant statistical characteristics of the EEG signal, such as power spectrum or amplitude distribution or other characteristics derived from these. Methods using correlation analysis [13], frequency analysis [10], reversed correlation analysis [14], and the fitting of time-series models [12,21,25] belong to this category.
2. Wave patterns. This yields essentially the frequency of occurrence and the form of particular, characteristic patterns [11,17].

The essential prerequisite for applying the first alternative is that characteristics be constant in time. Long samples (of the order of tens of seconds) are used for analysis. In the second alternative any wave form is allowed; however, to render the analysis practical the

wave forms must belong to only few and restricted classes. This means that attention is focused on transient properties and short samples are used. For these reasons EEG's with slowly changing characteristics are often treated as piece-wise stationary and the theory investigates the effects of limited sample lengths [24]. In addition, at least two methods have been developed that do not assume (piece-wise) stationarity and operate on samples that are not short; viz, those using "complex demodulation" [18] and parameter-tracking by a Kalman estimator [7].

Thus, the various descriptions imply different ways to process an EEG signal for analysis and also yield different kinds of information about the EEG. Therefore, it is generally difficult to compare different methods with respect to performance, and such comparisons are few. Mainly, a difficulty is that a common basis must be found, relating elements in different classes of models. A natural common basis for stationary EEG's, used in Ref. 15, is the power spectrum.

However, the estimated power spectra arrived at by any method are necessarily estimates of smoothed versions of an imaginary, "true" spectrum, and the principles for smoothing differ. In frequency analysis methods smoothing is defined by means of frequency or lag "windows" [4]; in a model-fitting method the postulated model structure defines the smoothing implicitly, and no window is needed. Also, analysis methods usually contain design parameters to be set by the user, such as filters, windows and model orders, and their values may have a significant effect on the performance. A fair comparison would require that one first optimize these parameters. A second, difficult to satisfy, requirement for a fair comparison is that test cases must not favor any of the methods to be compared.

This paper attempts to compare two methods that use time-series models: a maximum likelihood identification method of a mixed autoregressive and moving-average time-series, proposed by Åström and Bohlin [1], and similar to that applied by Zetterberg [21], and the well-known method of least-squares identification of an autoregressive series, applied to EEG analysis by Gersch [12] and by Fenwick, et al. [25]. These two methods are sufficiently related to make a comparison feasible and yet, sufficiently different to make it interesting.

The models arrived at by the two methods have different forms and therefore cannot be compared directly. However, form need not be physically relevant. The following, physically relevant characteristics can be evaluated for both methods and are used for the comparison:

- Power spectrum of the model.
- Model quality, as measured by the residual variance, or equivalently by the rms value of the error incurred when each model is used to predict the EEG signal one sampling interval ahead.
- Computing effort, measured as the number of repetitions of a common set of algebraic operations (accumulation of lagged products) approached for long samples. This makes the measure independent of the particular computer and the particular programming used.

Essential design parameters, affecting all three characteristics, are the model orders, which are defined differently for the two methods. However, it appears that in both cases orders do not affect spectra appreciably, if they are high enough. This makes a comparison of models feasible.

Also, for a given EEG sample and each method it is possible to establish a highest order such that the model quality does not improve significantly for still higher orders. This suggests two ways to compare computing effort: For a number of test cases either 1) use the significant orders for both methods or, 2) use the significant order for one of the methods (the maximum likelihood method, say) and set the order of the other method such that model qualities coincide. Evaluate and compare computing efforts.

Thus, a fair comparison between maximum-likelihood and generalized least-squares methods, when applied to EEG signals, is made possible by the following fortunate circumstances:

1. The models have comparable power spectra.
2. Design parameters can be set by objective means, thus fixing both the degree of smoothing and the amount of computing for each method.
3. The set of arithmetic operations dominating the computing for long samples is common to both methods.

In practice a number of things reduce the discriminating power of a comparison:

- Results of order tests that use physical data are generally somewhat ambiguous. This is a consequence, of course, of the fact that a physical signal cannot be trusted to have a well-defined order.
- The generalized least-squares method has other design parameters besides the order; that is, it uses a prefilter. The latter affects the order substantially. An attempt to compute the optimal prefilter leads to the so-called modified generalized least-squares method [9], which is not evaluated in this paper. Fortunately, it appears that the prefilter does not affect the model quality appreciably (if the order is high enough), and that in practice it is sufficient to choose a prefilter heuristically and use it for all samples.
- The result depends on the particular test case. Averaging generally requires many test cases.

For the comparison carried out in this paper only four EEG samples are used for test cases. However, the discriminating power appears to be sufficient since the results are reasonably clear in all cases. Spectra agree well, and the generalized least-squares method needs substantially less computing than the particular maximum-likelihood algorithm used, because iteration is not needed, and in spite of the fact that higher orders are needed instead.

2. Theory

Zetterberg [21,22] proposed the following class of models for a description of stationary parts of an EEG signal $y(t)$, with zero mean and sampled with the interval h:

$$y(t) + a_1 y(t-h) + \cdots + a_n y(t-nh) = \lambda[e(t) + c_1 e(t-h) + \cdots + c_m e(t-mh)]\,. \quad (1)$$

Here $\{e(t)\}$ is a sequence of uncorrelated, normal random variables with zero means and unit variances. $a_1, \cdots, a_n, c_1, \cdots, c_m$, are parameters, m and n are integers defining the order of the model, and λ is a scaling factor.

It is shown in Ref. 21 that the form (1) results from superimposing signal components of three kinds, generated by zero-, first-, and second-order difference equations driven by uncorrelated sequences:

1. White noise.
2. Random signals of low-pass type, characterized by power and noise bandwidth.
3. Random signals of band-pass type, characterized by power, frequency, bandwidth, and skewness.

The motivation for adopting the form (1) is the fact that its spectrum can be decomposed into such components, each having a description that is easy to interpret physically and identify as one of the conventional α, β, δ or θ rhythms.

In Ref. 19 Wennberg and Zetterberg demonstrated experimentally, using records from healthy subjects, that models of the fifth order ($m = n = 5$) would normally describe adequately such samples that were manually selected as free of artifacts. The identification principle used was that of maximum likelihood [1]. Generally, the likelihood function is a function of the unknown parameters in a model. It may be interpreted as measuring the likelihood that any given model could have produced the actually observed signal. The particular model corresponding to the maximum of that function is therefore of interest and is known as the maximum-likelihood model.

Zetterberg used a new algorithm for finding the maximum of the likelihood function [21]. This algorithm differs essentially from the one proposed in Ref. 1 in two respects: 1) the likelihood is expressed in terms of sample autocorrelations instead of the sample values, and 2) a different optimization algorithm is used for finding the parameter estimates.

The use of sample autocorrelations in place of the data sample is believed to be advantageous in cases where a hardware correlator is available, and, possibly, even when the analysis is done exclusively by a digital computer, since the original data sample has to be processed only once for an approximate evaluation of the likelihood of alternative models. However, the efficient optimization scheme of Refs. 1 and 5 cannot be used, since derivatives are not readily available. Thus, the algorithm based on sample autocorrelations can be expected to offer fewer computations per iteration step, but also to need more steps to find the estimate. The method reported in this paper uses a modification of the Newton-Raphson algorithm described in Ref. 5; it is not based on sample autocorrelations.

Setting $c_1 = \cdots = c_m = 0$ in (1) yields the autoregressive series applied to EEG signals in [12] and [25]. To compensate for the loss of degrees of freedom in the model, the order n has to be increased, but the point is that estimating the remaining coefficients $a_1, \cdots, a_n, \lambda$ is done by a least-squares algorithm, which does not need iteration. A slight modification is that of using constant but nonzero c_i values; i.e., the c_i values are the same for all samples. This leads to the so-called generalized least-squares algorithm, needing only slightly more computing for a given order n. The point is that a good choice of c coefficients may reduce the necessary order.

• 2.1 Power spectra

Let $\mathcal{M}$ be the class of models represented by (1), where both coefficient arrays a and c as well as m and n, are used as characteristics. Accepting the assumption that the class $\mathcal{M}$ is adequate for modeling stationary EEG signals does not immediately make it clear that a smaller class $\mathcal{M}'$ can also be used, where the value of c is common to all samples and a, alone, is used to characterize an individual sample. However, a weakly stationary random signal is uniquely characterized by its spectrum. Therefore, identifying the signal by a model of the class $\mathcal{M}$ means approximating its spectrum by the spectral function associated with (1):

$$\lambda^2 \left| \frac{C(\exp 2\pi i f h)}{A(\exp 2\pi i f h)} \right|^2, \quad (2)$$

where

$$A(z) = 1 + a_1 z + \cdots + a_n z^n,$$
$$C(z) = 1 + c_1 z + \cdots + c_m z^m. \quad (3)$$

Since (2) is a symmetric function it is always possible to express the spectrum as a ratio of polynomials in cos $(2\pi f h)$. It follows that (2) can approximate any continu-

ous nonnegative spectrum in the range $fh \in (0, 0.5)$. However, the same condition holds for a single polynomial, and, therefore, also for a model in $\mathcal{M}'$ (which means approximating the signal spectrum by the ratio of a fixed polynomial $|C'|^2$ and a variable one $|A'|^2$) as long as both the spectrum and $|C'|^2$ are continuous and positive. Since the C' polynomial in $\mathcal{M}'$ is fixed, it may not coincide with that in $\mathcal{M}$. The A' polynomial in $\mathcal{M}'$ is generally of higher order than that in $\mathcal{M}$; the extra factors have the functions of both 1) approximating the factors in $|C|^{-2}$ that are not approximated adequately by factors in $|C'|^{-2}$, and 2) compensating for the factors in $|C'|^2$ that do not fit any of the factors in $|C|^2$.

It follows that the choice of C' should not affect the degree of fit of the model; however, a well-designed C' polynomial may reduce the order of the A' polynomial to n. On the other hand, a bad choice may increase the order by m at most.

• 2.2 *Computing effort*

To derive the amount of computing needed to fit a model of the class $\mathcal{M}$ or $\mathcal{M}'$ to data it is necessary to investigate the algorithms in some detail. Both result from an application of the maximum-likelihood principle. This principle assigns a model, within a postulated parametric class, to a given data sample by selecting the parameter vector $\hat{\theta}$ that minimizes the loss function [1]

$$V(\theta) = \frac{1}{2} \sum_{t=1}^{N} \epsilon^2(t|\theta) , \tag{4}$$

where the "residuals" $\epsilon(t|\theta)$ satisfy the model

$$y(t) = F(z^{-1}|\theta)\epsilon(t|\theta), \qquad t = 1, \cdots, N . \tag{5}$$

Here, $F(z^{-1}|\theta)$ is a postulated, analytic function of the backwards-shift operator z^{-1} and the parameters. It defines the class of models.

For both classes $\mathcal{M}$ and $\mathcal{M}'$

$$F(z^{-1}|\theta) = C(z^{-1})/A(z^{-1}), \tag{6}$$

but

$\theta = \{a_1, \cdots, a_n, c_1, \cdots, c_m\}$ in $\mathcal{M}$, and

$\theta = \{a_1, \cdots, a_n\}$ in $\mathcal{M}'$.

In $\mathcal{M}$ the minimum is reached by a second-order search method. In $\mathcal{M}'$ (the generalized least-squares method) $V(\theta)$ is a quadratic function and a "search" converges in one step. However, in both cases the computations that dominate for large sample lengths N consist of evaluating first- and second-order partial derivatives of $V(\theta)$. Other computations (inverting the second-order derivative matrix, evaluating the spectrum, etc.) do not grow with N and will not be included in the measure of computing effort.

The partial derivatives are obtained from the relations [1]

$$\begin{aligned} \delta V(\theta) &= \sum_{i=1}^{m+n} \frac{\delta V}{\delta \theta_i} \delta\theta_i \\ &= \sum_{i=1}^{n} \delta a_i \sum_t \epsilon(t) y_c(t-i) \\ &\quad - \sum_{j=1}^{m} \delta c_j \sum_t \epsilon(t)\epsilon_c(t-j) , \end{aligned} \tag{7}$$

$$\begin{aligned} \delta^2 V(\theta) &= \sum_{i,j=1}^{m+n} \frac{\delta^2 V}{\delta\theta_i \delta\theta_j} \delta\theta_i \delta\theta_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \delta a_i \delta a_j \sum_t y_c(t-i) y_c(t-j) \\ &\quad + 2\sum_{i=1}^{n}\sum_{j=1}^{m} \delta a_i \delta c_j \sum_t [-y_c(t-i)\epsilon_c(t-j) \\ &\qquad - \epsilon(t) y_{cc}(t-i-j)] \\ &\quad + \sum_{i=1}^{m}\sum_{j=1}^{m} \delta c_i \delta c_j \sum_t [\epsilon_c(t-i)\epsilon_c(t-j) \\ &\qquad + 2\epsilon(t)\epsilon_{cc}(t-i-j)] , \end{aligned} \tag{8}$$

where

$$\begin{aligned} &C(z^{-1})\, y_c\,(t) = y(t), \\ &C(z^{-1})\, \epsilon_c\,(t) = \epsilon(t) = A(z^{-1})\, y_c\,(t), \\ &C(z^{-1})\, y_{cc}\,(t) = y_c\,(t), \text{ and} \\ &C(z^{-1})\, \epsilon_{cc}\,(t) = \epsilon_c\,(t). \end{aligned} \tag{9}$$

Terms containing δc_j vanish in $\mathcal{M}'$.

For large N the total computing effort needed to perform an r-order search is generally $J = N_r M$, where M is the number of iterations and N_r is the computing effort needed to evaluate loss and loss derivatives up to rth order; $r = 0$ if only the loss function has to be evaluated, $r = 1$ for gradient methods, and $r = 2$ for second-order methods. A useful, incomplete second-order method is based on omitting the y_{cc} and ϵ_{cc} terms in (8) [5]. Denote the corresponding computing effort by $N_2{}^*$. The number of steps M is generally a stochastic variable, which depends on the data and on the search strategy, while the computing effort per step N_r is a function of the order and the sample length only.

2.2.1 *Computing effort per step*

An inspection of Eqs. (4), (7), (8), and (9) reveals that evaluating loss and loss derivatives involves only two kinds of computation; viz,

1. Evaluating residuals ϵ and the auxiliary sequences y_c, y_{cc}, ϵ_c, ϵ_{cc}. The latter are obtained as solutions x_c of a common linear difference equation $C(z^{-1})x_c(t) = x(t)$

with $x(t) = y(t)$, $y_c(t)$, $\epsilon(t)$, $\epsilon_c(t)$, while $\epsilon(t) = A(z^{-1})y_c(t)$.

2. Evaluating a number of cross-correlations between y, ϵ, and the auxiliary sequences.

Also, the operation in both tasks 1 and 2 comprises forming sums of products, where the factors are arranged sequentially in two arrays in the computer memory. The computing effort per step in the search can therefore be measured in terms of the number of repetitions of one basic operation that comprises a fixed sequence of elementary operations: address modification, loading, multiplication, accumulation, and storing. This number is independent of machine, programming language, fixed- or floating-point arithmetic, etc. Moreover, the number is independent of the hill-climbing strategy. It is therefore a suitable measure of computing effort per step in the search.

Counting operations in the formulas (7)–(9) yields, for $n > 0$ and $n \geq m$:

$$\text{In } \mathcal{M}, \quad N_0 = (m + n + 1)N\,,$$

$$N_1 = (3m + 2n + 1)N\,,$$

$$N_2{}^* = (5m + 4n)N\,,$$

$$N_2 = (10m + 5n - 2)N\,; \tag{10}$$

$$\text{In } \mathcal{M}', \quad N_2{}' = (m + 3n + 1)N\,. \tag{11}$$

Notice that these formulas establish a relation between the computing needed to evaluate derivatives and that of loss only. This suggests that one may choose a second measure of computing effort, $J/(m + n + 1)N$, which may be called the "equivalent number of function evaluations."

The alternative measure makes it feasible to compare experimentally the performances of different-order search strategies for the minimum of the particular loss function $V(\theta)$, independently of which computer is used, and independently of the way in which loss and derivatives (if any) are evaluated.

A point in favor of the higher-order search strategies is the efficient way in which derivatives are evaluated; one gets the $m + n$ first-order derivatives with only $(3m + 2n + 1)/(m + n + 1)$ times the effort of evaluating loss only, and the $(m + n + 1)(m + n)/2$ second-order derivatives with only $(10m + 5n - 2)/(m + n + 1)$ times that effort. For example, with $m = n = 5$ a zero-order search strategy is superior to the second-order method only if it needs less than 6.6 times as many steps in the search. Using the incomplete second-derivative matrix during part of the search reduces the limit further, down to 4.1 times. Notice, however, that these are asymptotic figures, valid for long samples. A fixed term should be added to N_r that depends only on the strategy and not on N, and this term probably favors the less sophisticated strategies.

2.2.2 The number of steps

Two things besides the search strategy affect substantially the total number of iterations M for a given sample: the stopping rule and round-off errors. For a strategy yielding quadratic convergence—such as Newton-Raphson's—the remaining error after stopping is small, when and only when, the latest loss reduction is small. However, this is not necessarily true for the hill-climbing scheme used in this paper. A number of modifications are employed in situations where the Newton-Raphson algorithm can otherwise be expected to fail due to unfavorable local properties of the loss function. The modifications generally ensure convergence, but occasionally ruin the quadratic convergence rate.

Also, for a flat minimum, round-off errors may disturb the ideal performance of the search and, in particular, ruin the function of any stopping rule based on loss reduction, since this reduction is normally much smaller than the loss itself. If short word lengths (16 bits) are used, as is the case in this paper, the effect of round-off errors on the number of steps is noticeable.

However, for the sake of comparison it would be unsatisfactory to have a number M that would depend essentially on programming and on the level of round-off errors. Therefore, in the tests no stopping rule has been used, but the search has been allowed to go on until round-off errors preclude further loss reduction. This allows a definition of M as the smallest number of steps k after which no significant improvement in loss is achieved during the remainder of the search; i.e.,

$$q(\theta^k,\theta^\infty) = N[V(\theta^k) - V(\theta^\infty)]/V(\theta^\infty) \tag{12}$$

$$\leq \text{constant} = 1\,,$$

where θ^j is the jth iterate.

This definition is based on the following consideration:

Let $\Delta V(\theta',\hat{\theta}) = V(\theta') - V(\hat{\theta})$ be the reduction in loss, when going from an arbitrary point θ' to the minimum $\hat{\theta}$. In particular, if θ' is equal to the true parameter point θ_0, then for large N,

$$q(\theta_0,\hat{\theta}) = N\Delta V(\theta_0,\hat{\theta})/V(\hat{\theta})\,,$$

and is χ^2-distributed with $m + n$ degrees of freedom; hence, it has the mean $m + n$ [20]. Since $\hat{\theta}$ does not deviate significantly from θ_0, a relative loss reduction $q(\theta^i,\theta^j)$ much smaller than $m + n$ is never significant.

It follows that the definition (12) is reasonable as long as the constant is smaller than $m + n$. The particular value one is, of course, arbitrary and gives a slightly conservative definition of M; one could have stopped

even earlier without getting a significantly inferior estimate. A second requirement is that the effect of round-off errors be much smaller than $m + n$. In the applications the effect has been in the range from 0.1 to 1.

Notice, however, that (12) cannot be used for a stopping rule. The problem of an efficient stopping rule in the presence of substantial round-off errors has not been satisfactorily solved. Because of this it is recommended that floating-point or double-precision arithmetic be used. This will save much trouble.

In $\mathcal{M}'$ the number of steps is M_2'. Ideally, M_2' is equal to one, since $V(\theta)$ is quadratic in $\mathcal{M}'$. However, in practice it may be advantageous to iterate also in this case for the following reason: The second-derivative matrix to be inverted may become nearly singular. It is therefore advantageous always to add a small diagonal matrix before inverting [6]. However, this causes a small systematic error in the estimate, which must be suppressed by iteration. Also, effects of round-off errors are generally reduced by iteration. Normally, it is sufficient for the value of M_2' to be in the range from 1 to 3.

• *2.3 Model quality*

Degree of fit to the data is measured by the residual variance

$$\hat{\lambda}^2 = \frac{2}{N} V(\hat{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \epsilon^2(t|\hat{\theta}) . \tag{13}$$

As long as $n \ll N$, the residuals ϵ may be interpreted as the part of the signal that cannot be predicted one step ahead and therefore cannot be explained in terms of the signal's past history [1]. A total correlation coefficient would be given by the formula $\rho = [1 - \hat{\lambda}^2]^{\frac{1}{2}}$. However, the use of correlation coefficients often gives high values even for quite substantial residuals (because of the root extraction), and may therefore be misleading as a measure of fit.

Increasing the order n generally results in a more detailed model spectrum and a better fit. Even if there is a time-invariant "true" spectrum $|C_0/A_0|^2$ of finite order n behind the sample, some details in the model spectrum become "fictitious" for orders above n, and emerge due to chance alone. If a true spectrum does not exist, or if its structure is different from that of the model, e.g., $C' \neq C_0$, then even for infinitely long samples and large orders the model spectrum will change with the order. In the present application there is probably no "true" order, and possibly not even a "true" spectrum that can be estimated. All one can reasonably assume is that for low orders, changes are mainly systematic and independent of sample length, while for high orders, changes are mainly random and due to a finite sample length.

This does not necessarily mean that larger orders will theoretically produce larger estimation errors, but large orders have the practical disadvantages that 1) computing effort increases and 2) the model spectrum will exhibit a number of spurious details, which may easily be taken for real ones. In a medical application one may easily observe spectral phenomena that seem to indicate physiological effects, when indeed there are none. For this reason it may be better to accept the alternative disadvantage, that of excessive smoothing of the spectrum.

A well-known principle for determining a sufficient order $\hat{n}$ is the following [3]:

Select $\hat{n}$ as the lowest order such that none of the loss reductions obtained for $n > \hat{n}$ contradicts the hypothesis that a true spectrum exists and is of the form $|C/A|^2$ and of order $\hat{n}$.

Apart from the theoretical objections raised above, this has the disadvantage that the result may depend on the extension of the test, in particular to how high an order one cares to pursue it. For any order increase rejected as not significant, there may always be a higher order, where loss reduction will be significant. Still, the approach is customary for testing order.

A standard hypothesis test is based on relative loss reduction [2]: Let $\hat{\theta}^\nu$ be the point of minimum of V, where ν parameters are used in the search. For models in $\mathcal{M}$, $\nu = m + n$, while in $\mathcal{M}'$, $\nu = n$. Then for large N,

$$q_{\mu\nu} = N[V(\hat{\theta}^\nu) - V(\hat{\theta}^\mu)]/V(\hat{\theta}^\mu) \tag{14}$$

is χ^2-distributed with $\mu - \nu$ degrees of freedom, provided ν and μ are both at least sufficiently large to describe the data sample adequately. Conversely, if $q_{\nu\mu}$ is significantly larger than $\mu - \nu$, this contradicts the hypothesis that ν parameters are sufficient. Thus the normalized loss reduction $q = N\,\Delta V/V = N\,\Delta\hat{\lambda}^2/\hat{\lambda}^2$ is useful for judging the significance of both premature stopping and of an order increase.

3. Application

Test data have been provided by Wennberg [19] in the form of tape-recorded EEG signals. Four signals were sampled with a frequency of 62.5 Hz, and for each signal a segment of $N = 1250$ data points was selected and analyzed using an IBM 1130 computer. Four test samples are shown in Figs. 1(a)–1(d). Samples #1, #2, and #4 have been selected from the material in Ref. 19 and represent healthy subjects. Sample #3 is from unpublished material and represents an abnormal subject.

A large number of models in $\mathcal{M}$ and $\mathcal{M}'$ were fitted to the four samples, using different order parameters. Generally, $m = n$ for models in $\mathcal{M}$, and $m = 9$ for models in $\mathcal{M}'$. In $\mathcal{M}$ multiple minima of the loss function were found, yielding further models. These ambiguities are discussed in greater detail in Ref. 6.

In $\mathcal{M}'$ the C' polynomial is fixed. It is well known that

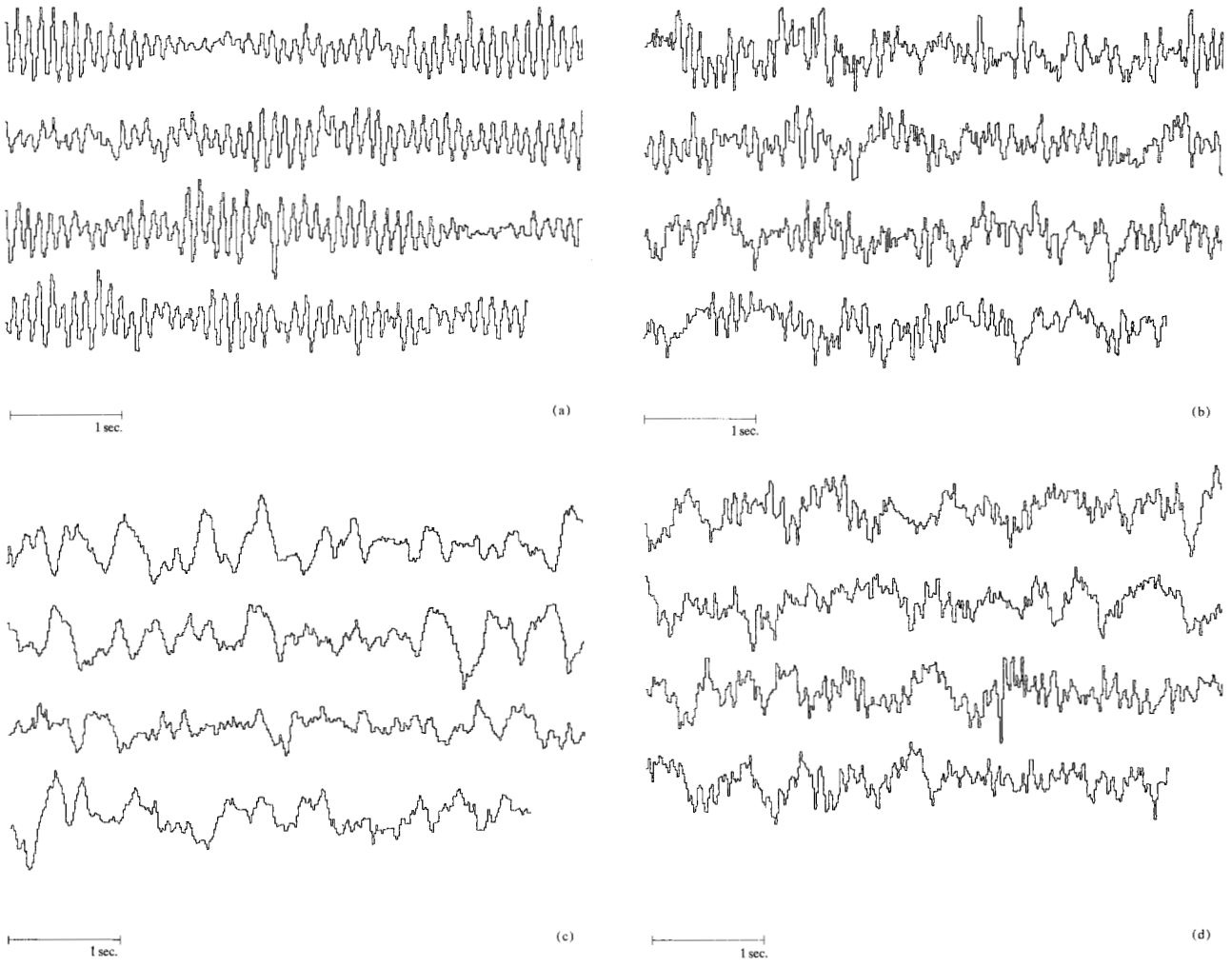

Figure 1 Computer-generated 1250-point plots of tape-recorded EEG signals sampled at a rate of 62.5 Hz. (a) Sample #1; (b) Sample #2; (c) Sample #3; (d) Sample #4. Sample #3 is from an abnormal subject; the others are from normal subjects.

determining C' is equivalent to choosing a prefilter of the form $C'^{-1}(z^{-1})$. Heuristically, the purpose of prefiltering is normally to reduce high-frequency noise or, more generally, to suppress the parts of the sample spectrum where no signal power is expected. For EEG signals this is the case for frequencies higher than about 15 Hz, with the exception of a range centered on 20 Hz, where so-called beta activity, if any, is located. Therefore a filter C'^{-1} with the (somewhat arbitrary) transmission characteristic shown in Fig. 2 has been applied.

One might suspect that this filter would cause frequency components in the model spectrum to appear where one wants to find them a priori (viz, from 0 to 15 Hz and around 20 Hz), irrespective of the frequency distribution of the signal. However, prefiltering is compensated for by the factor $|C'|^2$ appearing in the model spectrum (2). The choice of C' influences only the weights implicitly attached to the various frequencies by the model fitting procedure. It has been confirmed by experiments that the particular choice in Fig. 2 does not affect the final result, only the order needed to achieve the result [6]. However, this does not hold for all possible choices of C'. In practice one has to see to it that $|C'|^{-2}$ has a substantial value—more than 10 percent, say, of its maximum—for all frequencies where the EEG has nonnegligible power. Otherwise the result will depend on C'. See also the discussion in Subsection 2.1.

For reference, models in $\mathcal{M}$ are labelled #ML*i.n* (for the maximum-likelihood method), and those in $\mathcal{M}'$ are labelled #GL*i.n* (for the generalized least-squares method), where *i* indicates the sample and *n* the order. An extra digit may be added (see the following Subsection 3.2). Results of a comparison of methods are presented in the following subsection.

• *3.1 Agreement of spectra*

From the set of fitted models, those to be compared were selected as follows:

In $\mathcal{M}$: Choices of orders were based mainly on the results of order tests as outlined in Subsection 2.3. Multiple minima caused some problems, since it was not possible to discriminate among them by means of the values of $\hat{\lambda}$ only. It happened that the lowest minimum did not always correspond to the physically most likely spectrum. Fortunately, the spectra corresponding to the various local minima differed substantially, making it possible to pick the ones showing the greatest likeness to the corresponding (less differing) models in $\mathcal{M}'$.

In $\mathcal{M}'$: Also in this case, choices had to be based on some engineering judgements. Considering the possible bases for a choice of order discussed in Subsection 2.3 (test, equal degree of fit, and likeness of spectra), the choice arrived at coincides with one based on equal degree of fit, with the single exception that for Sample #1 the order n was raised by one.

Power spectra of the selected models are plotted in Fig. 3. Spectra agree reasonably well; major characteristics have similar frequencies, powers and bandwidths. Agreement is less good for low frequencies, but it is, reasonable to accept the conclusion that the latter differences are random and due to the finite sample length, since low frequencies are generally difficult to estimate accurately. On the other hand, the very small beta-frequency at 20 Hz in Sample #1 is significant in spite of the fact that its power is only about 1 percent of the total signal power.

• *3.2 Comparison of computing effort*

There is a great number of possible algorithms for fitting models in $\mathcal{M}$. The one evaluated below and compared with the generalized least-squares algorithm for models in $\mathcal{M}'$ is a modified Newton-Raphson hill-climbing algorithm. It has two search phases, besides an initialization phase. The scheme is described in more detail in Ref. 6; what is interesting in this context is that the two phases employ the incomplete and the complete second-derivative matrix respectively. Hence, the total computing effort is, from (10) with $m = n$:

$$J = J^0 + 9n\,N\,M_2{}^* + (15n - 2)\,N\,M_2\,, \qquad (15)$$

where J^0 corresponds to the initialization phase. Two different initialization phases were tried, on occasion leading to different local minima.

$$J^0 = \begin{cases} (3n+1)N & \text{for "standard" initialization} \\ (7n+2)N & \text{for "special" initialization}\,. \end{cases}$$

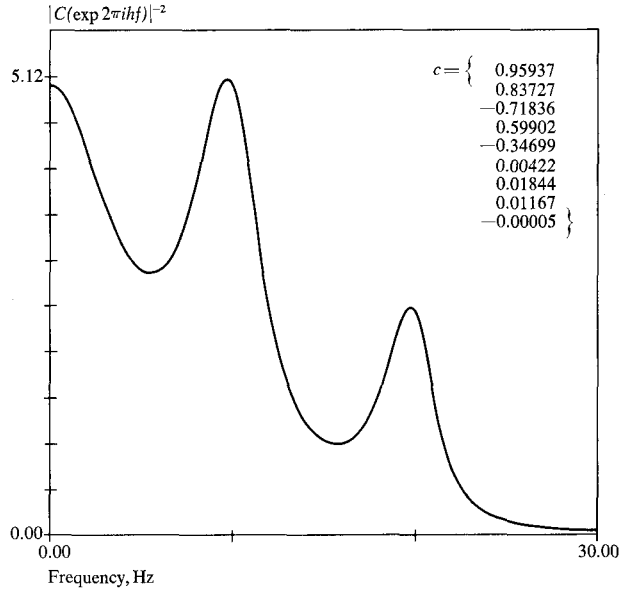

Figure 2 Transmission characteristic of filter C'^{-1}.

"Special" initialization is indicated by an extra digit added to the label #ML*i.n.*l. Again details are given in Ref. [6]. Generally, "special" initialization shortens the search phases. However, it turns out also to have the adverse effect of increasing round-off errors substantially, and it is even possible that the rapid convergence is partly due to the fact that the level of round-off errors is reached early. This level is about the same as the constant in (12) and one magnitude higher than the level resulting from "standard" initialization, but still well below the value $m + n$ where errors begin to be statistically significant (see 2.2.2). However, the margin is uncomfortably small.

The numbers of steps, $M_2{}^*$ and M_2, taken in the two search phases are given in Table 1. The search may terminate in either phase; in both cases the event of "termination" has been determined by the inequality (12).

For models in $\mathcal{M}'$ the following formula is used ($m = 9$):

$$J = M_2{}'\,(3n + 10)N \qquad (16)$$

with $M_2{}' = 2$.

Table 1 gives the results of an evaluation of computing effort in terms of the number of basic operations per sample point J/N and also in terms of the equivalent number of function evaluations J_{fe}. The table includes all the models in $\mathcal{M}$ that were fitted to the data, except those corresponding to false minima and those resulting from particular, favorable starting values for the itera-

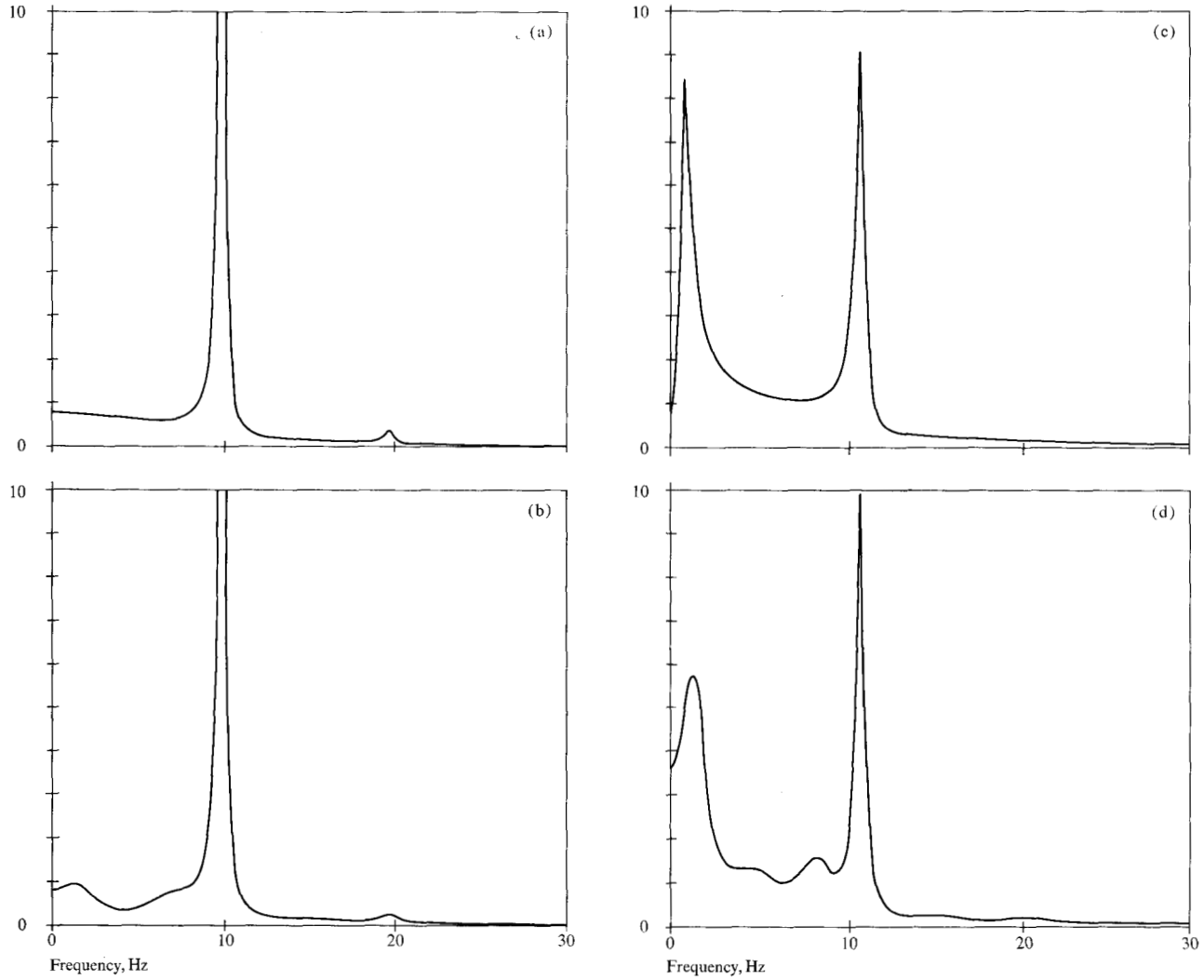

Figure 3 Power spectra. (a) Model ML1.5; (b) Model GL1.13; (c) Model ML2.7.1; (d) Model GL2.17; (e) Model ML3.5; (f) Model GL3.10; (g) Model ML4.5; (h) Model GL4.11. The model spectra to be compared are vertically adjacent.

Table 1 Comparison of computing effort.

Model No.	M_2*	M_2	$\hat{\lambda}^2$	*J/N*	J_{fe}	*Model No.*	$\hat{\lambda}^2$	*J/N*	J_{fe}
ML1.5	3	8	0.1587	735	66.4	GL1.12	0.1505	92	4.2
ML1.7	3	1	0.1577	314	20.9	GL1.13	0.1573	98	4.3
ML2.5	3	1	0.4045	224	20.4	GL2.11	0.4044	86	4.1
ML2.7.1	4	0	0.3990	303	20.2	GL2.17	0.3988	122	4.5
ML3.5	2	9	0.0972	763	69.4	GL3.10	0.0971	80	4.0
ML3.7.1	2	0	0.0970	177	11.8	GL3.11	0.0967	86	4.1
ML3.9.1	2	1	0.0967	360	18.9	GL3.11	0.0967	86	4.1
ML4.5	8	8	0.2994	960	87.3	GL4.11	0.2988	86	4.1
ML4.7	16	0	0.2978	1030	68.7	GL4.18	0.2977	148	5.3

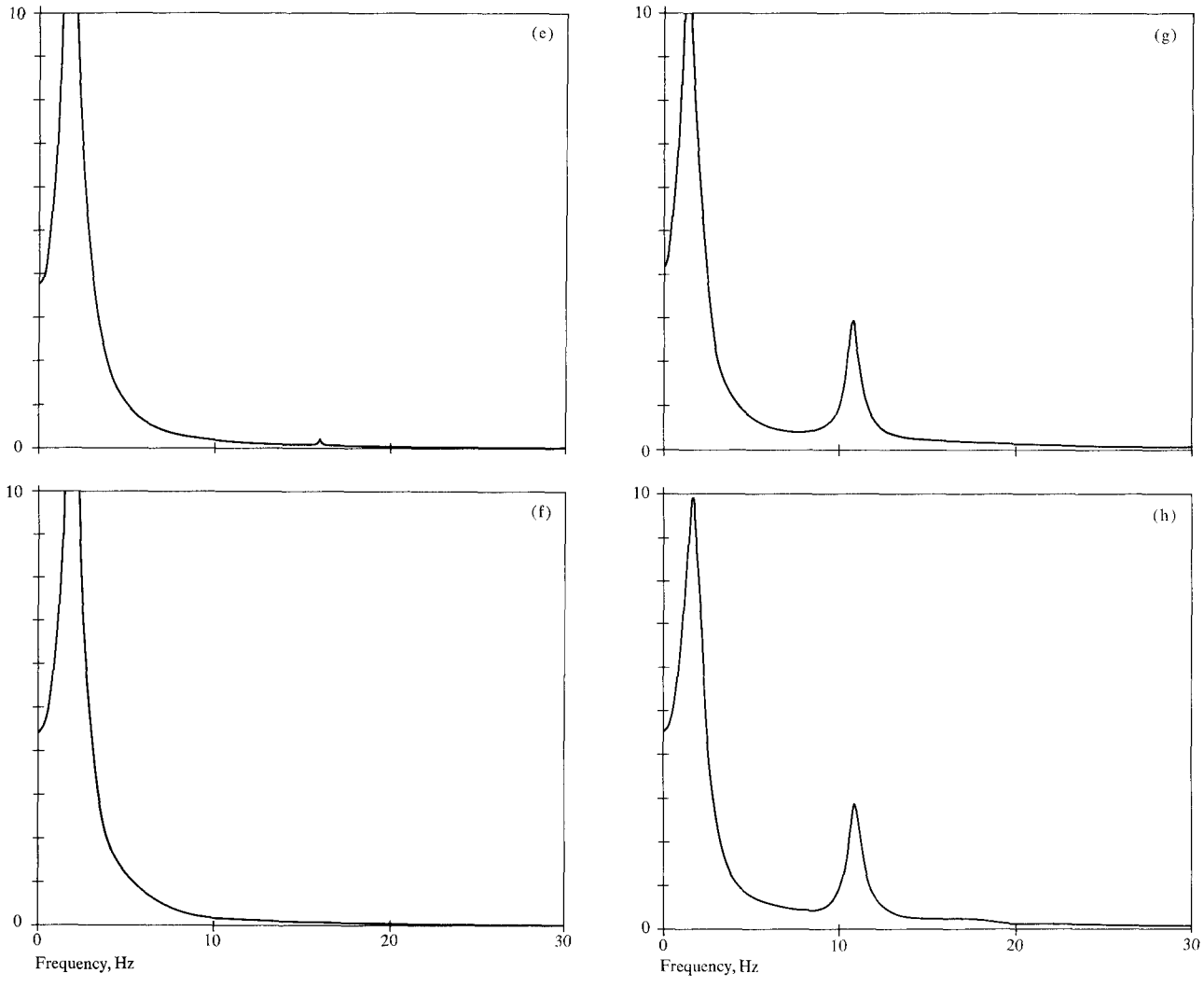

Figure 3 (Continued.)

tion. Thus, four models in $\mathcal{M}$ have unnecessarily high orders and one has an insufficient order. The corresponding models in $\mathcal{M}'$ have been selected so as to have a comparable measure of fit, $\hat{\lambda}^2$.

Comments on Table 1:

- The generalized least-squares method clearly needs less computing effort than the maximum likelihood method to evaluate loss and loss derivatives. Values of the computing measures, J/N and J_{fe}, for the generalized least-squares models are approximately 0.1 to 0.5 of the values for the comparable maximum-likelihood models. The computing effort needed to invert the second-order derivative matrix is not included. However, this matrix is not much larger for models in $\mathcal{M}'$; on the average 14, compared to 11 for models in $\mathcal{M}$. Also, this portion of the computing effort does not grow with the sample length.
- As might be expected, the computing effort for the hill-climbing scheme varies greatly from case to case. However, contrary to what might be expected, the computing effort, when measured in equivalent number of function evaluations, J_{fe}, shows no tendency to increase with the model order. (Orders are indicated by the second digits in the model labels.)

The alternative way of fitting models in $\mathcal{M}$, used in Ref. 21, takes an entirely different approach: several autocorrelations are first computed from the data. The loss function is then evaluated as a function of these autocorrelations, and the model is found by a zero-order search method.

In this case the part of the computing effort that grows with sample length is that of evaluating the autocorrelations. It requires a computing effort J/N equal to the number of autocorrelations needed to evaluate the loss function. Ref. 21 reports the computation of 100 autocorrelations, which yields a clearly smaller value of J/N than for the ML method used in this paper, and approximately equal to that of the GL method. However, this does not include the hill-climbing part. The latter, even if it does not grow with the sample length, still constitutes a substantial, possibly dominating part of the computing effort for the sample sizes used in practice. Unfortunately the computer times cannot be compared directly, since different computers have been used. However, one still arrives at the conclusion that, compared with the alternative ML algorithm in Ref. 21, the GL method is faster.

Conclusions

For the purpose of comparing two methods of modeling stationary EEG signals, maximum likelihood (ML) and generalized least-squares (GL) methods have been applied to four selected samples of EEG signals. The results indicate that models can be obtained using the GL method that are comparable to those obtained by the more general ML method at the cost of only a slight increase in number of parameters, i.e., 10 to 17 parameters for GL models compared to 10 to 14 for ML models.

Computing effort per sample point has been evaluated for both methods in terms of the number of a common set of basic operations. This number is 2 to 10 times higher for the ML method. In addition, the ML method has a number of practical disadvantages when applied to EEG signals:

- Multiple minima may exist. A standard initialization phase does not always yield the right minimum. Also, it is not a simple task to select the right minimum from among a set of local minima.
- Round-off errors (with 16-bit words) appear to cause more trouble with the ML than with the GL method.
- Intermediate, unstable models may be obtained in the search, requiring a substantial amount of additional computations.

Generally, the GL method appears to be superior to the ML method used in this paper for spectral analysis of stationary EEG signals.

However, it must be stressed that, strictly, the comparison of computing effort is valid only for the particular search strategy used in this paper. Although the strategy is believed to be good, there might exist another, far better one. Also, the comparison is confined to the four EEG samples investigated. However, it is believed that, numerically, the samples are representative of stationary EEG signals and, furthermore, it seems likely that the conclusion reached here holds in general for random signals that have a small number of spectrum peaks.

The conclusion that the GL method is superior is not necessarily valid when there are many peaks, as in line spectra, since the orders of the models then become much higher. Neither need it hold for identification of input-output systems, since the presence of inputs increases the order of a least-squares model more than it does that of its counterpart. A comparison would therefore be less favorable to the least-squares method.

Acknowledgment

The author expresses his gratitude to Dr. A. Wennberg of Karolinska Hospital, Stockholm for providing the data for this investigation, and to Prof. L. H. Zetterberg of the Royal Institute of Technology, Stockholm, who manages the joint study on EEG analysis.

References

1. K. J. Åström and T. Bohlin, "Numerical identification of linear dynamic systems from normal operating records," *Proc. IFAC Symposium on Self-Adaptive Control Systems*, Teddington, 1965. Also, IBM Nordic Laboratory Report TP18.159, Lidingö, Sweden.
2. K. J. Åström, "On the Achievable Accuracy in Identification Problems," *Preprints of IFAC Symposium on Identification in Automatic Control Systems*, Prague, June 1967.
3. K. J. Åström and P. Eykhoff, "System Identification, a Survey," *Automatica* **7,** 123 (1971).
4. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*", Dover, New York, 1958.
5. T. Bohlin, "On the Maximum Likelihood Method of Identification, *IBM J. Res. Develop.* **14,** 41 (1970).
6. T. Bohlin, "Analysis of Stationary EEG-signals by the Maximum-Likelihood and Generalized Least-Squares Methods," IBM Nordic Laboratory Report TP 18.200, Lidingö, Sweden, 1971.
7. T. Bohlin, "Analysis of EEG-signals with Changing Spectra," IBM Nordic Laboratory Report TP 18.212, Lidingö, Sweden, 1971.
8. M. Brazier, "Computer Techniques in EEG-analysis," *J. EEG & Clin. Neurophys., Supplement No. 20*, 1961.
9. D. W. Clarke, "Generalized Least Squares Estimation of the Parameters of a Dynamic Model," *Preprints of IFAC Symposium on Identification in Automatic Control Systems*, Prague, June 1967.
10. G. Dumermuth and H. Flühler, "Some Modern Aspects in Numerical Spectrum Analysis of Multichannel Electroencephalographic Data," *Med. & Biol. Eng.* **5,** 319 (1967).
11. B. G. Farley, "Recognition of Patterns in the EEG," *J. EEG & Clin. Neurophys., Supplement No. 20*, 1961.
12. W. Gersch, "Spectral Analysis of EEG's by Autoregressive Decomposition of Time Series," *Math. Biosciences* **7,** 205 (1970).
13. O. M. Grindel, "The Significance of Correlation Analysis for Evaluation of the EEG in Man," *Proc. Symposium on the Mathematical Analysis of the Electrical Activity of the Brain*, Erivan, USSR, 1964; Harvard University Press, Cambridge, Mass., 1968.
14. E. Kaiser and I. Petersen, "Automatic Analysis in EEG," *Acta Neurologica Scandinavica* **42,** Suppl. 22 (1966).
15. W. S. van Leeuwen, "Comparison of EEG Data Obtained with Frequency Analysis and with Correlation Methods," *J. EEG & Clin. Neurophys., Supplement No. 20*, 1961.

16. M. N. Livanov and V. S. Rusinov, *Mathematical Analysis of the Electrical Activity of the Brain*, Harvard University Press, Cambridge, Mass. 1968.
17. I. A. Peimer, "On the Application of Computing Techniques to the Investigation of Short-time Non-periodic Processes in the Electroencephalogram," *Proc. Symposium on the Mathematical Analysis of the Electrical Activity of the Brain*, Erivan, USSR, 1964; Harvard University Press, Cambridge, Mass. 1968.
18. D. O. Walter, and A. B. Brazier, "Advances in EEG Analysis," *J. EEG & Clin. Neurophys., Supplement No. 27*, 1968.
19. A. Wennberg and L. H. Zetterberg, "Application of a Computer-based Model for EEG Analysis," Techn. Report 40, Division of Telecommunication Theory, Royal Institute of Technology, Stockholm, Sweden, 1970; *J. EEG & Clin. Neurophys.* **31,** 457 (1971).
20. S. Wilks, *Mathematical Statistics*. Wiley, London, 1962.
21. L. H. Zetterberg, "Estimation of Parameters for a Linear Difference Equation with Application to EEG Analysis," *Math. Biosciences* **5,** 227 (1969).
22. L. H. Zetterberg, "Analysis of a Large-sample Procedure for Estimating Parameters in a Linear Difference Equation," Technical Report No. 26, Division of Telecommunication Theory, Royal Institute of Technology, Stockholm, Sweden, 1969.
23. L. H. Zetterberg and K. Ahlin, "Engineering Aspects of EEG Analysis," Eurocon 71, Lausanne, Switzerland, Oct. 1971. Technical Report 49, Division of Telecommunication Theory, Royal Institute of Technology, Stockholm, Sweden.
24. L. D. Meshalkin and T. M. Efremova, "Estimation of Spectra of Physiological Processes over Short Intervals of Time," *Proc. Symposium on the Mathematical Analysis of the Electrical Activity of the Brain*, Erivan, USSR, 1964; Harvard University Press, Cambridge, Mass., 1968.
25. P. B. C. Fenwick, P. Michie, J. Dollimore and G. W. Fenton, "Mathematical Simulation of the Electroencephalogram Using an Autoregressive Series," *Biomedical Computing* **2,** 281 (1971).

Received May 31, 1972

The author was with the IBM Nordic Laboratory, Lidingö, Sweden, and is now at the Institutionen för Reglerteknik of the Kungli. Tekniska Högskolan, Stockholm.