# A Review of the Parameter Estimation Problem of Fitting Positive Exponential Sums to Empirical Data

Jöran Petersson and Kenneth Holmström

Applied Optimization and Modeling Group (**TOM**) [1]

Center of Mathematical Modeling
Department of Mathematics and Physics
Mälardalen University
P.O. Box 883, S-721 23 Västerås, Sweden

Research Report in MATHEMATICS / APPLIED MATHEMATICS
Technical Report IMa-TOM-1997-08

5 August 1998

**Abstract**

Exponential sum models are used frequently: In heat diffusion, diffussion of chemical compounds, time series in medicine, economics, physical sciences and technology. Thus it is important to find good methods for the estimation of parameters in exponential sums. In this paper we review and discuss results from the last forty years of research. There are many different ways of estimating parameters in exponential sums and model a fit criterion, which gives a valid result from the fit.

We find that a good choice is a weighted two-norm objective function, with weights based on the maximum likelihood (ML) criterion. If the number of exponential terms is unknown, statistical methods based on an information criterion or cross validation can be used to determine the optimal number.

It is suitable to use a a hybrid Gauss-Newton and Quasi-Newton algorithm to find the unknown parameters in the constrained weighted nonlinear least squares problem formulated using a ML objective function. The problem is highly ill-conditioned and it is crucial to find good starting values for the parameters. To find the initial parameter values, a modified Prony method or a method based upon rewriting partial sums as geometrical sums seems promising. The latter method deserves further investigation for the use as an initial value algorithm.

[1] The TOM home page is **http://www.ima.mdh.se/tom**.

# 1 Introduction

In empirical sciences, there is an increasing use of mathematical models to describe various phenomena in society. One typical situation is a time series of vector measurements $(\mathbf{t}, \mathbf{y})$, i.e. $(t_j, y_j),\ j = 1, ..., n$. In this paper we consider the problem to determine parameters $\mathbf{a} = [a_1, ..., a_p]$ and $\mathbf{b} = [b_1, ..., b_p]$ and the number of terms $p$ in a model of exponential type:

$$f(t) = \sum_{i=1}^{p} a_i \exp(-b_i t). \tag{1}$$

The parameters are determined to fit the empirical data observations $y$ by e.g. using a weighted least-squares criterion

$$F(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n} (w_j (f(t) - y_j))^2. \tag{2}$$

Figure 1 and Figure 2 illustrates the approximation problem for the empirical Steyn and Wyk series [85]. The figures show the data together with the optimal approximating exponential model with $p = 1$, $p = 2$ and $p = 3$. Dependent on the weighting either two or three terms seem to be optimal. More details on this particular problem is given in our paper [68].

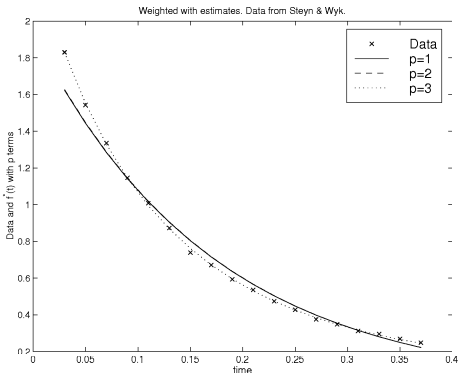


Figure 1: The empirical Steyn and Wyk data series and the approximating exponential model for one, two and three terms. The residuals are weighted with the data values $(w_j = 1/y_j, j = 1, ..., n)$. Three terms seem to be optimal.

In this review, we discuss some ways of attacking problem 2. Although some authors claim to have achieved successful results using easy methods, their claim have been disputed strongly by others. The problem is difficult and we think there is a need for more efficient and reliable algorithms.

In **Section 2** we present some graphical methods. They have been used before the event of computers to find solutions and later on also to find acceptable starting values for nonlinear least squares (NLLS) algorithms. In **Section 3** we mention some results for Prony methods. The classical Prony method is generally regarded as not satisfactory, but modified versions are reported as effective. **Section 4** treats a solution method, which we call generalized interpolation. Classical interpolation methods use one equation for each

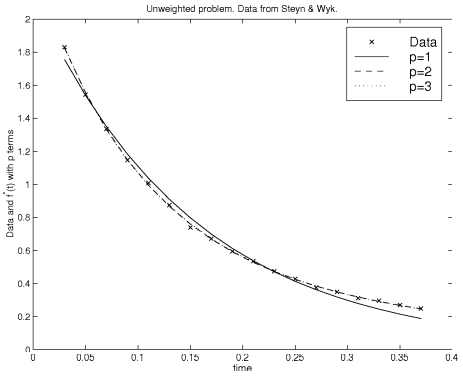


Figure 2: The empirical Steyn and Wyk data series and the approximating exponential model for one, two and three terms. The residuals are unweighted ($w_j = 1, j = 1, ..., n$). Two terms seem to be enough.

data point. Here we discuss a method that divides the data points into $2p$ subgroups of partial sums where $2p$ is the number of parameters to be estimated. Then each partial sum is used to form one equation. For data equidistant in time the partial sums can be rewritten as geometrical sums and the equations become simpler. This method seems promising and we discuss our results in this section. We presented preliminary results for this method at the International Symposium of Mathematical Programming 1997 and in [65]. Further developments are described in a series of papers [68, 67, 66, 70, 69]. **Section 5** presents some statistical aspects such as Monte-Carlo methods, information theory, the choice of weights in weighted least squares estimation and autocorrelation. Monte-Carlo methods seems to be of less use, but an information criterion looks helpful in determining the number of terms. In nonlinear parameter estimation using a weighted least squares criterion, weights $w_j = 1/y_j$ or $w_j = 1/f(t_j)$ are recommended.

**Section 6** compares some transform techniques that have been used for this problem. Some authors are sceptical to transform methods. In **Section 7** we study some iterative methods based upon removing and adding terms $a_i \exp(-b_i t)$ by clustering and splitting techniques. These methods automatically determine the number of exponential terms $p$ in (1). Such methods is of use if the number of exponential terms is unknown. **Section 8** treats optimization algorithms for nonlinear least squares problems. We discuss both Gauss-Newton type methods, hybrid methods and adaptive structured secant methods. Structured secant methods are especially designed for nonlinear least squares methods. In **Section 9** we study some theoretical results from approximation theory and catastrophe theory. It explains why noise may transform an extremum to a saddle points and vice versa. In **Section 10** the relation to mathematical system theory is discussed. Finally we draw some conclusions in **Section 11** and give suggestions for further work.

# 2 Graphical methods

Graphical methods depend upon properties, which are "easy to watch". Their results are often not very accurate, but can be used as an initial guess for more sophisticated iterative methods.

Graphical methods are discussed by Steyn and Wyk [85] and Wiscombe and Evans [88]. The main idea in the parameter identification is: Plot the curve $(t, \ln y)$. Study if the tail stabilizes along a straight line. Then the slope of the tail is $-b_1$. Find $a_1$ as a linear least-squares estimate using the tail. Now one term $a_1 \cdot \exp(-b_1 t)$ in (1) is identified. Next, compute $\widetilde{y} := y - a_1 \cdot \exp(-b_1 t)$ and repeat the procedure on the plot $(t, \ln \widetilde{y})$ to get $-b_2$. Find new values of all $a_i$ by simultaneous linear least-squares fit using all the so far obtained $b_i$.

Steyn et al. [85] does not go this far. They treat only the cases with one or two terms identified "by hand". Wiscombe and Evans [88] reports that Avrett and Hummer [8] and Hunt and Grant [40] has made something like the procedure above. They report that this method usually fails when tried on the problem of identifying more than two terms.

An explanation to this failure is found when studying a Taylor expansion. First write $y$ in the form

$$y = \sum_{i=1}^{p} a_i \cdot \exp(-b_i t) = a_1 \cdot \exp(-b_i t) \sum_{i=1}^{p} \frac{a_i}{a_1} \cdot \exp((b_1 - b_i) t). \tag{3}$$

Then make a logarithmic transformation

$$\ln y = \ln a_1 - b_i t + \ln \left( 1 + \sum_{i=2}^{p} \frac{a_i}{a_1} \cdot \exp((b_1 - b_i) t) \right). \tag{4}$$

Make a Taylor expansion of the last expression in parenthesis around $t = s$:

$$\ln y = \ln a_1 - b_i t + \frac{\sum_{i=2}^{p} \frac{a_i}{a_1} (b_1 - b_i) \cdot \exp((b_1 - b_i) s)}{\left( 1 + \sum_{i=2}^{p} \frac{a_i}{a_1} \cdot \exp((b_1 - b_i) s) \right)} (t - s) + O\left((t - s)^2\right). \tag{5}$$

If the components are not well resolved, $\exp((b_1 - b_i) s)$ will not be small and if a slow component $b_1$ has a small amplitude, it will be distorted by a big factor $\frac{a_i}{a_1}$. Thus the method has its severe limitations. Wiscombe et al. [88] reports that the error is at best 1%.

# 3 The Prony method

In the Prony method the wanted solution $f(t) = \sum a_i \exp(-b_i t)$ is interpreted as a solution to a homogeneous ordinary differential equation $0 = P(D) y$. Then $-b_i$ will be roots to the characteristic polynomial. It is common to illustrate the method by assuming equidistant time points $t_j$ and then reparametrize the exponential sum as $f(t_j) = \sum c_i u_i^j$, which is a solution to a homogeneous difference equation

$$r_2 y_{j+2} + r_1 y_{j+1} + r_0 y_j = 0. \tag{6}$$

We illustrate the method for two terms. The details are given in [36]. Denote $f_j = c_1 u_1^j + c_2 u_2^j$ and multiply $f_{j,j+1,j+2}$ with coefficients $r_{0,1,2}$ for different indices $j$ in the

difference equation. For example:

$$\begin{cases} r_0 f_0 = r_0 (c_1 + c_2) \\ r_1 f_1 = r_1 (c_1 u_1 + c_2 u_2) \\ r_2 f_2 = r_2 (c_1 u_1^2 + c_2 u_2^2) \\ r_0 f_1 = r_0 (c_1 u_1 + c_2 u_2) \\ r_1 f_2 = r_1 (c_1 u_1^2 + c_2 u_2^2) \\ r_2 f_3 = r_2 (c_1 u_1^3 + c_2 u_2^3) \end{cases} \tag{7}$$

Sum the two groups together and note that the left hand side is the difference equation (6). Choose $r_2 = 1$ and formulate it as a Hankel system of equations.

$$\begin{cases} 0 = r_0 f_0 + r_1 f_1 + r_2 f_2 \\ 0 = r_0 f_1 + r_1 f_2 + r_2 f_3 \end{cases}$$

$$\Leftrightarrow$$

$$\begin{pmatrix} -f_2 \\ -f_3 \end{pmatrix} = \begin{pmatrix} f_0 & f_1 \\ f_1 & f_2 \end{pmatrix} \begin{pmatrix} r_0 \\ r_1 \end{pmatrix} \tag{8}$$

Solving this with respect to $r_{0,1}$ gives a solution to the difference equation (6). Find $u_{1,2}$ by solving the characteristic equation of (6). Then find $c_{1,2}$ for example as the linear least squares solution of (7). The Prony method can be generalized to higher order exponential sums and also modified to an overdetermined system. The overdetermined variant is named the Pisarenko method, autocorrelation analysis or the covariance method.

Early implementations suffered from problems satisfying constraints $\mathbf{b} \in R^n$ and $\mathbf{a} \geq \mathbf{0}$. This is overcome by algorithms for nonnegative least squares (NNLS) for $\mathbf{a}$ and nonlinear programming methods for getting real $\mathbf{b}$. Raschke and Stucke [72] was the first to solve the NNLS problem for the Prony method. They developed an empirical method.

Marple [54] compares the Pisarenko (autocorrelation analysis) and Prony (least-squares estimate) methods. The examined case is a sum with components of the type $f_j(x) = \exp(\mathrm{b}_j x)\sin(\gamma_j x)$. The Prony method is reported to have two advantages: It has lower computing complexity and it gives fewer spurious components.

In 1977 Wiscombe et al. [88] investigated the Prony method rather thoroughly. The method needs good initial values to cope with three terms. In fact, the original Prony method is inconsistent even as the number of observations increases to infinity [44].

But since then Osborne and others in a series of articles have developed modified Prony algorithms for several problems satisfying a linear homogeneous difference equations, giving the maximum likelihood estimate. In [59] Osborne shows how to overcome the numerical difficulties in the Prony method. This he extends to rational fitting in [60] and proves by a convergence theorem in [61]. In [83, 62, 63, 44, 53, 64] Osborne and others tries to explain the reported good results from [60] and show some consistency, convergence and asymptotic stability properties with stronger results for damped sinusoids then pure exponentials. The algorithms are scale dependent and objective function reweighting is the preferable choice [44].

In the recent article [64], Osborne and Smyth report that the modified Prony algorithm is insensitive to initial values. They make numerical experiments with data generated by $f(t) = 0.5 + 2\exp(-4t) - 1.5\exp(-7t)$ with added noise from different distributions, four

choices of standard deviations and different sample sizes. They find their implementation of the modified Prony method superior to Levenbergs method and only little affected by different types of added noise. But sometimes their Prony algorithm converges to a damped sinusoid.

# 4 Methods based upon geometrical sums

A method based upon geometrical sums is primarily developed for equally spaced data $t_j$, but Steyn and Wyk [85] mention that interpolation could be used to give equal distances by introducing artificial data points. The method should not be used with too few data points. Della Corte, Buricchi and Romano [26] use linear interpolation to augment the number of data points. With this trick they get more reliable results, i.e. the risk of failures in the fit is diminished. In their data series, this was of extra importance since they worked with both positive and negative intensities in the powers. Then false peaks could show up when having too large space between data points. The use of linear interpolation does not seem natural for this kind of decay curves. In a context of transform techniques, Smith et al. [82] discusses linear interpolation on semilog data. They also report that one should not try to smooth the curve by using several data points. It can deteriorate the result as several intensity components are involved and thus change the shape of the interpolated curve over larger time intervals. They report that Shapiro [81], [80] tried interpolation by curve fit which take into account the curvature. This was not more successful then log-linear fit. In the context of transform methods, Smith et al. [82, page 477] also discusses cubic interpolation and splines for interpolation. Cubic interpolation can introduce false peaks (for transform methods). They have not tried splines.

Cornell [21] has another approach of forming a geometrical sum. If the data is equally spaced in time, $t_j = t_0 + d \cdot (j-1)$, then

$$f(t_j) = \sum_{i=1}^{p} a_i \cdot \exp(-b_i t_j) = \sum_{i=1}^{p} \widetilde{a}_i \cdot u_i^j. \tag{9}$$

Here $\widetilde{a}_i = a_i \cdot \exp(-b_i \cdot (t_1 - d))$ and $u_i = \exp(-b_i d)$. The form $\sum_{i=1}^{p} \widetilde{a}_i \cdot u_i^j$ in (9) is called an exponential polynomial. From this Cornell derives an analytical expression for one term case $p = 1$ and present a numerical procedure for the case $p = 2$. For $p = 1$, the expected value $E\left[\sum_{j=1}^{n} y(t_j)\right] = \sum_{j=1}^{n} \widetilde{a} u^j$ is a geometrical sum.

Set $q = n/2$, or $q = (n-1)/2$ if $n$ is odd, and partition into

$$S_1 = \sum_{j=1}^{q} \widetilde{a} u^j = \widetilde{a} u \left(1 - u^q\right) / (1 - u) = \Sigma_1 \tag{10}$$

and

$$S_2 = \sum_{j=q+1}^{n} \widetilde{a} u^j = \widetilde{a} u \left(u^q - u^n\right) / (1 - u) = \Sigma_2. \tag{11}$$

Then we have two equations which are linear in $\widetilde{a}$ and nonlinear in $u$. For several terms,

a set of equations are derived: Partition the data into

$$S_k = \sum_{j=(k-1)q+1}^{kq} y_j, \ k = 1, ..., 2p, \tag{12}$$

where e.g. $S_2 = y_{q+1} + y_{q+2} + ... + y_{2q}$ and let $r = n - n \bmod 2p$ giving $2p$ partial sums of equal length $q = r/(2p)$. Rewrite the geometrical sums and solve the equations

$$\Sigma_k - S_k = 0, \ k = 1, ..., 2p. \tag{13}$$

They are linear in the vector $\widetilde{\mathbf{a}}$ and nonlinear in the vector $\mathbf{u}$. Cornell finds the equations too complex to be solved except for the case $p = 1$. Steyn and Wyk [85] compare this method with two other methods. One based on overlapping partial sums by Agha [1], and the regression-difference method by Shah [79]. Steyn and Wyk give some guidelines in getting an initial good estimate and recommend how many data points to interpolate when needed. They conclude that the method of Agha is the most reliable initial estimate method of these three. The best overview of methods based on geometrical sums is also found in their paper. We have improved the method of Cornell for the case $p = 2$ by finding an analytical expression for $\mathbf{u}$. Furthermore we have developed a numerical algorithm which extends the method to the cases $p = 3, 4$. Preliminary results for our methods is presented in [65]. It has been further developed in [68, 69] and applied to similar exponential sum problems in [67, 66].

# 5 Statistical aspects

In this section we study the statistical properties of the exponential sum fitting problem. One method using jack-knife estimates is examined by Duncan [29]. He concludes that it is "marginally effective in establishing confident regions for moderate ($\approx 24$) samples". But in nonlinear regression, the samples are often of small size. The article also refers to Beale (1960), who constructed a measure of nonlinearity, which quantifies when a linear approximation can be done.

Burstein [17] has tried a Monte-Carlo method which he calls 'algorithm of root variation'. He chooses a random vector $\mathbf{u}$ with $p$ entries in the interval $(0, 1)$. A linear least-squares approximation is used to compute the linear coefficients (amplitudes) $\mathbf{a}$. To compare different amplitudes he uses a criterion of fit, e.g. standard deviation. The set corresponding to the best fit is chosen. He discusses some improvements: Form an expanding list by adding a new result if it is "the best result so far". This list is used to extract the best result. Furthermore he uses the same algorithm in the transformed space for Laplace transformed data series and truncated series expansions.

Another point of view is "which criterion of fit give most reliable results?" Price [71] discusses this problem. The least squares (LS) criterion $\sum (y_j - f(t_j))^2$ for measuring the errors in absolute distance means that a big $y_j$ and a big $f(t_j)$ could give a big absolute error $y_j - f(t_j)$ although it is small in percentage and a small $y_j$ with small error $y_j - f(t_j)$ could be big in relative measure. For decaying curves like exponential sums this is "not

fair" for the tail of the time series. It would be more realistic measuring the relative distance, either weighting with the data giving

$$\sum \left( (y_j - f(t_j))/y_j \right)^2 \tag{14}$$

or

$$\sum \left( (y_j - f(t_j))/f(t_j) \right)^2 . \tag{15}$$

The equation (15) is the maximum likelihood (ML) criterion derived by Price [71]. It is simpler to use weighted least squares (wLS) than ML. And ML does not take into account the fact that systematic errors may not be proportional to $\mathbf{y}$. Price [71] makes a comparison between wLS and ML for a time series following the Poisson distribution and finds ML to be the better. Statisticians prefer ML because it often gives more effective estimates.

The number of exponential terms is often not known in advance. A kind of Occam's razor is needed. This problem is treated in information theory. If few terms are used, the fit to data will not be very good, but the model is easy to handle. If many terms are used, the fit to data will be very good, but the model is clumsy. It may be even unrealistic, because the curve fits the random errors in the data points (Ljung [52, page 417]). To avoid this overfitting problem, measures of information distance have been developed. The information distance has a physical interpretation as a measure of negative entropy. Ljung [52, page 421] suggests a pragmatic formulation of the information criterion:

$$W_n^0 (\theta, M, Z^n) = V_n (\theta, Z^n) + U_n (M) . \tag{16}$$

Here $\theta$ is the parameters to be determined, $M$ is the model space (exponential sums in our case) and $Z^n$ is the discrete space of the $n$ data points. $V_n$ is a measure of the fit of the curve to data and $U_n$ is a penalty term for the complexity of the model. $V_n$ can be chosen as a prediction-error criterion $V_n = \frac{1}{n} \sum_{j=1..n} l (\epsilon (t_j, \theta), t_j, \theta)$ where $\epsilon (t, \theta) = y (t) - \widehat{y} (t, \theta)$. As a minimizer of the prediction-error, Ljung [52, page 175] suggests $l (\epsilon (t, \theta), t, \theta) = \frac{1}{2} \epsilon^T \Lambda^{-1} \epsilon$. The matrix $\Lambda$ is the covariance matrix. Assuming the observation errors $\epsilon$ are uncorrelated, the covariance matrix is a diagonal matrix. For the ML-criterion $\Lambda = diag (\widehat{y}^2 (t_j), j = 1..n)$ is the estimated (co-)variance and thus

$$V_n = \sum_{j=1}^{n} \left( \frac{y_j - f(t_j)}{f(t_j)} \right)^2 . \tag{17}$$

It corresponds to a minimized variance (and a weighted least squares with ML-weights).

The term $U_n$ should measure the complexity of the model. How to chose $U_n$ is subjective. Ljung [52, page 421] has two suggestions, either the Akaike criterion [3, 4]

$$U_n (M) = \frac{\dim \theta}{n} \tag{18}$$

or the Rissanen criterion [74]

$$U_n (M) = \frac{\log n}{n} \dim \theta. \tag{19}$$

The goal of the Akaike criterion is to find a system description that gives the smallest mean-square error. The goal of the Rissanen criterion is to achieve the shortest possible description of data.

A common choice of the function $W_n^0(\theta, M, Z^n)$ is the Akaike Final Prediction-Error Criterion (FPE), first described by Akaike 1969 [2],

$$FPE(\dim\theta) = \frac{1+(\dim\theta)/n}{1-(\dim\theta)/n} \cdot \frac{1}{2n}\sum_{j=1}^{n}\left(y_j - f(t_j)\right)^2, \tag{20}$$

which reflects the prediction-error variance that one will obtain, on the average, when the model is applied as a predictor to other data sets than those used for the identification.

# 6 Transform techniques

Transforms $\widetilde{y} = \int_0^\infty yg(t)\,dt$ is another class of methods, which is a main tool in analyzing frequency spectra. They work well for signals consisting of sines and cosines, but how do they work for signals without that kind of structure? Decaying exponential sums has a lack of structure like peaks or valleys. They contain simply a descending slope. The literature is concerned about three major tasks: Which transforms are more suited or less suited for the problem? When planning measurements, should we use linearly or exponentially distant data points? Is transforms superior to other methods such as optimizing least squares objective functions?

Examples of transform techniques are the Orthonormal, Laplace, Gardner, $\Gamma$ and $\kappa^2$ transform. They are briefly discussed below.

Smith et al. [82] study three types of transforms. The first one is the orthonormal exponential transform. In the interval $[0,\infty]$ an orthonormal (ON) set of basis functions is $\Phi_m(t) = \sum_{k=1}^{m} c_{km}\exp(-kt)$, where

$$c_{km} = (-1)^k (k+m+1)!\cdot m!/\left[k!\,(m-k)!\,(k+1)!\,(m+1)!\right].$$

Coefficients $B_m = \int_0^\infty \Phi_m(t)\,f(t)\,dt$ is computed and the original function has the spectrum

$$f(t) = \sum_{m=1}^{\infty} B_m\Phi_m(t) = \sum_{m=1}^{\infty} D_m\exp(-kt). \tag{21}$$

The conclusion from two single component decays $\exp(-2t)$ and $\exp(-2.1t)$ is that only integer components $k$ are resolved. This method is of use only when searching among a discrete set of decay constants, which can be numbered. Next the inverse Laplace transform is investigated. The problem here is that the integral $f(t) = \int_{r-i\infty}^{r+i\infty}\widetilde{f}(z)\exp(tz)\,dz$ is performed in the complex plane, while $\widetilde{f}(z)$ is known only on the real axis. Smith et al. reports that a number of techniques (e.g. approximation $\widetilde{f}(z)$ with Jacobi polynomials) has been tried. They found that this method failed even when analyzing a single component. The third transform is defined as a function

$$G(\exp(-x)) = \Gamma^{-1}\left(\frac{\Gamma(\exp(x)\,f(x))}{\Gamma(\exp(-x)\exp(-\exp(x)))}\right) \tag{22}$$

of Fourier transforms $\Gamma$ on exponential expressions. When this was presented by Gardner et al. 1959, the computer resources were too limited for the numerical processes. In [82] Smith et al. investigated if this transform could be of use for exponential sums, because it has been tried with success for some other types of functions. There are three things to worry about: The infinite Fourier transform (FT) must be approximated by a Discrete FT (DFT), computational noise in the transforms and finally errors due to interpolation and approximation.

They assumed that the measurements were made at equal time distances in an interval of six to ten times longer than the time constant. They found out that to analyze a single component decay, an interval of less than 22 time constants made the signal to noise ratio $S/N$ poor. To overcome this, some different methods of extrapolations are discussed. The draw-back is that they introduce noise ($S/N$ decreases). Smith et al. [82, page 477] report of an earlier examination of the effect of prefiltering. Prefiltering will damp fast intensities. On the other hand it will greatly magnify the quotient $S/N$. This is important because even noise of small amplitude significantly decreased the quality of the spectrum. To interpolate they used different polynomials, cubic least squares fit to the logarithm of the data. The final conclusion is that the Gardner transform is not yet (1976) mature for high resolution multicomponent analyze even though they use prefiltering and post filtering. A verdict from Wiscombe and Evans 1977 [88] mention disappointing results from investigations made by Arking, Grossman and Domoto and claims that "Gardner methods are hopelessly ill suited to exponential sum fitting of transmissions".

Other transforms have been derived by Nakamura et al. in [56, 57]. Nakamura and Takahashi derived a formula for the inverse $\kappa^2$ transform. The formula is an infinite sum including Laguerre polynomials. A quite complex but at least direct formula. These two authors have also, together with Kodama, derived a formula for the $\Gamma$ transform defined as

**Definition 1** . $G(\alpha,\beta)=\int_0^\infty \frac{t^{\alpha-1}}{\beta^\alpha\Gamma(\alpha)}\exp(-t/\beta)\,f(t)\,dt$, $\alpha>0,\beta>0$.

This formula can be approximated as

$$G(\alpha,\beta)\approx\sum_{j=1}^{K}Q_j/\left(1+\beta x_j/T\right)^\alpha. \tag{23}$$

Here $x_j$ is the $j$th zero of the $K$th Laguerre polynomial and $Q_j$ is determined by solving a system of linear equations $f(t_h)\approx\sum_{j=1}^{K}Q_j\exp(-x_jt_h/T)$, $h=1,2,...,K$. Instead of using Laguerre polynomials, the quadrature can also be performed by using Legendre, Hermite or Chebyshev polynomials. The accuracy of the estimated parameters does not always increase with smaller steps in $\alpha,\beta$ and $T$. They compared the accuracy of the parameters estimated by the $\Gamma$ transform with parameters estimated with the Prony method. Their method based upon the $\Gamma$ transform gave better results.

Another one who has investigated transforms with exponentially spaced data is Steyn [84]. He uses the "graphical method" by Steyn and van Wyk [85] to determine initial values and then uses a Gardner-type method based on DFT. Together with this, filtering and smoothing is used. This produces a spectrum, from which peaks can be identified by

hand. To make a judgement of the peaks, Steyn [84] divides the height of a peak with its width at half height. This favors narrow peaks compared to error ripples.

Planning the placement of the measurements is not unimportant. Bertero et al. [9, 10] determines an efficient placing of data points $t_j$. Their prerequisite is a fixed number of linearly spaced data. They derive the placing of $t_j$ which minimizes the condition number of singular value inversions and report that this would further reduce the number of sample points required for inversion of the Laplace transform. They report problems with edge effects during the inversion procedure and discuss how to overcome these difficulties.

Kammler [46] studies theoretically the Laplace transform for this problem. Define the set $V_p := \{f : f \in R,\ f(t) = \sum_{i=1}^{p} a_i \exp(-\mathrm{b}_i t)\}$ and the auxiliary function

$$\xi(\mathrm{b}) = \int_0^\infty \exp(-\mathrm{b}t)\, d\mu(t).$$

Choose either the measure

$$d\mu(t) = dt \tag{24}$$

or

$$\mu(t) = \sum_{j=1}^{n} w_j \chi(t - t_j), \tag{25}$$

where $\chi(t)$ is the step function.

Compute the Laplace transform of both the data $y(t)$ and the approximating function $f(t)$, i.e. define

$$\begin{array}{ll} \widetilde{Y}(\mathrm{b}) = \int_0^\infty \exp(-\mathrm{b}t)\, y(t)\, d\mu(t) & \mathrm{b} > L, \\ \widetilde{F}(\mathrm{b}) = \int_0^\infty \exp(-\mathrm{b}t)\, f(t)\, d\mu(t) & \mathrm{b} > L, \end{array} \tag{26}$$

where $L = \inf\{\mathrm{b} : \xi(\mathrm{b}) < \infty\}$ and also $L < 2 \cdot \mathrm{b}_i,\ i = 1, ..., p$. The choice of the measure (24) gives the value $L = 0$ and a continuous $L_2$ approximation problem, whereas the measure (25) gives the value $L = -\infty$ and a weighted discrete $L_2$ approximation problem.

In the paper the investigation is restricted to the case of monotonic $y$ on $(0, +\infty)$. The Generalized Aigrain-William (GAW) equations are defined as

$$\left\{ \begin{array}{lcll} \widetilde{Y}(\mathrm{b}_i) & = & \widetilde{F}(\mathrm{b}_i), & i = 1, ..., p \\ \widetilde{Y}'(\mathrm{b}_i) & = & \widetilde{F}'(\mathrm{b}_i), & i = 1, ..., p \end{array} \right. \tag{27}$$

Kammler proves the following practical result:

**Theorem 1** . *There exists a best approximation $\widetilde{F} \in V_p$ solving the GAW equations (27).*

Kammler suggests the dual algorithm of Rice to solve the GAW equations, using repeated solutions of nonlinear Hermite interpolation problems. One numerical example is presented, and the algorithm converges nicely up to $p = 6$ terms. He also comments on other approaches:

- Solving the GAW equations numerically demands good initial values to work well, as well as a low order $p$.

- Using the common choice of minimizing $\|f(t,\mathbf{a},\mathbf{b})-y(t)\|^2$ by eliminating the linear parameters $\mathbf{a}$ gives notoriously difficult problems already for $p=2,3$ and next to impossible problems for $p>3$.

Varah [87] makes an empirical continuation of the theoretical work of Kammler. Varah gives examples of numerical results and some explanations of it. Two sides of the problem are studied. A continuous subproblem (C) and a discrete one (D):

$$\begin{cases} I(\mathbf{a},\mathbf{b})=\frac{1}{2}\int_0^\infty \left(y(t)-f(t:\mathbf{a},\mathbf{b})\right)^2 dt & \text{(C)} \\ I(\mathbf{a},\mathbf{b})=\frac{1}{2}\sum_{j=1}^n \left(y_j-f(t_j:\mathbf{a},\mathbf{b})\right)^2 & \text{(D)} \end{cases} \tag{28}$$

The Laplace transform of the first order conditions of the continuous subproblem (C) gives the GAW-equations. As a comparison, a solution for the classic Lanczos and Osborne data is made. The Lanczos data are generated in time steps of $\Delta t=0.5$ by the exponential sum

$$f(t)=0.0951\exp(-t)+0.8607\exp(-3t)+1.5576\exp(-5t).$$

But Varah chooses $\Delta t=0.1$ instead and $0\leq t\leq 3.2$. The Lanczos- and Osborne data are presented with some plots of $\min_a I(\mathbf{a},\mathbf{b})$ as a function of $\mathbf{b}$. When solving the continuous subproblem (C) in (28), he derives the Hessian for the GAW equations and finds a part of it to be the Hilbert matrix, which is known to be ill-conditioned. Also the eigenvalues are close to that of the minimum point and besides all this, the extrema and saddles are multiple.

Braess has studied this approximation problem in several papers [11, 12, 13, 15, 14]. He refers to Hobby and Rice who introduced the idea of a $\gamma$ polynomial, defined as

**Definition 2** . *Let $\gamma(b,t)$ be an extended totally positive Kernel defined on $B\times T$, a rectangle in $R^2$. Then a $\gamma$ polynomial $F$ of order $p$ is defined as $F(t)=\sum_{i=1}^p a_i\cdot\gamma(b_i,t)$, $a_i\in R$, $b_i\in B$.*

As an example he gives $\gamma(b,t)=(1-bt)^{-1}$ on $(-1,1)\times[-1,1]$. From this he derives the minimization problem $r(a,b)=\|f-a/(1-bt)\|^2$, which is the same problem that appears when solving the Aigrain-Williams equations discussed above. Braess [11] shows that the solution is not unique in the $\|\cdot\|_2$-norm for these rational approximation problems.

If several terms $p$ are involved, the expression $a/(1-bt)$ extends to a rational expression $\sum_i a_i/(1-b_it)=P/Q$ where the degrees of $P$ and $Q$ are $\partial P$ and $\partial Q$ with $\partial P<\partial Q$. In [12] Braess discusses such rational $L_2$-approximation. He shows that critical points are either isolated or belong to one-dimensional analytical submanifolds for $\partial Q=1,2$. The consequence of this is that local best approximations may hinder an iterative algorithm to find a global minimum.

In [13] Braess discusses Chebyshev approximation by $\gamma$ polynomials. The numerical consequences are that difficulties in the classical algorithms for treating spline functions with free knots may be overcome by the regularization procedure. He uses complicated Morse theory to establish this.

# 7 Iteratively adding and removing terms

Another type of method has been developed by Gustafson [35], Kaijser [45], Ruhe [76] and Evans et al. [30]. Kaijser studies a fit of log-normal functions and not an exponential sum, but the idea is the same and Kaijser has got inspiration and advice from Gustafson. The methods are stepwise nonlinear regression methods and the basic algorithmic steps are:

1. Make a grid with $N$ points in the space of $\mathbf{b}$ and compute the corresponding $N$ optimal values for $\mathbf{a}$ subject to $\mathbf{a} \geq \mathbf{0}$. When determining $\mathbf{a}$, algorithms for nonnegative least squares (NNLS) can be used. Compute the distance between data points and the exponential sum (1) with estimated parameters for some chosen measure of distance, e.g. the weighted sum of squares.
2. Cluster close terms $a_i \cdot \exp(-b_i t)$ with an "averaging formula", in such a way that the distance does not increase too much.
3. Use an algorithm for nonlinear least squares optimization to compute the optimal parameters $\mathbf{a}$ and $\mathbf{b}$ in the chosen norm.
4. If step 3 results in a decreased distance, repeat the clustering step. If the distance is increased and it is the first iteration, then go to step 1 and make a finer grid, else stop.

Gustafson suggests the grid in step 1 to be chosen by an auxiliary function $s = g(b)$. This function is strictly decreasing. Two examples are $g(b) = \exp(-b)$ and $g(b) = 1/(1+b)$. The advantage of an auxiliary function is that $s \in [0,1]$ for $b \geq 0$, which is an easier interval to survey. Use this to make a grid in $s$-space $[0,1]$ and then compute $b_i = g^{-1}(s_i)$.

In Gustafsson [35] all details of the clustering algorithm is given. He also gives the advice that clustering should be made in such a way that "the distance does not increase too much". Kaijser [45] gives another description. He makes the following two definitions:

**Definition 3** . *A vector* $\mathbf{a} = [a_1, a_2, ...a_N] \in R_+^N$ *is said to be singular if most coordinates are equal to zero.*

**Definition 4** . *A singular vector* $\mathbf{a}$ *is said to have the* $m$-mode-property *if the non-zero components can be grouped, in a natural way, into* $m$ *different subgroups.*

He gives a rule of thumb for the $m$-mode-property. "If there is no background information, the natural way to define a subgroup is to require that all components in a subgroup shall have indices which differ at most 5, say, or by some other small number (Small in comparison with the dimension N)." This idea is expressed as a formula by Ruhe [76] and discussed later in this section.

Kaijser omits the nonlinear optimization in step 3, while Gustafson suggests Gauss-Newton for this. Ruhe and Wedin [78] presents three algorithms for separable nonlinear least squares of Gauss-Newton type. They separate the linear variables $\mathbf{a}$ from the nonlinear variables $\mathbf{b}$ and optimize in $\mathbf{a}$ keeping $\mathbf{b}$ constant and then optimize in $\mathbf{b}$ keeping

**a** constant. The first optimization is simple, it reduces to a linear least squares problem. In the second optimization, different methods to correct the residual vector and Jacobian matrix are presented. Two of the methods give superlinear convergence; the same convergence rate as the ordinary Gauss-Newton method. The third, and most simple, algorithm has only linear convergence. All algorithms converge almost quadratically for almost compatible problems. Thus are these algorithms very suitable for the exponential fitting problem. One of the algorithms is implemented in our nonlinear program package NLPLIB TB [38], which is part of the optimization environment TOMLAB [39].

In 1970 Cantor and Evans [18] wrote a paper on exponential sum fitting for equally distant data points [18]. In 1977 Wiscombe and Evans rewrote this paper [88]. They said that [18] seems to have been forgotten due to being theoretically quite advanced. In a third paper by Evans, Gragg and LeVeque [30] the algorithm is extended to nonequidistant time spacing and together with the use of divided differences overcome numerical ill-conditioning when computing **a**. We first present some theory from [18] and then the algorithms in [88] and [30].

Cantor and Evans use the exponential polynomial form (9) for equally distant data, $\widetilde{f}(j) = \sum_{i=1}^{p} \widetilde{a}_i u_i^j$. In the following we, for simplicity, make no difference between $\widetilde{a}$ and $a$ and assume $n+1$ data points $0, 1, ..., n$. Let $\mathbf{a} > \mathbf{0}$ and $u_i \in S = [0, 1]$ . Then $\widetilde{f}(j)$ is called a "positive exponential polynomial on $S$ of degree $p$". If ordered in the form $0 < u_i < u_{i+1} < 1$, the representation is unique and is called "reduced form". Denote the space $\mathbf{F} := \left[\widetilde{f}(0), \widetilde{f}(1), ..., \widetilde{f}(n)\right]$ and define $C_n := \left\{\mathbf{F} : \widetilde{f} \text{ is positive exponential polynomial on } S\right\}$, i.e. the smallest cone of all sequences $(1, u, u^2, ..., u^n)$. Based on these definitions they derive two existence theorems. The existence of a unique representation is proved in

**Theorem 2** . *If* $\mathbf{y} := \{y_j\}_{j=0}^{n} \in C_n$ *is on the boundary of* $C_n \Rightarrow$ *then there exists a unique positive exponential polynomial* $\widetilde{f}$ *of degree* $\leq n$ *on* $S$*, representing* $\mathbf{y}$.

Denote $\mu(e) = |e|^{r-1} sgn(e)$.

**Definition 5** . $P_f(x, \mathbf{y}, r) := r \sum_{j=0}^{n} w_j \mu\left(\widetilde{f}(n) - y_j\right) x^j$.

The notation is $P_f(x, \mathbf{y}, r) = P_f(x)$, if $\mathbf{y}$ and $r$ are understood. Using this definition it is easy to see that $P_f(x) = \partial\left(\|\mathbf{F} - \mathbf{y}\|_r\right)^r / \partial a_i$, and we have

**Theorem 3** . *If* $1 < r < \infty$ *then the following equivalence is true:*
*A positive exponential polynomial* $\widetilde{f}$ *on* $S$ *is a best approximation to* $\mathbf{y}$ *in the norm* $\|\cdot\|_r \Leftrightarrow$ $P_f(x) \geq 0 \ \forall x \in S$ *and each* $P_f(u_i) = 0$.

$\widetilde{f}$ is a strong best approximant to $\mathbf{y}$ if the following statement holds: If $\widetilde{f}$ is a best approximant in the norm $\|\cdot\|_\infty$, and, in addition $0 \in S \Rightarrow w_0 \left\|\widetilde{f}(0) - y_0\right\|_\infty = \|\mathbf{F} - \mathbf{y}\|_\infty$.

**Theorem 4** . *There exists a strong best approximant* $\widetilde{f}$ *to* $\mathbf{y}$. *If* $\|\mathbf{F} - \mathbf{y}\|_\infty \neq 0$, *then* $\widetilde{f}$ *is unique.*

The algorithm developed in [18] and [88] builds upon the optimality criterion for the residual polynomial in Theorem 3:

$$\text{Best fit} \Leftrightarrow \begin{cases} P(u_i) = 0 \quad \forall i & \text{a)} \\ P_f(x) \geq 0 \quad \forall x \in S & \text{b)} \end{cases} \tag{29}$$

Criterion (29 a) is the common least-squares criterion for $\mathbf{a}$. The idea of the method is to jump between finding the linear coefficients $\mathbf{a}$ in (29 a) and solving (29 b) to obtain a new best $\mathbf{b}$. The algorithm is of stepwise regression type and starts with no terms and iteratively adds and removes terms. The idea is different from that of Gustafson [35], which starts with a lot of candidates and iteratively coalesce them. To minimize $P_f(x)$ they use a special algorithm, which is not presented in their paper [88]. For the optimizing in $x$, one could just form a grid in the $S$-space and make a direct computation of $P_f(x)$ to find its minimum. They claim that in practice $u_i$ tend to cluster close to $u = 1$ and thus an equidistant distribution of the search points is not needed. A finer grid around old $u_i$ can be used. The reason why they do not use a root search with derivatives involved, is that such a search is notoriously ill-conditioned for high degree polynomials. The degree in $u$ becomes equal to $n$, one less than the number of measurement points.

The linear optimization is solved using a divided difference technique. They use a factorization and orthogonalization process. The dropping of terms is done using a nonlinear least squares algorithm. The least desirable term is determined and dropped by looking for the first element passing zero while passing through the convex combinations of old and new values of $\mathbf{a}$. Let $\mathbf{a}^{old}$ denote the old $a_i$-values, together with the new term $a_{p+1} = 0$. Let $\mathbf{a}^{new}$ be the last computed $\mathbf{a}$, possibly with some negative entries. The studied convex combination is $\mathbf{a}^{old}(1-\beta) + \mathbf{a}^{new}\beta$, searching from $\beta = 0$ towards $\beta = 1$. Drop the first zero term and then compute a new least-squares solution.

Based on empirical tests, they have developed a set of convergence criteria:

$$\begin{cases} \text{i)} & \frac{R_{k-1}-R_k}{R_{k-1}} < \epsilon_1 & \text{Decrease of the total residual in a step.} \\ \text{ii)} & P_k(u_0) \geq -\epsilon_2 & \text{Residual polynomial almost positive.} \\ \text{iii)} & |u_0 - u_i| < \epsilon_3 & \text{New terms close.} \\ \text{iv)} & \text{Too many iterations} & \text{Emergency stop.} \end{cases} \tag{30}$$

**Coalescence:** At the end of the algorithm, close $u_i$ are replaced by a single term. The method used find new parameter estimates by solving a set of nonlinear equations with the Newtons method. They report that the loss of accuracy is usually tolerable and with 14 significant digits computations, they coped up to 150 data points. With 29 significant digits computations, they coped up to 250 data points. A typical error was within 0.01% to 0.001% and the number of terms were between 3 and 8. They find their method good at recovering constructed data. In fact "excellent ability" is reported. They also made a preliminary test with inverse Laplace transform using Laguerre interpolation but found it totally unsatisfactory. In [30] the algorithm is extended to the case of non equidistant data points. Also a weighted norm is used in the least squares fitting.

**Two algorithms by Ruhe.** The first algorithm by Ruhe [77] is a stepwise nonlinear regression algorithm. Define the inner product $(\mathbf{x}, \mathbf{y})_W = \mathbf{x}^T\mathbf{W}\mathbf{y}$ with $\mathbf{W} = diag\left(1/y_j^2\right)$

and the exponential vector

$$e(b) = \begin{pmatrix} \exp(-bt_1) \\ : \\ \exp(-bt_n) \end{pmatrix}. \tag{31}$$

Ruhe considers two kinds of fitting criteria, either weighted least squares (Equation (14)) or maximum likelihood (Equation (15)). He states the convex cone optimality criterion

$$\rho(b) \leq 0,\ b \in [\alpha, \beta], \tag{32}$$

where $\rho(b) = (y - y^*, e(b))_W$ is the discrete Laplace transform of the residual $r = y - y^*$.

The algorithm to fulfil (32) makes a grid in $b \in [\alpha, \beta]$. If (32) is fulfilled, then stop. Otherwise add the value $b_*$ to the vector $\mathbf{b}$, which maximizes $(y - y^*, e(b))_W$. Find the vector $\mathbf{a}$ corresponding to the new $\mathbf{b}$ using a nonnegative least squares algorithm (NNLS). Then solve a weighted nonlinear least squares problem (or a maximum likelihood approximation problem) to find the optimal values of the parameters $(\mathbf{a}, \mathbf{b})$ by a Gauss-Newton algorithm. Ruhe uses a primitive cross validation method to find the optimal number of terms, which sometimes indicates too few terms. In such cases he recommends a more general cross validation method.

In [76] Ruhe analyses the sensitivity to data, which is detected as close terms in $b_i$ and by looking at the size of $a_i$. Defining the matrix $\mathbf{E}(\mathbf{b}) = [e(b_1), ..., e(b_p)]$, the residual vector $\mathbf{r}$ with elements $r_j = \sum_{i=1}^{p} a_j \exp(-b_j t_j) - y_j$ is written as $\mathbf{r} = \mathbf{E}(\mathbf{b})\mathbf{a} - \mathbf{y}$ and $\mathbf{y}^* = \mathbf{E}(\mathbf{b})\mathbf{a}$. The second new main algorithm by Ruhe has three phases. The Phase I algorithm gives the maximal number of terms:

1. Make a grid in $b \in [\alpha, \beta]$ giving $[b_1, ..., b_N]$.
2. Solve $\underset{a \geq 0}{min} \|\mathbf{E}(\mathbf{b})\mathbf{a} - \mathbf{y}\|_W$ using a NNLS-algorithm.
3. Condense clusters with the formula $b_i^{(0)} := \sum_{i=r}^{k} a_i b_i / \sum_{i=r}^{k} a_i$ where $r$ and $k$ are defined from $a_{r-1} = 0$, $a_r \cdot a_{r+1} \cdot a_{r+2} \cdot ... \cdot a_k > 0$, $a_{k-1} = 0$. The result is a vector $\mathbf{b}$ with $p$ terms.

The clustering formula is the same as in [35] and the choice of terms to cluster is based on the *m-mode-property* presented in [45]. The finer the grid, the more computational effort needed and the more ill-conditioned is the matrix $\mathbf{E}(\mathbf{b})$. On the other hand, the *1-mode-property* is nothing to strive for. Ruhe [75, page 17] recommends three terms for each final exponent as ideal. For $n$ data points, the usual result is $p \leq (n+1)/2$ and most often a much smaller $p$.

In Phase II a nonlinear least squares algorithm finds the optimal parameter values with $p$ terms. If using a weighted least squares fit criterion, a separable algorithm is used with the vector $\mathbf{b}$ from Phase I as initial values. If ML-weighting is used, a separable algorithm can not be used as the vector $\mathbf{a}$ is no longer linear in the residual vector. In this case the vector $\mathbf{b}$ from Phase I is used as initial values together with a vector $\mathbf{a}$ obtained by solving a linear least squares problem using a NNLS algorithm.

The Hessian has a $2 \times 2$-block structure in the variables $\mathbf{a}$ and $\mathbf{b}$. Ruhe uses a Marquardt-like method: He sets the $[1, 2]$ and $[2, 1]$-block to zero and replaces the diagonal element

in the $[2, 2]$-block with their absolute values. Ruhe motivates the method with three arguments: The problem can be solved as an augmented least-squares problem; a nonpositive definite Hessian is avoided; quadratic convergence close to a minimum is achieved.

Phase III investigates if the current $p$ is optimal by checking criterion (32). Ruhe also plots the transformed residual $\rho(b)$ on $[\alpha, \beta]$. If the transformed residual is nonpositive, an optimum is found. Otherwise go to Phase I.

Tests were made on some artificial and some empirical data series. There were some problems with convergence to local minima. A plot of the transformed residual showed to be positive sometimes.

# 8 Minimizing the least-squares objective by optimization

In this section we consider the problem of minimizing the nonlinear least squares objective by standard optimization techniques. This problem (NLLS)) is formulated as to minimize a function

$$\min s(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{n} r_j^2(\mathbf{x}) = \frac{1}{2}\mathbf{r}^T\mathbf{r}. \tag{33}$$

As we have seen in previous sections some kind of nonlinear least squares problem is a subproblem in many of the more advanced algorithms for the exponential fitting problem. One problem, which is much more prominent for the exponential fitting problem, is the choice of initial values for the optimization. This problem is often overlooked. We first discuss some of the references that have used standard numerical optimization code and then give some general views of the pros and cons of different algorithms for the nonlinear least squares problem.

Steyn and Wyk [85] present some initial value techniques (see Section 2 and 4) and then compare two implementations of nonlinear least squares algorithms for three special sets of initial values. A Gauss-Newton routine named BMDP3R and a Marquardt-routine in the program SHARE Program Library Agency. The two implementations were approximately equal, maybe with some differences dependent on the implementations. They preferred the Marquardt method due to better implementation and that this method is less sensitive to initial values.

Jennrich and Bright [42] also use the Gauss-Newton routine BMPD3R. Their data originates from a two-compartment model of sulfate in the body of a baboon. From the slope of the logarithm of the first part and the last part of the time series, they pick out an initial $\mathbf{b}$. A variety of other initial values were also tried with the same result. Thus the initial values for the Gauss-Newton method were not critical in this case.

In the early seventies Kirkegaard and Eldrup [47] developed a Fortran program POSITRONFIT implementing a separable nonlinear least squares algorithm to estimate positron lifetime spectra in nuclear physics. The algorithm is a semi-linear Marquardt type of algorithm. The function to be fitted is a sum with constant background $a_0$, defined as

$f_j = a_0 + \sum_{i=1}^{p} F_{i,j}$ of decaying exponentials

$$F_{i,j} = a_i \left[ Y_{i,j} - Y_{i,j+1} - \operatorname{erf}\left( \left( t_j - T_0 \right) / \sigma \right) + \operatorname{erf}\left( \left( t_{j+1} - T_0 \right) / \sigma \right) \right] \tag{34}$$

where

$$Y_{i,j} = \exp\left( -\mathrm{b}_i \left( \left( t_j - T_0 - \frac{1}{4} \mathrm{b}_i \sigma^2 \right) \right) \right) \cdot \left( 1 - \operatorname{erf}\left( \frac{1}{2} \mathrm{b}_i \sigma - \left( t_j - T_0 \right) / \sigma \right) \right). \tag{35}$$

The noise is assumed to be Gaussian. The unknown parameters to be determined are

$$\mathbf{a} = [a_0, a_1, ..., a_p] \quad \text{and} \quad \mathbf{b} = [T_0, \mathrm{b}_1, ..., \mathrm{b}_p]. \tag{36}$$

The program converges rapidly in most cases and has been used by many research groups. In their tests the program convergenced in about fifteen iterations to the same point for different initial values, except for a few cases where the background was mixed with a very long lifetime. Only a few cases gave more than three terms due to the variance, which did not decrease compared to three terms. But two terms were normally insufficient to give a good fit. Later they extended the program [48] for non-Gaussian errors. Problem with sensitivity for fast components were reported. Very short lifetimes are reasonably resolved, but the results are sensitive to input data.

Since the development of the basic Gauss-Newton and Levenberg-Marquardt type of algorithms, new types of optimization algorithms have been developed for the NLLS problem. Combining such an algorithm with a separable technique, a powerful algorithm for the exponential fitting problem is obtained. We now discuss in more general terms algorithms for nonlinear least squares.

The derivatives of (33) are $\nabla s = \mathbf{J}^T \mathbf{r} = \mathbf{g}$ and $\nabla^2 s = \mathbf{J}^T \mathbf{J} + \sum_{j=1}^{n} r_j(\mathbf{x}) \mathbf{S}_i(\mathbf{x}) = \mathbf{H}$. We denote the two parts of the Hessian $\mathbf{C} = \mathbf{J}^T \mathbf{J}$ and $\mathbf{S} = \sum_{j=1}^{n} r_j(\mathbf{x}) \mathbf{S}_i(\mathbf{x})$. The problem is that $\mathbf{S}$ is usually computationally expensive. The Jacobian $\mathbf{J}$ may also be expensive to compute and it is common to make errors in the analytic derivation and coding of this matrix. Therefor it is an advantage if computer algebra or automatic differentiation techniques are possible to use. From an optimization point of view, the nonlinear least squares problems are often divided into two or three classes:

**Definition 6** . *Zero residual problem (Z): The residual in a (local) optimum $x^*$ is zero.*

**Definition 7** . *Small residual problem (S): The residual in a (local) optimum $x^*$ is small and $s$ is close to linear.*

**Definition 8** . *Large residual problem (L): The residual in a (local) optimum $x^*$ is large or $s$ is strongly nonlinear.*

The problem of fitting an exponential sum, is classified as (L) due to its nonlinearity. As the number of terms $p$ increases, the exponential fitting problem gets more and more ill-conditioned, which demands more and more on the numerical implementation and properties of the algorithm used. A short overview of some quasi-Newton methods for getting a search direction $\mathbf{d}$ and their performance are:

- The standard DFP and BFGS methods lack some of the second order information in $\mathbf{C}$ and $\mathbf{S}$ and are not locally quadratic convergent. BFGS gives superlinear but not quadratic convergence for (Z).

- More suited for minimizing sum of squares is the Gauss-Newton method (GN): solve $\mathbf{Cd} = -\mathbf{g}$. Near a local optimum of the kind (Z) we have $\mathbf{r} \approx \mathbf{0}$ and thus $\mathbf{H} \approx \mathbf{C}$. The GN method has local quadratic convergence and behaves well. For (S) it has quick linear convergence but for (L) it behaves poorly (see [27, page 225]).

- If the Jacobian $\mathbf{J}$ is strongly rank deficient, the system $\mathbf{Cd} = -\mathbf{g}$ has not a well-defined solution. The Levenberg-Marquardt method (LM) overcome this by adding a diagonal matrix $k\mathbf{I}$ or $\mathbf{D}$ so that the system $(\mathbf{C}+k\mathbf{I})\,\mathbf{d} = -\mathbf{g}$ or $(\mathbf{C}+\mathbf{D})\,\mathbf{d} = -\mathbf{g}$ get a unique solution. This give a robust method, but on the other hand it might behave more like steepest descent and get a slow convergence. For problems (L) this Hessian approximation does not take into account the second part $\mathbf{S}$ of the Hessian. Nazareth [58] presents some results from Meyer (1970) and McKeown (1975). By studying the spectral radius $\rho$ of $\mathbf{C}^{-1}\mathbf{S}$ in a local optimum $x^*$, he concludes only linear convergence for GN and LM if the spectral radius is not small. (See also [27, page 233] or [32, page 116]).

- Another way to handle the ill-conditioning of the Jacobian is to use some kind of regularization, like subspace minimization [51]. We have implemented this technique in the nonlinear least squares solvers in NLPLIB TB [37], and this works well for the exponential fitting problem.

- Hybrid methods. As GN has good properties for (Z) and (S) and BFGS has good properties for general NLP, one can invent a criterion for switching between GN and BFGS. This strategy is used by Al-Baali and Fletcher [6] and Fletcher and Xu [31]. The critical point is when to switch between the two algorithms in a hybrid method. Fletcher simply suggests a GN-step when $s^k - s^{k+1} \geq t \cdot s^k$ with e.g. $t = 0.2$ ([32, page 117]).

- Structured secant methods: Another way is to invent a QN-updated approximation $\widetilde{\mathbf{S}}$ for the $\mathbf{S}$-part of the Hessian and use $\mathbf{H} \approx \mathbf{C}+\widetilde{\mathbf{S}} = \mathbf{B}$. Such methods has been discussed in [28], [27], [89] and [41].

The motivation for structured secant methods differs. Some hope for a better convergence property than LM. Huschens worries about the switch criterion in hybrid methods. As GN has local quadratic convergence for (Z), it is also desirable to get an approximation matrix $\widetilde{\mathbf{S}}$ which approaches zero in the case of a zero residual problem (Z). This is obtained by either sizing $\mathbf{B} = \mathbf{C}+\sigma\widetilde{\mathbf{S}}$ with some $\sigma$ or using $\widetilde{\mathbf{S}} = \|\mathbf{r}\|\,\mathbf{A}$ as an adaptive scaling. The aim of these methods could be viewed as finding a hybrid method with continuous switch.

In recent years parallel computing algorithms using distributed systems for the NLLS problem has been investigated. In [19] Coleman and Plassmann propose a parallel LM algorithm for solving NLLS. In [20] they analyze the efficiency of this method. They find the experimental results to be "essentially full efficiency when the row size is sufficiently larger than the number of processors".

# 9 Some theoretical results

Cromme has written some papers on approximation in Chebyshev norm. In [22] he shows that 'local strong unicity' is sufficient to give good numerical behavior of the algorithms. He investigates the usefulness of Chebyshev-approximation also for exponential sums and the use of strong local unicity. Cromme works in the same spirit in [25]. A theorem about strong unique local best approximations and regularity in $\mathbf{a}$ for a problem of minimizing $\|f - F(\mathbf{a})\|_\infty$ implies quadratic convergence for a sequence of iterations in this norm.

In [23] Cromme gives a basis for convergence analysis for a large class of iterative methods. He studies the notion "strong uniqueness" and prove a theorem on convergence order. He gives some examples of applications. One of these is the exponential sum problem. To overcome problems of separating close terms, he propose a second order method to be used.

In [24] and [25] Cromme derives a differential characterization. He reports that it retains differentiability and regularity even for coalescing terms. This helps to derive necessary and sufficient conditions for local best approximations. It could be of use in the development of numerical procedures.

In 1981 Braess writes an introduction to the critical point theory [15] and some consequences of it [14]. In [15] he discusses $C^1$-manifolds, Strong unique local best approximations (l.b.a), the Haar-property ( i.e. Spanning a Haar subspace under some conditions) and descending flow (described as continuous variant of Newton iteration).

In [14] Braess theorem on unique Haar-embedded manifolds is applied to the exponential sum problem. The main result of this paper is that the number of solutions are bounded by $\frac{1}{4\sqrt{n}} 2^{2n/3} \leq c_n \leq n!$ Thus it is no idea to try to find all local best approximations in a general case.

An empirical illustration is given by Van den Bos in [86]. He finds that a small number of data points may give fewer terms then expected. He gives an explanation to some of this phenomena in terms of catastrophe theory. His hope is that future research (after 1980) will enable experimental design to avoid what he calls structural differences, i.e. the fit gives fewer terms then is hidden in the source of the data points. To explain this phenomena start by defining a weighted sum of functions $g(t)$ as

$$f(t, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p} a_i \cdot g(t, b_i). \tag{37}$$

The function $g(t)$ can be exponential type, Gaussian peaks etc. The normal equations for the residual square sum

$$\frac{1}{2} \sum_{j=1}^{n} [y_j - f(t_j, \mathbf{a}, \mathbf{b})]^2 \tag{38}$$

is

$$\begin{cases} \sum_{j=1}^{n} [y_j - f(t_j, \mathbf{a}, \mathbf{b})] \, g(t, b_i) & = & 0, & i = 1, ..., p \\ \sum_{j=1}^{n} [y_j - f(t_j, \mathbf{a}, \mathbf{b})] \, a_i \frac{\partial g(t, b_i)}{\partial b_i} & = & 0, & i = 1, ..., p. \end{cases} . \tag{39}$$

If we have a solution (stationary point) $(\mathbf{a}^*, \mathbf{b}^*)$ to the normal equations and split one term into two: $a_i = a_{i1} + a_{i2}$, $b_i = b_{i1} = b_{i2}$, we get a bigger set of variables $\left(\mathbf{a}^{*'}, \mathbf{b}^{*'}\right)$,

which is also a solution to the normal equations. This can be viewed as a bifurcation point in the fitting process. A numerical example by Van Den Bos [86] is illustrated below. Set $f(t) = 0.7\exp(-t) + 0.3\exp(-0.8t)$, use equidistant data $t = \{0.4, 0.8, ..., 4.0\}$ and distort the exact data with errors $e_j$ so that $y_j = f(t_j) + e_j$. Plot the sum of squared residuals (38) with and without errors as a function of $b_{1,2}$, shown in Figure 3 and Figure 4, respectively.

Van Den Bos studied three standard deviations in twenty experiments and found that sometimes the saddle point has turned into a unique minimum or that the absolute minimum and the second one have changed place with respect to the line $b_1 = b_2$. That the three stationary points have become one, he explains by making a Taylor expansion around some point $b_1 = b_2 = b$ and then making the variable transformation $b_1' = b_1 - b$, $b_2' = b_2 - b$. By diagonalizing the Hessian of the least squares function in this point $b$ he finds conditions for the types of the stationary points. This explains why the stationary points become one. To explain why the two minima are interchanged, he uses catastrophe theory and views it as a cusp point, which defines a bifurcation set. As a simple example we illustrate with errors $e_1 = 0.005, e_2 = 0.005$ and the other errors set to zero.

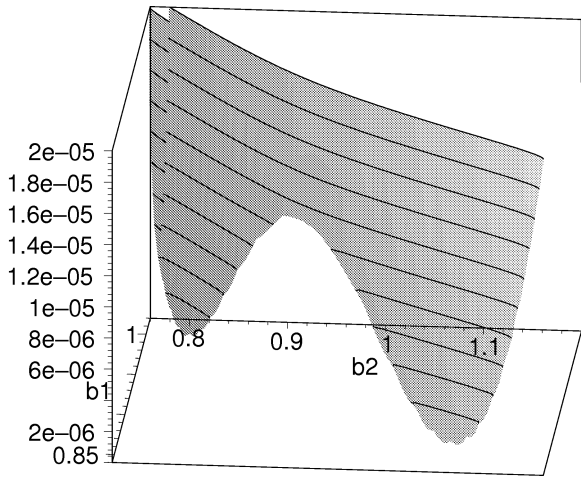


Figure 3: Van Den Bos example. The undisturbed squared norm of the least squares residual as a function of the parameters $b_1$ and $b_2$. The figure show two minima and one saddle point.

Without distortion there are minima in $(1.0, 0.8)$ and $(0.874, 1.096)$ and a saddle point in $(0.933, 0.933)$. The first minimum is absolute. Note that the saddle is on the line $b_1 = b_2$. The disturbed data give a unique minimum. The bifurcation is between 0.0035 and 0.004 for the errors, where the plot turn from a saddle as in Figure 3 to a bowl as in Figure 4.

Structural stability is also discussed by Jongen and Weber [43]. They present a result on necessary and sufficient conditions for structural stability of a constrained optimization problem defined as

$$\begin{array}{l} \min f(x) \\ \\ h_i(x) = 0,\ i \in I \quad (\text{ set of indices}) \\ g_j(x) \leq 0,\ j \in J \quad (\text{ set of indices}). \end{array} \tag{40}$$

They define equivalent optimization problems as the existence of continuous mappings

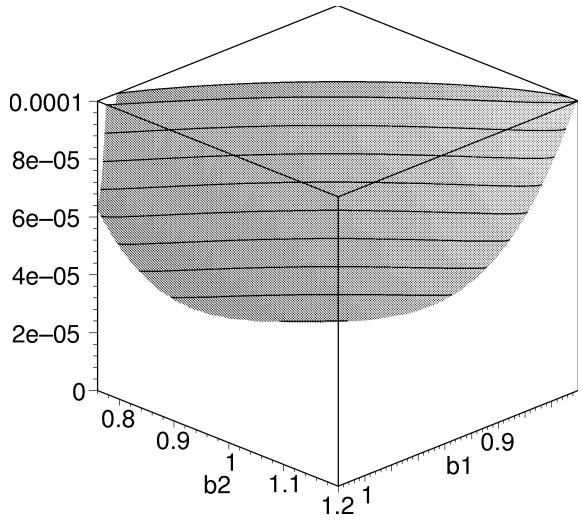


Figure 4: Van Den Bos example. The disturbed squared norm of the least squares residual as a function of the parameters $b_1$ and $b_2$. The function now has one unique minima. The banana shaped valley looks narrow due to the small vertical scaling. With equal vertical and horizontal scaling the figure will be much more flat.

with some homeomorphic properties and define *Structurally stable* as equivalence between two problems in a $C^2-$neighborhood. A third definition is a constraint qualification:

**Definition 9** . *The Mangasarian-Fromovitz constraint qualification (MFCQ) holds if the following two conditions (MF1, MF2) are satisfied:*

*MF1: The gradient vectors* $Dh_i(x)$ *of the equalities in (40) are linearly independent.*

*MF2: There exists a vector* $\xi \in R^N$ *satisfying both*

$$Dh_i(x)\,\xi = 0,\ i \in I$$

*and*

$$Dg_j(x)\,\xi > 0,\ j \in J_0(x),$$

*where* $J_0(x) = \{j : g_j(x) = 0\}$.

The main theorem proved is

**Theorem 5** . *An optimization problem (40) with compact feasible set is structurally stable* $\Leftrightarrow$ *The conditions C1, C2 and C3 are fulfilled. The conditions are:*

*C1: The MFCQ condition is satisfied at every feasible point.*

*C2: Every Karush-Kuhn-Tucker point is strongly stable in the sense of Kojima [49].*

*C3: Different Karush-Kuhn-Tucker points have different function values.*

It seems that condition C3 is not fulfilled, because numerically, the goodness of fit seems to coincide for different points according to [86]. So the problem is probably not structurally stable. Jongen has written other articles on the subject, e.g. [34].

# 10 Relations to mathematical systems theory

Rewriting a linear homogeneous differential equation as a square system $\dot{\mathbf{x}} = \mathbf{Ax}$ and summing the parts together by $\mathbf{y} = \mathbf{Cx}$ exemplifies the relation between the exponential sum problem (2) and the partial realization problem of identifying matrices $\mathbf{A}, \mathbf{C}$ in a linear control problem with a finite data set. If there is an external linear input $\mathbf{u}$ with $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$, a transfer function $\mathbf{W}(\mathbf{z}) = \mathbf{C}(\mathbf{zI} - \mathbf{A})^{-\mathbf{1}}\mathbf{B}$ can be defined as a matrix fraction description (MFD). There is much written literature in this field. Here follows a brief outline.

In [73] Rissanen developed an iterative algorithm by letting the numerical rank of a Hankel matrix $\mathbf{A}$ determine the order of the system. This is the same matrix as in (8). In his thesis on MFD-theory, Kung [50] developed algorithms for determining the greatest common divisor of two polynomial matrices and extended some results for systems of one independent variable to the two variable case. As a tool in systems theory, Brockett [16] showed that sequences of at least McMillan degree $k$ is a differentiable manifold. A result which simplifies some proofs in systems theory. Gragg studied the correspondence between infinite Hankel matrices, their associated formal Laurent series $f(z) = \sum_{j=0}^{\infty} f_j z^{j-1}$ and real sequences $\{f_j\}_{j=1}^{\infty}$. On the other hand, if the numbers $f_j$ are regarded as moments, the theory of the moment problem can be applied. A classical book on this is the one by Akhiezer [5]. Gragg and Lindquist [33] studied the scalar partial realization problem from a numerical linear algebra point of view. Another article pointing out the interdependence between the different mathematical branches is the one by Martin and Clyde [55]. They showed that the question of observability is equivalent to the existence of a solution to a boundary value problem. In the case where the linear differential equation has multiple roots, the coefficients $a_i$ in the exponential sum model (1) become polynomials. From a mathematical systems theory point of view Ammar et al. developed an algorithm for this problem in [7].

# 11 Conclusions and further work

The exponential fitting problem is a tough one and although very much research has been done for the problem more efficient and reliable algorithms should be developed. Given good enough starting points, it seems like a stepwise regression algorithm is preferable. To find the optimal parameters given the number of exponential terms $p$, a constrained non-linear least squares algorithm with robust performance should be run. The currently best choice is probably a hybrid Gauss-Newton and Quasi-Newton algorithm. An information criterion is then used to determine the optimal number of terms $p$.

It is essential to have good starting values for the nonlinear parameter estimation. To find good starting values we have found the idea of geometrical sums, described in Section 4, to be fruitful and an alternative to a modified Prony algorithm. Work on finding good starting values should be continued and used in a forth-coming general, robust exponential fitting algorithm.

# References

[1] M. Agha. A direct method for fitting linear combinations of exponentials. *Biometrics*, 27:399–413, 1971.

[2] Hirotugu Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, 20:425–439, 1969.

[3] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F.Csaki, editors, *Proceedings from 2nd Int. Symposium on Information Theory, Budapest, Supp. to Problems of Control and Information Theory*, pages 267–281, 1972.

[4] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, December 1974.

[5] N. L. Akhiezer. *The classical moment problem and some related questions in analysis.* Oliver and Boyd, 1965.

[6] M. Al-Baali and R. Fletcher. Variational methods for non-linear least squares. *J. Oper. Res. Soc.*, 36:405–421, 1985.

[7] G. Ammar, W. Dayawansa, and C. Martin. Exponential interpolation: Theory and numerical algorithms. *Applied Mathematics and Computation*, 41:189–232, 1991.

[8] E. Avrett and D. Hummer. *Mon. Notic. Roy. Astron. Soc.*, 130:865, 1965.

[9] M. Bertero, P. Boccacci, and E. R. Pike. On the recovery and resolution of exponential relaxation rates from experimental data ii - the optimum choice of experimental sampling points for laplace transform inversion. *Proc. Roy. Soc.*, 393(51), 1984.

[10] M. Bertero, P. Brianzi, and E. R. Pike. On the recovery and resolution of exponential relaxation rates from experimental data: Laplace transform inversions in weighted spaces. *Inverse Problems*, 1:1–15, 1985.

[11] Dietrich Braess. On the nonuniqueness of monosplines with least $L_2$-norm. *Journal of Approximation Theory*, 12(1):91–93, September 1974.

[12] Dietrich Braess. On Rational $L_p$-Approximation. *Journal of Approximation Theory*, 18(2):136–151, October 1976.

[13] Dietrich Braess. Chebyshev approximation by $\gamma$-polynomials. III. On the number of best approximations. *Journal of Approximation Theory*, 24(2):119–145, October 1978.

[14] Dietrich Braess. Global analysis and nonlinear approximation and its application to exponential approximation. 2. Applications to exponential approximation. In Zvi Ziegler, editor, *Approximation Theory and Applications*, pages 39–63. Technicon, Haifa, Israel, Academic press, May 5-June 25 1980 1981.

[15] Dietrich Braess. Global analysis in nonlinear approximation and its application to exponential approximation. 1. The uniqueness theorem for Haar-Embedded manifolds. In Zvi Ziegler, editor, *Approximation Theory and Applications*, pages 23–37. Technicon, Haifa, Israel, Academic press, May 5-June 25 1980 1981.

[16] R. Brockett. The geometry of the partial realization problem. In *Proceedings of the 1978 IEEE Conference on Decision and control*, pages 1048–1052, 1978.

[17] Joseph Burstein. Approximations by exponentials, their extensions & differential equations. metric press, Boston, 1997.

[18] David G. Cantor and John W. Evans. On approximation by positive sums of powers. *SIAM Journal on Applied Mathematics*, 18(2):380–388, March 1970.

[19] Thomas F. Coleman and Pauls E. Plassmann. Solution of nonlinear least squares problems on a multiprocessor. In G. A. van Zee and J. G. G. van de Vorst, editors, *Lecture Notes in Computer Science 384, Parallel Computing*, pages 44–60. Springer-Verlag, 1988.

[20] Thomas F. Coleman and Pauls E. Plassmann. A parallel nonlinear Least-Squares solver: Theoretical analysis and numerical results. *SIAM Journal on Scientific and Statistical Computing*, 13(3):771–793, May 1992.

[21] Richard G. Cornell. A method for fitting linear combinations of exponentials. *Biometrics*, pages 104–113, 1962.

[22] Ludwig Cromme. Eine Klasse von Verfahren zur Ermittlung bester nichtlinearer Tschebyscheff-Approximationen. *Numerische Mathematik*, 25:447–459, 1976.

[23] Ludwig Cromme. Strong uniqueness - a far-reaching criterion for the convergence analysis of iterative procedures. *Numerische Mathematik*, 29:179–193, 1978.

[24] Ludwig Cromme. A unified approach to differential characterizations of local best approximations for exponential sums and splines. Technical report, Lehrstühle für Numerische und Angewandte Mathematik, Universität Göttingen, University of Berkeley, California, May 1981.

[25] Ludwig J. Cromme. Regular $C^1$-parametrizations for exponential sums and splines. *Journal of Approximation Theory*, 35(1):30–44, May 1982.

[26] M. Della Corte, L. Buricchi, and S. Romano. On the fitting of linear combinations of exponentials. *Biometrics*, 30:367–369, June 1974.

[27] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Englewood Cliffs, N. J., 1983.

[28] J. E. Dennis, Jr, D. M. Gay, and R. E. Welch. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7(3):348–368, 1981.

[29] George T. Duncan. An empirical study of jackknife-constructed confidence regions in nonlinear regression. *Technometrics*, 20(2):123–129, May 1978.

[30] John W. Evans, William B. Gragg, and Randall J. LeVeque. On least squares exponential sum approximation with positive coefficients. *Mathematics of Computation*, 34(149):203–211, 1980.

[31] R. Fletcher and C. Xu. Hybrid methods for nonlinear least squares. *IMA Journal of Numerical Analysis*, 7:371–389, 1987.

[32] Roger Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, New York, 2nd edition, 1987.

[33] W. Gragg and A. Lindquist. On the partial realization problem. *Linear Algebra Applications*, 50:277–319, 1983.

[34] J. Guddat and H. Th. Jongen. Structural stability in nonlinear optimization. *Optimization*, 18:617–631, 1987.

[35] S.-Å. Gustafson. A computational scheme for exponential approximation. *Z. Angew. Math. Mech*, 61:284–287, 1981.

[36] F. Hilebrand. *Introduction to Numerical analysis*. McGraw-Hill, 1956.

[37] Kenneth Holmström. NLPLIB TB 1.0 - A MATLAB Toolbox for Nonlinear Optimization and Parameter Estimation. Technical Report IMa-TOM-1997-02, Department of Mathematics and Physics, Mälardalen University, Sweden, 1997.

[38] Kenneth Holmström. The NLPLIB Toolbox for Nonlinear Programming in MATLAB. *Annals of Operations Research*, Modeling Languages and Approaches, 1998. Submitted.

[39] Kenneth Holmström. TOMLAB - An Optimization Development Environment in MATLAB. *Annals of Operations Research*, Modeling Languages and Approaches, 1998. Submitted.

[40] G. Hunt and I. Grant. *J. Atmospheric Sci.*, 26:963, 1969.

[41] J. Huschens. On the use of product structure in secant methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 4(1):108–129, February 1994.

[42] Robert I. Jennrich and Peter B. Bright. Fitting systems of linear differential equations using computer generated exact derivatives. *technometrics*, 18(4):385–392, November 1976.

[43] H. Th. Jongen and G.-W. Weber. Nonlinear optimization, characterization of structural stability. *Journal of Global Optimization*, 1(1):47–64, 1991.

[44] M. Kahn, M. S. Mackisack, M. R. Osborne, and G. K. Smyth. On the consistency of Prony's method and related algorithms. *Journal of Computational and Graphical Statistics*, 1:329–349, 1992.

[45] Thomas Kaijser. A simple inversion method for determining aerosol size distributions. Technical Report C 30231-E, FOA, Huvudavdelning 3, 581 11 Linköping, September 1981.

[46] David W. Kammler. Least squares approximation of completely monotonic functions by sums of exponentials. *SIAM Journal on Numerical Analysis*, 16:801–818, 1979.

[47] Peter Kirkegaard and Morten Eldrup. Positronfit: A versatile program for analysing positron lifetime spectra. *Computer Physics Communications*, 3:240–255, 1972.

[48] Peter Kirkegaard and Morten Eldrup. Positronfit extended: A new version of a program for analysing positron lifetime spectra. *Computer physics communications*, 7:401–409, 1974.

[49] M. Kojima. Strongly stable stationary solutions in nonlinear programs. In S.M. Robinsson, editor, *Analysis and computation of fixed points*. Academic Press N.Y., 1980.

[50] S. Y. Kung. *Multivariable and multidimensional Systems: Analysis and Design*. PhD thesis, Department of Electrical Engineering, Stanford University, June 1977.

[51] Per Lindström and Per-Åke Wedin. Methods and Software for Nonlinear Least Squares Problems. Technical Report UMINF-133.87, Inst. of Information Processing, University of Umeå, Sweden, 1988.

[52] Lennart Ljung. *System Identification - Theory for the User*. Prentice Hall, Englewood Cliffs, N. J., 1987.

[53] M. S. Mackisack, M. R. Osborne, and G. K. Smyth. A modified Prony algorithm for estimating sinusoidal frequencies. *Journal of Statistical Computation and Simulation*, 49:111–124, 1994.

[54] Larry Marple. Spectral line analysis by Pisarenko and Prony methods. Technical Report CH1379-7/79/0000-0159, Advent systems INC, 1183 Bourdeaux Dr. Sunnyvale. Ca94086, 1979.

[55] C. Martin and J. Smith. Approximation, interpolation and sampling. *Contemporary Mathematics*, 68:227–252, 1987.

[56] Masahiko Nakamura and Takashi Takahashi. Inversion of the chi-square transform, 1978. Translated paper from the original Japanese paper.

[57] Masahiko Nakamura, Takashi Takahashi, and Shinzo Kodama. Parameter estimation of a linear combination of exponential decays by a method of the gamma transformation - a numerical investigation. *The Transactions of the IECE of Japan*, E 63(6):500–501, June 1980.

[58] Larry Nazareth. Some recent approaches to solving large residual nonlinear least squares problems. *SIAM*, 22(1):1–11, January 1980.

[59] M. R. Osborne. A class of nonlinear regression problems. In R. S. Anderssen and M. R. Osborne, editors, *Data Representation*, pages 94–101. University of Queensland Press, 1970.

[60] M. R. Osborne. Some special nonlinear least squares problems. *SIAM Journal on Numerical Analysis*, 12(4):571–592, 1975.

[61] M. R. Osborne. Nonlinear least squares - the Levenberg algorithm revisited. *Journal of Australian Mathematical Society*, B19:343–357, 1976.

[62] M. R. Osborne and G. K. Smyth. An algorithm for exponential fitting revisited. In J. Gani and M. B. Priestly, editors, *Essays in Time Series and Allied Processes: Papers in Honour of E. LJ. Hannan*, pages 419–430. Sheffield: Applied probability trust, 1986.

[63] M. R. Osborne and Gordon K. Smyth. A modified Prony algorithm for fitting functions defined by difference equations. *SIAM Journal on Scientific and Statistical Computing*, 12:362–382, 1991.

[64] M. R. Osborne and Gordon K. Smyth. A modified Prony algorithm for exponential function fitting. *SIAM Journal of Scientific Computing*, 16(1):119–138, 1995.

[65] Jöran Petersson and Kenneth Holmström. Fitting of Exponential Sums to Empirical Data. Technical Report IMa-TOM-1997-05, Department of Mathematics and Physics, Mälardalen University, Sweden, 1997. Presented at the 16th International Symposium on Mathematical Programming, Lausanne, Switzerland, August 24-29, 1997.

[66] Jöran Petersson and Kenneth Holmström. Identifying parameters and model order for two classes of exponential sum fitting problems. Technical Report IMa-TOM-1998-05, Department of Mathematics and Physics, Mälardalen University, Sweden, 1998.

[67] Jöran Petersson and Kenneth Holmström. Initial values for a class of exponential sum least squares fitting problems. Technical Report IMa-TOM-1998-04, Department of Mathematics and Physics, Mälardalen University, Sweden, 1998.

[68] Jöran Petersson and Kenneth Holmström. Initial values for the exponential sum least squares fitting problem. Technical Report IMa-TOM-1998-01, Department of Mathematics and Physics, Mälardalen University, Sweden, 1998.

[69] Jöran Petersson and Kenneth Holmström. Initial Values for Two Classes of Exponential Sum Least Squares Fitting Problems. Technical Report IMa-TOM-1998-07, Department of Mathematics and Physics, Mälardalen University, Sweden, 1998.

[70] Jöran Petersson and Kenneth Holmström. Methods for Parameter Estimation in Exponential Sums. In Arne Løkketangen, editor, *Proceedings from the 5th Meeting of the Nordic Section of the Mathematical Programming Society*, Molde, 1998. Division of Operations Research, Molde University. To be published.

[71] P. F. Price. A comparison of the Least-Squares and maximum likelihood estimators for counts of radiation quanta which follow a poisson distribution. *Acta Cryst.*, A35:57–60, 1979.

[72] E. Raschke and U. Stucke. *Beitr. Physik Atmos.*, 46:203, 1973.

[73] J. Rissanen. Recursive identification of linear sequences. *SIAM Journal of Control*, 9:420–430, 1971.

[74] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[75] Axel Ruhe. Least squares fitting by positive sums of exponentials. Technical Report UMINF-70.78, Inst. of Information Processing, University of Umeå, Sweden, 1978. Revised February 29, 1980.

[76] Axel Ruhe. Fitting empirical data by positive sums of exponentials. *SIAM Journal on Scientific and Statistical Computing*, 1(4):481–498, 1980.

[77] Axel Ruhe. Fitting empirical data by positive sums of exponentials. *Compstat*, pages 622–628, 1980.

[78] Axel Ruhe and Per-Åke Wedin. Algorithms for Separable Nonlinear Least Squares Problems. *SIAM Review*, 22:318–337, 1980.

[79] B. K. Shah. Obtaining preliminary estimates to fit two-term exponential model to blood concentration data. *Journal of Pharmaceutical Sciences*, 62:1208–1209, 1973.

[80] R. Shapiro. Information loss and compensation in linear interpolation. *Journal of Computational Physics*, 10:65–84, 1972.

[81] R. Shapiro. Smoothing, filtering and boundary effects. *Ref. Geophys. and Space Phys.*, 8:359–387, 1972.

[82] M. R. Smith, S. Cohn-Sfetcu, and H. A. Buckmaster. Decomposition of multicomponent exponential decays by spectral analytic techniques. *Technometrics*, 18(4):467–482, November 1976.

[83] G. K. Smyth. *Coupled and Separable Iterations in Nonlinear Estimation.* PhD thesis, Canberra, Australian National University, 1985.

[84] H. S. Steyn. Fitting linear combinations of exponential decays by using Fourier analysis. In *Compstat*, pages 615–621, Vienna, 1980. Physica-Verlag.

[85] H. S. Steyn and J. W. J. van Wyk. Some methods for fitting compartment models to data. Technical report, Wetenskaplike bydraes van die pu vir cho, Potchefstroomse Universiteit vir CHO, 1977.

[86] A. Van den Bos. A class of small sample nonlinear least squares problems. *Automatica*, 16:487–490, 1980.

[87] J. M. Varah. On fitting exponentials by nonlinear least squares. *SIAM Journal on Scientific and Statistical Computing*, 6:30–44, 1985.

[88] W. J. Wiscombe and J. W. Evans. Exponential-Sum fitting of radiative transmission functions. *Computational Physics*, 24(4):416–444, August 1977.

[89] Hiroshi Yabe and Toshihiko Takahashi. Factored quasi-Newton methods for nonlinear least squares problems. *Mathematical Programming*, 51(1):75–100, July 1991.