MASSACHUSETTS INSTITUTE OF TECHNOLOGY
HAYSTACK OBSERVATORY
WESTFORD, MASSACHUSETTS 01886

18 November 1996

Telephone: 508-692-4764
Fax: 617-981-0590

To: Millimeter-wave VLBI Group

From: Alan E.E. Rogers AEER

Subject: Model fitting using non-linear weighted least squares

1] Least squares is equivalent to maximum likelihood

The estimation of parameters from data corrupted by Gaussian noise by weighted least squares yields the "most likely" or maximum likelihood estimate. This can be shown by using Bayes rule:

$$\max_i P\left[S_i|X\right] = \max_i P\left[S_i\right] p\left(X|S_i\right)/p(X)$$

where
S_i = i^{th} choice of parameter vector
X = measurement vector = S+N
N = noise vector
S = true parameter vector
P[ ] = probability
p = probability density

Since p(X) is not a function of i (using $\max_i$ to denote the index in multidimensional space which maximizes the quantity which follows)

$$\max_i P\left[S_i|X\right] = \max_i P\left[S_i\right] p_n\left(X - S_i\right)$$
$$= \max_i P\left[S_i\right] e^{-|X-S_i|^2} \quad \text{for Gaussian noise}$$

So that $\max_i P\left[S_i|X\right] = \min_i \left|X - S_i\right|^2$

(see Detection; Estimation and Modulation Theory - part 1 Harry L. Van Trees, Wiley 1968)

2] Linear weighted least squares

If the measurements depend on the parameters in a linear manner

$$X = AS + N$$

where X = measurement vector
S = parameter vector
N = noise vector
A = steering, design, sensitivity or observation coefficient matrix (different names depending on field of study)

In this case we can minimize the sum of weighted errors squared (this sum being known as chi-squared)

$$\chi^2 = \varepsilon^T w \varepsilon$$

where $\varepsilon = X - A\hat{S}$

to obtain the best estimate of S, denoted by $\hat{S}$ (and T designates the transpose of a matrix)

from $\hat{S} = \left(A^T wA\right)^{-1} A^T wX.$

The elements of the weight matrix are given by

$$W_{ij} = 1/\sigma_i^2 \qquad i = j$$

$$= 0 \qquad i \neq j$$

for uncorrelated colored noise with sigma σ_i.
It can be shown that

$$\chi^2 = M - L$$

where M = number of elements in measurement vector
L = number of elements in parameter vector

provided the data can be perfectly characterized by L parameters.

3] Non-linear weighted least squares

In the case of finding the most likely image from measurements of the correlated flux on each baseline and closure phase there is no nice linear transformation from image parameters to measurements. Some parameters, like the overall flux are linearly related but most are not and we have to resort to a brute force exhaustive search in order to minimize the sum of weighted errors squared. For example if we choose to model the image using 2 elliptical Gaussian components we have the following parameters:

1] Total flux of both components
2] Relative flux of 2nd component

3]   Angular size of major axis of component 1
4]   Angular size of major axis of component 2
5]   Axial ratio of ellipse of component 1
6]   Axial ratio of ellipse of component 2
7]   Position angle of major axis of component 1
8]   Position angle of major axis of component 2
9]   Angular distance between components
10]   Position angle of separation of components

Since the measurements of correlated flux are only linear in the first parameter (closure phase being unaffected by this parameter) we need to perform a 9 dimensional exhaustive search to globally minimize the remaining 9 parameters.

The flux ratio f can be found from

$$f = \sum w_i m_i a_i \,/\, \sum w_i m_i^2$$

where m_i = model visibility amplitude
and a_i = observed visibility amplitude.

4] Gaussian model

An elliptical Gaussian image has an analytic visibility

model:

$$e^{-\left\{\left(u^2+v^2\right)w^2\left(\cos^2(\theta)+\sin^2(\theta)/a^2\right)/2\right\}}$$

where u,v are projected baseline components (rad/mas)
w major axis (mas) (FWHM = 2 ln (2) w)
a axial ratio (major/minor)
θ = P.A. of ellipse - P.A. of baseline.

5] Computational efficiency

In order to minimize the computational task it is advantageous to organize the nested loops of the multidimensional search so that the most complex calculations (especially those involving transcendental functions) by done in the outer loops. In addition the quantization (or grid size) needs to be as large as possible without significant chance of missing the global minimum of χ^2. Following a global search, a fine search can be done on each parameter in an iterative manner. An efficiently coded program can compute χ^2 for 100 visibility amplitudes and an equal number of closure phase in about 200 microseconds/model (bench marked on dopey Oct 96).

6] Ambiguities in the model

Even if we find a global minimum in multidimensional χ^2 space there is no guarantee that the model is unique. Other models may fit the data almost equally well. As a means of accessing the uniqueness of the best fit model we can determine the minimum of χ^2 for other models at a given "distance" from the best fit model. An ideal distance metric might be the inverse of the normalized correlation between models. Unfortunately computing this metric is very time consuming and a simpler metric like

$$d = \left(\sum w_i d_i^2\right)^{1/2}$$

is more practical. For this metric, the distance in each dimension is just the number of grid points away from the best fit model. A weight being assigned to each dimension in proportion to its importance.

7] Errors in the parameters

The +/- one sigma error in a given parameter is given by the range of that parameter for which

$$\Delta\chi^2 = \chi^2 - \chi^2_{\min} \leq 1$$

given a global search for a minimum χ^2 over all other parameters. (An excellent discussion of this and other properties of χ^2 is in Numerical Recipes by Press et al)

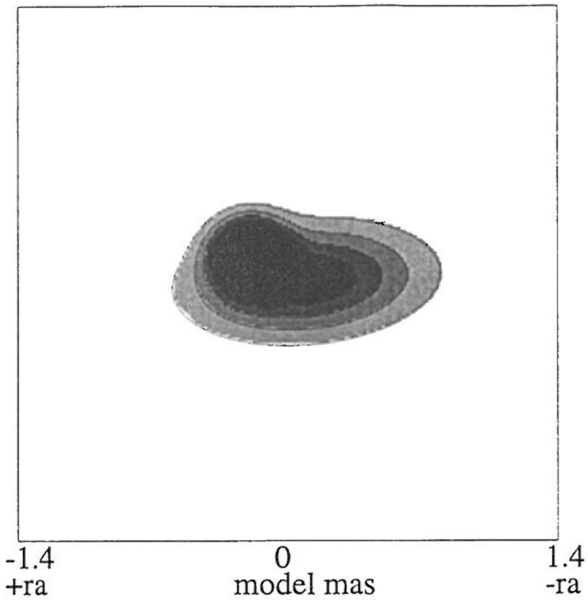
In practice it is easier to vary one parameter a small amount δ from the best fit, search for $\chi^2_{\min}$ in the other parameters and assume that close to the best fit $\Delta\chi^2$ varies quadratically with δ

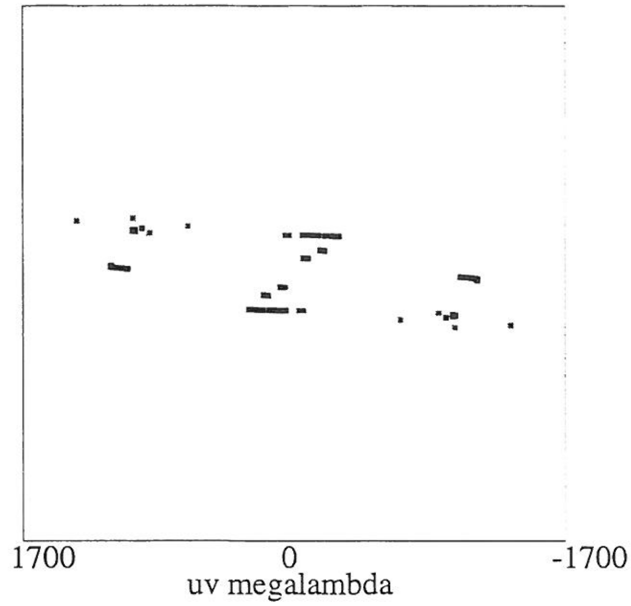
If the observations are linear in the parameters the errors are given by the square root of the diagonal elements of the covariance matrix

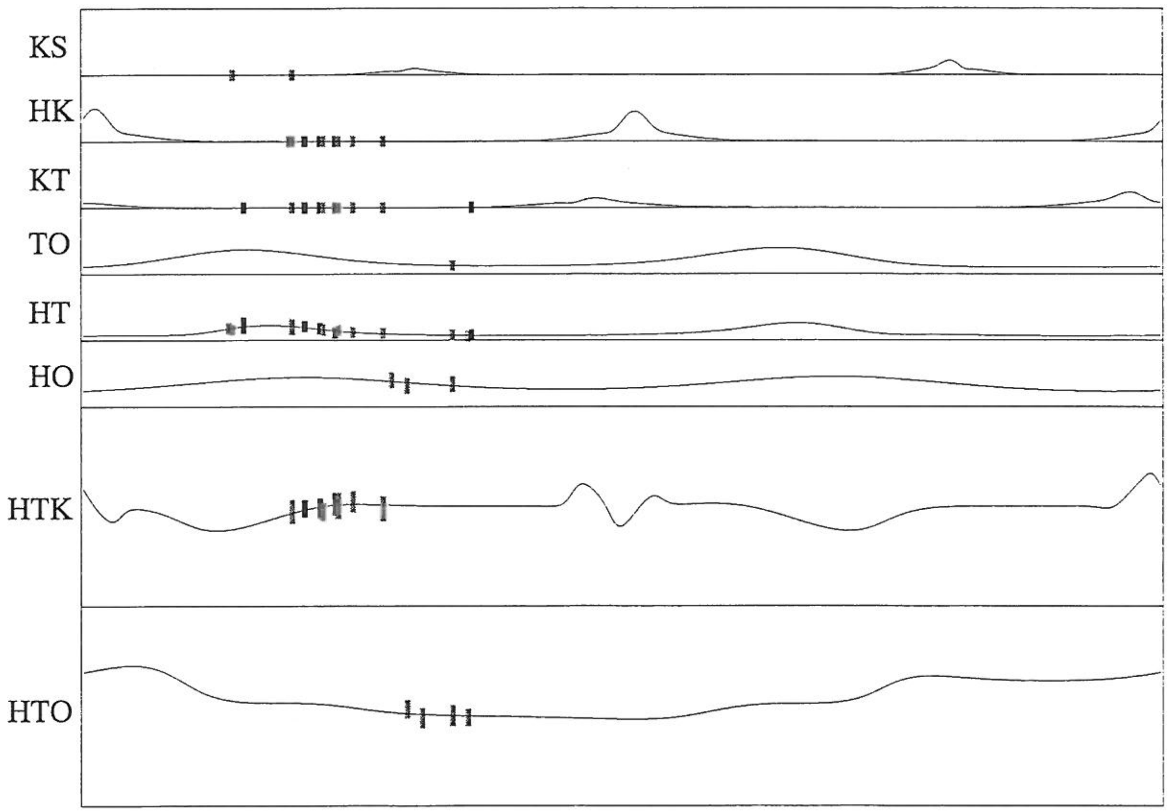
$$C = \left(A^T w A\right)^{-1}$$

8] Adequacy of model

If χ^2 / degree of freedom is less than one the model may be too complex and simpler model may fit the data. On the other hand if χ^2 / degree of freedom is greater than one the number of model parameters may have to be increased to adequately fit the observations. A useful test is to vary the number of model parameters and check the χ^2 dependence.







model fits 0-24 GHA hrs

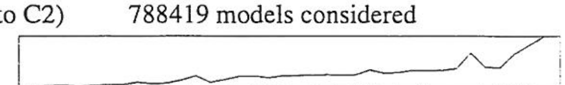
3C273B expt# 2673 dataspan: 96:118:05:00:00 to 96:120:04:40:00 (mk3/4 phase defn.)

Component 1: 2.535 +/- 0.04 Jy 0.166 +/- 0.003 masFWHM(maj.axis) 1.05 +/- 0.21 axial_ratio(maj/min) 83 +/- 4 P.A.(maj.axis)

Component 2: 4.993 +/- 2.26 Jy 0.538 +/- 0.080 masFWHM(maj.axis) 2.11 +/- 0.56 axial_ratio(maj/min) -85 +/- 9 P.A.(maj.axis)

Component separation: 0.320 +/- 0.07 mas -114 +/- 6 P.A.(C1 to C2) 788419 models considered

chi-squared/degree_of_freedom= 1.367 mode=11



chi-squared vs model distance metric (log scale)

biases: H -0.01 O -0.01 T 0.00 K -0.01 S -0.05

CMVA

Mon Nov 18 08:20:05 1996 model.data prg version: 1.2

9] Sample results from model fitting program

Figure 1 shows simulated data (for times at which there is real data) with SNR of 10 in both the amplitudes and closure phases. Sites are

K = Haystack
S = Onsala
H = hat creek
T = kitt peak
O = OVRO

The 2 Gaussian component model fits the observations with reduced χ^2 close to unity. The model is fairly unique as determined by the increasing χ^2 with model distance. Thus there is no other significantly different 2 component Gaussian model which fits the observations without substantial increase in χ^2.

10] Data weighting

The data is weighted in proportion to the SNR^2. However a parameter allows the relative weighting of the closure phases and amplitudes to be changed.