Gaia
DPAC
Data Processing & Analysis Consortium

A general Maximum-Likelihood algorithm for model fitting to CCD sample data

prepared by:	L. Lindegren
approved by:	
reference:	GAIA-C3-TN-LU-LL-078-01
issue:	1
revision:	0
date:	2008-11-12
status:	Issued

Abstract

A general Maximum-Likelihood algorithm is described for fitting arbitrary models to CCD samples, where data are modelled as a Poisson process plus (gaussian) readout noise. It can be applied for example to 1D (LSF) and 2D (PSF) centroiding, but also for estimating the parameters of the LSF, PSF or CDM models.

Document History

Issue	Revision	Date	Author	Comment
D	0	2008-10-28	LL	First draft
1	0	2008-11-12	LL	A few corrections and clarifications added after review by U. Bastian

1 Introduction

An earlier technical note (Lindegren, SAG-LL-032) described an algorithm for centroiding on one-dimensional (1D) sample data. Although that algorithm has been used quite a lot for centroiding on AF data, it is not suitable for generalization e.g. to two-dimensional (2D) PSF fitting or to even more complex models including for example charge distortion effects.

The general principle for Maximum-Likelihood (ML) fitting of arbitrary models in the presence of poissonian noise is however quite simple and can be formulated in a general framework which is independent of the precise model. In this way it should be possible to use the same fitting procedure for 1D and 2D profile fitting to CCD sample data, as well as for more complex fitting (e.g. for estimating the parameters of the LSF model). This note provides the mathematical basis for this framework. Additive noise, including CCD readout and discretization noise, are formally handled as described in the Appendix.

2 Model of sample data

The basic input for the estimation procedure consists of *data* and a parameterized *model*. The estimation procedure will adjust the model parameters until the predicted data agrees as well as possible with observed data. At the same time it will provide an estimate of the covariance matrix of the estimated parameters and a measure of the goodness-of-fit. The ML criterion is used for the fit, which in principle requires that the probability distribution of the data is known as a function of the model parameters. In practice the simplified noise model derived in the Appendix is used; this is believed to be accurate enough and leads to simple and efficient algorithms.

Let $\{N_k\}$ be the sample data, $\boldsymbol{\theta} = \{\theta_i\}$ the model parameters, and $\{\lambda_k(\boldsymbol{\theta})\}$ the sample values predicted by the model for given parameters. Thus, if the model is correct and $\boldsymbol{\theta}$ are the true model parameters, we have for each k

$$\mathrm{E}(N_k) = \lambda_k(\boldsymbol{\theta}) \tag{1}$$

Based on the noise model discussed in the Appendix, we have in addition

$$\mathrm{Var}(N_k) = \lambda_k(\boldsymbol{\theta}) + r^2 \tag{2}$$

where r is the standard deviation of the readout noise. More precisely, the adopted probability density function (pdf) for the random variable N_k is given by (31).

It is assumed that N_k, λ_k and r are all expressed in electrons per sample (not in arbitrary AD units, voltages, or similar). In particular, N_k is the sample value after correction for bias and gain, but including dark signal and background. The readout noise r is assumed to be known; it is never one of the parameters to be estimated by the methods described in this note.

The functions $\lambda_k(\boldsymbol{\theta})$ are in principle defined by the various source, attitude and calibration models, including the LSF, PSF and CDM models. The set of parameters included in the vector $\boldsymbol{\theta}$ varies depending on the application. For example, in the 1D image centroiding algorithm $\boldsymbol{\theta}$ may consist of just two parameters representing the intensity and location of the image; in the LSF calibration process, $\boldsymbol{\theta}$ will contain the parameters (e.g. spline coefficients) defining the LSF for a particular class of stars; and so on. The intensity model $\lambda_k(\boldsymbol{\theta})$ is left completely open here; the only thing we need to know about it is the number of free parameters, $n = \dim(\boldsymbol{\theta})$.

3 Maximum Likelihood estimation

3.1 The log-likelihood function and likelihood equations

Given a set of sample data $\{N_k\}$, the ML estimation of the parameter vector $\boldsymbol{\theta}$ is done by maximizing the likelihood function

$$L(\boldsymbol{\theta}|\{N_k\}) = \prod_k p(N_k|\lambda_k(\boldsymbol{\theta}), r) \tag{3}$$

where $p(N|\lambda, r)$ is the pdf of the sample value from the adopted noise model. Mathematically equivalent, but more convenient in practice, is to maximize the log-likelihood function

$$\ell(\boldsymbol{\theta}|\{N_k\}) = \sum_k \ln p(N_k|\lambda_k(\boldsymbol{\theta}), r) \tag{4}$$

Using the modified poissonian model, Eq. (31), we have

$$\ell(\boldsymbol{\theta}|\{N_k\}) = \text{const} + \sum_k \left[(N_k + r^2)\ln\left(\lambda_k(\boldsymbol{\theta}) + r^2\right) - \lambda_k(\boldsymbol{\theta})\right] \tag{5}$$

where the additive constant absorbs all terms that do not depend on $\boldsymbol{\theta}$. (Remember that r is never one of the free model parameters.) The maximum of (5) is obtained by solving the n simultaneous likelihood equations

$$\frac{\partial \ell(\boldsymbol{\theta}|\{N_k\})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{6}$$

Using (5) or (32) these equations become

$$\sum_k \frac{N_k - \lambda_k(\boldsymbol{\theta})}{\lambda_k(\boldsymbol{\theta}) + r^2}\,\frac{\partial \lambda_k}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{7}$$

3.2 Iterative solution of the likelihood equations

Since the equations in (7) are non-linear, they must be solved by iteration. Given an initial estimate $\boldsymbol{\theta}^{(0)}$, successive updates $\Delta\boldsymbol{\theta}^{(m)}$ and improved estimates $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \Delta\boldsymbol{\theta}^{(m)}$ are calculated for $m = 0, 1, \ldots$, such that $\boldsymbol{\theta}^{(m)}$ converges toward the ML solution $\hat{\boldsymbol{\theta}}$. In the mth iteration we have

$$\boldsymbol{\delta}^{(m)} \equiv \left(\frac{\partial \ell(\boldsymbol{\theta}|\{N_k\})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} \neq \mathbf{0} \qquad (8)$$

and we now seek an update $\Delta\boldsymbol{\theta}^{(m)}$ that will reduce $\boldsymbol{\delta}$ to zero. Using the Taylor expansion

$$\boldsymbol{\delta}^{(m+1)} \simeq \boldsymbol{\delta}^{(m)} + \left(\frac{\partial^2 \ell(\boldsymbol{\theta}|\{N_k\})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} \Delta\boldsymbol{\theta}^{(m)} \qquad (9)$$

and putting the right member to $\mathbf{0}$, we get the update

$$\Delta\boldsymbol{\theta}^{(m)} = -\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|\{N_k\})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} \boldsymbol{\delta}^{(m)} \qquad (10)$$

Note that the Hessian $\boldsymbol{H} = \partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is a symmetric $n \times n$ matrix, while $\Delta\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are n-dimensional vectors or $n \times 1$ matrices. If the parameter estimation is a well-posed problem, then $-\boldsymbol{H}$ is positive definite and there is a unique ML solution.

Dropping the superscript $^{(m)}$ for the moment, the elements of $\boldsymbol{H}$ are

$$\begin{aligned} H_{ij} \equiv \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} &= \frac{\partial}{\partial \theta_i} \sum_k \frac{N_k - \lambda_k}{\lambda_k + r^2} \frac{\partial \lambda_k}{\partial \theta_j} \\ &= \sum_k \left[\left(\frac{-1}{\lambda_k + r^2} - \frac{N_k - \lambda_k}{(\lambda_k + r^2)^2}\right) \frac{\partial \lambda_k}{\partial \theta_i} \frac{\partial \lambda_k}{\partial \theta_j} + \frac{N_k - \lambda_k}{\lambda_k + r^2} \frac{\partial^2 \lambda_k}{\partial \theta_i \partial \theta_j}\right] \end{aligned} \qquad (11)$$

Taking the expectation of this expression, while assuming that we are close to the true solution so that $\mathrm{E}(N_k) \simeq \lambda_k$, we find the much simpler approximation

$$\mathrm{E}(H_{ij}) \simeq -\sum_k \frac{1}{\lambda_k + r^2} \frac{\partial \lambda_k}{\partial \theta_i} \frac{\partial \lambda_k}{\partial \theta_j} \qquad (12)$$

which does not involve the second derivatives of λ_k. As discussed e.g. in Press et al. (2007, Ch. 15.5), it is often favourable for the stability of the iteration process *not* to include the second derivatives when calculating the Hessian. Introducing the symmetric, positive definite matrix $\boldsymbol{A}$ with elements

$$A_{ij} = \sum_k \frac{1}{\lambda_k(\boldsymbol{\theta}) + r^2} \frac{\partial \lambda_k}{\partial \theta_i} \frac{\partial \lambda_k}{\partial \theta_j} \qquad (13)$$

and the vector of right-hand sides

$$\delta_i = \sum_k \frac{1}{\lambda_k(\boldsymbol{\theta}) + r^2} \frac{\partial \lambda_k}{\partial \theta_i} \qquad (14)$$

the linear system to be solved in each iteration is then simply

$$\boldsymbol{A}^{(m)} \Delta\boldsymbol{\theta}^{(m)} = \boldsymbol{\delta}^{(m)} \qquad (15)$$

The iterations will normally converge quickly, if the initial estimate is reasonable. Calculating a reasonable initial estimate in a robust way may however be one of the trickiest part of the solution.

4 Covariance and goodness-of-fit of the ML solution

Having obtained the ML solution as described above, we also want to calculate an estimated covariance of the solution vector $\hat{\boldsymbol{\theta}}$ and some goodness-of-fit statistic of the solution. For the covariance we use the Cramér–Rao limit given by the inverse of the matrix $\boldsymbol{A}$ after convergence:

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \simeq \boldsymbol{A}^{-1} \tag{16}$$

A useful goodness-of-fit statistic, analogous to χ^2, can be obtained by means of a well-known theorem about the asymptotic distribution of the likelihood ratio statistic (Kendall & Stuart, 1979, Ch. 24.7). In the present context, the theorem can be formulated as follows.

For the given data, let $L(\boldsymbol{\lambda})$ be the likelihood function written in terms of the vector of intensity values, $\boldsymbol{\lambda} = \{\lambda_k\}$, with $K = \dim(\lambda)$ being the number of data points. Furthermore, let Λ be the K-dimensional space of all possible intensities – essentially the half-space of $\mathbb{R}^K$ for which all $\lambda_k \geq 0$ – and let Λ_0 be the subspace of Λ permitted by the intensity model $\boldsymbol{\lambda}(\boldsymbol{\theta})$. For a non-degenerate problem, the dimension (rank) of Λ_0 is $n = \dim(\boldsymbol{\theta})$. Obviously, the ML estimate $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\lambda})$ over Λ_0. If we remove the constraints imposed by the intensity model, and allow any $\boldsymbol{\lambda} \in \Lambda$ (this is sometimes called the full model, or the saturated model), we can in general increase the likelihood. Now consider the likelihood ratio statistic

$$t = \frac{\sup\limits_{\boldsymbol{\lambda}\in\Lambda_0} L(\boldsymbol{\lambda})}{\sup\limits_{\boldsymbol{\lambda}\in\Lambda} L(\boldsymbol{\lambda})} \tag{17}$$

which must clearly be the range $0 \leq t \leq 1$. According to the above-mentioned theorem, $D = -2\ln t$ has approximately the χ^2 distribution with $K - n$ degrees of freedom. The statistic D is sometimes called the *deviance* of the fit.

From (5) the log-likelihood of the ML estimate is

$$\sup_{\boldsymbol{\lambda}\in\Lambda_0} \ell(\boldsymbol{\lambda}) = \text{const} + \sum_k \left[(N_k + r^2) \ln \left(\lambda_k(\hat{\boldsymbol{\theta}}) + r^2 \right) - \lambda_k(\hat{\boldsymbol{\theta}}) \right] \tag{18}$$

The log-likelihood for the saturated model is maximized by putting $\lambda_k = N_k$ for every k. This gives the log-likelihood

$$\sup_{\boldsymbol{\lambda}\in\Lambda} \ell(\boldsymbol{\lambda}) = \text{const} + \sum_k \left[(N_k + r^2) \ln \left(N_k + r^2 \right) - N_k \right] \tag{19}$$

The deviance of the ML fit is -2 times the difference between (18) and (19), or

$$D = -2 \sum_k \left[(N_k + r^2) \ln \left(\frac{\lambda_k(\hat{\boldsymbol{\theta}}) + r^2}{N_k + r^2} \right) + \left(N_k - \lambda_k(\hat{\boldsymbol{\theta}}) \right) \right] \tag{20}$$

For a perfect fit ($N_k = \lambda_k$) we have $D = 0$; otherwise $D > 0$. As previously stated, D is expected to have approximately the χ^2 distribution with $K - n$ degrees of freedom.

It is readily seen that the ML estimation is equivalent to minimizing D. Therefore, D is a better measure of the goodness-of-fit of the ML estimate than the traditionally computed χ^2,

$$\chi^2 = \sum_k \frac{\left(N_k - \lambda_k(\hat{\boldsymbol{\theta}})\right)^2}{\lambda_k(\hat{\boldsymbol{\theta}}) + r^2} \tag{21}$$

In most cases D and χ^2 should not be too much different.

5 Implementation

The implementation of the algorithm described here is in principle straightforward. For maximum flexibility, it is important to have a simple and general interface between the model and the fitting procedure. Basically, the model only needs to know how to calculate λ_k and $\partial\lambda_k/\partial\boldsymbol{\theta}$ for any sample. However, in practice the fitting procedure needs to have access to a lot of auxiliary information, such as the geometry of the samples and the readout noise, and this requires additional interfaces that will be specific to each application. Also the provision of a robust initial estimate may be quite difficult and application dependent.

As an example, we have implemented and successfully tested a simple generic algorithm for 1D (LSF) and 2D (PSF) centroiding on CCD samples.

6 References

Kendall, M., Stuart, A., 1979, *The advanced theory of statistics, Volume 2, 4th Edition*, Griffin, London

Lindegren, L., 2000, *Centroiding on GAIA CCD sample data*,
`SAG-LL-032`,
URL `http://www.rssd.esa.int/llink/livelink/open/356853`

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007, *Numerical Recipes: The Art of Scientific Computing, 3rd Edition*, Cambridge University Press

Appendix: Noise model

A1. Photon noise (poissonian model)

The incoherent detection of light is a stochastic process in which the number of detected photons, K, to very good accuracy follows the discrete Poisson distribution

$$\Pr(K|\lambda) = \frac{\lambda^K}{K!}\,e^{-\lambda}\,, \quad K = 0,\ 1,\ 2,\ \ldots \tag{22}$$

where the so-called intensity parameter $\lambda \geq 0$ equals the expected (or mean) value of K:

$$\mathrm{E}\left[K\right] = \lambda \tag{23}$$

A well-known property of the Poisson distribution is that the variance of K equals the mean value[1]

$$\mathrm{Var}\left[K\right] = \mathrm{E}\left[(K-\lambda)^2\right] = \lambda \tag{24}$$

A2. Including the readout noise

The sample data value N resulting from the CCD readout process, after correction for bias and gain, is formally expressed in units of detected photons (or electrons), just like K, and contains a noise component corresponding to the Poisson process, i.e., with a variance that equals the mean value. However, N is not necessarily an integer value, and it contains additional noise from the readout process, amplification and AD conversion. The additional readout noise may be modelled as a continuous gaussian process with zero mean and a standard deviation r, although the actual noise has a more complex character due to the discretization etc. The readout noise, r, is independent of the intensity and is expressed on the same scale as the sample data, i.e., in electrons. Thus,

$$N = K + g \tag{25}$$

where $K \sim \mathrm{Poisson}(\lambda)$ and $g \sim \mathrm{N}(0, r^2)$ are independent stochastic processes with parameters λ and r. The mean value and variance of the compound process are given by[2]

$$\mathrm{E}(N) = \lambda\,, \quad \mathrm{Var}(N) = \lambda + r^2 \tag{26}$$

The probability density function (pdf) of N is the convolution of the Poisson distribution with the normal distribution, viz.:

$$p(N|\lambda, r) = \frac{a^{-\lambda}}{r\sqrt{2\pi}} \sum_{K=0}^{\infty} \frac{\lambda^K}{K!} \exp\left(-\frac{(N-K)^2}{2r^2}\right) \tag{27}$$

[1] This and higher moments of the Poisson distribution follow from the easily proven expression for the factorial moments, $\mathrm{E}\left[K(K-1)(K-2)\cdots(K-n+1)\right] = \lambda^n$, where $n = 1,\ 2,\ \ldots$.

[2] The additivity rule $\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y)$ holds strictly for any uncorrelated random variables x and y, whatever their distributions, provided that the variances exist.

Unfortunately this pdf is not well suited for ML estimation, which involves the product of $p(N|\lambda, r)$ for all the data samples, cf. (3). We therefore want to approximate it by a simpler distribution better suited for ML. The obvious choices are a gaussian or a modified poissonian distribution, in either case with a mean value and variance matching (26).

The gaussian model

Using the gaussian model for the compound process (25) we have

$$p(N|\lambda, r) = \frac{1}{\sqrt{2\pi(\lambda + r^2)}} \exp\left(-\frac{(N-\lambda)^2}{2(\lambda + r^2)}\right) \tag{28}$$

where the mean value and variance obviously agrees with (26). In order to compute the likelihood equations we need the partial derivative of $\ln p(N|\lambda, r)$ with respect to λ. For the gaussian model we find

$$\frac{\partial \ln p(N|\lambda, r)}{\partial \lambda} = \frac{N-\lambda}{\lambda + r^2} + \frac{(N-\lambda)^2 - (\lambda + r^2)}{(\lambda + r^2)^2} \tag{29}$$

The modified poissonian model

In this case we assume that $N + r^2$ follows the Poisson distribution with intensity parameter $\lambda + r^2$; thus

$$\mathrm{Pr}(N|\lambda, r) = \frac{(\lambda + r^2)^{N+r^2}}{(N + r^2)!}\, e^{-\lambda - r^2} \tag{30}$$

It is readily seen that also in this case the mean value and variance of N agrees with (26). In principle (30) is a discrete distribution where $N + r^2$ can only have non-negative integer values, i.e., $N = -r^2, -r^2 + 1, -r^2 + 2, \dots$. However, by analytical extension we may construct a continuous probability density function

$$p(N|\lambda, r) = \text{const} \times \frac{(\lambda + r^2)^{N+r^2}}{\Gamma(N + r^2 + 1)}\, e^{-\lambda - r^2} \tag{31}$$

valid for any real value $N \ge -r^2$. The multiplicative constant, whose function is to normalize the integral of the pdf, is assumed to be independent of λ. For the partial derivative we then find for the modified poissonian model

$$\frac{\partial \ln p(N|\lambda, r)}{\partial \lambda} = \frac{N-\lambda}{\lambda + r^2} \tag{32}$$

Thus, although the gaussian model appears to be the more straightforward approach to modelling the compound noise process, it turns out that the modified poissonian leads to considerably simpler likelihood equations; compare (29) and (32). The latter model is therefore adopted as the basis for the ML estimation.