ASM data handling by maximum-likelihood

GDAAS-FA-01
F. Arenou
Revision: 2.4

June 26, 2004

Abstract

A method for estimating the centroiding and flux with the Gaia Astrometric Sky Mapper (ASM) data or any other CCD window is developed. It relies on a maximum-likelihood approach similar to the one adopted in the Astrometric field but extended in several respects. It allows the use of a PSF instead of a LSF, the determination of background, the handling of numerically binned samples, of multiple stellar systems, or to incorporate prior knowledge of the parameters. Implementation details concerning robustness or alternative models are also added. The formulation is general enough to allow its use for CCD data of other GAIA instruments such as AF11, and thus provides a consistent method in terms of reusability of software code.

1 Introduction

The centroiding algorithm used in GDAAS for the analysis of the patches observed in the Astrometric field (AF) is a maximum-likelihood (ML) estimation developed in GAIA-LL-32[Lin00]. It is thus tempting to use the same method for the analysis of the Astrometric Sky Mappers (ASM) windows, in the 1-02-C Task of the GDAAS Phase II study [Lin03].

Several important differences between AF and ASM need however to be taken into account:

- ASM data need a PSF instead of a LSF. In this document, the way to obtain the PSF is not discussed. The PSF is simply assumed here to represent the *effective* PSF which includes the 2×2 pixel analog binning of the samples, calibrations, smearing, etc. Note that the binning makes this PSF undersampled, and we thus expect a systematic error as a function of the sub-pixel position.

- The sky background on the observed star is one of the parameters to determine as it is not given by the measurements in the rest of the AF, which combine the sky background from the two telescopes: only an upper limit can be known in each ASM. The computed "background" will in fact be the sum of the actual sky background, straylight, residues from bias subtraction, etc.

- A special ASM windowing design has been adopted for faint objects, denoted WA0T25 in [HAM+03]: at each corner of the patch, the sample transmitted on ground is in fact the sum of 4 read samples which have been numerically binned on-board. This binning should thus be taken into account in the estimation method.

2 The estimation method

We would like to estimate the location and flux of the object(s) the most precisely. The best precision which can theoretically be achieved by an estimate is given by the right-hand side of the Fréchet (or Rao-Cramér) inequality; if an estimate has this minimal precision, it is called Minimum variance bound (MVB). It can be shown that the ML method gives the MVB estimate, if it exists. This is also one of the motivations to use an approach based on ML or MAP (Sect. 2.5). The ML estimate is however not unbiased, except asymptotically.

The fundamental principles of the estimation have been presented in [Lin00], to which the reader could refer.

2.1 Notations

We mostly adopt the notations used in [Lin00] and expand them to reflect our case:

- the index i or k and coordinates x or ξ_0 refer to the along-scan (AL) position in sample units
- the index j or l and coordinates y or η_0 refer to the across-scan (AC) position in sample units
- (ξ_0, η_0) is the true location of the object on the window
- (i, j) is the location of a transmitted sample, with integer values, assuming (1,1) at the center of the lower left sample
- $C(x, y)$ is the approximation of the effective PSF, which is assumed to be centred at $(0, 0)$.
- $C_{ij} = C(i - \xi_0, j - \eta_0)$ is its value on sample (i, j), given the true location (ξ_0, η_0) of the object
- $C'^x_{ij} = \frac{\partial C}{\partial x}(i - \xi_0, j - \eta_0)$, $C'^y_{ij} = \frac{\partial C}{\partial y}(i - \xi_0, j - \eta_0)$, $C''^{xx}_{ij} = \frac{\partial^2 C}{\partial x^2}(i - \xi_0, j - \eta_0)$, $C''^{yy}_{ij} = \frac{\partial^2 C}{\partial y^2}(i - \xi_0, j - \eta_0)$ and $C''^{xy}_{ij} = \frac{\partial^2 C}{\partial x \partial y}(i - \xi_0, j - \eta_0) = C''^{yx}_{ij}$ are its first and second derivatives with respect to AL and AC coordinates
- n_{ij} is the observed number of counts in e^- in the sample (i, j)
- N is the total number of e^- of the object
- b is the background per sample in e^-
- R is the total noise, including read-out noise, quantization noise, etc. in e^-
- s_{ij} is the expectation of the number of counts in e^- in a normal sample (i, j), t_{ij} being its analogue in a numerically binned sample
- T_{ij} is the number of read and numerically-binned samples in the (i, j) sample ($T_{ij} = 4$ in the corners of the ASM patch)
- K index refers to the K-th component in a multiple stellar system
- $\mathbf{p} = (p_m)$ is the vector of parameters to be determined by ML
- $\mathbf{n} = (n_{ij})$ is the vector of observables

In the ASM, or more generally in any windowing which includes numerically binned samples, a modification of the basic estimation model is needed. We thus define the two sets of samples, those which are not binned, noted $\mathscr{A}$, and those which have been numerically binned on-board, noted $\mathscr{B}$. According to [HAM$^+$03]:

- For stars fainter than $G = 16$, the set $\mathscr{B}$ is $\{(1,1),(1,5),(5,1),(5,5)\}$ and $\mathscr{A} = \{1 \leq i \leq 5, 1 \leq j \leq 5\} - \mathscr{B}$.
- For stars $12 \leq G \leq 16$, the set $\mathscr{B} = \varnothing$ and $\mathscr{A} = \{1 \leq i \leq 8, 1 \leq j \leq 8\} -$ $\{(3,1),(4,1),(5,1),(6,1),(3,8),(4,8),(5,8),(6,8),$ $(1,3),(1,4),(1,5),(1,6),(8,3),(8,4),(8,5),(8,6)\}$

2.2 The single star model

The 4 parameters to be determined by ML are $\mathbf{p} = (p_m) = (\xi_0, \eta_0, \ln N, b)$, where we also follow [Lin00] in determining $\ln N$ instead of N in order to avoid negative values during iterations. We allow b to be negative, which may occur if the calibrations are incorrect – or simply because the error on b is much larger than b itself.

It is assumed that the offset R^2 has been added to the observed counts. This allows to use a Poisson law of expectation R^2 for the distribution of the read-out noise [1] so that the total counts per sample follows a Poisson process.

On a given sample (i,j), the expectation of these counts, $E[n_{ij}]$, will be either

$$s_{ij} = R^2 + b + e^{\ln N} C_{ij} \tag{1}$$

if $(i,j) \in \mathscr{A}$, or a sum for each binned sample in $\mathscr{B}$:

- $\sum_{k=0}^{1}\sum_{l=0}^{1}\left(R^2 + b + NC_{kl}\right)$ if $(i,j) = (1,1)$
- $\sum_{k=5}^{6}\sum_{l=0}^{1}\left(R^2 + b + NC_{kl}\right)$ if $(i,j) = (5,1)$
- $\sum_{k=0}^{1}\sum_{l=5}^{6}\left(R^2 + b + NC_{kl}\right)$ if $(i,j) = (1,5)$
- $\sum_{k=5}^{6}\sum_{l=5}^{6}\left(R^2 + b + NC_{kl}\right)$ if $(i,j) = (5,5)$

which, noting $D_{ij} = \sum_{k(i),l(j)} C_{kl}$, can be written more generally

$$t_{ij} = \sum_{k(i),l(j)} \left(R^2 + b + e^{\ln N} C_{kl}\right) = T_{ij}R^2 + T_{ij}b + e^{\ln N} D_{ij} \tag{2}$$

2.3 Double and Multiple stars

Although it would not be realistic to expect determining the parameters of more than two stars, it is indicated for completeness how to handle multiple systems.

The parameters for the K-th component in a multiple system are $(\xi_{0K}, \eta_{0K}, \ln N_K)$, and the vector of parameters to be estimated is thus $\mathbf{p} = (b, \xi_{01}, \eta_{01}, \ln N_1, \ldots, \xi_{0K}, \eta_{0K}, \ln N_K, \ldots)$. The index K will often not be indicated for legibility. In particular, N, C_{ij} and D_{ij} will refer to N_K, C_{ijK} and D_{ijK}. The expected number of counts in sample (i,j) is

$$\begin{aligned} s_{ij} &= R^2 + b + \sum_K e^{\ln N_K} C_{ijK} \\ t_{ij} &= T_{ij}R^2 + T_{ij}b + \sum_K e^{\ln N_K} D_{ijK} \end{aligned}$$

[1] Which is not the correct distribution. Moreover, the effect of the quantization (large gain, uniform distribution assumed) is not negligible: for instance the quantization error is larger than the Poisson noise on the sky background...; one purpose of this study is to check how this approximation and the others impact on the results.

when the sample is respectively a normal or a binned sample.

2.4 The likelihood

With $\mathbf{n} = (n_{ij})$ as observables and $\mathbf{p} = (p_m)$ as parameters to determine, the log-likelihood of our patch is

$$\ln \mathscr{L}(\mathbf{n}|\mathbf{p}) = \sum_{(i,j)\in A} [n_{ij} \ln s_{ij} - s_{ij} - \ln(n_{ij}!)] + \sum_{(i,j)\in B} [n_{ij} \ln t_{ij} - t_{ij} - \ln(n_{ij}!)] \tag{3}$$

We then compute the score vector, i.e. the vector whose components are the partial derivatives of the log-likelihood with respect to the parameters $\mathbf{p}$. The score should thus be 0 when the log-likelihood is maximum:

$$\frac{\partial \ln \mathscr{L}}{\partial \xi_{0K}} = N_K \sum_{(i,j)\in A} (1 - \frac{n_{ij}}{s_{ij}}) C'^{x}_{ijK} + N_K \sum_{(i,j)\in B} (1 - \frac{n_{ij}}{t_{ij}}) D'^{x}_{ijK} = 0 \tag{4}$$

$$\frac{\partial \ln \mathscr{L}}{\partial \eta_{0K}} = N_K \sum_{(i,j)\in A} (1 - \frac{n_{ij}}{s_{ij}}) C'^{y}_{ijK} + N_K \sum_{(i,j)\in B} (1 - \frac{n_{ij}}{t_{ij}}) D'^{y}_{ijK} = 0 \tag{5}$$

$$\frac{\partial \ln \mathscr{L}}{\partial (\ln N_K)} = -N_K \sum_{(i,j)\in A} (1 - \frac{n_{ij}}{s_{ij}}) C_{ijK} - N_K \sum_{(i,j)\in B} (1 - \frac{n_{ij}}{t_{ij}}) D_{ijK} = 0 \tag{6}$$

$$\frac{\partial \ln \mathscr{L}}{\partial b} = -\sum_{(i,j)\in A} (1 - \frac{n_{ij}}{s_{ij}}) - \sum_{(i,j)\in B} T_{ij} (1 - \frac{n_{ij}}{t_{ij}}) = 0 \tag{7}$$

The Hessian matrix, i.e. the matrix containing the second derivative of the log-likelihood with respect to the parameters

$$\mathbf{H} = \left(\frac{\partial^2 \ln \mathscr{L}}{\partial p_m \partial p_n} \right)$$

measures the curvature of the log-likelihood at the point where it is evaluated. The Fisher information matrix $\mathbf{F}$ is the expectation of $-\mathbf{H}$, with n_{ij} the random variables here. The usefulness of these matrices comes from the fact that $\mathbf{F}^{-1}$ is the variance-covariance matrix of the parameters to determine, and $\mathbf{H}$ will be used for the iterations (Sect. 3.1). $\mathbf{F}$ is usually approximated by $-\mathbf{H}$ evaluated at the value of the parameters which maximise the likelihood. Noting δ_{KL} the Kronecker symbol for stars K and L, the components of $-\mathbf{H}$ are:

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \xi_{0K} \partial \xi_{0L}} = \begin{array}{l} N_K N_L \sum_{(i,j)\in A} \left(\frac{n_{ij}}{s_{ij}^2} C'^{x}_{ijK} C'^{x}_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{s_{ij}}) C''^{xx}_{ij} \right) + \\ N_K N_L \sum_{(i,j)\in B} \left(\frac{n_{ij}}{t_{ij}^2} D'^{x}_{ijK} D'^{x}_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{t_{ij}}) D''^{xx}_{ij} \right) \end{array} \tag{8}$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \xi_{0K} \partial \eta_{0L}} = \begin{array}{l} N_K N_L \sum_{(i,j)\in A} \left(\frac{n_{ij}}{s_{ij}^2} C'^{x}_{ijK} C'^{y}_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{s_{ij}}) C''^{xy}_{ij} \right) + \\ N_K N_L \sum_{(i,j)\in B} \left(\frac{n_{ij}}{t_{ij}^2} D'^{x}_{ijK} D'^{y}_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{t_{ij}}) D''^{xy}_{ij} \right) \end{array} \tag{9}$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \xi_{0K} \partial (\ln N_L)} = -N_K N_L \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C_{ijL} C'^{x}_{ijK} - N_K N_L \sum_{(i,j)\in B} \frac{n_{ij}}{t_{ij}^2} D_{ijL} D'^{x}_{ijK} \tag{10}$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \xi_{0K} \partial b} = N_K \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C'^{x}_{ijK} + N_K \sum_{(i,j)\in B} T_{ij} \frac{n_{ij}}{t_{ij}^2} D'^{x}_{ijK} \tag{11}$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \eta_{0K} \partial \eta_{0L}} = N_K N_L \sum_{(i,j)\in A} \left(\frac{n_{ij}}{s_{ij}^2} C'^y_{ijK} C'^y_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{s_{ij}}) C''^{yy}_{ij} \right) + N_K N_L \sum_{(i,j)\in B} \left(\frac{n_{ij}}{t_{ij}^2} D'^y_{ijK} D'^y_{ijL} + \delta_{KL} \frac{1}{N} (1 - \frac{n_{ij}}{t_{ij}}) D''^{yy}_{ij} \right) \quad (12)$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \eta_{0K} \partial (\ln N_L)} = -N_K N_L \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C_{ijL} C'^y_{ijK} - N_K N_L \sum_{(i,j)\in B} \frac{n_{ij}}{t_{ij}^2} D_{ijL} D'^y_{ijK} \quad (13)$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \eta_{0K} \partial b} = N_K \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C'^y_{ijK} + N_K \sum_{(i,j)\in B} T_{ij} \frac{n_{ij}}{t_{ij}^2} D'^y_{ijK} \quad (14)$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial (\ln N_K) \partial (\ln N_L)} = N_K N_L \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C_{ijK} C_{ijL} + N_K N_L \sum_{(i,j)\in B} \frac{n_{ij}}{t_{ij}^2} D_{ijK} D_{ijL} \quad (15)$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial (\ln N_K) \partial b} = N_K \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} C_{ijK} + N_K \sum_{(i,j)\in B} T_{ij} \frac{n_{ij}}{t_{ij}^2} D_{ijK} \quad (16)$$

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial b^2} = \sum_{(i,j)\in A} \frac{n_{ij}}{s_{ij}^2} + \sum_{(i,j)\in B} T_{ij}^2 \frac{n_{ij}}{t_{ij}^2} \quad (17)$$

2.5 Prior information

Up to now, it was assumed that the parameters had to be determined by the observations $\mathbf{n}$ alone. In practice, some prior estimation of several parameters may be available. A typical example is the background determination, where an on-board determination may provide an estimation of background on a global scale by interpolation between values obtained in e.g. 32×32 pixels areas, while the window where this ML estimation is done gives a local scale background only, and these two values may be considered as uncorrelated in practice.

Moreover, some other prior information about the distribution of the parameters may be known. To incorporate this knowledge, bayesian inference can be applied. Noting $\mathbf{p_1}$ the set of parameters for which nothing is known beforehand, and $\mathbf{p_2}$ the set of parameters for which we had obtained an estimate $\hat{\mathbf{p}}_\mathbf{2}$, we have:

$$\mathscr{P}(\mathbf{p_1}, \mathbf{p_2} | \mathbf{n}, \hat{\mathbf{p}}_\mathbf{2}) \propto \mathscr{F}(\mathbf{n}, \hat{\mathbf{p}}_\mathbf{2} | \mathbf{p_1}, \mathbf{p_2}) \pi(\mathbf{p_1}, \mathbf{p_2}) \quad (18)$$

$$\propto \mathscr{L}(\mathbf{n} | \mathbf{p_1}, \mathbf{p_2}) \mathscr{G}(\hat{\mathbf{p}}_\mathbf{2} | \mathbf{p_2}) \pi(\mathbf{p_1}, \mathbf{p_2}) \quad (19)$$

where the likelihood $\mathscr{L}$ is given by Eq. 3, $\mathscr{G}$ is the error distribution of our estimation of $\mathbf{p_2}$ and the prior is π. Instead of maximising $\mathscr{L}$, a maximum a posteriori estimation (MAP) of $\mathscr{P}$ would now be used.

It is reasonnable to assume that $\hat{\mathbf{p}}_\mathbf{2}$ has been obtained by ML, or any other estimation asymptotically normally distributed [AKS99, 18.16]. We also assume a non-informative prior π for the parameters in what follows.

Consequently, the MAP estimate is given by $\mathrm{argmax}_{(\mathbf{p_1}, \mathbf{p_2})} \ln \mathscr{P}(\mathbf{p_1}, \mathbf{p_2} | \mathbf{n}, \hat{\mathbf{p}}_\mathbf{2})$ with

$$\ln \mathscr{P}(\mathbf{p_1}, \mathbf{p_2} | \mathbf{n}, \hat{\mathbf{p}}_\mathbf{2}) = \ln \mathscr{L}(\mathbf{n} | \mathbf{p_1}, \mathbf{p_2}) - \frac{1}{2} (\hat{\mathbf{p}}_\mathbf{2} - \mathbf{p_2})^T \mathbf{\Sigma}^{-1} (\hat{\mathbf{p}}_\mathbf{2} - \mathbf{p_2}) + \text{constant} \quad (20)$$

where $\mathbf{\Sigma}$ is the variance-covariance matrix of the $\hat{\mathbf{p}}_\mathbf{2}$ estimates. The first derivatives are

$$\frac{\partial \ln \mathscr{P}}{\partial \mathbf{p_1}} = \frac{\partial \ln \mathscr{L}}{\partial \mathbf{p_1}} \quad (21)$$

$$\frac{\partial \ln \mathscr{P}}{\partial \mathbf{p_2}} = \frac{\partial \ln \mathscr{L}}{\partial \mathbf{p_2}} + \mathbf{\Sigma}^{-1} (\hat{\mathbf{p}}_\mathbf{2} - \mathbf{p_2}) \quad (22)$$

The second derivatives with respect to $\mathbf{p_2}$ are $\frac{\partial^2 \ln \mathscr{P}}{\partial \mathbf{p_2}^2} = \frac{\partial^2 \ln \mathscr{L}}{\partial \mathbf{p_2}^2} + \mathrm{Diag}(\mathbf{\Sigma}^{-1})$ while the other second derivatives are the same as those of $\ln \mathscr{L}$.

In particular, when the background is the only parameter for which some estimate $\hat{b}$ with formal error $\sigma_{\hat{b}}$ is available, Eqs. 7 and 17 should simply be substituted by Eqs. 23 and 24 respectively:

$$\frac{\partial \ln \mathscr{P}}{\partial b} = -\sum_{(i,j)\in A}(1-\frac{n_{ij}}{s_{ij}}) - \sum_{(i,j)\in B} T_{ij}(1-\frac{n_{ij}}{t_{ij}}) + \frac{\hat{b}-b}{\sigma^2_{\hat{b}}} = 0 \tag{23}$$

$$-\frac{\partial^2 \ln \mathscr{P}}{\partial b^2} = \sum_{(i,j)\in A}\frac{n_{ij}}{s^2_{ij}} + \sum_{(i,j)\in B} T^2_{ij}\frac{n_{ij}}{t^2_{ij}} + \frac{1}{\sigma^2_{\hat{b}}} \tag{24}$$

The above equations allow for instance to estimate the parameters:

- with AF11 windows, using the prior estimation of the sum of background from the sky mappers of telescopes 1 and 2
- with ASM windows, using the on-board large scale background estimation
- for multiple systems in AF windows, when the parameters on secondary components have been estimated in another close or overlapping window.

Although a non-informative prior has been used, ASM data reduction could take into account that the background lies within a $[b^-, b^+]$ interval, where $b^- = 0$ and b^+ is limited by the sum of telescopes background as determined in AF11, would be natural choices. The impact on the confidence interval is not mentioned here.

3 Implementation

3.1 Iterations

Hopefully, the on-board detection algorithm, results from other data reduction algorithms, or results from previous iterations will provide a good enough starting value of the parameters. Then, the solution is found by iterations using a Newton-Raphson method, as in [Lin00]. If we note $\mathbf{p}^{(I)} = \left(p_m^{(I)}\right)$ the vector of parameters obtained at iteration I, the improved value at next iteration is computed with

$$\mathbf{p}^{(I+1)} = \mathbf{p}^{(I)} + \mathbf{F}^{-1}\mathbf{S}^{(I)} \tag{25}$$

with score vector $\mathbf{S}^{(I)}$ given by Eq. 4 to 7 and $\mathbf{F}$ the approximation of the Fisher matrix given by Eq. 8 to 17. This matrix should be symmetric and definite positive and will be inverted using a Cholesky decomposition.

However, the inversion turns out to be impossible in several cases, either due to outlying measurements or to a bad $p^{(0)}$ initial value. In this case, it is preferable to use a Whittaker-Robinson method (Newton with a fixed derivative), that is

$$\mathbf{p}^{(I+1)} = \mathbf{p}^{(I)} + \mathbf{S}^{(I)} \tag{26}$$

instead of Eq. 25.

The iterations are stopped when $||\mathbf{p}^{(I+1)} - \mathbf{p}^{(I)}||$ is small enough. The covariance matrix is then approximated by $\mathbf{F}^{-1}$ evaluated at the final value of the parameters; in particular the formal precision on a parameter of index m is $\approx \sqrt{F^{-1}_{mm}}$.

3.2 Variable number of parameters

Provisions should be taken in the implementation to have a variable number of parameters to estimate.

For instance, we may have better ξ_0 and $\ln N$ estimates from the AF, and the background can be negligible and fixed to some upper value coming from the sum of the contributions from the two telescopes. This would then give one parameter only to determine, η_0, although Sect. 2.5 allows to still determine the other parameters using their prior information. Limiting the number of parameters in a first step may however prove useful to ensure convergence.

Conversely we may have more parameters to compute if there are several stars in the ASM patch. This means that the various parameters should be indexed, determining only those which are indicated upon calling the method, and fixing the values of the others.

In terms of implementation, it is logical to consider each star k, $1 \leq k \leq K$ as an object with three parameters to determine, $(\xi_{0k}, \eta_{0k}, \ln N_k)$. By analogy, it could be convenient to consider the background as an object also, the object of index 0, with three parameters being respectively the gradient AL, the gradient AC and the value of background flux.

3.3 Robustness

Cosmic rays or solar protons impact are likely to be present in the ASM windows which represent an area of more than 100 pixels. Compared to the data in the rest of the astrometric field, ASM will mostly contain impacts of low energy particles as the AF1 confirmation will reject the others – except in the less frequent case where both ASM and AF1 receive particles of comparable energy at the same position.

The "random" character of these particles in terms of inclination and energy ranges make them hard to model, and it is thus preferable to consider affected samples as outlying measurements. It will be needed to build a table of good and bad samples, the former being used in the above analysis, and the latter excluded from the ML process after having been detected the following way.

The expected number of counts s_{ij} or t_{ij} will be computed on sample (i, j). The probability to observe n_{ij} or more counts is $1 - \sum_{n=0}^{n_{ij}-1} \frac{e^{-s_{ij}} s_{ij}^n}{n!}$. In view of the non-discrete value of n_{ij} after gain multiplication or bias and flatfield corrections, and the generally large values of n_{ij}, it is probably preferable to use a Gaussian approximation, and to flag the sample (i, j) if

$$\Delta_{ij} = \frac{n_{ij} - s_{ij}}{\sqrt{s_{ij}}} \text{ or } \Delta_{ij} = \frac{n_{ij} - t_{ij}}{\sqrt{t_{ij}}}$$

is larger than some threshold. A large threshold during the first iterations will probably be needed.

At each iteration, Δ_{ij} will be computed on each sample, and the flagged sample with the largest Δ_{ij} will be rejected. Outliers will be rejected one by one until no one remains and until convergence is ensured.

Under the same approximation that Δ_{ij} follow a standard normal distribution, then the sum of its squares will follow a χ^2 distribution,

$$\sum_{ij} \Delta_{ij}^2 \rightsquigarrow \chi^2(f)$$

where f is the degree of freedom, i.e. the difference between the number of used samples and the number of parameters to determine. This χ^2 will be given as output of the analysis. Alternatively, a Gaussian Goodness of Fit (GOF) can be computed.

3.4 Truncation

Censorship or truncation should be explicitly taken into account in the above model. There is one simple case, which is the pixel or sample saturation, with S the value of the maximum signal which can be handled. In this case, the expected number of counts is

$$s_{ij} = \mathrm{Inf}(R^2 + b + e^{\ln N} C_{ij}, S)$$

The non-linearity of the CCD is more complex to handle, as it would need to change the equations above (NC_{ij} should be substituted by another expression). Accounting for this effect would however be beyond the scope of the GDAAS II study.

3.5 Alternative models

During the first iteration of the Global Iterative Solution, it will not be known whether two objects detected on-board in one ASM window represent two components of a double star or one star and a low energy cosmic ray. This example illustrates that several alternative models can compete to describe the content of an ASM window, and a way to differentiate between models with different number of parameters should be provided.

One way to achieve this purpose is a likelihood-ratio test. Noting $\mathscr{L}_m$ the likelihood computed with m unknown parameters, does another model with n supplementary parameters provide a better fit? It can be shown that, asymptotically,

$$-2 \ln \frac{\mathscr{L}_m}{\mathscr{L}_{m+n}} \rightsquigarrow \chi^2(n)$$

The null hypothesis that $\mathscr{L}_m$ is sufficient to describe the observed data will thus be rejected if the left-hand side of the expression above is larger than the $\chi^2(n)$ corresponding to some probability level fixed in advance.

Acknowledgements

Michaël Bos, Fabien Chéreau, Didier Pelat and Noël Robichon are thanked for discussions and corrections on the adopted notations. The matrix inversion software is a port to Java of a Fortran code written by François Mignard... for the Hipparcos data reduction.

Versions

- Revision: 2.2, August 29, 2003: first version
- Revision: 2.4 , June 2004: added bayesian section

References

[AKS99] Stuart A., Ord J. K., and Arnold S. *Kendall's Advanced Theory of Statistics, classical inference & the Linear Model*, volume 2A. Oxford University Press Inc., New York, 1999.

[HAM+03] E. Høg, F. Arenou, S. Mignot, C. Babusiaux, D. Katz, and C. Jordi. Scientific requirements for the on-board processing. Technical Report GAIA-CUO-117, -, 2003. http://wwwhip.obspm.fr/gaia/obd/CUO-117.pdf.

[Lin00] L. Lindegren. Centroiding on GAIA CCD sample data. Technical Report SAG-LL-032, Lund Observatory, 04-Oct-2000.

[Lin03] L. Lindegren. Algorithms for gdaas phase ii – definition. Technical Report GAIA-LL-044, Lund Observatory, 2003.