Astron. Astrophys. 340, L35–L38 (1998)

ASTRONOMY AND ASTROPHYSICS

LETTER

Letter to the Editor

On the uncertainty of distances derived from parallax measurements

J. Kovalevsky

CERGA, Observatoire de la Côte d'Azur, Av. Copernic, F-06130 Grasse, France

Received 1 September 1998 / Accepted 18 October 1998

Abstract. The problem of determining the distances and their uncertainties from the measurements of parallaxes has no straight forward solution. In contrast to Bayesian-like approaches, a direct approach is possible assuming that the probability density function (pdf) in relative distance is cut for negative distances. This gives an asymmetrical pdf which is difficult to use and overestimates variances. It is proposed to fit it by a Gaussian pdf which minimizes the deviation from the actual pdf for the central interval, and the weighted deviation from most of the pdf. The characteristics of this Gaussian is given as functions of the parallax and its rms.

Key words: distances – parallaxes – stars – statistics

1. Introduction

A major contribution of the Hipparcos mission was to provide an unprecedented large number of independently determined accurate trigonometric parallaxes (ESA, 1997). For about 20 000 stars, the parallax is determined to better than 10% and for 30 000 others, to better than 20%. Various comparisons with otherwise estimated parallaxes of distant stars or clusters have shown no hint of regional or global systematic errors exceeding 0.1 mas. The possible effect of correlations in small fields, not exceeding a few percents within 2-3 degrees field as demonstrated by Lindegren (1989), makes no significant exception. However for small regions of the sky (like the case of clusters) may need some special consideration.

This large amount of data opens the use of star distances in large statistical investigations in stellar kinematics or astrophysical applications without calling for additional assumptions on the probability distribution function of the parameters to be determined. Indeed, in the absence of a sufficient number of independent determinations, and particularly in using photometric or spectroscopic parallaxes such a Bayesian approach has been widely used in the past. Let me mention the pioneering work by Lutz and Kelker (1973), who assumed that stars are uniformly distributed in space so that the number of stars within an interval of distances between r and $r + dr$ is

$$N(r)dr = 4\pi r^2 dr,$$

This leads to a law of distribution of the possible true values of the parallax ϖ_0 given the observed value ϖ

$$g\left(\varpi_0\right) = \frac{\sqrt{2\pi}}{\sigma\varpi_0^4}\exp\left(\frac{-\left(\varpi-\varpi_0\right)^2}{2\sigma^2}\right)d\varpi_0, \quad (1)$$

where σ is the standard deviation of the trigonometric parallax ϖ. It is not a probability density function (pdf), because the integral does not converge for infinity.

Other constraints on the pdf have been used. To take an example, Hanson (1979) uses an a priori assumption that the number of stars per unit magnitude interval around M inside unit volume is

$$n(M) = 10^{AM+B},$$

so that the distribution of absolute magnitudes for stars of a given apparent magnitude m is

$$n(M, m) = 4\pi N(M) r^2 dr.$$

This led to a distribution law where the exponent of ϖ differs from 4 as in Lutz and Kelker, and depends on the values of A and B retained.

This remark is just to empharise the fact that Bayesian approaches provide distance evaluations that depend upon the *a priori* assumptions that one makes on the distribution of some parameters derived from the distance.

In what follows, it is intended to provide a working pdf for the distance that can be used, with the restriction of the end of the first paragraph but without any other a priori assumption of a Bayesian type. In addition, it can be used simply to compute uncertainties of derived quantities, in particular in dynamical investigations in the Galaxy, in which proper motions and radial velocities with their Gaussian pdf are associated in the evaluation of the space velocities. This is also the case for absolute magnitudes, provided that the photometric uncertainty is also Gaussian.

2. Distance estimation and its pdf

The uncertainties of parallaxes determined by Hipparcos obey a Gaussian law of the form

$$f\left(\varpi\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(\varpi-\varpi_0\right)^2}{2\sigma^2}\right), \quad (2)$$

Now, if one transform (2) by changing the independent variable from ϖ to $r = 1/\varpi$, the law of distribution of r is proportional to

$$f'(r) = \frac{1}{\sqrt{2\pi}\sigma r^2}\exp\left(\frac{-(r-r_0)^2}{2\sigma^2 r^2 r_0^2}\right), \quad (3)$$

where σ is, as in (2), the standard deviation of the observed parallaxes and $r_0 = 1/\varpi_0$. For lack of denomination, we shall refer to it as *classical*. The problem is that although $f'(r)$ is a pdf, the calculation of the second moments does not converge, a problem dealt in particular by Smith and Eichhorn (1996). This is even more true for the transform of Lutz and Kelker (LK) representation given by formula (1) and which can be inverted directly since no a priori knowledge is assumed, and becomes with the same notations:

$$g'(r) = \frac{\sqrt{2\pi}}{\sigma} r^2 \exp\left(\frac{-(r-r_0)^2}{2\sigma^2 r^2 r_0^2}\right). \quad (4)$$

To analyse these expressions, we introduce a new variable

$$x = \frac{r - r_0}{r_0},$$

which represents the relative deviation from the true (hence unknown) distance defined as $r_0 = 1/\varpi_0$, and a parameter

$$\alpha = \sigma/\varpi_0 = \sigma r_0,$$

in which the uncertainty of the observed parallax is introduced. We obtain respectively for $f'(r)$ and $g'(r)$

$$F(x) = \frac{F_0}{(1+x)^2}\exp\left(\frac{-x^2}{2\alpha^2(1+x)^2}\right), \quad (5)$$

$$G(x) = G_0(1+x)^2\exp\left(\frac{-x^2}{2\alpha^2(1+x)^2}\right), \quad (6)$$

where one may chose F_0 and G_0 in such a way that the maximums of the functions are scaled to 1.

The maximums are obtained by differentiating (5) or (6) with respect to x and writing that the derivative is equal to zero. One obtains, for the classical $F(x)$, the relation

$$x^2 + x\left(2 + \frac{1}{2\alpha^2}\right) + 1 = 0.$$

The maximum corresponds to the larger solution of this equation:

$$x = \left(-2 - \frac{1}{2\alpha^2} + \sqrt{\left(2 + \frac{1}{2\alpha^2}\right)^2 - 4}\right)/2. \quad (7)$$

In the case of the LK distribution, one has

$$x = \left(-2 + \frac{1}{2\alpha^2} - \sqrt{\left(2 - \frac{1}{2\alpha}\right)^2 - 4}\right)/2. \quad (8)$$

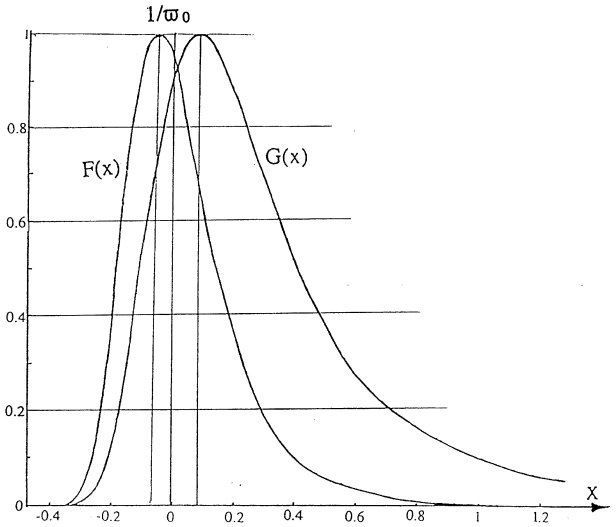

Fig. 1. Comparison of the scaled functions $F(x)$ and $G(x)$ for $\alpha = 0.2$. Note the position of the maximums in comparison with $x = 0$, corresponding to $r = r_0$

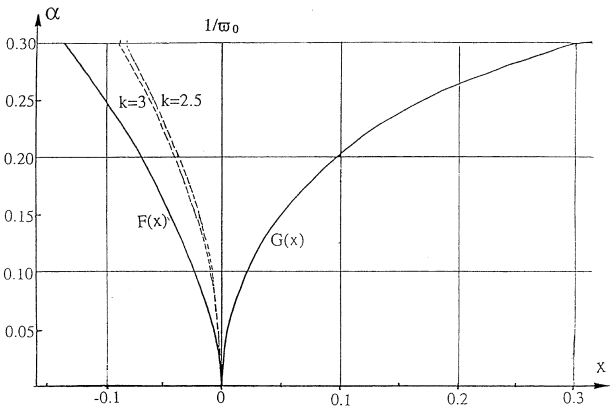

Fig. 2. Values of the maximums of $F(x)$ and $G(x)$ as functions of α and of the Gaussian representations with k =2.5 and 3, in relation with α

Replacing x in (5) and (6) respectively by (7) and (8), and imposing the value 1, one gets F_0 and G_0. The scaled (in this way) functions, $F(x)$ and $G(x)$ are represented in Fig. 1. One sees that while in the LK solution, the maximum corresponds to a distance larger than r_0, the $F(x)$ distribution gives a maximum smaller than r_0 (see Fig. 1). Within the assumptions behind each of these distributions, the maximum corresponds to the most probable value. It is tempting to use it as an estimate of the distance. This has been frequently done for the LK distribution and I believe this is dangerous. If α is small, both converge to a unique value, equal to zero (distance $r_0 = 1/\varpi_0$). But for large values of α, there are increasing differences which depend upon the underlying assumptions.

3. A probability distribution function for distances

As such, none of the functions $F(x)$ and $G(x)$ satisfy the conditions required to express uncertainties in distances. The LK

Table 1. Characteristics of the probability distribution function $F^*(x)$

α	$x_{\max}$	rms	Values of x for					
			$-3\,\sigma$	$-2\,\sigma$	$-\sigma$	$+\,\sigma$	$+\,2\,\sigma$	$+\,3\,\sigma$
0.010	−0.0002	0.0100	−0.0292	− 0.0196	−0.0099	0.0101	0.0204	0.0310
0.030	−0.0018	0.0302	−0.0829	−0.0566	−0.0291	0.0309	0.0638	0/0993
0.050	−0.0050	0.0511	−0.1309	−0.0909	−0.0476	0.0526	0.1111	0.1773
0.075	−0.0110	0.0786	−0.1842	−0.1304	−0.0698	0.0811	0.1764	0.2918
0.100	−0.0192	0.1084	−0.2314	−0.1666	−0.0909	0.1111	0.2499	0.4309
0.125	−0.0294	0.1414	−0.2735	−0.1999	−0.1111	0.1428	0.3331	0.6038
0.150	−0.0414	0.1786	−0.3112	−0.2307	−0.1304	0.1764	0.4283	0.8242
0.200	−0.0693	0.2734	−0.3759	−0.2856	−0.1666	0.2500	0.6663	1.5192
0.250	−0.1010	0.4557	−0.4295	−0.3332	−0.2000	0.3334	1.0011	3.1584
0.300	− 0.1348	0.8079	−0.4746	−0.3749	−0.2307	0.4299	1.5210	28.1909

distribution $G(x)$ is not a pdf. In the case of $F(x)$, one may note that it has a second maximum for a negative value of r (cf. formula 7 with a negative sign before the square root), which is physically a nonsense. Let us remark however that $F(-1)$, corresponding to $r = 0$ is null as well as its first four derivatives. Since negative values of r are in any case prohibited, one may consider a new function $F^\star(x)$ such that

$$F^*(x) = F(x) \text{ for } x \geq -1\,,$$
$$F^*(x) = 0 \text{ for } x < -1\,.$$

The integral

$$I = \int_{-\infty}^{+\infty} F^*(x)dx$$

is finite and can be computed as a function of α. The new function F_1 defined by

$$F_1(x) = F^\star(x)/I,$$

is a probability density function for which one may compute the second moment, and define a variance and a rms. The values found are given in Table 1.The values of x for different limits of probability are also given. This is not a fully satisfactory solution because of its numerical character. These parameters of the pdf may be computed by a simple software. Its advantage over Smith and Eichhorn (1996) solutions is that it needs no adjustable parameter.

4. A Gaussian approximate pdf

The asymmetric shape of the pdf is a drawback when it is necessary to combine uncertainties with other parameters associated with Gaussian pdf. In addition, a consequence is that the mode does not correspond to the expected value of x. The following is an attempt to represent the pdf F_1 by a Gaussian distribution and analyse additional errors in the evaluation of the uncertainty.

Let us consider the Gaussian pdf

$$H(x) = \frac{1}{\sqrt{2\pi}s}\exp\left[\frac{-(x-\mu)^2}{2s^2}\right], \quad (9)$$

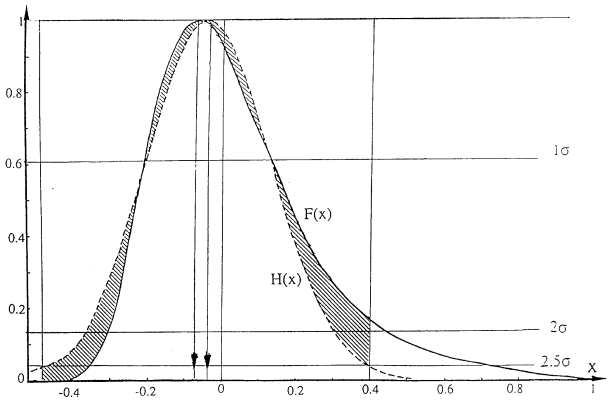

Fig. 3. Example of the fit of $F(x)$ by the Gaussian $H(x)$ for $k = 2.5$ and $\alpha = 0.2$. The shaded area corresponds to the part of the plane that enters in the computation of Q

with a maximum at $x = \mu$ and an rms denoted s. The problem is to determine s and μ in such a way that $H(x)$ is as close as possible to $F^*(x)$. One must chose two conditions to determine them. The suggestion is to equalize to zero the integral

$$J = \int_A^{\mathrm{B}} (F_1(x) - H(x))\,\mathrm{W}(x)\,dx, \quad (10)$$

where $W(x)$ is some weight function. The first condition is to force $H(x)$ to be very close to $F^*(x)$ in the interval $(\mu-s, \mu+s)$ within which, the probability to get the true value is 0.6826. It writes:

$$\int_{\mu-s}^{\mu+s} (F_1(x) - H(x))\,dx = 0. \quad (11)$$

In this case, the weight function $W(x)$ is taken equal to 1. The second condition is chosen in such a way that it encompasses the major significant part of $F^*(x)$ and gives more weight to the most probable values of x. One may take as the weight function either $F^*(x)$ or $H(x)$. Actually the difference between the two possibilities is small, so that I retained the Gaussian. Two sets of limits of integration were tested:

$$A_2 = \mu - ks \;\; ; \;\; B_2 = \mu + ks,$$

Table 2. Characteristics of the best Gaussian fit of $F_1(x)$

	k = 3			k = 2.5		
α	μ	s	Q	μ	s	Q
0.010	−0.0001	0.0100	0.0102	−0.0001	0.0100	0.0073
0.030	−0.0011	0.0299	0.0288	−0.0009	0.0299	0.0230
0.050	−0.0028	0.0494	0.0488	−0.0024	0.0495	0.0370
0.075	−0.0062	0.0732	0.0755	−0.0055	0.0732	0.0521
0.100	−0.0100	0.0959	0.0953	−0.0099	0.0958	0.0702
0.125	−0.0172	0.1170	0.1197	−0.0154	0.1170	0.0824
0.150	−0.0243	0.1365	0.1348	−0.0220	0.1366	0.1002
0.200	−0.0420	0.1704	0.1765	−0.0390	0.1704	0.1230
0.250	−0.0632	0.1971	0.1992	−0.0598	0.1972	0.1443
0.300	−0.0876	0.2172	0.2219	−0.0836	0.2172	0.1571

with k = 2.5 and 3 (respective probabilities of 0.9864 and 0.9974). The condition has the form

$$\int_{\mu-ks}^{\mu+ks} (F_1(x) - H(x))\, H(x)\, dx = 0. \quad (12)$$

Solving simultaneously (11) and (12) gives a univocal solution for μ and s. In addition, a quality factor $\mathcal{Q}$ is defined so as to describe the surface of the part that is not common to the pdf, weighted by $H(x)$

$$\mathcal{Q} = \int_{\mu-ks}^{\mu+ks} |\mathrm{F}_1(x) - H(x)|\, H(x)\, dx.$$

This surface is illustrated in Fig. 3 in a particular case. The results obtained are summarized in Table 2. The values of μ as a function of α are also shown in Fig. 2 and can be compared with the value of the maximums of $F(x)$ (or $F_1(x)$). Rather than interpolating Table 2, and for the convenience of users, the following expressions were computed for μ:

$$\mu(k = 2.5) = -0.036\alpha - 0.87\alpha^2, \quad (13)$$

$$\mu(k = 3) = 0.019\alpha - 1.34\alpha^2 + 1.018\alpha^3. \quad (14)$$

One may note that s is not sensitive to k. It can be expressed as

$$s = 1.056\alpha - 0.80\alpha^2 - 1.00\alpha^3, \quad (15)$$

and since $r = r_0(1 + x)$, one has

$$\sigma_\rho = r_0 s.$$

5. Conclusions

Among the two methods presented in this paper to represent the pdf of the distance determined from the parallax $F_1(x)$ has two disavantages.

- The distribution is strongly asymmetrical and can be used only numerically in later computations.
- In deriving the rms from the variances, the unprobable large values of x play a major role, increasing the rms in a way that can be considered as unrealistic.

The Gaussian representation $H(x)$ of $F_1(x)$ has the advantage of being usable as such in further evaluations of uncertainties of quantities depending among others on the observed parallaxes. One may remark from Table 2 that the quality of fit worsens rather quickly when k increases. For this reason, our suggestion is to use the solution with k=2.5 which encompasses 98.64% of the probability distributions and formulae (13) and (15). One may remark also that with this Gaussian pdf, the position of the maximum correspond closely to the mean between the maximum given by $F(x)$ and $x = 0(r = 1/\varpi_0)$. It is however to be noted that its derivation used the true rather than the observed value of the parallax. This may introduce an additional uncertainty, particularly for large α.

Finally, the tables presented show that the solutions degrade for large values of $\alpha(\alpha > 0.2)$. It is not reasonnable to use any pdf for $\alpha > 0.3$. Then, the determination of the distance from parallaxes looses its physical significance, and photometric determinations are definitely better. In any case, cutting a sample at some value of α may introduce a bias, so that users are invited to ascertain that the sample satisfies the conditions of applicability of the method.

Acknowledgements. I thank Drs F. Arenou and F. Mignard for enlightening discussions, and Dr X. Luri for constructive comments..

References

ESA, 1997, in *The Hipparcos Tycho Catalogues*, ESA Publ. SP-1200, June 1997, Noordwijk

Hanson, R.B., 1979, Monthly Not. R.A.S., 186, 875-896

Lindegren, L., 1989, in *The Hipparcos Mission: pre-launch status*, ESA Publ. SP-1111, Noodwijk, 3, 311–323

Lutz, T.E. and Kelker, D.H., 1973, Publ. Astron. Soc. of the Pacific, 85, 573–578

Smith, H. Jr. and Eichhorn, H., 1996, Mon. Not. R. Astron. Soc., 281, 211–218