Bayesian approach for matching multiple stellar observations

Gyöngyi Kerekes\textsuperscript{1}, Tamás Budavári\textsuperscript{2,1}, István Csabai\textsuperscript{1}

\textsuperscript{1}Eötvös Loránd University, Department of Physics of Complex Systems, Pázmány P. sétány 1/A, Budapest, 1117, Hungary
\textsuperscript{2}Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA

E-mail: kerekes@complex.elte.hu, budavari@jhu.edu

Abstract. The cross-identification of sources in separate catalogs is one of the most basic tasks in observational astronomy. Recently Budavári & Szalay (2008) formulated the problem in the probability theory, and laid down the statistical foundations of an extendable methodology. An application of the same Bayesian approach to stars is presented that, we know, can measurably move on the sky, and it is shown how to associate their observations. Models are studied on a sample of stars in the Sloan Digital Sky Survey, which allow for an unknown proper motion per object, and their improvements are shown over the simpler static model. The new models and conclusions are directly applicable to the upcoming surveys, whose data sets will be most likely dominated by stars in the region of the Galactic Plane.

1. Introduction

Today the most important astronomical studies use catalog merging as a basic step, so that measurements at different wavelengths by potentially separate instruments and telescopes can be combined. Hence, the cross-identification of sources is becoming an outstanding issue. The case of stars is especially a hard problem. The ones that move between observations are difficult to merge into multicolor objects even within a single survey. Today this poses a significant challenge for the next-generation photometric surveys, such as PanSTARRS and LSST, that plan to cover the Galactic Plane.

In the recent work of (Budavári & Szalay, 2008) a general probabilistic formalism was introduced that is extendable to arbitrarily complex models. The Bayesian hypothesis testing clearly separates the contributions of different types of measurements, e.g., the position on the sky or the colors of the sources, yet, naturally combines them into a coherent method. In this paper, we go beyond the simple case of stationary objects, and study the cross-identification of point sources that move on the sky. Most importantly we focus on stars that, we know, can be significantly offset between the epochs of observations. Although we only have loose constraints on their proper motions in general, this prior knowledge is enough to revise our static models, and work out the Bayesian evidence of the matches.

2. The Bayes Factor

The positional accuracy is characterized by a probability density function (hereafter PDF) on the celestial sphere. In a given model \( M \), this \( p(x|r,M) \) function tells us where to expect \( x \).
detections of an object that is at its true location \( \mathbf{r} \). We use 3-dimensional unit vectors for the positions on the sky, e.g., the aforementioned \( \mathbf{x} \) and \( \mathbf{r} \) quantities.

For hypothesis testing Bayes-factor is used. Let \( H \) denote the hypothesis that assumes that all measured positions, \( D \) are observations of the same object, and \( K \) denote its complement, i.e., any one or more of the detections might belong to a separate object. By definition, the Bayes factor is the ratio of the likelihoods of two hypotheses we wish to compare

\[
B(H,K|D) = \frac{p(D|H)}{p(D|K)}
\]

that are calculated as integrals of their parameter spaces. If we assume that there is a single object behind the observations, we can integrate over its unknown proper motion and position to calculate

\[
p(D|H) = \int d\mathbf{r} \int d\mu p(\mathbf{r}, \mu|H) \prod_{i=1}^{n} p_{i}(\mathbf{x}_{i}|\Delta t_{i}, \mathbf{r}, \mu, H) \tag{2}
\]

where the joint likelihood of \( H \) given the data is written as the product of the independent components and \( p(\mathbf{r}, \mu) \) is the prior on the parameters, which is the subject of the following section. In the complement hypothesis the integral separates into the product of

\[
p(D|K) = \prod_{i=1}^{n} \int d\mathbf{r}_{i} \int d\mu_{i} p(\mathbf{r}_{i}, \mu_{i}|K) p_{i}(\mathbf{x}_{i}|\Delta t_{i}, \mathbf{r}_{i}, \mu_{i}, K) \tag{3}
\]

3. Prior for stars

Based on the basic properties of conditional densities, we can write the prior as the product

\[
p(\mathbf{r}, \mu|H) = p(\mathbf{r}|H)p(\mu|\mathbf{r}, H) \tag{4}
\]

where the first term is the prior on the position, and the second term describes the possible proper motions as a function of location and optionally other properties. To derive the \( p(\mu|\mathbf{r}, H) \) prior, we choose to study the ensemble statistics of stars instead of approaching the problem with an analytic model, as the formulas are difficult to derive and the analytic approximations might miss subtle details of the relation.

We study the properties of stars in the Sloan Digital Sky Survey catalog archive that also contains accurate proper motion measurements from the recalibrated United States Naval Observatory (USNO) B1.0 Catalog (Munn et al., 2004). For the analysis, we pick stars from the Stripe 82 dataset where multiple observations are available in 300 square degrees of a narrow declination range between \( \pm 1.26^\circ \) (Adelman-McCarthy et al., 2008). After rejecting saturated and faint sources, the number of stars is around 100,000. This size does not allow for a high-resolution determination of the prior, hence we analyze additional simulation data.

The current state-of-the-art Besançon models are used (Robin et al., 2003) that match the SDSS distributions well. Assuming four different stellar populations in the Milky Way, using Poisson equation and collisionless Boltzmann equation with some observed parameters, (and fitting parameters to the dynamical rotation curve) they compute the number of stars of a given age, type, effective temperature and absolute magnitude, at any place in the Galaxy. 700,000 stars were generated with Besançon models using large-field equatorial coordinates.

We separate the dependence of the prior on the different proper motion components. Since the Stripe 82 dataset in SDSS contains sources only in a narrow declination range, \( -1.26^\circ < \delta < +1.26^\circ \), we can safely neglect the dependence on declination in Equation (4) for the purpose of this study, thus

\[
p(\mu_{\alpha}, \mu_{\delta}|\mathbf{r}) \approx p(\mu_{\delta}|\alpha)p(\mu_{\alpha}|\mu_{\delta}, \alpha) \tag{5}
\]
Figure 1. Illustration of the first term of prior (5) on the footprint of SDSS Stripe 82. As one moves into the direction of galactic plane (upwards and downwards on the figure), the probability distribution gets narrower since the velocity dispersion of stars decreases.

To achieve a better signal-to-noise behaviour across the entire parameter space, we do not use a uniform grid but varying bin sizes that follow the asinh() in both parameters. This way one can have higher resolution bins where more data are available and wider bins in the tail. With this transformation $p(\mu_6 | \alpha)$ becomes an asymmetric shape roughly centered on $-0.8\mas/\mathrm{yr}$. Figure 1 shows the detailed shape of the prior. The second term of Equation (5) is constructed similarly but in even higher dimensions. It is difficult to visualize a 4-dimensional PDF, hence, in Figure 2, we plot slices of the prior at various $\alpha$ values. To evaluate the density for a given position and proper motion, we use linear interpolation on these grids.

4. Results

As the Stripe 82 was observed repeatedly from 1998 to 2005, we can obtain multiepoch observations to test our method. We choose randomly a dozen of stars with different proper motions observed at different epochs. For each star we get on the average 20 epochs, from which many were observed with small time shifts. The longest time intervals between the epochs is on the average 6.5 years, we divide it to 3 roughly equal parts. Thus we get 4 observations of each star with much the same time intervals between them. For the test stars the ObjID marked in Table 1 is the ObjID of PhotoPrimary table in SDSS DR6.

In order to calculate the integrals in (2) and (3), we use Monte-Carlo method. The program generates uniformly distributed positions (normal vectors on a sphere) and velocity vectors in a reasonable range. In theory one should integrate for the whole celestial sphere and to infinity in proper motion, hence one can bound all the relevant parameters easily to reduce the computational need. The calculated Bayes-factors are plotted on Figure 3. Applying the means of probability theory, we also calculate the conditional expectations of proper motions for the test stars in case of 4 epochs. On the Figure 4 one can see that using only the observed positions and the information from priors the calculated and the reference values are very close to each other (inside the margin of error).
Figure 2. Illustration of the second term of prior (5). The prior is a 4 dimensional PDF, cuts are shown at different RAs. The same effect can be detected as on the Figure 1 - in the cases of $RA = 57.5^\circ$ and $RA = 307.5^\circ$ the distribution is a bit sharper than in the other two cases.

Table 1. Weight of Evidence for Models with no PM and with PM of test stars in case of 2, 3 and observations as a function of USNO proper motions

| ObjID                  | $\mu$ [mas/yr] | 2 observations static | 2 observations motion | 3 observations static | 3 observations motion | 4 observations static | 4 observations motion |
|------------------------|----------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 587731186192679018     | 2              | 12.60                  | 12.58                  | 25.32                  | 25.31                  | 38.04                  | 37.97                  |
| 587730847423987981     | 8              | 12.60                  | 12.59                  | 25.26                  | 25.22                  | 37.91                  | 37.86                  |
| 587730847429427304     | 13             | 12.59                  | 12.59                  | 25.25                  | 25.21                  | 38.11                  | 38.08                  |
| 587731173305876571     | 18.6           | 12.59                  | 12.59                  | 25.15                  | 25.12                  | 37.99                  | 37.98                  |
| 587731186187763779     | 19             | 12.55                  | 12.56                  | 25.20                  | 25.23                  | 37.97                  | 37.88                  |
| 587731173305614418     | 40             | 12.59                  | 12.59                  | 25.21                  | 25.22                  | 37.47                  | 37.64                  |
| 58801550926875010      | 98             | 11.93                  | 11.90                  | 22.51                  | 22.86                  | 31.10                  | 34.56                  |
| 588015509288813878     | 143            | 11.64                  | 11.70                  | 19.62                  | 22.20                  | 25.92                  | 34.41                  |
| 588015509268201645     | 163            | 11.44                  | 11.56                  | 18.09                  | 21.84                  | 21.09                  | 34.22                  |
| 588015509271805995     | 196            | 11.32                  | 11.38                  | 10.67                  | 21.48                  | 6.46                   | 33.88                  |
| 58801550927342731      | 255            | 10.02                  | 10.08                  | 7.54                   | 20.04                  | -5.76                  | 32.60                  |
| 588015509273378938     | 257            | 9.41                   | 9.82                   | 5.44                   | 20.44                  | -5.48                  | 32.49                  |
| 587730847426740272     | 300            | 9.43                   | 9.63                   | 10.23                  | 20.04                  | -1.85                  | 31.60                  |
| 588015509283930154     | 555            | 0.40                   | 4.12                   | -24.19                 | 4.14                   | -129.09                | -18.95                 |
Figure 3. Illustration of the weight of evidence as a function of proper motion in case of 2, 3 and 4 observations. Proper motion is shown on a logarithmic scale. Open circles represent the static model demonstrated in (Budavári & Szalay, 2008). The crosses show the values calculated with the empirical prior introduced in this study. For reference, the $W = 0$ horizontal line shows the theoretical limit above which the observations support the hypothesis of the match. The error shown on the figure comes from the uncertainty in position measurements.

Figure 4. Proper motions calculated for the test stars using the formula of conditional expected value: $E(\mu) = \int dr \int d\mu \, p(r, \mu|H) \prod_{i=1}^{n} p_i(x_i|\Delta t_i, r, \mu, H)$.
5. Conclusions
In the models presented here we intended to show that the application of more realistic prior distributions within the Bayesian framework can improve cross-identification of moving astronomical objects. In accord with our expectations, we found that using an empirical prior of the proper motion would assign larger observational evidence to the match. The dependence of the quality of these cross-identifications was studied as a function of separation in time (and space) as well as using multi-epoch observations. The SDSS Stripe 82 sample provided a good test set with 2–4 detections at different times with a few years in between. We found that, even though the 2-epoch data sets benefit from the proper motion model, and the 3-4-epoch observations essentially recover the right associations even for fast-moving stars.

An even more complicated model can include magnitude or color dependence which should improve weight of evidence. Prior could and should be different for every source based on the brightness and color: \( p(\mu_\alpha, \mu_\delta | \alpha, C) = p(\mu_\delta | \alpha) p(\mu_\alpha | \mu_\delta, C) \) The method can be applied in the following way: first finding objects (galaxies, quasars, etc.) using the static model and for remainders (supposed to be stars) the above-mentioned iterations can be run.

Acknowledgements
The authors would like to acknowledge the use of the online tools of the Besançon collaboration to obtain simulated stars for this study and thank Rosemary Wyse for her invaluable insights and help with stellar model of the Galaxy. G.K. and I.C. acknowledge support from NKTH: Polányi. T.B. acknowledges support from the Gordon and Betty Moore Foundation via GBMF 554.

References
Adelman-McCarthy J K et al. 2008 ApJS 175 297-313
Budavári T, Malik T, Szalay A S, Thakar A R and Gray J 2003 ASP Conf. Ser. Vol. 295 Astronomical Data Analysis Software and Systems XII, eds. H E Payne, R I Jedrzejewski and R N Hook (San Francisco: ASP), p.31
Budavári T et al. 2004 ASP Conf. Ser. Vol. 314 Astronomical Data Analysis Software and Systems XIII, eds. F Ochsenbein, M Allen and D Egret (San Francisco: ASP) p.177
Budavári T and Szalay A S 2008 ApJ 679 301
Fisher R, 1953, Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences Vol. 217 No. 1130., pp.295–305
Munn et al. 2004 AJ 127 3034
Robin A C, Reyl C, Derriere S and Piccaud S 2003 A&A 409 523
York D G et al. 2000 AJ 120 1579