Efficient calculation of covariances for astrometric data in the Gaia Catalogue

Berry Holl, Lennart Lindegren, and David Hobbs

Abstract For users of the Gaia astrometric catalogue it will be essential to have access to the covariance between any pair of astrometric parameters when computing quantities that combine multiple catalogue parameters. The computation and storage of the full covariance matrix for the expected 5×10^9 astrometric parameters ($\sim 10^8$ TeraByte) is however expected to be infeasible considering near-future storage and floating-point capabilities. In this paper we explain (without going into the mathematical details) how it might be practically feasible to estimate the covariance between any pair of source parameters in a computationally efficient way, from a reduced amount of data instead ($\sim$2 TeraByte). We also include two examples, explaining how to practically compute the covariance for the average parallax of a star cluster, and the acceleration of the solar system barycentre in a cosmological frame.

1 Introduction

The forthcoming ESA space astrometry mission Gaia, will provide the most comprehensive and accurate catalogue of astrometric data for galactic and astrophysical research in the coming decades. For roughly 1 billion stars, quasars and other point like objects (hereafter called 'sources') the five astrometric parameters (position, parallax and proper motion) will be determined. These parameters will not be perfect: every derived parameter has an error, ultimately resulting from the combination of a very large number of microscopic stochastic processes. The actual errors in the Gaia catalogue are of course unknown, but can nevertheless be statistically characterized, and in two forthcoming papers (1; 2) we derive and study the error properties based on a simplified least-squares formulation of the astrometric solution.

For most applications it is sufficient to consider the first and second moments of the errors, i.e., the expected values (biases), variances (or standard errors), and covariances (or correlation coefficients). We assume that the biases are negligible,

Berry Holl, Lennart Lindegren, and David Hobbs
Lund Observatory, Lund University, Box 43, SE-22100 Lund, e-mail: berry@astro.lu.se, lennart@astro.lu.se, david@astro.lu.se

and therefore concentrate on the second moments, which are most generally described by the covariance matrix. For an end user of the catalogue, knowledge of the covariances is needed when estimating the uncertainty of quantities that combine more than one astrometric parameter (see Sect. 4 for some examples). Therefore, tools need to be developed to allow the efficient computation of the variance of any scalar quantity y calculated from the N astrometric parameters in the catalogue $\mathbf{x} = (x_1, \ldots x_N)$. We can generally formulate this as $y = f(\mathbf{x})$. Assuming that f is linear for small errors, the variance of y is given by

$$\mathrm{Cov}(y) = \sigma_y^2 = \left(\frac{\partial y}{\partial \mathbf{x}}\right)' \mathbf{C} \left(\frac{\partial y}{\partial \mathbf{x}}\right) = \sum_i \sum_j \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} C_{ij}, \tag{1}$$

with $\mathbf{C} = [C_{ij}] = \mathrm{Cov}(\mathbf{x})$, and C_{ij} the covariance between astrometric parameters x_i and x_j. More generally, we may want to characterize the errors of M different scalar quantities calculated from the astrometric parameters, i.e., $\mathbf{y} = (y_1, \ldots, y_M)$. Introducing the $M \times N$ Jacobian matrix $\mathbf{J}$ of the partial derivatives $J_{mj} = \partial y_m / \partial x_j$, we have in analogy with Eq. (1)

$$\mathrm{Cov}(\mathbf{y}) = \mathbf{J}\mathbf{C}\mathbf{J}'. \tag{2}$$

1.1 Efficient computation of the quadratic form

It should be noted that although $\partial y / \partial \mathbf{x}$ is formally a vector of length N ($\sim 5 \times 10^9$ for the entire Gaia Catalogue), most of the elements are zero, meaning that only a subset of the N astrometric parameters are needed to compute y. Let us denote by n the number of *active* astrometric parameters, i.e., for which $\partial y / \partial x_i \neq 0$. To evaluate Eq. (1) we obviously do not need the full $N \times N$ matrix $\mathbf{C}$, but only the submatrix of size $n \times n$ corresponding to the active parameters. Taking into account the symmetry of $\mathbf{C}$ this involves $n(n-1)/2$ non-redundant elements C_{ij}. For the more general case of Eq. (2), the size of the submatrix of $\mathbf{C}$ corresponds to the n non-zero rows of $\mathbf{J}$. An important point to note is that in typical computations involving many stars, M is usually much smaller than n. Consequently, the number of non-redundant elements in $\mathrm{Cov}(\mathbf{y})$ that we want to compute, that is $M(M-1)/2$, is *very* much smaller than the $n(n-1)/2$ non-redundant elements in $\mathbf{C}$ that enter into Eq. (2).

The goal of this paper is to explain (without going into the mathematical details) how it might be possible to estimate $\mathrm{Cov}(\mathbf{y})$ in a computationally efficient way, avoiding the intermediate stage of evaluating a very large number of elements C_{ij} from the covariance matrix of the astrometric solution.

The question we want to address can be formulated quite simply: For a given astrophysical problem we are given a list of the n active astrometric parameters relevant for the problem, and the corresponding partial derivatives of the output data (i.e., the non-zero rows of $\mathbf{J}$). How can we evaluate Eq. (2) in a way that is both accurate and computationally feasible? This question has two parts: accuracy and feasibility. The accuracy depends on a number of simplifying assumptions and approximations that will be discussed elsewhere; here we are mainly concerned with the practical feasibility of the computation.

2 Covariance model for Gaia astrometry

Before we describe our model let us first answer the question why one would need a model for computing the covariances in the first place. Let us assume for a moment that we have been able to compute (or estimate) the source covariance matrix for a full Gaia solution. Given that the final solution will contain 10^9 sources the data volume of the full matrix would be $\sim 10^8$ TeraByte (TB), which seems a totally impractical amount of data to store and query efficiently. In (4) it was actually found to be infeasible to invert the full normal matrix for Gaia considering current and near-future available storage and floating-point capabilities. But independent of the question if we could populate the full table by other means, or if such storage space could be available at the time the final catalogue will come out, it is clearly desirable that the covariance between any pair of source parameters can be computed from a reduced amount of data (e.g. the final catalogue values themselves complemented with some additional observation statistics).

2.1 A practical model

As Gaia is a 'self-calibrating' mission, not only the astrometric source parameters, but also other 'nuisance' parameters will be estimated from the observations. We will neglect the influence of the instrument calibration, but will include attitude calibration as it may have very local influence across on the sky, which could render their disentanglement more difficult (cf. 3, Sect. 1.4.6).

The astrometric parameters in the vector $\mathbf{x}$ are naturally grouped according to the sources, with (usually) 5 parameters per source, corresponding to the two positional components, the parallax, and the two proper motion components. From here on, indices like i and j in the above expressions will refer to the sources, rather than the astrometric individual parameters, so that x_i is the subvector of the 5 astrometric parameters for source i, and C_{ij} is the 5×5 submatrix block containing the covariances between the astrometric parameters of the two sources i and j.

For the estimation of the astrometric and attitude parameters it is demonstrated in (1) that the source covariance matrix block between source i and j can be recursively expanded as

$$C_{ij} = C_{ij}^{(1)} + C_{ij}^{(2)} + ... + C_{ij}^{(p)} + ... \tag{3}$$

Elements in $C^{(1)}$ are the covariances resulting from only estimating the source parameters from the observations assuming that the attitude is known. Since there is no coupling between the sources, only the diagonal elements $i = j$ are non-zero. Elements in $C^{(2)}$ are the covariances resulting from only estimating the attitude parameters and propagating those covariances back to the sources covariance estimation. Because each source is on average observed during 72 field-of-view transits (meaning that it is coupled to the attitude parameters at those transit times), this second covariance term will be non-zero for sources that are observed together at least once (meaning that they have at least one attitude parameter in common). Going to

higher terms, we find that the coupling between source and attitude parameters goes recursively deeper:

(i) any *odd* term ($p = 1,3,\ldots$) depends on the sources that have observations at any of the attitude intervals involved in the previous term,

(ii) any *even* term ($p = 2,4,\ldots$) depends on the attitude parameters that are coupled to any of the sources involved in the previous term.

Of course the actual strength of the coupling depends on how many observations were in common between the source/attitude parameter at each step, which is not further discussed here; see (1) for further details.

For the practical computation of a covariance element C_{ij} (up to any term p) we can use the recursive structure to combine the required data from a much reduced amount of model input data (described in the next section) without the need to compute and invert the full underlying normal matrix first.

2.2 Model input data

The expansion model described in the previous section allows us to approximate the covariance between any pair of astrometric parameters to any level of accuracy from the following data per source and field-of-view transit:

(i) partial derivatives of the along-scan observations with respect to the source parameters (typically 5),

(ii) observation time,

(iii) combined weight of the observations.

As each source will on average have 72 field-of-view transits this results in 504 numbers per source. Uncompressed these data take up $\sim$2 TB for one billion sources and can populate lookup tables which use about the same amount of space. Note that in this way we need to store about 10^8 times less information than would be needed for the full covariance matrix.

3 Connectivity between source and attitude parameters

One important question that arises when considering the above recursive structure for computing covariance elements is how much connections there actually are between the source and attitude parameters for each term. To illustrate this, we have computed the connections resulting for two different sources, namely at position (0° , 0°) and (0°, 50°) in ecliptic coordinates. Based on the nominal scanning law for Gaia, this corresponds to positions on the sky which are rather poorly and overabundantly sampled, with 64 and 186 field-of-view transits over 5 years, respectively.

3.1 Gaia-like simulation data

To test the connectivity for each term in Eq. (3) we initialize our covariance model with artificially generated data for a 5 years mission between 2014 and 2019, for

a set of 196,608 sources distributed in a uniform grid over the sky following a HEALPix map (5) of depth 7 giving a typical source separation of 0.46°. Since the field-of-view size of Gaia is about $0.7 \times 0.7^\circ$ this is a reasonable spatial sampling. We sample the attitude with a 60 second interval (resulting in 2,629,800 attitude intervals). A typical field-of-view transit consists of 10 observations during 45 seconds, making this a reasonable time sampling as well. The covariance model and the observation generator are part of our simulation software AGISLab.

3.2 Connectivity results

In Fig. 1 we show for both positions on the sky to which sources they are connected in successively higher terms. Note that in both cases the source is connected to all other sources within only three steps, demonstrating the high level of entanglement of the astrometric solution which makes it well-conditioned.

In Fig. 2 we show for both positions on the sky, to which attitude intervals they are connected in successively higher terms. We plot only the first six months of the full five years of attitude since the scanning law will give a similar attitude filling for subsequent half year periods; even so the time resolution of the plot is not high enough to show all individual attitude intervals. Therefore we give a histogram of the number of connected attitude intervals for each plot bin of 0.18 days (containing 263 attitude intervals).

These figures illustrate the connectivity of a given source (i) with itself, and are relevant for computing the diagonal block element C_{ii} of the covariance matrix. In order to compute C_{ij} for $i \neq j$, only the common connections are relevant.

4 Example variance computations

4.1 Mean parallax of stars in a cluster

An obvious, but very useful property of objects in a cluster is that their distance is (almost) the same, therefore allowing the mean cluster distance to be estimated by averaging over the parallaxes of the individual stars. As correlations between the stars at small angular separation are expected in the Gaia catalogue (see 6), is is necessary to do a proper covariance computation to determine the statistical uncertainty of the cluster distance. For n stars the mean parallax will be $y = (\varpi_1 + ... + \varpi_n)/n$ with $\partial y/\partial \varpi_i = n^{-1}$. Using Eq. (1) we find then

$$\sigma_y^2 = n^{-2} \sum_{i \in n} \sum_{j \in n} C_{ij} \tag{4}$$

The variance of the mean parallax can therefore be computed at successive higher approximations (p) by considering the connections between all possible pairs (i, j) of stars in the cluster, including (for $p > 1$) the 'indirect' connections via common attitude intervals and other stars.

4.2 Acceleration of the solar system barycentre

As an example involving the combination of astrometric data from many objects scattered over the sky, we take the determination of the acceleration of the solar system barycentre in a cosmological frame. Such an acceleration is produced by asymmetries in the distribution of masses around the solar system at different length scales, and is seen as an apparent 'streaming' motion of cosmological objects (mainly quasars) due to the changing stellar aberration. The main expected acceleration is caused by the mass of the Galaxy within the solar circle and amounts to about 2×10^{-10} m s^{-2} directed towards the Galactic centre, and the observable effect is that the quasars will appear to have a streaming motion towards the Galactic centre, with an amplitude of 4 μas yr^{-1}. However, deviations from this could be produced by local irregularities of the mass distribution, and it is therefore interesting to measure the effect.

Based on data for n quasars, the weighted least-squares estimate of the acceleration vector $\mathbf{a}$ (with 3 elements) is just a linear combination of the $2n$ observed proper motion components, and the partial-derivative matrix $\mathbf{M}$ therefore has three columns with non-negative elements only in the $2n$ rows corresponding to the quasar proper motions. Since n is large, the number of terms to consider even for $p = 1$ is quite large, and it may not be feasible to compute it as accurately as for a problem with fewer sources.

5 Conclusion

In order to estimate the covariances of arbitrary functions of the astrometric data, we propose to use a recursive algorithm based on structural data (about how the sources and attitude intervals are connected, and the observation weights) that can be stored relatively compactly. The accuracy of the estimates depends on the level to which the recursions are taken, and are ultimately limited by available computing power. The practical implementation and testing of this algorithm is an ongoing project.

Acknowledgements This work was supported by the European Marie-Curie research training network ELSA (MRTN-CT-2006-033481). LL and DH acknowledge support by the Swedish National Space Board.

References

[1] B. Holl, L. Lindegren, D. Hobbs, A&A **in preparation** (2011)
[2] B. Holl, L. Lindegren, D. Hobbs, A&A **in preparation** (2011)
[3] F. van Leeuwen, *Hipparcos, the New Reduction of the Raw Data* (Astrophysics and Space Science Library Vol. 350, 2007)
[4] A. Bombrun, L. Lindegren, B. Holl, S. Jordan, A&A **516**, A77+ (2010). DOI 10.1051/0004-6361/200913503
[5] K.M. Górski, E. Hivon, A.J. Banday, B.D. Wandelt, F.K. Hansen, M. Reinecke, M. Bartelmann, ApJ **622**, 759 (2005). DOI 10.1086/427976
[6] B. Holl, D. Hobbs, L. Lindegren, in *IAU Symposium*, *IAU Symposium*, vol. 261, ed. by S. A. Klioner, P. K. Seidelmann, & M. H. Soffel (2010), *IAU Symposium*, vol. 261, pp. 320–324. DOI 10.1017/S1743921309990573

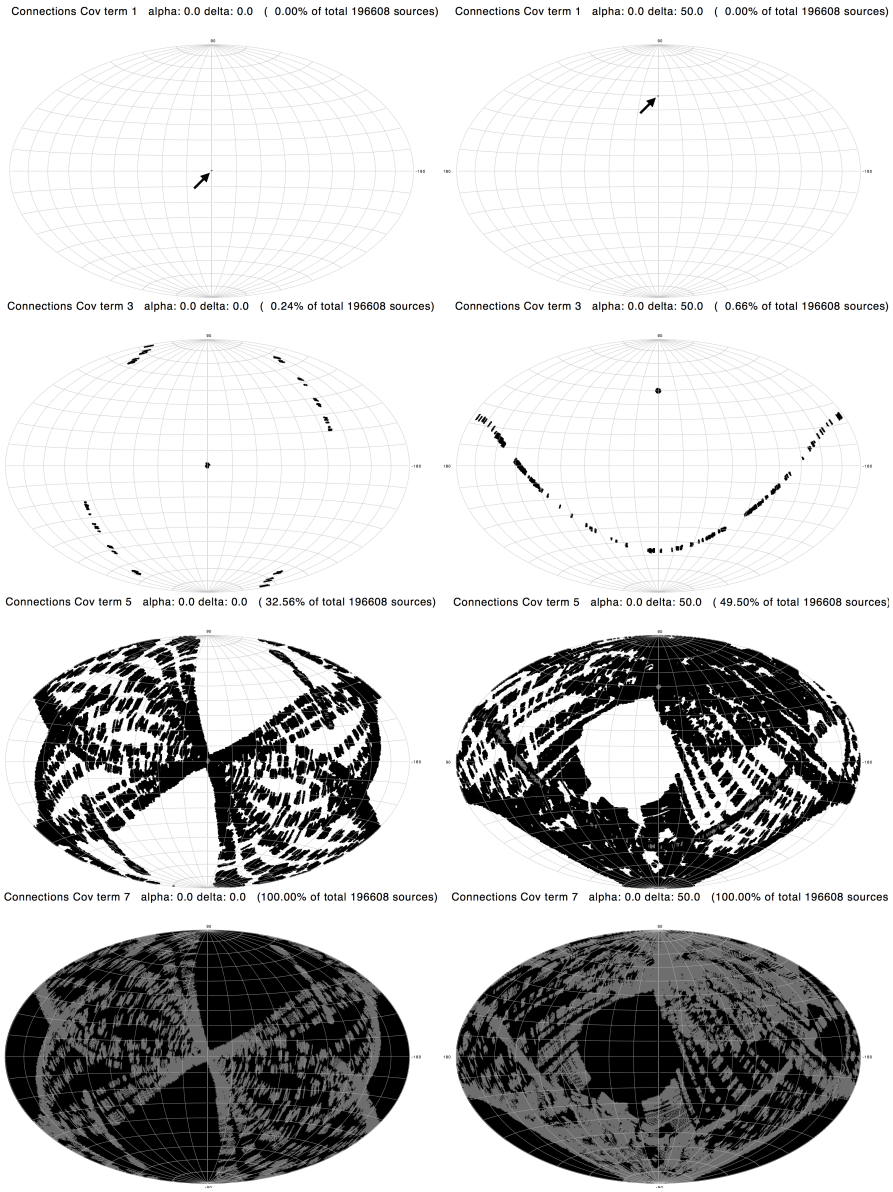

Fig. 1 Left: a source at ecliptic position (0° , 0°), showing its connection to other sources for each odd term in Eq. (3). We assign a black color to the sources that are new with respect to the previous term (the sources from previous terms are shown in gray). Right: the same for a source at ecliptic position (0°, 50°). In both cases the source is connected to all other sources on the sky within 3 steps. When computing the covariance between these two sources, one would need to consider, at each term, the intersection between the corresponding left and right graphs. Maps are plotted in International Celestial Reference System (ICRS) coordinates, centred on (0°, 0°).

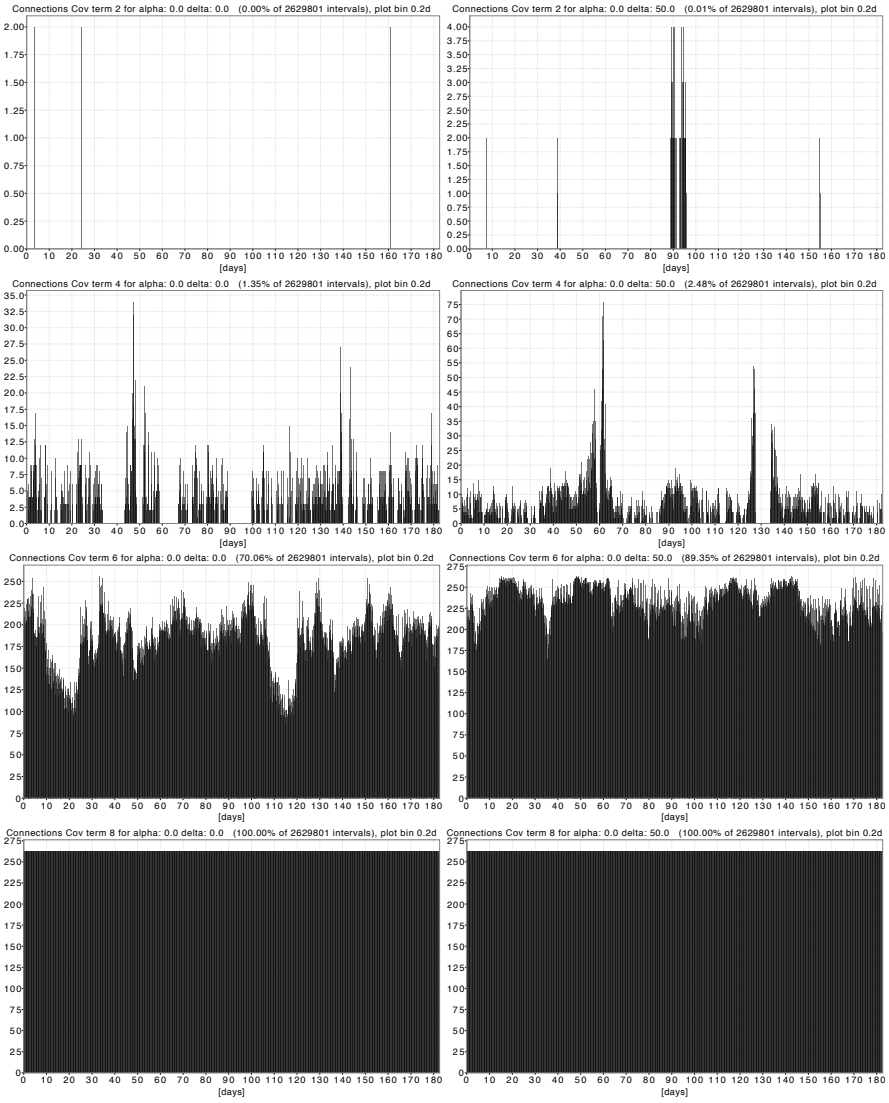

Fig. 2 Left: for a source at ecliptic position (0° , 0°) we show the first six months of the full five years of attitude. For each even term in Eq. (3) the number of connected attitude intervals is shown in a histogram, having a bin size of 0.18 days (containing 263 attitude intervals). Right: the same for a source at ecliptic position (0°, 50°). In both cases the attitude intervals in which the source was observed are connected to all other attitude intervals within 3 steps.