1. Introduction

The compilation of catalogs with the positions of celestial sources is among the most established tasks of astronomy. Instrumental progress has allowed astronomers to survey the sky at an increasingly greater depth across a wide range of the electromagnetic spectrum. The availability of such catalogs brings about the problem of cross-matching as a study of source populations and individual celestial objects requires the identification of physically related measurements listed across different catalogs.

Examples of pertinent astronomical catalogs include the HIPPARCOS and Tycho catalogs of optical point sources (ESA 1997), the infrared ALLWISE catalog (Cutri et al. 2021), or the NRAO VLA Sky Survey (Condon et al. 1998) with discrete sources in the radio regime. We refer to McKaelian (2016) for a more comprehensive review. Of particular interest, with regard to the context of stellar physics, are the data collections of the Gaia satellite (Gaia Collaboration 2016; Brown et al. 2018) and eROSITA instrument (Predehl et al. 2021). Gaia has been carrying out the most comprehensive optical survey of the sky thus far, with more than a billion positions already measured. Similarly, the eROSITA instrument onboard the Spectrum Roentgen Gamma (SRG) satellite carries out the deepest yet X-ray survey of the sky. Clearly, we expect Gaia and eROSITA to observe different but potentially overlapping source populations, so that the matching problem naturally arises.

The angular separation in the sky is crucial in addressing the matching problem as a common position is a necessary condition for a common physical origin. Given two catalogs of positions and uncertainties, adopting the closest available sources from both catalogs can be a viable matching procedure as long as confusion by unrelated sources with random positional proximity remains weak (e.g., Condon et al. 1975; de Ruiter et al. 1977).

Expressions for the likelihoods and their ratios for the geometric matching problem (but not necessarily limited to it) have been presented, for instance, by Richter (1975), Prestage & Peacock (1983), and Sutherland & Saunders (1992). Fioc (2014) has presented an extensive probabilistic treatment of the matching problem and Brand et al. (2006), Budavári & Szalay (2008), and Salvato et al. (2018) have studied the matching problem from a specifically Bayesian viewpoint. For an historical overview of the problem, we refer to Salvato et al. (2018).

In our paper, we study the geometric matching problem between two catalogs of positions and uncertainties using nearest-neighbor distributions and we highlight their relation to point processes. We state the problem in Sect. 2 and introduce spatial point processes and nearest-neighbor distributions in Sect. 3. In Sect. 4, we discuss the nearest-neighbor distributions of sources with the same celestial origin and the use of their densities to determine the fraction of such sources in the catalogs in Sect. 5. We present applications to mock and real-world data in Sect. 6 and we summarize our results in Sect. 7.

2. Posing the problem

We considered two catalogs with their positions and uncertainties. The distribution of the positional measurements is assumed to be Gaussian and isotropic, so that the angular distance between the measured and true source position follows a Rayleigh distribution with a density of:

\[
f_R(r; \sigma) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right),
\] (1)
as specified by the Rayleigh parameter $\sigma$; we note that the relevant properties of this distribution are given in Appendix A. For the sake of clarity, we assume that one is a catalog, $C$, of X-ray sources, such as the one produced by eROSITA and the other one a catalog, $X$, of potential optical counterparts, such as the one produced by Gaia. This specification only serves to offer more clarified terminology. Throughout, we use the index $i$ with $1 \leq i \leq N_C$ to refer to quantities from the catalog of potential counterparts and index $j$ with $1 \leq j \leq N_X$ to those from $X$ wherever possible. Consequently, the catalogs contain the positions $(p_{C,i}$ and $p_{X,j})$ and the Rayleigh parameters ($\sigma_{C,i}$ and $\sigma_{X,j}$).

In the following, we deal with the geometric relation between the sources in $C$ and $X$. For any source in $X$, we call the one in $C$ with the smallest angular distance to it the first neighbor, the one with the second smallest angular distance to it the second neighbor, and so forth. While many of the entries in both $C$ and $X$ may correspond to celestial objects with entries exclusively present in one of the catalogs, a certain fraction could originate from the same direction in the sky and, thus, likely coming from the same celestial object such as a star or galaxy. In this case we call the catalog entries “associated” and the respective entry in $C$ the “associated counterpart” for the source in $X$. The remaining entries in $C$ are called “unassociated”.

On the assumption that any two positions $p_{C,i}$ and $p_{X,j}$ from $C$ and $X$ are associated, their mutual distance follows a Rayleigh distribution, characterized by the parameter

$$\sigma_{ij} = \sqrt{\sigma_{C,i}^2 + \sigma_{X,j}^2}.$$  

Clearly, any pair of sources with positions, $p_{C,i}$ and $p_{X,j}$, separated by a distance on the order of $\sigma_{ij}$ represents a candidate association, however, the positional proximity may also be the result of a chance alignment. We call the fraction of X-ray sources with associations in $C$, the catalog association fraction, $f$. While any X-ray source must be associated with only one potential counterpart at most, we assume in principle potential counterparts may be associated with more than one X-ray source. This corresponds to the “several-to-one” hypothesis formulated by Fioc (2014) and is practically equivalent to the “one-to-one” hypothesis with unique counterparts if the spatial density of sources in $X$ is low compared to that in $C$, so that the sets of candidate associations for the X-ray sources are mutually exclusive. This convention enables to consider the calculations for individual X-ray sources independent and may also be justified on physical grounds in cases, where the resolution of $C$ is comparably poor (Fioc 2014).

Our problem consists of estimating the catalog association fraction and the geometric association probability for the individual pairs of entries. To address this problem, a statistical model of the distribution of sources in the sky is required.

### 3. Spatial point processes

For our purposes, a (spatial) point process is a method to stochastically distribute points on the celestial sphere or, likewise, the plane approximating it locally (i.e., $\mathbb{R}^2$). A comprehensive overview over point processes can be found in Daley & Vere-Jones (2003); discussions with a more astrophysical focus are presented by Babu & Feigelson (1996) and Kerscher (2001).

General point processes have a representation in the form of a probability generating functional, which can be expanded with respect to spatial correlation functions (Kerscher 2001; Daley & Vere-Jones 2003), producing processes with successively more intricate structures, which can be used to model the distribution of sources in the sky. In the following, we briefly introduce the spatial Poisson and Gauss-Poisson processes, which can be used to model the distribution of catalog sources. It is assumed that the distributions and spaces are well-behaved throughout.

#### 3.1. Spatial Poisson process

The spatial Poisson process is the most fundamental one in the sense that it represents the first term in the expansion of the probability generating functional with no spatial correlation between any of its points (Kingman 1993; Daley & Vere-Jones 2003). If the distribution of sources in some region $A$ with area, $|A|$, of the sky adheres to a spatial Poisson process with (constant) intensity $\lambda$, which here represents the spatial source density, the number of sources, $n$, in that region follows a Poisson distribution with expected value $\lambda|A|$ and probability mass function,

$$P(n, \lambda|A|) = \frac{(\lambda|A|)^n}{n!} \exp(-\lambda|A|),$$

which is the probability to obtain exactly $n$ points. Their positions are uniformly distributed within $A$. If the intensity does not depend on the position, the process is called homogeneous. The sum of two homogeneous Poisson processes with intensities $\lambda_{1,2}$ is again a homogeneous Poisson process with an intensity of $\lambda_{1} + \lambda_{2}$.

We use $Z$ to denote the hypothesis that a set of $n$ points with positions $p_{1..n} \equiv \{p_1, \ldots, p_n\} | A$ in $A$ was produced by a homogeneous spatial Poisson process, so that there are zero associated counterparts among them. Then, the likelihood of the realization is expressed as (Daley & Vere-Jones 2003):

$$L_{\text{HP}}(p_{1..n} | A, Z) = \lambda^n \exp(-\lambda|A|).$$

#### 3.2. Nearest-neighbor distribution of the homogeneous Poisson process

Point processes can be characterized by their nearest-neighbor distribution. The latter is determined by the probability to find the closest point within a given distance around an arbitrarily chosen reference point in the process; the corresponding distributions can also be defined for the second, third, ... (etc.) closest neighbor.

If a set of points in $A$ was generated by a homogeneous spatial Poisson process, the distance distribution to the nearest neighbors (ignoring edge effects) has been explicitly derived, for instance, by Thompson (1956). The joint density of the distances $r_{1..n} \equiv \{r_1, \ldots, r_n\}$ to the $n$ nearest neighbors from any point in the process is expressed as:

$$P(r_{1..n} | Z) = \exp\left(-\pi \lambda r^2 \right) (2\pi \lambda)^n r_1 \cdots r_n,$$

where $r_1 < r_2 < \ldots < r_n$.\footnote{The likelihood is the probability density of the data conditional on the model. While it is a probability density in the data (i.e., its integral is normalized), it is usually considered a function of the model parameters, with respect to which it is not a probability density. Therefore, constant (and frequently difficult to obtain) factors are often dropped from specifications of the likelihood.}

The marginal density of the distance distribution to the $m$-th neighbor ($m \geq 1$) is given by:

$$f_m(r_m | Z) = 2 (\pi \lambda)^m r_m^{m-1} \left( \frac{m-1}{m} \right)! \exp\left(-\pi \lambda r^2 \right).$$

\footnote{In this paper, $P(r|H)$ and its analogous expressions are used to indicate $P(R \in [r, r + dr]|H)$, where $R$ is a random variable and $[r, r + dr]$ is the interval of its values.}
3.3. The Gauss-Poisson process

The Gauss-Poisson process is obtained by retaining the first and second term in the expansion of the general probability generating functional. For our purposes, it is most suitably interpreted as a Poisson cluster process (Kerscher 2001). Clusters consist of either one point, in the case of which we call the centers “bare” or, otherwise, there are two points that call for the cluster to be dubbed “complete”. While the locations of all cluster centers is governed by a Poisson process, the relative positioning of points forming complete clusters is given by a probability distribution such as a Gaussian. The points of the complete clusters are thus associated in our terminology. Therefore, the source distribution of the union of the two catalogs \( X \) and \( C \) can be modeled as a Poisson cluster process. Sources without associated counterpart are represented by bare cluster centers, adhering to a Poisson process, and those with an associated counterpart are modeled by complete clusters. In the following (Sect. 4), we discuss some consequences for the nearest-neighbor distribution and how it can be used to identify sources with associated counterpart.

4. Nearest-neighbor distribution for X-ray sources with an associated counterpart

In the following, we assume that the sky density of unassociated counterparts in a large enough circular region, \( \xi_j \), around some particular source at position \( p_{X,j} \) in \( X \) is a constant \( \eta_{X,j} \) and that:

\[
\sigma_{X,j}^2 = \sqrt{\sigma_{X,j}^2 + \sigma_{\xi,j}^2},
\]

is also constant for all \( i \) referring to neighbors in \( \xi_j \) (see Fig. 1 for a sketch). Conditional on the premise, \( E \), that there exists exactly one associated counterpart, the distance of which is a realization of the Rayleigh process, we derive the probabilities for the individual neighbors to be the associated counterpart in Sect. 4.1 and give marginal and joint probability densities for the angular distances of all neighbors in Sects. 4.2 and 4.3. In the following, we drop the \( X \) and \( j \) indices for better readability.

4.1. Marginal probability for being the associated counterpart

The expectation value for the number of unassociated counterparts within a circular region of radius, \( r \), around the X-ray source is given by \( \lambda(r) = \pi \eta r^2 \). If the \( k \)-th neighbor is associated with the X-ray source, exactly \( k - 1 \) unassociated counterparts are closer to it. Therefore, the probability, \( P_{\text{assoc}}(k \mid E) \), that the \( k \)-th neighbor is the associated counterpart, given that there is an associated counterpart (\( E \)), is obtained by

\[
P_{\text{assoc}}(k \mid E) = \int_0^\infty f_k(r, \sigma) P(k - 1, \lambda(r)) \, dr = \int_0^\infty \frac{r^k \exp\left(-\frac{r^2}{2\sigma^2}\right)}{k!} \, dr = \frac{q^k}{q(q + 1)^k} \quad \text{with} \quad q = 2 \pi \eta r^2,
\]

where \( q \) is a confusion parameter with a value of twice the expected number of unassociated counterparts in the error circle, and \( P(n, \lambda(r)) \) is given by Eq. (3).

The probabilities \( P_{\text{assoc}}(k \mid E) \) and \( P_{\text{assoc}}(k + 1 \mid E) \) are related by the recursive relation:

\[
P_{\text{assoc}}(k + 1 \mid E) = \left(\frac{q}{q + 1}\right) P_{\text{assoc}}(k \mid E).
\]

4.2. Marginal distance distributions

The marginal distance distributions of the nearest neighbors for the spatial Poisson process are well known (Eq. (6)). We now proceed to give the marginal distributions on the premise, \( E \), that there is one associated counterpart with a Rayleigh distance distribution around the X-ray source under consideration.

If the \( n \)-th neighbor is the associated counterpart (event \( e_n \)), exactly \( n - 1 \) unassociated counterparts must precede it in the circular region encompassed by its radius vector. Therefore, we can write

\[
P(r_n, e_n \mid E) = \mathcal{P}(n - 1, \pi \eta r_n^2) \exp\left(-\frac{r_n^2}{2\sigma^2}\right) = \frac{r_n^{n-1}}{(n-1)!} \frac{q}{2\sigma^2} \exp\left(-\frac{r_n^2}{2\sigma^2}\right) \left(\frac{q}{q + 1}\right).
\]

Finally, if the \( n \)-th neighbor is unassociated and placed at a smaller distance than the associated counterpart (event \( e_{n,a} \)), \( n - 1 \) unassociated counterparts must precede it, while the associated counterpart must be located farther out (see Eqs. (A.4) and (A.5) for relevant integrals of the density of the Rayleigh distribution) and, therefore,

\[
P(r_n, e_{n,b} \mid E) = \mathcal{P}(n - 1, \pi \eta r_n^2) \left(2\pi \eta r_night) \exp\left(-\frac{r_n^2}{2\sigma^2}\right) = \frac{2r_n^{n-1}}{(n-1)!} \left(\frac{q}{2\sigma^2}\right) \times \exp\left(-\frac{r_n^2}{2\sigma^2}\right)
\]

Finally, if the \( n \)-th neighbor is unassociated and located at a larger distance than the associated counterpart (event \( e_{n,d} \)), then the \( n - 2 \) unassociated as well as the associated counterpart must precede it. Thus, we write

\[
P(r_n, e_{n,d} \mid E) = \mathcal{P}(n - 2, \pi \eta r_n^2) \left(2\pi \eta r_night) \left(1 - \exp\left(-\frac{r_n^2}{2\sigma^2}\right)\right) = \frac{r_n^{n-3}}{(n-2)!} \left(\frac{q}{2\sigma^2}\right)^{n-1} \exp\left(-\frac{r_n^2}{2\sigma^2}\right)
\]
This latter expression is only meaningful for \( n > 1 \), in which case we write:

\[
P(r_n | E) = P(r_n, e_n | E) + P(r_n, z_{an}, | E) + P(r_n, z_{an}, b | E)
\]

\[
= \frac{r_n^{2n-3}}{(n-2)!} \left( \frac{q}{2\sigma^2} \right)^{n-1} \exp \left( -\frac{r_n^2}{2\sigma^2} (q + 1) \right)
\times \left[ \frac{r_n^2}{(n-1)\sigma^2} (1 + q) + 2 \left( \exp \left( \frac{r_n^2}{2\sigma^2} \right) - 1 \right) \right]
\]

for \( n > 1 \).

For the specific case of the first-neighbor distance, we find another Rayleigh distribution with density:

\[
P(r_1 | E) = \frac{(q + 1)r_1}{\sigma^2} \exp \left( -\frac{r_1^2}{2\sigma^2} (1 + q) \right).
\]

which may be compared with the distribution of the first neighbor distance in the absence of an associated counterpart (Eq. (6)), with its density written as a function of \( q \) :

\[
P(r_1 | Z) = \frac{qr_1}{\sigma^2} \exp \left( -\frac{qr_1^2}{2\sigma^2} \right).
\]

The first neighbor in the presence of an associated counterpart is on average closer than the associated counterpart by a factor of \( \sqrt{1/(q + 1)} \).

In Fig. 2, we show examples of the marginal densities of the first three neighbor distances under hypothesis E (Eq. (14)) and Z (Eq. (6)) for the specific set of parameters \( \eta = 0.02 \) arcsec\(^{-2}\) and \( \sigma = 2' \), which yields \( q = 0.5 \). Evidently, the densities of the distributions with associated counterpart are narrower, favoring smaller distances and that of the third neighbor under E is close to that of the second neighbor under the hypothesis Z.

### 4.3. Joint distance distribution

The product rule allows us to write the joint probability of the \( n + 1 \) first distances, \( r_{1..n+1} \), as

\[
P(r_{1..n+1} | E) = P(r_{n+1} | r_{1..n}, E) \times P(r_{1..n} | E).
\]

If the joint probability density of the first \( n \) distances is known along with the conditional distribution of the next distance, \( r_{n+1} \), the joint density for the distribution of all distances can be constructed iteratively.

As seen for the marginal distributions (Sect. 4.2), the conditional distribution of distance \( n + 1 \) can be constructed by considering three distinct cases, distinguished by the location of the associated counterpart with respect to the considered one at \( n + 1 \). Specifically, the associated counterpart may be found among the neighbors preceding the one with index \( n + 1 \) (hypothesis \( e_{an+1} \)), it may just be that one (hypothesis \( e_{n+1} \)), or it may be a later one (hypothesis \( e_{n+1} \)), such that:

\[
P(r_{n+1} | r_{1..n}, E) = P(r_{n+1}, e_{an+1} | r_{1..n}, E) + P(r_{n+1}, e_{n+1} | r_{1..n}, E) + P(r_{n+1}, e_{n+1} | r_{1..n}, E).
\]

If the associated counterpart is presumed to be among the first \( n \) neighbors \( (e_{an+1}) \), it is clear that neighbor \( n + 1 \) must be an unassociated counterpart and (Thompson 1956)

\[
P(r_{n+1} | e_{an+1}, r_n, E) = \frac{2\pi\eta r_{n+1}}{\sigma^2} \exp \left( -\frac{\pi\eta}{2\sigma^2} \left( r_{n+1}^2 - r_n^2 \right) \right)
\]

\[
= \frac{qr_{n+1}}{\sigma^2} \exp \left( -\frac{q}{2\sigma^2} \left( r_{n+1}^2 - r_n^2 \right) \right).
\]

Here, the exponential term represents the Poisson probability to have no unassociated counterpart within the annulus limited by the radii \( r_n \) and \( r_{n+1} \) and \( 2\pi\eta r_{n+1} \) is the probability of finding one unassociated counterpart in the annulus limited by \( r_{n+1} \) and \( r_n + dr_{n+1} \).

Assuming that the first \( n \) neighbors are random counterparts \( (e_{an+1}) \), neighbor \( n + 1 \) may either be the associated counterpart itself, in which case:

\[
P(r_{n+1}, e_{an+1} | e_{an+1}, r_n, E) = \frac{r_{n+1}}{\sigma^2} \exp \left( -\frac{(q + 1)}{2\sigma^2} \left( r_{n+1}^2 - r_n^2 \right) \right)
\]

or the associated counterpart may be an even farther neighbor \( (e_{n+1}) \), and then:

\[
P(r_{n+1}, e_{n+1} | e_{an+1}, r_n, E) = \frac{qr_{n+1}}{\sigma^2} \exp \left( -\frac{(q + 1)}{2\sigma^2} \left( r_{n+1}^2 - r_n^2 \right) \right).
\]

hence, summing Eqs. (20) and (21),

\[
P(r_{n+1} | e_{n+1}, r_n, E) = \frac{(q + 1)r_{n+1}}{\sigma^2} \exp \left( -\frac{(q + 1)}{2\sigma^2} \left( r_{n+1}^2 - r_n^2 \right) \right).
\]

By an iterative application of the product rule and Eq. (21), we find the joint density of the first \( n \) distances, \( r_{1..n} \), and the hypotheses that they all belong to unassociated counterparts:

\[
P(r_{1..n}, e_{n+1} | E) = P(r_1, e_{1..n}, r_{n+1}, e_{n+1} | E)
\]

\[
= q^{r_1} r_1 \ldots r_n \frac{2\pi\eta}{\sigma^2} \exp \left( -\frac{q + 1}{2\sigma^2} r_n^2 \right); \]

in Appendix C, we carry out the start of the iteration explicitly. Because \( P(r_{1..n}, e_{n+1} | E) = P(r_{1..n}, e_{n+1}, E)P(e_{n+1} | E) \) (see Sect. 4.1), we then find

\[
P(r_{1..n} | e_{n+1}, E) = \frac{(q + 1)^{r_1} ... r_n}{\sigma^2} \exp \left( -\frac{q + 1}{2\sigma^2} r_n^2 \right).
\]
Fig. 3. Conditional probability density of $r_2$ given $r_1$ for three values for $r_1$, assuming $\eta = 0.02$ arcsec$^{-2}$ and $\sigma = 2''$.

$P(r_{n+1}, r_{1:n} \mid E) = P(r_{1:n} \mid E) \times [P(r_{n+1}, e_{2:n+1} \mid r_{1:n}, E) + P(r_{n+1}, e_{cn+1} \mid r_{1:n}, E)],$ (25)

where

$P(r_{n+1}, e_{2:n+1} \mid r_{1:n}, E) = P(r_{n+1} \mid r_{1:n}, E)P(e_{2:n+1} \mid r_{1:n}, E),$ (26)

and the last factor may be written as

$P(e_{cn+1} \mid r_{1:n}, E) = 1 - P(e_{2:n+1} \mid r_{1:n}, E),$ (28)

which allows us the specify the second summand in Eq. (25). Thus, the joint probability density of the distances can be constructed iteratively.

In Fig. 3, we show, as an example, the density of the conditional probability $P(r_2 \mid r_1, E)$, assuming the parameters $\eta = 0.02$ arcsec$^{-2}$ and $\sigma = 2''$. In Eq. (B.1), we give a closed expression for the joint probability density of the first two neighbor distances.

Similarly to the derivation of Eq. (23), we obtain the joint density of distances of the $n$ nearest neighbors and the probability that the $m$-th neighbor with $1 \leq m \leq n$ is, in fact, the associated counterpart,

$P(r_{1:n}, e_m \mid E) = \frac{q^{n-1} \cdots q_1}{\sigma^{2n}} \exp\left(-\frac{q + 1}{2\sigma^2} r_1^2 \right) \cdots \exp\left(-\frac{q}{2\sigma^2} (r_n^2 - r_m^2)\right),$ (29)

and, therefore, also

$P(e_m \mid r_{1:n}, E) = \frac{P(r_{1:n}, e_m \mid E)}{P(r_{1:n} \mid E)}.$ (30)

As a note of caution regarding marginalization, we state that

$P(r_{1:n} \mid E) = \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E) + P(r_{1:n}, e_{cn} \mid E),$ (31)

which can be approximated by

$P(r_{1:n} \mid E) \approx \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E),$ (32)

if the term $P(r_{1:n}, e_{cn} \mid E)$ is small. This is the case if sufficiently many neighbors are considered, so that $n$ is large, which will be relevant shortly.

4.4. Association probability

We now include the hypothesis $Z$, namely, that there is no associated counterpart for our X-ray source, into our discussion. Using Bayes’ theorem, we find:

$P(E \mid r_{1:n}) = \frac{P(E)P(r_{1:n} \mid E)}{P(E)P(r_{1:n} \mid E) + P(Z)P(r_{1:n} \mid Z)},$

where $P(E)$ is the a priori probability that source $j$ from $X$ under consideration here has an associated counterpart in $C$ and $P(r_{1:n} \mid Z)$ is given by Eq. (5). Utilizing marginalization and assuming that Eq. (32) holds, we write:

$P(r_{1:n} \mid E) = \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E),$ (34)

to obtain

$P(E \mid r_{1:n}) = \frac{P(E) \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E) + P(Z)P(r_{1:n} \mid Z)}{P(E) \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E)} = \left(1 + \frac{P(Z)P(r_{1:n} \mid Z)}{P(E) \sum_{m=1}^{n} P(r_{1:n}, e_m \mid E)}\right)^{-1}.$ (35)

As for the second summand, we revert to Eq. (29) to obtain:

$\left(\sum_{m=1}^{n} P(r_{1:n}, e_m \mid E)\right)^{-1} = \left(\sum_{m=1}^{n} \frac{1}{\eta} \exp\left(-\frac{r_m^2}{2\eta^2}\right)\right)^{-1}$ (36)

and, finally, we rewrite the expression in Eq. (35) in the form:

$P(E \mid r_{1:n}) = \frac{1}{1 + \frac{P(E)}{P(Z)} \left(\sum_{m=1}^{n} \frac{1}{\eta} \exp\left(-\frac{r_m^2}{2\eta^2}\right)\right)^{-1}} = \frac{P(E) \sum_{m=1}^{n} \frac{1}{2\sqrt{\pi} \eta^2} \exp\left(-\frac{r_m^2}{2\eta^2}\right)}{P(E) \sum_{m=1}^{n} \frac{1}{2\sqrt{\pi} \eta^2} \exp\left(-\frac{r_m^2}{2\eta^2}\right) + (1 - P(E))\eta},$ (37)

where we intentionally substituted $2\sqrt{\pi}/\eta\sqrt{\pi}$ for $q$ to demonstrate that the expression amounts to a comparison of the sum of Gaussian densities and the counterpart density $\eta$.

4.5. Consideration of the likelihood of the Gauss-Poisson process

Equation (37) can also be approached through the likelihood of the Gauss-Poisson process, a general expression of which was given by Daley & Vere-Jones (2003; their Eq. (7.1.6)). With a set $p_{0:n}$ of $n+1$ points, corresponding to positions within some region $A$ in the sky, and assuming a translation-invariant process, the likelihood expression is expressed as:

$P_{GP}(p_{0:n} \mid A) \propto \sum_{k=0}^{[\eta t^{1/2}]} \sum_{i} k_{1}^{\eta t^{1/2} - 2k} \times k_{2}(|p_{1,i} - p_{1,j}| \cdots |p_{n,i} - p_{n,j}|).$ (38)

Here, the constant $\kappa_{1}$ and the function $\kappa_{2}(\cdot)$ are the densities of the first and second so-called Khinchin measures, $|p_{i} - p_{j}|$ denotes the angular distance between points $p_{i}$ and $p_{j}$, the sum
over \( k \) is the over the possible number of complete clusters among the \( n + 1 \) points, and \([\cdot] \) indicates the floor function. For a given \( k \), there are a total of \((n + 1)!/(n + 1 - 2k)!2^k\) \( k \)-distinct possibilities to allocate the complete clusters and \( n + 1 - 2k \) bare centers to the \( n + 1 \) points. The second sum is over all conceivable index sets \( t \equiv \{ (t_{1,2}, t_{1,1}), \ldots, (t_{2,2}, t_{1,1}) \} \) consisting of \( k \) pairs of indices with \( t_{m+1} > t_{m,1} \), where \( 1 \leq m \leq k \) and \( t_{q+1,1} < t_{k,1} \), where \( 1 \leq q \leq k - 1 \), so that each complete cluster is defined by a unique pair of indices.

We now assume that \( p_0 \) corresponds to one of our X-ray sources and the remaining \( n \) points to its closest neighbors. Using marginalization, Eq. (38) can be written as:

\[
P_{\text{GP}}(p_{0,n} | A) = \sum_{i=0}^{\lfloor n+1/2 \rfloor} P_{\text{GP}}(p_{0,n}, k = i | A),
\]

and the individual summands can readily be identified as the inner sum in Eq. (38). In our problem, we are only interested in configurations with either zero or one complete cluster (i.e., \( k = 0 \) or \( k = 1 \)) and if a complete cluster exists, \( p_0 \) has to be part of it. With these assumptions, collective denoted by \( A \), we write:

\[
P_{\text{GP}}(k = 1 | p_{0,n}, A, A) = P_{\text{GP}}(p_{0,n}, k = 1 | A, A) = \frac{\kappa_1^{n-1} + \kappa_1^{n+1} \sum_{i=1}^{n-1} \kappa_2(r_i)}{\kappa_1^{n+1} + \kappa_1^{n+1} \sum_{i=1}^{n} \kappa_2(r_i)},
\]

where \( r_i = \| p_i - p_0 \| \). Comparison of Eq. (37) with Eq. (40) shows equivalence for the choices \( \kappa_1 = (1 - P(E)) \eta \) and \( \kappa_2(r) = \kappa_1 P(E) g(r) \), where \( g(r) \) is the Gaussian density. This corresponds to a Gaussian-Poisson process, with a density \((1 - P(E)) \eta \) of cluster centers, produced by a homogeneous Poisson process, a probability \( P(E) \) for a bare cluster center to be completed, and a mean density of \( \kappa_1(1 + P(E)) \). While this demonstrates a close relationship between the Gaussian-Poisson process and our problem, a complete integration into the point-process framework requires an adaptation of the Gaussian-Poisson process to more accurately reflect our settings.

### 4.6. Scale-free formulation of the problem

Thus far, we used the parameter \( q = 2\pi \eta \sigma^2 \) along with the Rayleigh parameter, \( \sigma \), to specify the probability densities. By using normalized angular distances:

\[
s_k = \frac{r_k}{\sigma}.
\]

the densities can be rewritten in a scale-free form by a change of variables. As an example, we carry out the change of variables explicitly for the density of the marginal distance distributions given in Eq. (14), which yields

\[
P(s_n | E) = \frac{s_n^{2n-3}}{(n-2)!} \left( \frac{q}{2} \right)^{n-1} \exp \left( -\frac{s_n^2}{2} \right) (q + 1)^{n-1} \times \left[ \frac{s_n^2}{(n-1)} (1 + q) + 2 \exp \left( \frac{s_n^2}{2} \right) - 1 \right]
\]

for \( n > 1 \).

The resulting distribution is completely determined by specifying \( n \) and \( q \). Therefore, all problems with a fixed product of \( \eta \) and \( \sigma^2 \) are equivalent in terms of the normalized angular distances.

### 5. Catalog association fraction

If the individual association prior probabilities, \( P(E_j) \), for the sources in \( X \) are unknown, the assumption that they are all the same appears reasonable. In this case, the value of the catalog association fraction, \( f \), is adopted as the prior such that \( f = P(E_j) = P(E) \) with \( 0 \leq f \leq 1 \). Consequently, \( fN_X \) becomes the a priori expected number of X-ray sources with an associated counterpart.

To determine the posterior distribution for the catalog association fraction, the scope needs to be widened from individual sources to a larger ensemble. To that end, we arrange the neighbor distances, \( r_{1,n} \), of the \( N_X \) X-ray sources into an offset matrix, \( \mathcal{R} \), such that

\[
\mathcal{R} = \begin{pmatrix}
r_{1,1} & \cdots & r_{1,j} & \cdots & r_{1,n} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
r_{N_X,1} & \cdots & r_{N_X,j} & \cdots & r_{N_X,n}
\end{pmatrix}.
\]

For example, the first column holds the distances to the first neighbor for every source in \( X \). The offset matrix represents the spatial structure of the catalog \( C \) around the sources in \( X \). We assume that the spatial counterpart densities, \( \eta_{X,j} \), valid in the regions \( C_j \) encompassing the \( n \) nearest neighbors around the sources in \( X \), are known as well as an effective parameter for the Rayleigh distribution, \( \sigma'_{X,j} \), which specifies the distance distribution of associated counterparts (Eq. (7)). For later reference, we combine the latter quantities into the set \( \mathcal{E} \equiv \{ \eta_{X,1}, \ldots, \eta_{X,N_x}, \sigma'_{X,1}, \ldots, \sigma'_{X,N_x} \} \).

#### 5.1. Likelihood of the sky

By definition, any source in \( X \) has either one or zero associated counterpart. Therefore, we can define the “likelihood of the sky”, namely, the probability of the offset matrix given the catalog association fraction and \( \mathcal{E} \), as a mixture of the distance distributions with and without an associated counterpart:

\[
P(\mathcal{R} | f, \mathcal{E}) = \int_{j} \left( f P(r_{nj} | E, \mathcal{E}) + (1 - f) P(r_{nj} | Z, \mathcal{E}) \right).
\]

Adopting a prior, \( P(f) \), for the catalog association fraction, we obtain its formal posterior distribution

\[
P(f | \mathcal{R}, \mathcal{E}) = \frac{P(f)P(\mathcal{R} | f, \mathcal{E})}{\int P(f)P(\mathcal{R} | f, \mathcal{E}) \, df},
\]

and, consequently, also the posterior association probability for any source \( j \) in \( X \):

\[
P(E_j | \mathcal{R}, \mathcal{E}) = \int P(f | \mathcal{R}, \mathcal{E}) P(E_j | f, \mathcal{R}, \mathcal{E}) \, df.
\]

Depending on the circumstances, the use of the entire offset matrix or a fraction of it may be useful.

#### 5.2. Using a single neighbor distance

The catalog association fraction and, consequently, association probabilities for individual sources can be estimated using only a single neighbor distance, which is equivalent to selecting a single column from the offset matrix.
which, by substituting the densities of the first neighbor distributions differ under the hypotheses $E$.

The densities of the marginal distance distributions of all neighbors are given by Eqs. (6) and (14). As all neighbor distance distributions are affected in the case of an association, the estimation works also if the considered neighbor distance does not correspond to that of the actually associated counterpart.

5.3. Choosing which neighbor to use

The densities of the marginal distance distributions of all neighbors differ under the hypotheses $E$ and $Z$, but not all to the same degree. One possibility for quantifying the difference between two probability densities is the Kullback-Leibler divergence (KLD; Kullback & Leibler 1951) defined as:

$$D_{\text{KL}}(P || Q) = \int p(x) \ln \left( \frac{p(x)}{q(x)} \right) dx,$$  \hspace{1cm} (49)

where $p(x)$ and $q(x)$ are the densities of the respective processes. The KLD is always positive, but not symmetric with regard to its arguments. It is the expected likelihood ratio for events which actually adhere to distribution $q(x)$ but are assumed to follow $p(x)$. According to Kullback & Leibler (1951), it is the “information in $x$ for discriminating between” the hypotheses that $x$ was drawn from $P$ or $Q$. Although there is no unequivocal reference scale to interpret the numerical value of the KLD, larger values imply a stronger difference between the distribution.

In Table 1, we list $D_{\text{KL}}(P(s_n | Z) || P(s_n | E))$, namely, the KLD between the densities of the normalized $n$th neighbor distances (Sect. 4.6) with and without an associated counterpart, for various values of the confusion parameter, $\zeta$. At the lowest considered value of $\zeta$, the first neighbor distribution provides the by far strongest lever to distinguish between sources with and without associated counterpart, while small values for the KLD are found for later neighbors and, thus, expected likelihood ratios close to one. This is because the associated counterpart tends to be much closer than the first unassociated neighbor (see Fig. 4). The difference is more pronounced than that between the densities of the distributions of the first and second unassociated neighbors, which would, for instance, be compared when looking at the second neighbor, assuming that the associated counterpart is the first neighbor (Fig. 2). When $\zeta$ increases, the dominance of the first-neighbor KLD weakens and is even reversed at the highest value tested here, $\zeta = 5$, which puts an expected 2.5 unassociated counterparts within the radius equal to the Rayleigh parameter, $\sigma$. This case already shows rather serious crowding. Therefore, if an individual neighbor-distance is to be used, it appears reasonable to adopt that of the first neighbor unless crowding is serious.

### Table 1. Kullback-Leibler divergence for marginal $n$th neighbor distance distributions with and without associated counterpart.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$q = 0.05$</th>
<th>$q = 0.5$</th>
<th>$q = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.955</td>
<td>0.901</td>
<td>0.018</td>
</tr>
<tr>
<td>2</td>
<td>0.421</td>
<td>0.327</td>
<td>0.029</td>
</tr>
<tr>
<td>3</td>
<td>0.230</td>
<td>0.212</td>
<td>0.036</td>
</tr>
<tr>
<td>4</td>
<td>0.158</td>
<td>0.154</td>
<td>0.040</td>
</tr>
<tr>
<td>5</td>
<td>0.120</td>
<td>0.119</td>
<td>0.041</td>
</tr>
<tr>
<td>10</td>
<td>0.055</td>
<td>0.055</td>
<td>0.038</td>
</tr>
</tbody>
</table>

Fig. 4. Densities of the distributions of normalized marginal distance of the first and second neighbor for the case $q = 0.05$.

5.4. Using multiple neighbor distances

When the first $n$ neighbor distances are considered simultaneously, their joint distributions under the hypotheses $Z$ and $E$ (Eq. (5) and Sect. 4.3) need to be considered:

$$P(R | f, E) = \prod_j \left( fP(r_j | E, E) + (1 - f)P(r_j | Z, E) \right)$$

$$= \prod_j \left( fP(r_j | E, E) + (1 - f)\exp(-\pi\eta r_j^2)/(2\pi\eta)^n r_1 \ldots r_n \right).$$  \hspace{1cm} (50)

Taking advantage of the relation in Eq. (36), and recalling that this requires that the number of neighbors considered is sufficient to justify neglecting the term $P(r_{1-n} | E, E)$ in Eq. (31), the following approximation of Eq. (50) can be obtained:

$$P(R | f, E) \approx \prod_j \left( f\sum_{n=1}^{\infty} \frac{1}{q} \exp(-\frac{r_j^2}{2\sigma^2}) + (1 - f) \right).$$  \hspace{1cm} (51)

6. Applications

In the following, we use the geometric model described above to estimate the catalog association fraction. First, we use mock data with high counterpart densities and, thus, strong source confusion, but with strict adherence to the underlying assumptions. Second, we estimate the stellar content of the eROSITA eFEDS X-ray catalog to demonstrate a real-world application.
Table 2. Posterior mean of catalog association fraction (in percent) along with 68% credibility interval using different likelihood representations.

<table>
<thead>
<tr>
<th>n</th>
<th>Eq. (47)</th>
<th>Eq. (50)</th>
<th>Eq. (51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.49 ± 1.22</td>
<td>20.49 ± 1.22</td>
<td>1.04 ± 0.04</td>
</tr>
<tr>
<td>2</td>
<td>19.22 ± 1.32</td>
<td>19.44 ± 1.15</td>
<td>1.88 ± 0.69</td>
</tr>
<tr>
<td>3</td>
<td>19.80 ± 1.51</td>
<td>19.79 ± 1.14</td>
<td>14.29 ± 1.17</td>
</tr>
<tr>
<td>4</td>
<td>19.93 ± 1.71</td>
<td>19.98 ± 1.14</td>
<td>18.14 ± 1.15</td>
</tr>
<tr>
<td>5</td>
<td>19.51 ± 1.92</td>
<td>20.04 ± 1.13</td>
<td>19.40 ± 1.14</td>
</tr>
</tbody>
</table>

6.1. Mock data

To test the likelihoods in the context of some fully controlled data set, we created a mock offset matrix comprising the simulated angular distances of 10000 hypothetical X-ray sources to their optical neighbors, assuming that 20% of the X-ray sources have an associated optical counterpart. Around all X-ray sources, we adopted a density of \( \eta = 0.01 \) arcsec\(^{-2} \) for the unassociated counterparts and a Rayleigh parameter \( \sigma' \) of 3\( '' \) for the distance distribution of the associated counterpart.

To generate the mock offset matrix, we first sampled angular distances for the five nearest unassociated counterparts for all X-ray sources from the neighbor distribution of the Poisson process (Eq. (5)). Subsequently, we selected 20% of the X-ray sources and added an angular distance for the associated counterpart, sampled from the Rayleigh distribution. We then sorted the neighbor distances to each of the 10 000 X-ray sources in ascending order and kept only the five smallest ones. These constitute the five columns and 10 000 rows of the mock offset matrix (Eq. (43)).

In this setting, 0.28 random sources are expected within a region of radius 3\( '' \) and one obtains a value of 0.57 for \( q \). Using Eq. (8), we find that if an associated counterpart exists, it is the first neighbor in 64% of the cases, the second neighbor in 23%, and the third neighbor in 8% of the cases.

In Table 2, we list the posterior mean of the catalog association fraction (Eq. (45)), obtained using various fractions of the offset matrix and, thus, forms of the likelihood of the sky. In particular, we apply Eqs. (47), (50), and (51). In the first case, we use only the \( n \)-th of the available five neighbor distances. In the second case, the first \( n \) distances are used simultaneously, and in the last case, the same \( n \) distances are used, assuming that Eq. (32) holds. If only the first neighbor distances are used, Eqs. (47) and (50) are equivalent by definition. While the use of Eq. (50) is theoretically always at least as good the other options, practical aspects such as the validity of the underlying assumptions or simply the availability of all required neighbors may have to be considered (Sect. 6.2).

In Table 2, one observes that Eq. (47) always yields a result consistent with the input, albeit with slightly increasing uncertainty as later neighbors, less likely to be associated, are used. Equation (50) also yields consistent results, but the uncertainty decreases the more counterpart distances are used in the estimation. This may be expected as more information is provided. The application of Eq. (51) yields consistent results only after four or five neighbor distances are considered because the marginalization prerequisite is not fulfilled with sufficient accuracy prior to that.

Using the point estimate of the catalog association fraction, we obtain the association probabilities for the mock sources. In Fig. 5, we show the distributions obtained by using only the first neighbor distance and the first five distances simultaneously. When using only the first neighbor, the association probability never exceeds a value of about ±0.4 beyond, which a sharp drop follows, because of the likelihood structure. Incorporating more neighbors produces a smoother distribution with higher maximum association probability. In any case, the existence of an associated counterpart can be excluded with reasonable confidence for a few hundred sources and the associated counterparts cannot be identified with great certainty, which is a consequence of the strong source confusion. Therefore, in this extreme situation, we can estimate the catalog association fraction with reasonable precision; however, we cannot say, with comparable certainty, which of the X-ray sources have an associated counterpart or which counterpart is the associated one if there is one.

6.2. Stellar content in the eFEDS field

The eROSITA Final Equatorial-Depth survey (eFEDS; Brunner et al. 2022), carried out in November 2019, covers about 140 square degrees. Efforts to identify counterparts for the point-like X-ray sources were presented by Salvato et al. (2022) and Schneider et al. (2022), the latter of which focused on identifying X-ray source with stellar (coronal) counterparts. The problem of identifying stellar coronal counterparts can be addressed by identifying matches in the Gaia catalog, which is essentially complete with respect to sources detectable in eFEDS. Also, the fact that some of the considered X-ray sources may, indeed, be spurious, that is, erroneously identified as sources, does not pose a problem for our method.

In our treatment of the data, we followed Schneider et al. (2022). In particular, we used the same criteria to screen the Gaia and eROSITA catalogs, which leaves us with a total of 27 056 eROSITA X-ray sources with a median Rayleigh parameter of 2.9\( '' \); the positional uncertainty of Gaia is negligible here. Combined with estimates of the local Gaia counterpart density around the X-ray sources (Schneider et al. 2022), we find a mean value of 0.01 for \( q \). It may be expected then, that 99% of the associated counterparts are, in fact, also the first neighbor to the corresponding X-ray source, making the problem of source confusion a much less severe issue here than in our mock problem (Sect. 6.1). In the following, we obtain posterior estimates of the catalog association fraction, using successively more complex models regarding the positional uncertainty and the density of potential counterparts.

First, we estimated the catalog association fraction with the same methodology as described in Sect. 6.1, taking the available estimates for the positional uncertainty and counterpart density at face value. The results are listed in Table 3; again, Eqs. (48)
and (50) are equal by definition if only the first neighbor distance is used. Using the first neighbor distance only, we obtained a catalog association fraction of $(7.54 \pm 0.19)\%$, which yields an estimate of $2040 \pm 51$ associated sources. Table 3 shows that using only the second or farther neighbor distances yields catalog association fractions of about 12%. We attribute this deviation to a violation of the underlying assumptions. Contrary to the assumptions, the spatial structure of the candidate optical counterparts likely shows correlation, in particular, due to the presence of resolved binary systems, the constituents of which are also plausible sources of coronal X-ray emission. In Fig. 6, we show the distribution of the angular distances of the second neighbors of X-ray sources with an association probability as seen in Table 3.

6.2.1. Evidence for binarity among X-ray detected stars

To investigate the hypothesis that binarity among X-ray detected stars and, therefore, correlations in the 3D space of the candidate counterpart distribution affects the estimated catalog association fraction, we examined the parallaxes and, thus, radial distances, $d_1$ and $d_2$, of the first and second neighbor in terms of their angular distance from the X-ray source in the sky. In Fig. 7, we show the value in parsecs of $|d| = |d_1 - d_2|$ for the first and second Gaia neighbor of X-ray sources with an association probability of either more than 50% (blue) and less than 10% (orange); with this probability being determined using only the second neighbor distance. Then 21% of the X-ray sources in the first sample, with high association probability, show a radial offset smaller than 10 pc between their two nearest Gaia neighbors, while this is the case for only 1% of the sources in the second sample. Since the median distance uncertainty is 11.6 pc in the high-probability sample, many of these systems are plausible physical binaries. We, thus, conclude that binarity is a confounding factor in the geometric analysis and, in the least, partially responsible for the differences in the catalog association fraction obtained primarily when individual neighbor distances are used (see Table 3, column Eq. (48)).
multiwavelength properties of X-ray emitters, regardless of their Galactic or extragalactic nature. By comparing our approach with that presented by Salvato et al. (2022), we find that for 2374 X-ray sources, the same most-likely associated Gaia counterpart was obtained, of which 2275 have a secure counterpart in the catalog of Salvato et al. (2022). In 2266 of these 2275 cases, Salvato et al. (2022) also found that the counterpart is Galactic.

In the context of the Salvato et al. (2022) catalog, the posterior number of X-ray sources with an associated counterpart can be estimated using the individual posterior association probabilities, obtained following the formalism of Budavári & Szalay (2008). In NWAY, this corresponding quantity is \( p_{\text{any,i}} \), which is the probability for X-ray source, \( i \), to have an associated counterpart in any of the comparison catalogs (of which there may be several in general).

If the \( p_{\text{any,i}} \) are considered the success rates of independent Bernoulli trials, the expectation value of the fraction of X-ray sources with an associated counterpart, called the posterior completeness fraction in NWAY, \( c_{\text{NWAY}} \), can be estimated. The posterior completeness fraction is comparable to our catalog association fraction. As the sum of the \( p_{\text{any,i}} \) follows a Poisson binomial distribution, and considering \( N_X \) independent Bernoulli trials, we obtain:

\[
 c_{\text{NWAY}} = \frac{1}{N_X} \sum_{i=1}^{N_X} p_{\text{any,i}} \left[ \pm \frac{1}{N_X} \sum_{i=1}^{N_X} \left(1 - p_{\text{any,i}}\right) \times p_{\text{any,i}} \right].
\]

Using only the \( p_{\text{any,i}} \) values of those X-ray sources in the catalog by Salvato et al. (2022) for which a Gaia source also considered by us is listed, we obtain \( c_{\text{NWAY}} = (68 \pm 8)\% \), which compares well with the value of 67\% obtained from our geometric association probabilities of the same X-ray sources. We note that the confidence in individual associations can significantly deviate, which is expected given that Salvato et al. (2022) incorporated more information than solely geometry.

7. Discussion and conclusion

Angular distance is a fundamental quantity to identify associated sources in the sky. We study the cross-matching problem between two catalogs of positions and uncertainties in the context of point processes and nearest-neighbor distributions. The probabilistic properties of the geometric matching problem have been discussed previously by several authors such as Brand et al. (2006), Budavári & Szalay (2008), and Fioc (2014), with different though partially overlapping methodologies, owing to the common problem.

Budavári & Szalay (2008) present a Bayesian probabilistic treatment of the cross-matching problem, which is the basis of the approach by Salvato et al. (2018). The authors derive Bayes factors for the hypotheses that position measurements from different catalogs originate from the same celestial source. While Budavári & Szalay (2008) also considered cases with three or more available catalogs and discussed the use of other properties than geometry (which is not part of our discussion), these authors did not explicitly discuss estimation of the catalog association fraction.

Brand et al. (2006) studied the specific problem of cross-matching X-ray sources observed with the Chandra X-ray telescope with optical sources detected by the NOAO Deep Wide-Field Survey (NDWFS). They presented a generally applicable probabilistic model of the distribution of sources in the sky (in their appendix). They consider the angular positional offsets between X-ray sources and their optical neighbors and, in addition to our treatment, the brightness distribution of potential optical counterparts, which we only take into account by limiting the brightness of the Gaia sources to physically plausible stellar candidate counterparts in our discussion (Sect. 6.2). The model by Brand et al. (2006) is formulated in terms of all possible associations between candidate optical counterparts and X-ray sources. The authors give expressions to calculate probabilities for individual candidate associations (their Eqs. (A.17) and (A.18)) as well as the formalism to obtain the posterior distribution of the catalog association fraction (Eq. (A.21)) and discuss the “mean completeness”, which is equivalent to our catalog association fraction.

Fioc (2014) studied the cross-matching problem of two astronomical catalogs of positions and uncertainties and present a detailed breakup of how associations between the catalogs can occur, with regard to the “several-to-one”, “one-to-several”, and “one-to-one” hypotheses are discussed; of these, we only considered the first. Fioc (2014) derived association probabilities for the candidate counterparts (see, e.g., their Eq. (50) for the several-to-one hypothesis) and constructed maximum-likelihood estimators for the quantity we call the catalog association fraction (see their Eq. (40)). The treatment by Fioc (2014) is not explicitly Bayesian, although the presented distributions could certainly be used to carry out a Bayesian analysis if suitable priors were adopted.

Although Brand et al. (2006), Budavári & Szalay (2008), and Fioc (2014) partly provided discussions that also address other aspects of the problem than geometric ones, none of them explicitly discuss it in the context of point processes or nearest-neighbor distributions. In particular, they have not produced expressions that would explicitly allow us to estimate the catalog association fraction and association probability, restricting the treatment to a limited number of angular neighbor distances such as that of the first one; it is always understood though that candidate associations between sources with overly large angular distances compared to the uncertainty are irrelevant (e.g., Fioc 2014, Sect. 5.2).

In our study, we derived the marginal and joint neighbor distributions for sources with an associated counterpart, the distance of which follows a Rayleigh distribution, each source being embedded in an environment of counterparts described by a spatial Poisson process. A mixture of the resulting distributions is an approximation of the neighbor distribution of the Gauss-Poisson process valid for low rates of complete clusters. We used the newly derived distributions to estimate the catalog association fraction based on the “likelihood of the sky”, in which the information from the entire ensemble of sources...
is combined. Subsequently, the association probabilities for the individual pairs of sources are calculated via their corresponding posterior distributions. Unless crowding is too strong, the first neighbor distance provides the strongest lever to distinguish between sources with and without an associated counterpart.

The catalog association fraction is recovered for simulated data even in cases of high source confusion, where the reliable identification of the individual associations remains broadly impossible. Application to the real-world example of the matching between the eROSITA eFEDS catalog and the Gaia catalog yields a catalog association fraction of \( (7.5 \pm 0.2)\% \) if the positional uncertainties of the eFEDS catalog are taken at face value and \( (8.3 \pm 0.2)\% \) if corrections to the positional uncertainty of the X-ray catalog are considered. In the latter case, our estimates of the correction terms are consistent with those published by Brunner et al. (2022). Then, Salvato et al. (2022) and Schneider et al. (2022) presented cross-matching efforts for the eFEDS field, taking geometry as well as non-geometric information into account, and using the formulations by Budavári & Szalay (2008) or Salvato et al. (2018). Insofar as comparisons can be made, our results are compatible with those published by Salvato et al. (2022) and Schneider et al. (2022). In this context, the catalog association fraction estimated here can be treated as a prior to be updated using further information.

Resolved stellar binaries violate the prerequisite of randomly distributed optical counterparts. Our analysis shows that these do indeed hinder the analysis. The theory of point processes, in principle, could be used to address this problem by constructing a more realistic geometrical model that incorporates spatial correlation among the candidate counterparts (e.g., Daley & Vere-Jones 2003).

Acknowledgements. We thank our referee Dr. Fioc for valuable and insightful comments, which greatly helped to improve our manuscript. S.C. acknowledges the support of the DFG priority program SPP 1992 “Exploring the Diversity of Extrasolar Planets” (CZ 222/5-1). P.C.S. acknowledges support through DLR 50 OR 2102 and 2205. This work is based on data from eROSITA, the soft X-ray instrument aboard SRG, a joint Russian-German science mission supported by the Russian Space Agency (Roskosmos), in the interests of the Russian Academy of Sciences represented by its Space Research Institute (IKI), and the Deutsches Zentrum für Luft- und Raumfahrt (DLR). The SRG spacecraft was built by Lavochkin Association (NPOL) and its subcontractors, and is operated by NPOL with support from the Max Planck Institute for Extraterrestrial Physics (MPE). The development and construction of the eROSITA X-ray instrument was led by MPE, with contributions from the Dr. Karl Remeis Observatory Bamberg & ECAP (FAU Erlangen-Nuernberg), the University of Hamburg Observatory, the Leibniz Institute for Astrophysics Potsdam (AIP), and the Institute for Astronomy and Astrophysics of the University of Tübingen, with the support of DLR and the Max Planck Society. The Argelander Institute for Astronomy of the University of Bonn and the Ludwig Maximilians Universität Munich also participated in the science preparation for eROSITA. The eROSITA data shown here were processed using the eSASS software system developed by the German eROSITA consortium. This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

References

ESA 1997, The HIPPARCOS and Tycho catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission (The Netherlands: Noordwijk), 1200
Kerscher, M. 2001, Phys. Rev. E, 64, 056109
Mickaelian, A. M. 2016, ASP Conf. Ser., 505, 203
Thompson, H. R. 1956, Ecology, 37, 391
Appendix A: Notes on the Rayleigh distribution

The Rayleigh distribution is the distribution of the distance about the origin of a 2D, isotropic Gaussian random variable. It is characterized by a single parameter, \( \sigma \), which is the standard deviation of the Gaussian along each coordinate axis.

\[
f_R(r; \sigma) = \frac{2\pi r}{2\pi \sigma^2} \exp\left( -\frac{x^2 + y^2}{2\sigma^2} \right) = \frac{r}{\sigma^2} \exp\left( -\frac{r^2}{2\sigma^2} \right) .
\]  

(A.1)

While the mode of the Rayleigh distribution is at \( \sigma \), its expected value is given by

\[
E[r] = \int_0^\infty r f_R(r; \sigma) \, dr = \sqrt{\frac{\pi}{2}} \sigma .
\]  

(A.2)

and the variance is expressed as

\[
V[r] = \int_0^\infty r^2 f_R(r; \sigma) \, dr - \frac{\pi}{2} \sigma^2 = \sigma^2 \left( 2 - \frac{\pi}{2} \right) .
\]  

(A.3)

The following two integrals of the density of the Rayleigh distribution are frequently used throughout this work:

\[
\int_0^r f_R(r'; \sigma) \, dr' = 1 - \exp\left( -\frac{r^2}{2\sigma^2} \right) ;
\]  

(A.4)

\[
\int_r^\infty f_R(r'; \sigma) \, dr' = \exp\left( -\frac{r^2}{2\sigma^2} \right) .
\]  

(A.5)

Appendix B: Expressions for the first two neighbors

The joint distance distribution for the distances of the first two neighbors under hypothesis \( E \) is given by:

\[
P(r_1, r_2 \mid E) = \frac{(q + 1)r_1}{\sigma^2} \exp\left( -\frac{(q + 1)r_1^2}{2\sigma^2} \right) \left( \frac{1}{1 + q} \frac{qr_2}{\sigma^2} \exp\left( -\frac{q}{2\sigma^2} (r_2^2 - r_1^2) \right) + \frac{q}{1 + q} \frac{(q + 1)r_2}{\sigma^2} \exp\left( -\frac{q + 1}{2\sigma^2} (r_2^2 - r_1^2) \right) \right) .
\]  

(B.1)

where \( r_2 > r_1 \). The probability that the first neighbor is the associated one is expressed as:

\[
P(e_1 \mid r_1, r_2, E) = \frac{1}{1 + (q + 1) \exp\left( -\frac{1}{2\sigma^2} (r_2^2 - r_1^2) \right)} .
\]  

(B.2)

Likewise, the probability that the second one is the associated one is given by

\[
P(e_2 \mid r_1, r_2, E) = \frac{1}{1 + q + \exp\left( \frac{1}{2\sigma^2} (r_2^2 - r_1^2) \right)}
\]  

(B.3)

and, consequently, the probability that the associated neighbor is beyond the second one is expressed as:

\[
P(e_{2,3} \mid r_1, r_2, E) = \frac{q}{1 + q + \exp\left( \frac{q}{2\sigma^2} (r_2^2 - r_1^2) \right)} .
\]  

(B.4)

Appendix C: Note on Eq. 23

In the following, we carry out the first step of the iteration leading to Eq. 23 specifically. From Eq. 10, we find that \( P(e_{1,2} \mid E) = \frac{q r_1}{\sigma^2} \exp\left( -\frac{r_1^2}{2\sigma^2} (1 + q) \right) \). Combined with Eq. 15, this yields:

\[
P(r_1, e_{1,2} \mid E) = \frac{q r_1}{\sigma^2} \exp\left( -\frac{r_1^2}{2\sigma^2} (1 + q) \right) .
\]  

(C.1)

Now we expand the hypothesis \( r_{1,2}, e_{2,3} \) and subsequently apply the product rule to obtain:

\[
P(r_{1,2}, e_{2,3} \mid E) = P(r_2, e_{2,3}) \mid (r_1, e_{1,2}) \mid E) = P(r_2, e_{2,3}) \mid (r_1, e_{1,2}) \mid E) \times P(r_1, e_{1,2} \mid E) .
\]  

(C.2)

Application of Eq. 23 then yields the next step of the iteration

\[
P(r_{1,2}, e_{2,3} \mid E) = \frac{q^2 r_1^2 r_2^2}{\sigma^4} \exp\left( -\frac{r_1^2}{2\sigma^2} (q + 1) \right) .
\]  

(C.3)