Draft version May 9, 2017
Typeset using LaTeX default style in AASTeX61

# DETECTING STRUCTURES OF ASTRONOMICAL SOURCES USING GRAPH-BASED SEEDED REGION GROWING

Minjie Fan,[1] Vinay L. Kashyap,[2] Andreas Zezas,[3] David A. van Dyk,[4] and Thomas C.M. Lee[1]

[1] *Department of Statistics*
*University of California, Davis*
*One Shields Avenue*
*Davis, CA 95616, USA*
[2] *Harvard-Smithsonian Center for Astrophysics*
*60 Garden Street*
*Cambridge, MA 02138, USA*
[3] *Physics Department*
*University of Crete*
*P.O. Box 2208, GR-710 03*
*Heraklion, Crete, Greece*
[4] *Statistics Section, Department of Mathematics*
*Imperial College London*
*180 Queen's Gate*
*London, SW7 2AZ, UK*

Submitted to ApJ

## ABSTRACT

This example manuscript is intended to serve as a tutorial and template for authors to use when writing their own AAS Journal articles. The manuscript includes a history of AASTeX and documents the new features in the previous version, 6.0, as well as the new features in version 6.1. This manuscript includes many figure and table examples to illustrate these new features. Information on features not explicitly mentioned in the article can be viewed in the manuscript comments or more extensive online documentation. Authors are welcome replace the text, tables, figures, and bibliography with their own and submit the resulting manuscript to the AAS Journals peer review system. The first lesson in the tutorial is to remind authors that the AAS Journals, the Astrophysical Journal (ApJ), the Astrophysical Journal Letters (ApJL), and Astronomical Journal (AJ), all have a 250 word limit for the abstract. If you exceed this length the Editorial office will ask you to shorten it.

*Keywords:* editorials, notices — miscellaneous — catalogs — surveys

Corresponding author: Minjie Fan
mjfan@ucdavis.edu

# 1. INTRODUCTION

# 2. METHODOLOGY

## 2.1. *Statistical Model*

Suppose that the observations are photons located in a bounded domain $D \subset \mathbb{R}^2$ with coordinates $\mathbf{x}_i =: (x_{1i}, x_{2i}), i = 1, \cdots, n$. They are assumed to originate from one of several sources or the background, where the former can be either point or extended sources. The exact origin of each photon, the number of sources, their locations and intensities are all unknown. We are particularly interested in the case when point sources lie within an extended source, since this is a challenging task for existing methods. Unlike Jones et al. (2015), the method to be proposed in this paper depends on the spatial information of photons, i.e., their locations only.

For simplicity, we assume that photons from the background are distributed uniformly in $D$. We also assume that photons from each source are distributed uniformly in the source by ignoring the point-spread function (PSF). Hence, the observed photons can be regarded as a realization of an inhomogeneous Poisson point process[1] with a piecewise constant intensity function. Let $\lambda(\mathbf{x}) \geq 0$, where $\mathbf{x} \in D$, denote the intensity function[2]. As a piecewise constant function, it can be represented as $\lambda(\mathbf{x}) = \sum_{k=0}^{K} \lambda_k \mathbb{1}_{S_k}(\mathbf{x})$, where $S_k$'s are disjoint subsets of $D$ such that $\bigcup_{k=0}^{K} S_k = D$, and $\lambda_k \geq 0$ for all $k$. Besides, $K$ denotes the number of sources, $k = 0$ corresponds to the background and $k = 1, \cdots, K$ correspond to the $K$ sources. Then $S_k$'s characterize the boundaries of the sources and the background, and $\lambda_k$'s measure their intensities.

Let $(\mathbf{X}_1, \cdots, \mathbf{X}_n, N)$ denote the inhomogeneous Poisson process, where $\mathbf{X}_i$ is the location of the $i$-th photon and $N$ is the number of photons, which are all random. Given the observations $(\mathbf{x}_1, \cdots, \mathbf{x}_n, n)$, the probability density function of the inhomogeneous Poisson process is

$$\begin{aligned} f(\mathbf{X}_1 = \mathbf{x}_1, \cdots, \mathbf{X}_n = \mathbf{x}_n, N = n) &= f(\mathbf{X}_1 = \mathbf{x}_1, \cdots, \mathbf{X}_n = \mathbf{x}_n | N = n) f(N = n) \\ &= \frac{\prod_{i=1}^{n} \lambda(\mathbf{x}_i)}{\left(\int_D \lambda(\mathbf{x}) d\mathbf{x}\right)^n} \frac{\exp\left(-\int_D \lambda(\mathbf{x}) d\mathbf{x}\right) \left(\int_D \lambda(\mathbf{x}) d\mathbf{x}\right)^n}{n!} \\ &= \frac{\prod_{i=1}^{n} \lambda(\mathbf{x}_i) \exp\left(-\int_D \lambda(\mathbf{x}) d\mathbf{x}\right)}{n!}. \end{aligned}$$

Write $\mathbf{S} = (S_0, \cdots, S_K)$ and $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_K)$. Then the log-likelihood function is

$$l(K, \mathbf{S}, \boldsymbol{\lambda} | \mathbf{x}_1, \cdots, \mathbf{x}_n, n) =: \log f = \sum_{i=1}^{n} \log \lambda(\mathbf{x}_i) - \int_D \lambda(\mathbf{x}) d\mathbf{x} - \log n!.$$

As a function of the model parameters given the data, it tells us what values of the parameters are supported by the data. The maximum likelihood estimates (MLE) of the parameters are the parameter values that maximize the log-likelihood. Plugging the expression of $\lambda(\mathbf{x})$ as a piecewise constant function into $l$, we have

$$l = \sum_{k=0, \#(S_k) \neq 0}^{K} \#(S_k) \log \lambda_k - \sum_{k=0}^{K} |S_k| \lambda_k - \log n!,$$

where $\#(S_k)$ and $|S_k|$ denote the number of photons in $S_k$ and the area of $S_k$, respectively. When $\#(S_k) = 0$, the summand $\#(S_k) \log \lambda_k$ is excluded from the sum. For fixed $S_k$'s, $l$ is maximized when

$$\lambda_k = \widehat{\lambda}_k := \#(S_k)/|S_k|.$$

Plugging $\widehat{\lambda}_k$'s into $l$, we reduce the log-likelihood to a profile log-likelihood of $S_k$, i.e.,

$$l_{\text{profile}}(K, \mathbf{S} | \mathbf{x}_1, \cdots, \mathbf{x}_n, n) = \sum_{k=0, \#(S_k) \neq 0}^{K} \#(S_k) \log \{\#(S_k)/|S_k|\} - n - \log n!. \tag{1}$$

[1] An inhomogeneous Poisson point process with an intensity function $\lambda(\mathbf{x})$ is characterized by the following stochastic property: for any collection of disjoint subsets of $D$, denoted by $B_j, j = 1, \cdots, m$, $\#(B_j)$'s are independent random Poisson variables with means $\int_{B_j} \lambda(\mathbf{x}) d\mathbf{x}, j = 1, \cdots, m$, where $\#(B_j)$ denotes the number of points in $B_j$.

[2] The intensity function has to be integrable over $D$, i.e., $\int_D \lambda(\mathbf{x}) d\mathbf{x} < \infty$.

Another way of modeling the data is by treating them as a mixture of uniform distributions, from which the same form of profile log-likelihood can be derived. Allard & Fraley (1997) considered a special case with only one source and the background.

### 2.2. *Model Selection Using Bayesian Information Criterion*

Usually the number of sources $K$ is unknown, and it cannot be estimated by maximizing the profile log-likelihood (1) since the profile log-likelihood can be made arbitrary large as $K$ increases. One way to resolve this issue is to add a penalty term to the profile log-likelihood to suitably penalize the complexity of the model. Aue & Lee (2011) used information theoretic model selection methods such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the minimum description length (MDL) principle to derive such a penalty; the latter two were shown to be consistent for image segmentation. However, the MDL principle is not well-defined in our case since there seems no way of encoding $S_k$'s in a continuous space without assuming any particular shapes for $S_k$'s. Thus, we use the BIC for model selection, which is defined as

$$\mathrm{BIC}(K, \mathbf{S}) = -2l_{\mathrm{profile}}(K, \mathbf{S}|\mathbf{x}_1, \cdots, \mathbf{x}_n, n) + p \log n,$$

where $p$ is the number of free/independent parameters in the model. The BIC estimates for $(K, \mathbf{S})$ are given by

$$(\widehat{K}, \widehat{\mathbf{S}}) = \underset{K \le K_{\max}}{\arg\max}\ \mathrm{BIC}(K, \mathbf{S}),$$

where $K_{\max}$ is the maximally allowed number of sources. This is apparently a non-parametric model and hence $p$ is not well-defined. Similar to Aue & Lee (2011), we can approximate the model by a parametric one through assuming a specific shape for $S_k$'s. For example, when the sources are close to ellipse-shaped, $p = 6K + 1$ because the number of parameters associated with each source is 6, and the background intensity is counted as one parameter. Thus, the BIC becomes

$$\mathrm{BIC}(K, \mathbf{S}) = -2 \log f(K, \mathbf{S}) + (6K + 1) \log n.$$

Another possible choice is specifying $p = K + 1$, the number of sources and the background, which to some extent reflects the complexity of the model (Magnussen et al. 2006), but totally ignores the shapes of the sources.

### 2.3. *Voronoi Tessellation*

Without additional constraints on $S_k$'s, the estimates of $S_k$'s obtained by minimizing the BIC are not well-defined for the following reasons: (i) $S_k$ with $\#(S_k) = 0$ can have arbitrary shape since it does not contribute to the BIC; and (ii) the BIC can be arbitrarily small by shrinking the areas of $S_k$'s towards zero since $|S_k|$'s are on the denominator. Thus, we consider a restricted class of $S_k$'s such that each of them consists of the Voronoi cells derived from the Voronoi tessellation of the data. In this case, each $S_k$ contains at least one photon.

The Voronoi tessellation of the observed photons uniquely partitions $D$ into $n$ convex cells, denoted by $C_i, i = 1, \cdots, n$, such that cell $C_i$ contains one and only one photon, say $\mathbf{x}_i$, and consists of all locations in $S$ closer to photon $\mathbf{x}_i$ than to any other photons. These cells are called Voronoi cells, and $\mathbf{x}_i$ is called the nucleus of $C_i$. To avoid border effects, we restrict the tessellation to Voronoi cells the vertices of which are all in $D$. Based on the Voronoi tessellation, Barr & Schoenberg (2010) introduced the Voronoi estimator $\widehat{\lambda}^{\mathrm{V}}(\mathbf{y}) =: 1/|C_i|$ for any location $\mathbf{y} \in C_i$. They showed that under certain conditions, the Voronoi estimator is approximately unbiased, and its sampling distribution is approximately inverse Gamma[3]; these results are further verified through simulation studies.

### 2.4. *Graph Segmentation Using Seeded Region Growing*

The dual graph of the Voronoi tessellation is called the Delaunay triangulation. Figure ? displays. Its vertices are exactly the photons, and its edges connect pairs of the Voronoi cells. Vertex $\mathbf{x}_i$ can be assigned the value of the Voronoi estimator, denoted by $\widehat{\lambda}_i^{\mathrm{V}}$, that estimates the intensity in Voronoi cell $C_i$. Based on the graph constructed by the Delaunay triangulation, the problem of estimating $S_k$'s is equivalent to that of graph segmentation, i.e., partitioning the graph into subgraphs such that the Voronoi cells corresponding to each of them form a source or the background.

[3] The probability density function of an inverse Gamma distribution is $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$, where $x > 0$, $\alpha$ and $\beta$ are the shape and rate parameters, respectively, and $\Gamma(\cdot)$ denotes the Gamma function.

Since we have assumed that the intensity function is piecewise constant, the vertices in each subgraph share similar values. Besides, the values in the subgraphs of the sources are significantly larger than those of the background.

A distinct advantage of treating this problem as graph segmentation is that we implicitly impose the constraint that each $C_i$ is connected. Traditional image segmentation[4] techniques can be adapted to the graph and used here. In particular, we propose a graph-based seeded region growing (SRG) method, which is very similar to the original SRG used for images except that the concept of "neighbors" is determined by the edges of the graph instead of neighboring pixels. The original SRG was proposed in Adams & Bischof (1994) and extended to several variants to deal with more complicated cases in Fan & Lee (2014). The graph-based SRG starts with identifying, either manually or automatically, a set of seeds from the graph. Each seed can be a single vertex or a set of connected vertices. For the moment, we assume that the seeds are perfectly specified, i.e., there is one and only one seed in each true source or the true background. The details of seed specification in practice are described in Section 3.1. Then the graph-based SRG grows these seeds into subgraphs by successively adding neighboring vertices to them. More specifically, at each time step, the method selects the pair of a growing subgraph $S$ and one of its neighboring vertex $i$ such that the following criterion is minimized

$$\delta(i, S) = \left| \log \widehat{\lambda}_i^{\mathrm{V}} - \log\{\#(S)/|S|\} \right|, \tag{2}$$

where we do not distinguish between the subgraph and its corresponding Voronoi cells so that the operators $\#(\cdot)$ and $|\cdot|$ are still meaningful for $S$. This criterion compares the the estimated log intensities of the subgraph and the vertex because taking the logarithm can magnify the contrast. Then the vertex in the pair with the smallest difference is added to the corresponding subgraph. The graph-based SRG finishes when all the vertices of the graph are assigned to one and only one subgraph. These subgraphs give the estimated $S_k$'s.

## 3. COMPUTATIONAL DETAILS

### 3.1. *Seed Specification*

We have assumed an ideal case for seed specification where there is one and only one seed in each true source or the true background. However, such perfect seed specification is not always possible without some source detection algorithm or human interaction. Nonetheless, we provide a simple but practical method for seed specification following Lee (2000). More specifically, we first overlay a regular grid with appropriate spacing on the domain $D$. A too dense grid would lead to an insufficient number of photons for seed specification, and a too sparse grid would miss some of the seeds that are supposed to be specified. Then for each grid point, we assign to the corresponding seed its closest $s$ photons in terms of the Euclidean distance that have not been assigned. The seed size $s$ can be increased to stabilize the initial estimates of the seed/growing subgraph intensities, especially when the signal-to-noise ratio (i.e., the ratio between the intensities of a source and the background) is low. Again, a too large $s$ would lead to an insufficient number of photons for seed specification.

The price for increasing $s$ is that some of the seeds may fall on true source boundaries and adversely affect subsequent processing. To overcome this, we take an additional step of seed rejection. For a typical Voronoi cell $C$ of a planar homogeneous Poisson point process with a constant intensity $\lambda$, we know that $\mathrm{E}(|C|) = 1/\lambda$ and $\mathrm{Std}(|C|) \approx 0.53/\lambda$ (Møller 1994, Chapter 4.2), where $|C|$ denotes the area of $C$. If a seed does not fall on the boundaries, then it can be regarded as a part of the Voronoi tessellation of a homogeneous Poisson point process. The areas of the cells in the seed should vary to the extent controlled by the standard deviation. More specifically, the areas have a very high probability of falling within the interval $1/\hat{\lambda} \pm \alpha \times 0.53/\hat{\lambda}$, where $1/\hat{\lambda}$ is the mean area of the cells in the seed, and $\alpha$ is a parameter determining the width of the interval, which is specified as 2 in our case. Thus, the range of the areas has a very high probability of being less than or equal to $2\alpha \times 0.53/\hat{\lambda}$, which is used as an ad-hoc threshold. If the actual range for a seed is larger than the estimated threshold, we reject the seed.

When the regular grid is not dense enough, seeds in point sources with small spreads are more likely to be missed than those in extended sources and the background. One remedy is to include additional seeds based on the fact that the intensities of point sources are significantly larger than those of extended sources and the background. More specifically, we first find vertices that are local maxima of the graph constructed by the Delaunay triangulation, in the sense that the value of the vertex (i.e., $\widehat{\lambda}_i^{\mathrm{V}}$) is larger than or equal to those of its closest $k$ vertices (including itself) in

[4] Image segmentation is the process of separating an image into several regions such that each region is composed of connected pixels with similar characteristics, such as similar pixel values.

terms of the Euclidean distance. Then for each found vertex, we assign to the corresponding seed its closest $s'$ vertices (including itself) in terms of the Euclidean distance.

### 3.2. *Subgraph Merging*

## REFERENCES

Adams, R., & Bischof, L. 1994, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16, 641

Allard, D., & Fraley, C. 1997, Journal of the American Statistical Association, 92, 1485

Aue, A., & Lee, T. C. M. 2011, The Annals of Statistics, 2912

Barr, C. D., & Schoenberg, F. P. 2010, Biometrika, 977

Fan, M., & Lee, T. C. 2014, IET Image Processing, 9, 478

Jones, D. E., Kashyap, V. L., & Van Dyk, D. A. 2015, The Astrophysical Journal, 808, 137

Lee, T. C. M. 2000, Journal of the American Statistical Association, 95, 259

Magnussen, S., Allard, D., & Wulder, M. A. 2006, Scandinavian journal of forest research, 21, 239

Møller, J. 1994, Lectures on random Voronoi tessellations, Vol. 87 (Springer Science & Business Media)