Computational Statistics and Data Analysis 92 (2015) 163–176

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A Bayesian mixture model to quantify parameters of spatial clustering

Martin Schäfer[a,*], Yvonne Radon[b], Thomas Klein[c], Sabrina Herrmann[d], Holger Schwender[a], Peter J. Verveer[e], Katja Ickstadt[d]

[a] *Mathematical Institute, Heinrich Heine University, Düsseldorf, Germany*

[b] *Department of Biomedical Sciences, School of Human Sciences, University of Osnabrück, Osnabrück, Germany*

[c] *Scottish Marine Institute Oban, Oban, United Kingdom*

[d] *Faculty of Statistics, TU Dortmund University, Dortmund, Germany*

[e] *Max Planck Institute of Molecular Physiology, Dortmund, Germany*

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 16 June 2015
Accepted 4 July 2015
Available online 13 July 2015

Keywords:
Cluster analysis
Nearest neighbor distances
Spatial statistics
Point processes
Matérn cluster process
Gamma mixture

ABSTRACT

A new Bayesian approach for quantifying spatial clustering is proposed that employs a mixture of gamma distributions to model the squared distance of points to their second nearest neighbors. The method is designed to answer questions arising in biophysical research on nanoclusters of Ras proteins. It takes into account the presence of disturbing metacluster structures as well as non-clustering objects, both common among Ras clusters. Its focus lies on estimating the proportion of points lying in clusters, the mean cluster size and the mean cluster radius without depending on prior knowledge of the parameters. The performance of the model compared to other cluster methods is demonstrated in a comprehensive simulation study, employing a specific new class of spatial point processes, the double Matérn cluster process. Further results and arguments as well as data and code are available as supplementary material.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

We propose a Bayesian mixture model for quantifying spatial clustering in the presence of non-clustering objects. The focus lies on estimating the proportion of points lying in clusters as well as the cluster size and radius. The approach is designed to answer questions arising in the context of cutting-edge biophysical research on nanoclusters of Ras proteins that cannot be answered by standard statistical approaches. It does not depend on prior knowledge of the parameters.

The most widely-used cluster approaches are likely hierarchical clustering (Johnson, 1967) and partitioning algorithms such as k-means or partitioning around medoids (PAM) (MacQueen, 1967; Kaufman and Rousseeuw, 1990). These approaches require the analyst to define a cutoff in a dendrogram, pre-define the number of clusters, or infer it by some heuristic. They also do not explicitly consider non-clustering objects, although both approaches can produce clusters of size one. These drawbacks have been tackled by more recent methods. Maitra and Ramler (2009), e.g., proposed a generalization of the k-means algorithm that explicitly considers scattered points. Some sophisticated grouping algorithms were proposed that only require specifying, e.g., a maximal cluster size (Scharl and Leisch, 2006), a minimal cluster size (Manley

[*] Correspondence to: Mathematical Institute, Heinrich Heine University, Universitätsstr. 1, D-40225 Düsseldorf, Germany. Tel.: +49 2118115526; fax: +49 2118110784.

E-mail address: schaefer@math.uni-duesseldorf.de (M. Schäfer).

http://dx.doi.org/10.1016/j.csda.2015.07.004
0167-9473/© 2015 Elsevier B.V. All rights reserved.

et al., 2008, relying on point trajectories over time) or both a minimal cluster size and an effective maximal cluster radius (Ester et al., 1996; Ankerst et al., 1999). Although such values appear somewhat easier to specify, they require important prior knowledge of the problem and their choices may strongly influence the results. Sensitivity analyses for the parameter choices, mandatory in all mentioned algorithms, cannot completely rule out an unwanted bias.

In model-based clustering, the distribution giving rise to the observations in a given parameter space is modeled by a mixture of distributions, usually Gaussian ones (McLachlan and Peel, 2000; Fraley and Raftery, 2007). Arbitrary parameter choices are avoided in this framework. While each cluster is generally represented by one mixture component, propositions have been recently made to model each cluster by several mixture components to give the model more flexibility (Baudry et al., 2010). Non-clustering observations, if considered, are generally viewed as noise rather than as a component of interest (see, e.g., Dasgupta and Raftery, 1998; Hennig and Coretto, 2008). A case can be made for fitting mixture models in a Bayesian framework (Frühwirth-Schnatter, 2006; Fritsch and Ickstadt, 2009). Among the advantages are a smaller susceptibility to problematic likelihood shapes such as local maxima or unboundedness, applicability in case of small sample sizes, and a straightforward generalization to an infinite number of mixture components (Lo, 1984). However, in the Ras application, fully model-based approaches directly modeling the protein locations are not able to identify all three parameters of interest (proportion of clustered proteins, mean cluster size and mean cluster radius) due to identifiability problems resulting from, e.g., dependences between cluster size and cluster radius.

We present GAMMICS (GAMma Mixtures for Inference on Cluster Structures), a novel compromise approach accomplishing inference for all three parameters. The core of the approach is a Bayesian model for the squared distances between proteins and their second nearest neighbors, classifying each point as clustered or non-clustered. Estimates for all parameters are obtained by incorporating some algorithmic aspects, inspired by the idea of density-based clustering originally advocated by the DBSCAN algorithm (Ester et al., 1996).

Our method is designed for the needs of biophysical research on clusters of the small GTPase Ras, a protein playing an important role in signal transduction at the plasma membrane. The spatial patterns of Ras proteins present complex and dynamic clustering behavior at the nano scale and are important for cell growth and, thus, for the development of tumors (Hancock, 2006). This makes them a target of biomedical research, an essential goal being to assess differences in the clustering behavior between distinct experimental conditions, such as healthy cells and tumor cells. A first step towards this goal is to quantify the parameters of the clustering behavior, as estimates of those parameters subsequently may be compared across different experimental conditions.

In biophysical literature, generally the K-function and derivatives have been used to estimate the mean cluster radius (see, e.g., Plowman et al., 2005; Tian et al., 2007; Kiskowski et al., 2009). However, the approach only allows to estimate this one parameter and, moreover, this approach is susceptible to severe biases. Experimental biophysical arguments based on gold-labeled particles have been used in some cases to derive estimates for the proportion of clustered proteins and the cluster size as well.

In a simulation study, our approach favorably compares to a K-function analysis, the DBSCAN algorithm, a Bayesian model-based cluster approach, as well as a mixture of the latter two, both in terms of the misclassification rate and the accuracy in parameter estimation.

This paper is organized as follows: In Section 2, we give a description of the experimental data as well as the simulation of point patterns via the double Matérn cluster process. In Section 3, we describe the GAMMICS method as well as the competing approaches. In Section 4, results of all approaches are compared, while in Section 5, some remaining challenges for future work are discussed.

2. Data

2.1. Experimental data

Ras proteins are small, measuring only 2–3 nm in diameter. In order to visualize Ras nanoclusters at the plasma membrane, fluorescence microscopy is employed. We focus on cells with an above normal expression level typical for cancer cells by overexpressing Ras tagged to a fluorescent protein. Due to the high expression level, proteins cannot be visualized and detected as single molecules by convenient fluorescence microscopy techniques. Therefore, photoactivated localization microscopy (PALM) and stochastic optical reconstruction microscopy (STORM) (Betzig et al., 2006) are used, in which photoswitchable fluorescent proteins allow a detection of single molecules at the necessary scale. Only a subset of molecules is activated by UV light at a time (Hess et al., 2006), avoiding an excess of overlap between light blobs indicating the presence of a protein. After a reasonable amount of time during which pictures are created every 100 ms, it can be expected that blobs representing a vast majority of Ras proteins in the cell have been recorded by a camera. Blobs may persist during several frames until they disappear. In some cases, proteins may be reactivated at later time points ('blinking').

In an effort to limit sources of noise, the Total internal reflection fluorescence technique (TIRF, Toomre and Manstein, 2001) ensures that only fluorophores in the basal membrane are activated, while those deeper inside the cell, i.e. those of no interest for the cluster phenomenons under investigation, are ignored.

The entire protein pattern of a cell is represented by a series of pixel images (1 pixel $= 107 \times 107$ nm^2) that first have to be denoised. Subsequently, protein coordinates need to be detected before their spatial pattern can be analyzed. Several methods have been proposed for each task. We employ the software rapidSTORM (Wolter et al., 2012) that disregards

Download English Version:

https://daneshyari.com/en/article/416426

Download Persian Version:

https://daneshyari.com/article/416426

Daneshyari.com