COMSTA: 6568 Model 3G pp. 1–13 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Penalized composite likelihoods for inhomogeneous Gibbs point process models

Jeffrey Daniel *, Julie Horrocks, Gary J. Umphrey

Department of Mathematics and Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada, N1G 2W1

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 22 December 2017
Accepted 23 February 2018
Available online xxxx

Keywords:
Gibbs point process
Composite likelihood
Regularization
Variable selection
Information criteria
Species distribution modelling

ABSTRACT

A novel general framework is presented for regularizing inhomogeneous Gibbs point process models via composite likelihood with convex penalty functions. Both penalized pseudolikelihood and a new approach based on penalized logistic composite likelihood are considered, and the selection properties and predictive performance of these two methods are evaluated in a simulation study. The use of composite information criteria for penalty tuning parameter selection is also investigated. A new criterion is proposed based on the extended regularized information criterion (ERIC), which outperforms other composite information criteria in simulations. In a species distribution modelling application, the new methods are compared to MAXENT, a popular software package that also fits regularized point process models. The models obtained using the new methods exhibit similar or better fit to the data than the MAXENT model while being sparser and more interpretable.

© 2018 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

1. Introduction

A spatial point pattern consists of a set of points representing the locations of objects or events within a spatial domain. The statistical modelling of spatial point pattern data is common in many fields, for instance in species distribution modelling in ecology (Renner et al., 2015; Funwi-Gabga and Mateu, 2012; Chakraborty et al., 2011; Warton and Shepherd, 2010). In this context, the points of the pattern represent occurrence records for a species or other taxon of interest. The point pattern is typically accompanied by spatial covariates representing environmental conditions, and the goal of analysis is to model the intensity of the species as a function of the environmental covariates in order to characterize the species' ecological niche and its geographic distribution within the study region (Elith and Leathwick, 2009; Franklin, 2010; Peterson et al., 2011).

A fundamental model for such purposes is the inhomogeneous Poisson process (IPP) (Cressie, 1993; Diggle, 2013; Baddeley et al., 2015), which assumes that points in the pattern do not interact. This independence assumption may not be appropriate; for example, if the point pattern represents a population of plants or animals, then points may be clustered due to seed dispersal or social aggregation, or points may exhibit inhibitive interaction due to competition for resources. Inhomogeneous Gibbs point processes are a broad class of models that account for both spatial inhomogeneity and interpoint interaction, and so are well-suited to modelling species–environment relationships in the presence of spatial dependence (Renner et al., 2015). IPP models can be fit via maximum likelihood, but Gibbs point process models in general have an intractable likelihood function. Instead, Gibbs point process models may be fit by maximizing a surrogate quantity called a composite likelihood (Varin et al., 2011), two examples of which are the pseudolikelihood of Besag (1977) and the logistic composite likelihood of Baddeley et al. (2014).

* Corresponding author.
E-mail address: jdaniel@uoguelph.ca (J. Daniel).

https://doi.org/10.1016/j.csda.2018.02.005
0167-9473/© 2018 Elsevier B.V. All rights reserved.

Please cite this article in press as: Daniel J., et al., Penalized composite likelihoods for inhomogeneous Gibbs point process models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.02.005.

When fitting inhomogeneous point process models to data there is the problem of deciding which covariates should be included in the final model. Regularization is an attractive procedure that performs variable selection and parameter estimation simultaneously. To achieve this, the model's (composite) likelihood is appended with a penalty function that acts to shrink small parameter values to zero. Maximizing the penalized (composite) likelihood function yields a model with only a few nonzero parameters. The degree of regularization and the sparsity of the resulting model is controlled by a tuning parameter. Typically, a regularization path of models is fit for a decreasing sequence of tuning parameter values, and the optimal model is chosen on the basis of some criterion. Since the introduction of the lasso (Tibshirani, 1996), regularization has become a very active area of research in statistics, and this interest has recently spread to spatial point process models.

MAXENT (Phillips et al., 2006; Phillips and Dudík, 2008; Phillips et al., 2017) is a software application for species distribution modelling that is very popular among ecologists, having been cited over 7000 times since 2006. This popularity is due to its superior performance relative to competing methods (Elith et al., 2006), which may be at least partially attributed to its implementation of regularization via the lasso. The model fit by MAXENT was recently shown to be equivalent to the IPP model and to logistic regression (Aarts et al., 2012; Fithian and Hastie, 2013; Renner and Warton, 2013), and this result motivated (Renner and Warton, 2013) to investigate regularization methods for IPP models and area-interaction models. Regularization of IPP models was also studied by Thurman and Zhu (2014) and Thurman et al. (2015). Yue and Loh (2015) presented an approach to regularizing IPP models and pairwise-interaction models using penalized pseudolikelihood.

Pairwise- and area-interaction models have been considered seperately in the regularized point process modelling literature, but there exist an array of other Gibbs point processes such as the hard core process and Geyer's saturation process (Geyer, 1999). Additionally, two or more Gibbs models of any kind may be combined to create hybrid models with complex interaction structures (Baddeley et al., 2013). What is needed is a unified approach to regularizing inhomogeneous Gibbs point process models with any interaction structure in order to afford modellers the greatest possible flexibility of model choice.

This paper presents a general framework for regularizing inhomogeneous Gibbs point process models via penalized composite likelihood. Our framework accommodates IPP models, pairwise-interaction models, area-interaction models, and indeed all finite inhomogeneous Gibbs point processes of exponential-family type (Geyer, 1999; van Lieshout, 2000; Møller and Waagepetersen, 2004). We also address some outstanding issues in regularized point process modelling that have not been adequately investigated in the literature. The first issue involves choice of model-fitting procedure. Maximum pseudolikelihood has been the standard method for fitting Gibbs point process models for many years, and all prior studies of regularized point process models have penalized the pseudolikelihood. This is different than MAXENT, which in its most recent implementation fits models via penalized logistic regression (Phillips et al., 2017). Although they are formally equivalent, in practice logistic composite likelihood has been found to obtain better parameter estimates than pseudolikelihood (Baddeley et al., 2014). We investigate the relative performance of these two fitting methods in a regularized setting to see if the advantage of logistic composite likelihood holds. The second issue concerns tuning parameter selection methods, which for IPP models include MAXENT's *ad hoc* procedure, cross-validation, and information criteria. Unfortunately, none of these methods are appropriate for Gibbs point process models fit via composite likelihood. Instead, we investigate the use of composite information criteria for tuning parameter selection, and we propose a new criterion, a composite analogue of the extended regularized information criterion (ERIC) (Hui et al., 2015).

The plan of the remainder of the paper is as follows. In Section 2 we define the exponential family of finite inhomogeneous Gibbs point process models to which our methods apply and provide some important examples. We also describe the pseudolikelihood and logistic composite likelihood approaches to fitting Gibbs point process models to data. In Section 3 we consider a general framework for regularizing inhomogeneous Gibbs models via penalized composite likelihood and show how regularization paths for such models may be efficiently computed using cyclical coordinate descent. We present several tuning parameter selection methods and propose the use of a hybrid estimation procedure to improve predictive performance. Section 4 presents two simulation studies: in the first we compare the performance of the tuning parameter selection methods in Section 3, and in the second we compare the performance of the penalized pseudolikelihood and penalized logistic composite likelihood approaches to regularization. In Section 5 we present an application of our methods to species distribution modelling, and in Section 6 we give some concluding remarks.

2. Inhomogeneous Gibbs point process models

2.1. Gibbs models for spatial point pattern data

In this section we briefly review the theory of Gibbs point processes based on the more comprehensive treatments in van Lieshout (2000) and Møller and Waagepetersen (2004). Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a spatial point pattern observed within a bounded spatial domain $D \subset \mathbb{R}^2$, and let $Z(u)$ be a vector of spatial covariates defined at each location $u \in D$. We treat $\mathbf{x}$ as a realization of a spatial point process X in D with intensity function $\lambda(u)$ such that

$$\mathbb{E}[N(X \cap A)] = \int_A \lambda(u) du,$$

where $N(X \cap A)$ is the number of points of X falling in any region $A \subseteq D$. We assume X to have probability density $f(\mathbf{x})$ with respect to the homogeneous Poisson point process with intensity $\lambda = 1$ such that the hereditary condition is satisfied: $f(\mathbf{x}) \geq 0$ implies $f(\mathbf{y}) \geq 0$ for all $\mathbf{y} \subseteq \mathbf{x}$. This defines the class of finite Gibbs point processes.

Please cite this article in press as: Daniel J., et al., Penalized composite likelihoods for inhomogeneous Gibbs point process models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.02.005.

UNCORRECTED PROOF

Download English Version:

https://daneshyari.com/en/article/6868698

Download Persian Version:

https://daneshyari.com/article/6868698

Daneshyari.com