Two-Dimensional Bayesian Information Criteria for Spatial Poisson Point Process (Case Study: Spatial Distribution Modeling of a Tree Species in Barro Colorado Island)

Sigit Dwi Prabowo[1], Achmad Choiruddin[2], Nur Iriawan[3]
{sigitdprabowo@gmail.com[1], choiruddin@its.ac.id[2*], nur_i@statistika.its.ac.id[3]}

Department of Statistics, Faculty of Science and Data Analystics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1,2,3]

Abstract. Species distribution modeling, where the distribution of specific species locations is connected to the environmental factors. Such data is called spatial point patterns, and the modeling is conducted based on the spatial point process. One central question is to select the best subset of such environmental factors that explain the best species distribution. Besides, the computational issue arises when numerous environmental factor is available. This paper focuses on developing a computational strategy to deal with variable selection through regularization methods for Poisson point process. In particular, two-dimensional Bayesian Information Criteria is proposed to select two types of tuning parameters. The first parameter plays the role of decreasing bias, and the second one improves the variances. Finally, the methodology is applied to tropical rain forest data in Barro Corrolado Island. The results show the adaptive elastic net regularization with the tuning parameters produces the best inhomogeneous poisson point process model.

Keywords: Two-Dimensional Bayesian Information Criteria, Regularization Methods, Species Distribution Modeling

1 Introduction

In ecology, one of the fundamental research areas is Species Distribution Modeling (SDM), where the aim is to explain the existence of a species associated with environmental factors [7]. Data on the existence of a species is included in the spatial point patterns data, and the modeling is conducted based on the spatial point process. The main focus in SDM is to choose the best covariate from the environmental factors that explain the distribution of the species.

In analyzing spatial point pattern data on SDM , the first step is etermining the intensity value. Intensity act as a first-order characteristic of spatial point processes and has been a major focus in many studies, mostly when spatial covariates rely on the estimated strength. For example, a survey of spatial modeling of tree species distribution in forests related to environmental factors [8]. A Poisson point process is a common spatial point process because computational implementation is technically simple, and the posisson likelihood function can be systematically derived. In order to obtain parameter estimation results, the poisson likelihood function was developed to suit these models with the data [2].

AECon 2020, December 19-20, Purwokerto, Indonesia

DOI 10.4108/eai.19-12-2020.2309177

The poisson likelihood function uses the regularization method to select the best covariate that describes the best model to determine the intensity value of the more optimal inhomogeneous Poisson point processes [3]. Many regularization methods have been studied previously, like ridge regression, LASSO, elastic net and other expantion, such as the regularization method that focuses on adaptive LASSO for the best model [8], then continued with research on the adaptive LASSO method that estimates value using cross-validation for optimal results [9]. Further investigation found that the adaptive Elastic-net regularization method with tuning parameter produces the best model compared to adaptive LASSO [3]. However, these studies only focus on estimating the first parameter , whereas, in the adaptive Elastic-Net regularization method, there is a second parameter that is not tuned so that the results of the estimation of the regularization method are not optimal by adjusting the parameters and where the first parameter plays a role in reducing bias and to increase the variance so that the best model is obtained. The criteria used to determine the optimal parameter tuning value are the two-dimensional Bayesian Information Criterion (BIC). BIC's advantage as the criteria for selecting the best model is that it provides a penalty for adding parameters and is suitable for large data sizes.

This paper will focus on developing computational strategies to handle variable selection through regularization methods for spatial point processes. We will select the parameters simultaneously using the two-dimensional Bayesian Information Criteria (BIC). Finally, this methodology's results were applied to the point data of the Beilschmiedia pendula Lauraceae tree species on the Barro Colorado Island.

2 Literature Review

2.1 Poisson Point Process

Let Y be a spatial point process on R^d. Let $W \subset \mathrm{R}^d$ be a compact set of Lebesgue measure $|W|$ which will play the role of the observation domain. A realization of Y, m representing number of locations observed points in W. Suppose Y has λ for intensity function and $\lambda^{(2)}$ for second-order product density [3].

A point process Y is a Poisson point process on W, if the following conditions are met:

1. if $C_1, C_2, \ldots \subseteq W$ are disjoint field, then $N(C_1), N(C_2), \ldots$ are independent variable randoms
2. for any bounded $C \subseteq W$, the number of point C, $N(C) \sim Poisson(\mu(C))$

Our research assumes that the function of intensity relies on a vector of parameters β, i.e. $\lambda(\beta)$, for the general spatial point processes model, as outlined in the introduction, maximum likelihood estimation is almost unfeasible. Instead of this approach, the Campbell formula offers an excellent tool to describe methods based on equations for estimating. In the form of spatial point processes, these approaches are now standard. The conventional parametric methods for evaluating β are obtained by maximizing the Poisson likelihood given respectively by:

$$\ell(\boldsymbol{\beta}) \approx \sum_{i=1}^{M} v_i \left\{ y_i \log \lambda(u_i; \boldsymbol{\beta}) - \lambda(u_i; \boldsymbol{\beta}) \right\} \qquad (1)$$

where $y_i = 1/v_i$.

2.2 Regularized estimating equations

The Newton Raphson algorithm used to maximize the penalized log-likelihood function can be done using the iteratively reweighted least-squares (IRLS) method, as $\ell(\boldsymbol{\beta})$ given by (1) is a concave function of the parameters. If the current estimate of the parameters is $\tilde{\boldsymbol{\beta}}$, using Taylor's expansion, we construct a quadratic approximation of the Poisson log-likelihood function [3]:

$$\ell(\boldsymbol{\beta}) \approx \ell_Q(\boldsymbol{\beta}) = -\frac{1}{2N}\sum_{i=1}^{N} \upsilon_i \left(y_i^* - z_i^T \tilde{\boldsymbol{\beta}}\right) + C(\tilde{\boldsymbol{\beta}}), \qquad (2)$$

where $C(\tilde{\boldsymbol{\beta}})$ is a constant, y_i^* are the working response values and v_i are the weights,

$$\upsilon_i = v_i \exp\left(\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}\right)$$

$$y_i^* = \boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}} + \frac{y_i - \exp\left(\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}\right)}{\exp\left(\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}\right)}$$

The regularized Poisson linear model works by first deciding a $\rho \in [\rho_{min}, \rho_{max}]$ lowering sequence, starting with a minimum value of ρ_{max} such that the entire vector $\hat{\beta} = 0$ works. An outer loop for comp $\ell_Q(\boldsymbol{\beta})$ at $\tilde{\boldsymbol{\beta}}$ is generated for every value of ρ. Secondly, to solve a penalized least square problem, a regularized technique is used.

$$\min_{\beta \in \mathrm{R}^p} \Omega(\beta) = \min_{\beta \in \mathrm{R}^p} \left\{ -\ell_Q(\boldsymbol{\beta}) + \sum_{k=1}^{p} p\rho_k \left(|\beta_k|\right) \right\} \quad (3)$$

Suppose we've got the $\tilde{\beta}_j$ calculation for $j \neq k$ $j, k = 1, 2, \ldots, q$. The method involved in partially optimizing (2) β_j relates to this

$$\min_{\beta \in \mathrm{R}^p} \Omega\left(\tilde{\beta}_1, \ldots, \tilde{\beta}_k, \tilde{\beta}_{k+1}, \ldots, \tilde{\beta}_q\right).$$

For example in the case, the penalized update by setting γ to 0 or 1 respectively for the adaptive elastic net is

$$\tilde{\beta}_k \leftarrow \frac{S\left(\sum_{i=1}^{N} \upsilon_i z_{ik}\left(y_i - \tilde{y}_i^{(k)}\right), \rho_k \gamma\right)}{\sum_{i=1}^{N} \upsilon_i {z_{ik}}^2 + \lambda_k (1-\gamma)} \qquad (4)$$

where $\tilde{y}_i^{(k)} = \tilde{\beta}_0 + \sum_{j \neq k} z_{ij} \tilde{\beta}_j$ is the fit value excluding the z_{ik} covariate contribution, and $S(z, \rho)$ is the operator of soft-thresholding with a value

$$sign(z)(|z|-\rho) = \begin{cases} z-\rho & if\ z>0\ and\ \lambda<|\rho| \\ z+\rho & if\ z<0\ and\ \lambda<|\rho| \\ 0 & \lambda \geq |\rho| \end{cases}$$

Update (4) $k = 1, 2, \ldots, p$ is replicated before convergence occurs. Methods of regularization for penalties are introduced in the `glmnet` R package [4]. For (4), $\gamma = 1$ set for adaptive lasso, while $0 < \gamma < 1$ set for adaptive elastic net.

3 Method

The spatial data used in this paper are secondary data sourced from Richard Condit et al.'s research on "Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years" regarding the location of Beilschmiedia pendula Lauraceae trees, which are located in an area of 50 hectares of forest, tropical rain Island Barro Colorado in Central Panama. Ninety-four covariates data related to environmental factors, consisting of 2 topological attributes, namely elevation and gradient; 13 soil nutrients, namely aluminium, boron, calcium, cooper, iron, potassium, magnesium, manganese, phosphorus, zinc, nitrogen, N(min) and pH; and 79 interactions between 2 soil nutrients and 2 topological points.

The method used in this paper is to estimate the parameters in the inhomogeneous Poisson point process model using the regularized maximum likelihood estimation. In regularized maximum likelihood estimation method, we will also do parameter selection for get the best model. In this method there are γ and ρ parameters that will be tuned by using two-dimensional BIC. The simulation study will then be carried out with several scenarios. Then we run the regularization method on the case of the Beilschmiedia pendula Lauraceae trees.

4 Result and Discussion

It is really worth noting that penalized procedures depend primarily on the tuning parameters ρ and γ in the adaptive elastic net, so that the choice of ρ and γ is also becoming an essential activity. The approximate value of γ from the maximum value of γ is close to 1 in the selection of γ for the adaptive elastic net system and the minimum value for γ is near 0. If the value γ selected is 1, the method switches to the adaptive lasso. In addition, the option of λ for the selected value is verified. The estimate with a huge value of ρ appears to have smaller variance, but larger biases, whereas the estimate with a small value of ρ contributes to zero biases, but greater variance. The trade-off between the biases and the variances results in an optimal option of ρ. A range of ρ values ranging from a maximum value of ρ, for which all penalized coefficients are zero to $\rho = 0$, is rational for choosing ρ. By fixing a path of ρ and γ, we select the tuning parameter ρ and γ which minimizes $BIC(\rho, \gamma)$, defined by

$$BIC(\rho, \gamma) = -2l_Q(\boldsymbol{\beta}) + s(\rho, \gamma)\log|W| \qquad (5)$$

where $s(\rho,\gamma)=\sum_{k=1}^{q}\mathrm{I}\left\{\hat{\beta}_j(\rho,\gamma)\neq 0\right\}$ is the number of selected covariates with coefficients of nonzero regression and $|W|$ is the volume of observation describing the sample size.

In simulation study, we make simulation with the spatial domain is W = [0, 1000] × [0, 500]. We centre and scale the 201 × 101 pixel images of elevation (x_1) and gradient of elevation (x_2) contained in the BPL datasets of spatstat library in R (R Core Team, 2016), and use them as two true covariates. In addition, we create scenarios to define extra covariates. We generate ninety two 201×101 pixel images of covariates as standard Gaussian white noise and denote them by $x_3,\ldots,x_{92}$. We define $\boldsymbol{z}(u)$ as the covariates vector. The regression coefficients for $z_3,\ldots,z_{92}$ are set to zero.

The mean number of points over the domain W , μ , is chosen to be 50, 500 and 1500. We set the true intensity function to be $\log\lambda(\beta_0)=\{\beta_0+\beta_1 z_1(u)+\beta_2 z_2(u)\}$, where $\beta_1=3$ represents a relatively large effect of elevation, $\beta_2=0,5$ reflects a relatively small effect of gradient, and β_0 is selected such that each realization has 50, 500 or 1500 points in average. With these scenarios, we simulate spatial point patterns from a Poisson point process using the `rpoispp` function in the spatstat package. For each of the three scenarios, we fit the intensity to the simulated point pattern realizations with 10 looping. In simulation, the regularization methods under the adaptive LASSO (AL) and adaptive elastic net (AENET) penalty were applied. For solving the estimation, the glmnet library in R was used. First for AENET, we choose value of γ with $0\le\gamma\le 1$. A quadratic approximation to the negative log-likelihood assessed in the current estimates was then generated for each value of ρ. Then, a method of regularization was introduced to solve a problem with penalized least squares. Finally, the $BIC(\rho,\gamma)$ was minimized to obtain ρ and γ.

Table 1. The selection of the regularization methods

μ	AL				AENET			
	x_1	x_2	x_3- x_{94}	x_1, x_2	x_1	x_2	x_3- x_{94}	x_1, x_2
50	0	0.5	0.9195652	0	0	0.5	0.9315217	0
500	1	1	0.9934783	1	1	1	0.9913043	1
1500	1	1	0.9956522	1	1	1	0.9956522	1

Table 1 shows the percentage of selection covariates of the regularization methods under various penalty functions in the simulation performance. The proportion of times when the actual covariates, elevation x_1, and gradient x_2 were correctly held in the selected model, and the average proportion of times when the noise covariates x_3 to x_{94} were correctly selected, are recorded for different μ values. Although the value of μ is small, the methods of regularization are not really good for the selected true covariate, but the noise covariate can be selected better than adaptive LASSO.

In application, censuses were performed in the 50-hectare area of the tropical moist forest of Barro Colorado Island, resulting in maps of tree species of Beilschmiedia pendula Lauraceae

[Hubbell et al., 2005]. It is of interest to know how the coexistence of the very large number of different tree species continues. The positions of 3,604 Beilschmiedia pendula Lauraceae (BPL) trees are of special interest to us. As a log-linear function with 2 topological attributes, 13 soil properties and 79 interactions between them as covariates, we model the intensity of BPL trees. To pick and estimate parameters, we apply the regularized Poisson probability with comparisons between adaptive LASSO and adaptive elastic net. Notice that all the covariates are based to observe which covariates have a relatively large impact on the intensity.

Table 2. Number of selected and non-selected covariates among 94 covariates

Method	Covariate Selected	Covariate Non selected	BIC
AL	56	37	40011.3
AENET	58	35	40019.3

Table 2 showed the number of covariates selected and not selected by each process. Regarding the method of regularization, the method of regularization with adaptive lasso selects 56 covariates. There are 58 covariates chosen in contrast to the adaptive elastic net methods. This implies that when the regularized Poisson probability is applied, adaptive LASSO selection and adaptive elastic net work almost equally.

5 Conclusion

Based on Poisson likelihood, we create regularized maximum likelihood estimation versions of estimating equations for estimate and select the paramater. For modeling the intensity of inhomogeneous Poisson point processes, our procedure may conduct covariate selection along with estimating it. We research the tuning parameter ρ and γ in adaptive elastic net method using BIC can make regularized Poisson likehood estimates more optimal than other regularization method.

In the simulation analysis, we carry out certain different parameters to observe the selection and prediction properties of the estimates. From the findings, we suggest applying the regularized Poisson likelihood combined with adaptive elastic net with tuning parameter when dealing with covariates that have a complex covariance matrix and when the point pattern looks very clustered. In its application to tree species of Beilschmiedia pendula Lauraceae data location, the regularized Poisson likelihood combined with adaptive elastic net with tuning parameter can estimate coefficient covaiate and choose the best covariate who have significant impact for the existence of a Beilschmiedia pendula Lauraceae tree species in the 50-hectare area of the tropical moist forest of Barro Colorado Island.

References

[1] Baddeley A, Rubak E, Turner R. Spatial Point Patterns. 2015. https://doi.org/10.1201/b19708.

[2] Berman M, Turner TR. Approximating point process likelihoods with GLIM. Insur Math Econ 1993;13:147. https://doi.org/10.1016/0167-6687(93)90845-g.

[3] Choiruddin A, Coeurjolly JF, Letué F. Convex and non-convex regularization methods for spatial point processes intensity estimation. ArXiv 2017;12:1210–55.

[4] Friedman J, Hastie T and Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22.
[5] Janine Illian, Antti Penttinen, Helga Stoyan DS. Statistical Analysis of Spatial Point Patterns. vol. 70. 2008. https://doi.org/10.1198/tech.2005.s318.
[6] Moller J, Waagepetersen RP. Statistical Inference and Simulation for Spatial Point Processes. 2003. https://doi.org/10.1201/9780203496930.
[7] Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, et al. Point process models for presence-only analysis. Methods Ecol Evol 2015;6:366–79. https://doi.org/10.1111/2041-210X.12352.
[8] Thurman AL, Zhu J. Variable selection for spatial Poisson point processes via a regularization method. Stat Methodol 2014;17:113–25. https://doi.org/10.1016/j.stamet.2013.08.001.
[9] Yue Y R and Loh J M. Variable selection for inhomogeneous spatial point process models. Can J Stat 2015;XX:1–18. https://doi.org/10.1002/cjs.