Generalizing regularized estimating equations for model selection to multiple clustered spatial point processes

*Xiaokang Liu**
Guanyu Hu

15 December 2018

Abstract

This project aims to generalize the method introduced in Thurman et al. (2015), which focus on one single clustered point process variable selection problem, to solve several clustered spatial point processes simultaneously and then conduct variable selection by group Lasso.

Introduction

The distribution pattern of one certain tree species in a forest can be modeled as a clustered spatial point process. There is a variety of factors which may affect the spatial distribution of trees, including the elevation, slope and concentration of minerals at the place where a tree stands. In point process, the first order property describes the distribution pattern of points, and it can be well characterized by intensity function. When we want to incorporate the influences from some geographical factors into the model, we can specify the intensity function as a log-linear function of the factors of interest. Furthermore, based on this log-linear form, it is possible to conduct variable selection to simplify the model and distinguish some crucial factors from some nuisance covariates. Thurman et al. (2015) proposed a method to achieve this goal by adding a penalty on the model complexity to the goodness-of-fit part to impose sparsity on the model, thus facilitate variable selection. In this project, we generalize the model of Thurman et al. (2015) to describe several clustered spatial point process simultaneously and select variables by combining information from all processes. Specifically, we model several clustered spatial point processes simultaneously based on the same set of factors, and then borrow strength from each process to get a small collection of variables that are influential to most of the processes. The model is established on the basis of independence between clustered spatial point processes condition on the predictors. The estimation and inference of the model are extentions from Thurman's paper, and it can be shown that we can write the model in a form of group lasso and solve it efficiently. Simulation studies are conducted to verify the power of variable selection of our model. In addition, we apply our model to the Barro Colorado Island data. We categorize different species of trees into several groups based on their genus. Then, for each genus we apply our method to find out the factors that are important to the growth of this kind of trees. Finally, we compare the located factors of different genus of trees to obtain the information that if different genus of trees have different flavor to its surrounding enviromental factors.

*xiaokang.liu@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

Model Formulation

Let $(\Omega, \mathcal{A}, P)$ denote a probability space and D is a spatial domain of interest. Let Y denote a mapping from $(\Omega, \mathcal{A}, P)$ to N_D, where N_D is a set of realizations y of Y and satisfy $y \cap A$ is a finite set of spatial points in D for any bounded Borel set $A \in D$. Then the mapping Y is said to be a spatial point process on D (Gaetan and Guyon 2010). For a point process Y, we can use the intensity function λ_Y to illustrate its spatial distribution. When we consider a model which allows existence of dependence among points, for example, the spatial point pattern data with clusters, we can use the Cox processes (Møller and Waagepetersen 2004) . In this project, we will focus on one example of Cox process, the Neyman-Scott process. Suppose we have a homogeneous Poission Process C which has a constant intensity function $\kappa > 0$, and let it be the parent process. Then given one event $c \in C$ from the parent process, we have a child process Y_c, which is a Poisson process with inhomogeneous intensity function $\lambda_c(s; \beta, w) = h(s - c; w) \exp\{x(s)'\beta\}$, where w is a parameter of the child process, $x(s) \in R^{p+1}$ is a vector of geographical features at position s, and $\beta = (\beta_0, \ldots, \beta_p) \in R^{p+1}$ is the corresponding coefficients vector. Then the Neyman-Scott process can be represented as $Y = \cup_{c \in C} Y_c$ with intensity function $\lambda(s; \beta) = \kappa \exp\{x(s)'\beta\}$. One classical example of Neyman-Scott process is the Thomas process which denotes

$$h(s - c, w) = \frac{1}{2\pi w^2} \exp\{-\frac{\|s - c\|_2^2}{2w^2}\},$$

where both s and c are locations. In Thomas process, the interaction parameter pairs $\theta = (\kappa, w)$ controls the spatial interactions among events. A small w leads to tighter clusters while a large w corresponding to looser clusters. When κ is small, there are fewer parent events and when κ is large, we have more parent events.

Model one single clustered point process

Let $y_1, \ldots, y_n$ denote the observed spatial point pattern data of Y in D. Note that y_i here denotes position in D. Let's first establish the likelihood function based on an inhomogeneous Poisson process, we have

$$l(\beta) = \sum_{i=1}^{n} \log \lambda(y_i; \beta) - \int_D \lambda(s; \beta) ds,$$

and its first order derivative with respect to β is

$$u(\beta) = \sum_{i=1}^{n} x(y_i) - \int_D x(s) \lambda(s; \beta) ds.$$

One way to obtain the estimates of β is to solve $u(\beta) = 0$, and it's also verified that $u(\beta) = 0$ is an unbiased estimating equation for β in a Cox process (Waagepetersen 2007). For improving computation efficiency, we focus on solving estimating equations to obtain estimates of β in this project. Denote the estimates of β from $u(\beta) = 0$ as $\widetilde{\beta}^{EE}$. However, as this model is based on the Poisson process which ignores the possible dependence among events, for regaining the efficiency, we consider the following weighted estimating equation (Guan and Shen 2010)

$$u(\beta, w) = \sum_{i=1}^{n} w(y_i) x(y_i) - \int_D w(s) x(s) \lambda(s; \beta) ds = 0,$$

which can be viewed as the first order derivative with respect to β of a weighted quasi-log-likelihood

$$l_w(\beta) = \sum_{i=1}^{n} w(y_i) \log \lambda(y_i; \beta) - \int_D w(s)\lambda(s; \beta)ds.$$

We denote $\widetilde{\beta}^{WEE}$ as the solution of $u(\beta, w) = 0$.

Next, we use quadrature approximation of the integral part in likelihood functions to help solve β (Baddeley and Turner 2000). At first, let's divide the domain of interest into a grid of rectangular pixels of number M, and let their centroids as dummy points. Thus, we have M dummy points apart from the sample points to help approximate the integral. For each pixel, it has a common area a. Then we can write the approximation form of the unweighted likelihood function as

$$l(\beta) = \sum_{i=1}^{n} \log \lambda(y_i, \beta) - \sum_{i=1}^{n+M} v_i \lambda(s_i, \beta),$$

where $v_i = a/n_i$ and n_i is the total number of sample points and dummy points in the pixel i. And similarly we have a weighted form

$$l_w(\beta) = \sum_{i=1}^{n} w_i \log \lambda(y_i, \beta) - \sum_{i=1}^{n+M} w_i v_i \lambda(s_i, \beta).$$

The performance of the quadrature approximation depends on a, when we have a finer devision the approximation will be better. Since these two approximation representations are similar to log-likelihood function for a weighted Poisson generalized linear model, we can solve them by using the a function 'glm' in R (Berman and Turner 1992).

Next, we're going to add a penalty on the model complexity to help conduct variable selection. In Thurman's paper, they use adaptive lasso penalty $\sum_{i=1}^{p} \gamma_i |\beta_i|,\ \gamma_i > 0$ to conduct model selection (Zou 2006) and get the following objective function

$$l_{pw}(\beta) = -l_w(\beta) + n \sum_{i=1}^{p} \gamma_i |\beta_i|$$

Let's denote the solutions for it to be $\widehat{\beta}^{WEE}$. In specific, they first use a Laplace approximation of $l_w(\beta)$ as

$$l_w^*(\beta) = (\beta - \widehat{\beta}^{(m-1)}) \frac{\partial l_w(\widehat{\beta}^{(m-1)})}{\partial \beta} + \frac{1}{2}(\beta - \widehat{\beta}^{(m-1)})' \frac{\partial^2 l_w(\widehat{\beta}^{(m-1)})}{\partial \beta \partial \beta'} (\beta - \widehat{\beta}^{(m-1)}),$$

where $\widehat{\beta}^{(m-1)}$ is the minimizer of $l_{pw}(\beta)$ at the $(m-1)$-th step. Then we rearrange $l_w^*(\beta)$ by writting

$$A'A = -\frac{\partial^2 l_w(\widehat{\beta}^{(m-1)})}{\partial \beta \partial \beta'},\ Y^* = (A^{-1})' \left\{ \frac{\partial l_w(\widehat{\beta}^{(m-1)})}{\partial \beta} - \frac{\partial^2 l_w(\widehat{\beta}^{(m-1)})}{\partial \beta \partial \beta'} \widehat{\beta}^{(m-1)} \right\}$$
$$X^* = Adiag(\gamma_i^{-1}),\ \beta^* = (\gamma_1 \beta_1, \ldots, \gamma_p \beta_p)'.$$

Then we have

$$l_w^*(\beta) = l_w^*(\beta^*) = -\frac{1}{2} \|Y^* - X^* \beta^*\|_2^2$$

and

$$l_{pw}(\beta) = \frac{1}{2} \|Y^* - X^* \beta^*\|_2^2 + n \|\beta^*\|_1,$$

which is exactly the form of a Lasso regression. As for the selection of γ_i, following Zou (2006) we can set $\gamma_i = \gamma \log(n)(n|\widetilde{\beta}_i^{WEE}|)^{-1},\ i = 1, \ldots, p$.

Model several clustered point processes together

Then, let's consider a multi-task learning setting. Suppose we want to model q many clustered spatial point processes $Y_1, \ldots, Y_q$ simultaneously on the same set of covariates $x_1(s), \ldots, x_p(s)$. Moreover, let's further assume these q processes are independent when given covariates. So the log-likelihood function for $Y_1, \ldots, Y_q$ given $x_1(s), \ldots, x_p(s)$ based on Poisson process can be written as

$$l(\beta^1, \ldots, \beta^q; Y_1, \ldots, Y_q) = l(\beta^1) + \ldots + l(\beta^q),$$

with

$$l(\beta^j) = \sum_{i=1}^{n_j} \log \lambda(y_{ji}; \beta^j) - \int_D \lambda(s; \beta^j) ds, \ j = 1, \ldots, q.$$

Then following the same procedure, we obtain

$$l_w^*(\beta^j) = l_w^*(\beta^{j*}) = -\frac{1}{2} \|Y_j^* - X_j^* \beta^{j*}\|_2^2$$

for each process. That is to say, for each process, we can obtain $Y_j^*, A_j, X_j^*, \beta_j^*$ and $\{\gamma_{ji}\}'s$. Then we can get the following formulation

$$\begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_q^* \end{bmatrix} = \begin{bmatrix} X_1^* & & \\ & \ddots & \\ & & X_p^* \end{bmatrix} \begin{bmatrix} \beta^{1*} \\ \beta^{2*} \\ \vdots \\ \beta^{q*} \end{bmatrix}$$

where $\beta^{j*} = (\beta_1^{j*}, \ldots, \beta_p^{j*})$. Since all processes share the same set of covariates, we know that in each β^{j*}, there is one element corresponding to the same covariate. When it aims to find out the covariates that are influential to most of the processes, we can change the penalty to a group penelty form. That is, collect all items in $\beta^{1*}, \ldots, \beta^{q*}$ that are corresponding to the same covariate to one group, which leads to p groups in our case. Then we add penalty on each group as a whole to conduct variable selection for all processes simultaneously. We can just use the sum of the ℓ_2-norm on each group as a penalty. The resulting model is the group lasso problem. Another thing that is different from Thurman's method is that, we have to adjust adaptive weights γ_{ij}'s. There are two modifications we can apply, one is by simply setting all $\gamma_{ij} = 1$, or we can further construct a new set of weights $g_i = \sqrt{\sum_{j=1}^q \gamma_{ij}^2}$ based on γ_{ij} as Zou (2005) and let the group Lasso penalty to be

$$n\gamma \sum_{i=1}^{p} g_i \sqrt{\sum_{j=1}^{q} (\beta_i^j)^2}.$$

Then let's look into the weights $w(s)$, if the process Y is m-dependent, i.e., $g(u-s) = 1$ when $\|u-s\| > m$, then we have

$$w(s) = \frac{1}{1 + \lambda(s) f(s)}, \ f(s) = K(m) - \pi m^2, \forall s.$$

The whole procedure to conduct estimation is as follows: based on the log-likelihood function, we use a quadrature approximation to approximate the integral part and then use 'glm' to solve for $\widetilde{\beta}^{EE}$. By using $\widetilde{\beta}^{EE}$, we get the value of weights and then follow the similar way to get $\widetilde{\beta}^{WEE}$. For several point process, we do the same thing to get $\widetilde{\beta}^{WEE}$ for each of them, and then use $\widetilde{\beta}^{WEE}$

as the initial value to fit the group Lasso model. Since in each iteration the selected variables may be different, so we have to update weights $w(s)$ at each iteration, then also adjust the group Lasso problem and get a new set of coefficients. This procedure will continue until some sort of convergence of coefficients is obtained. Moreover, the coefficient corresponding to the intercept in each process will remain unpenalized. For simplicity, in this paper we will only conduct the whole procedure once without iteration, and keep the results as the final estimates.

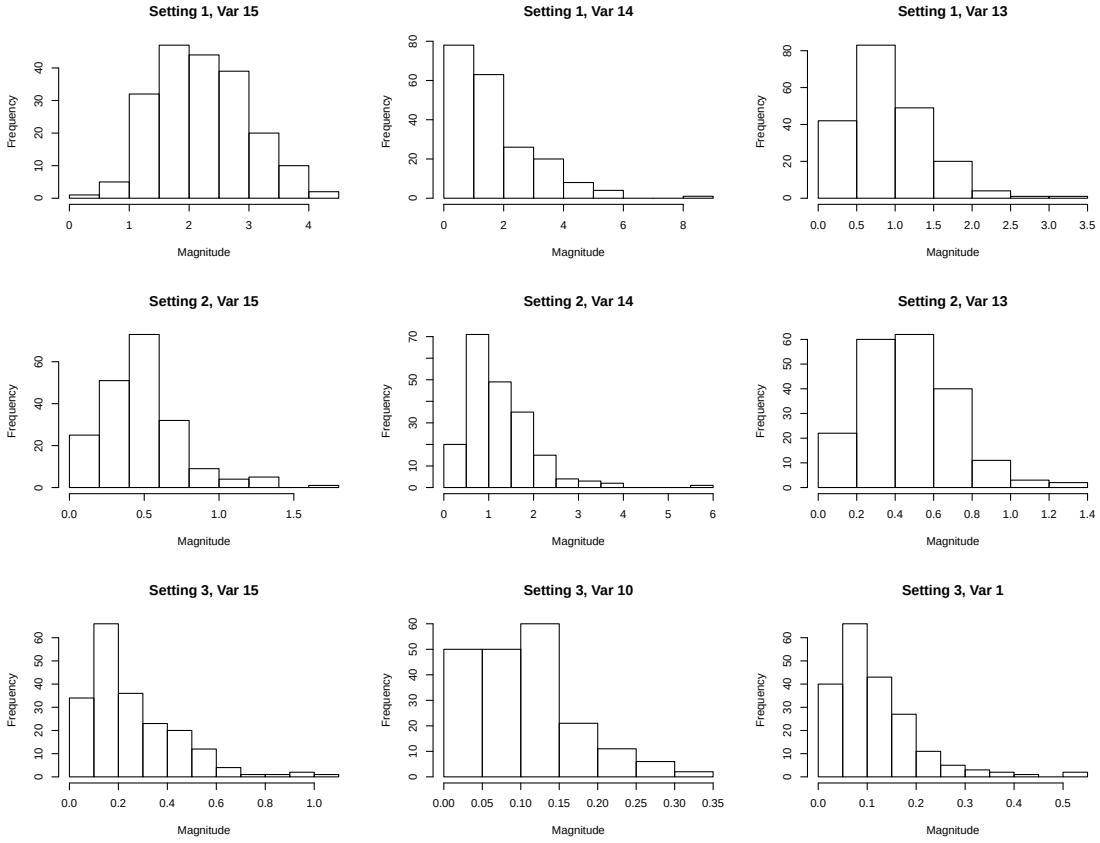
Simulation Study

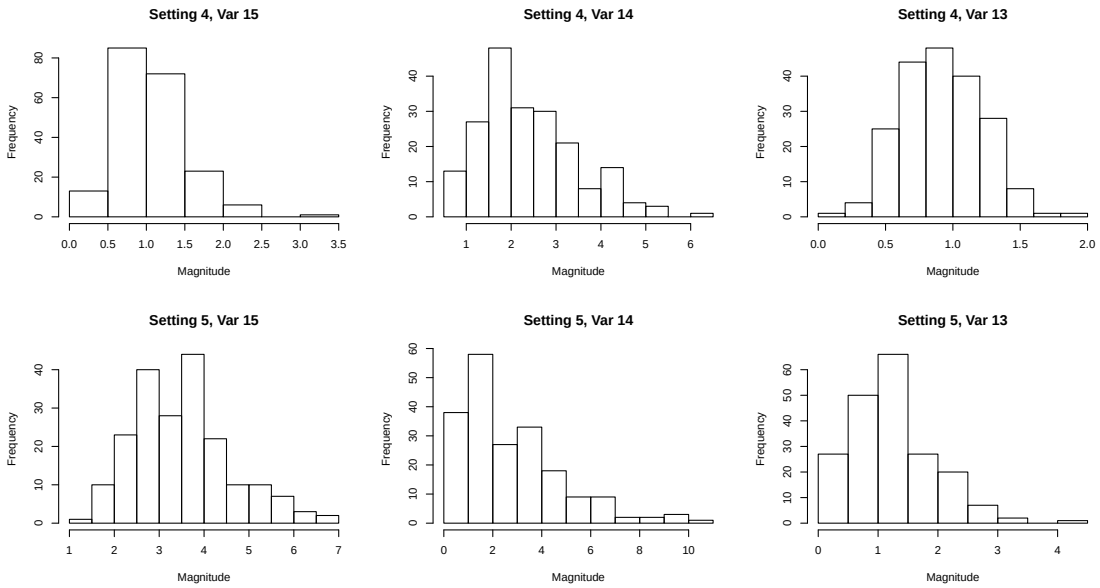
The simulation study has a setup similar to Waagepetersen (2007) and Thurman (2015). The spatial domain is $D = [0, 1000] \times [0, 500]$. From the R package 'spatstat', we get and standardize the elevation and slope data from the dataset 'bei' to be two covariates z_{14} and z_{15}. Also, we generate thirteen covariates z_j from standard normal as noise covariates, and all of them are standardized before fitting into the model. Then we add some multicollinearity among covariates into the data. That is, we define $\Sigma = V'V$ and $\Sigma[i, j] = \Sigma[j, i]) = 0.5^{|i-j|}, i, j = 1, \ldots, 15$, except $\Sigma[14, 15] = \Sigma[15, 14] = 0$, and then we take $x(s) = V'z(s)$ as the covariates we used in simulation study. That is, $x_1(s), \ldots, x_{13}(s)$ are noise predictors, and $x_{14}(s)$ is the elevation at s, $x_{15}(s)$ is the slope at location s. We consider $q = 5$ clustered point processes here. All processes share the same devision of D to conduct quadrature approximation. For each one of them, the coefficients are generated in a similar pattern.

- case 1: $\beta_{15,1} = 1.0,\ \beta_{15,2} = 1.1,\ \beta_{15,3} = 1.9,\ \beta_{15,4} = 1.5,\ \beta_{15,5} = 1.7$ and $\beta_{13,1} = 1.30,\ \beta_{13,2} = 1.20,\ \beta_{13,3} = 1.20,\ \beta_{13,4} = 1.60,\ \beta_{13,5} = 1.51$. The remaining covariates have no impact on the process.
- case 2: $\beta_{14,1} = 1.0,\ \beta_{14,2} = 1.1,\ \beta_{14,3} = 1.9,\ \beta_{14,4} = 1.5,\ \beta_{14,5} = 1.7$ and $\beta_{13,1} = 1.30,\ \beta_{13,2} = 1.20,\ \beta_{13,3} = 1.20,\ \beta_{13,4} = 1.60,\ \beta_{13,5} = 1.51$. The remaining covariates have no impact on the process.
- case 3: $\beta_{1,1} = 1.0,\ \beta_{1,2} = 1.1,\ \beta_{1,3} = 1.9,\ \beta_{1,4} = 1.5,\ \beta_{1,5} = 1.7$ and $\beta_{10,1} = 1.30,\ \beta_{10,2} = 1.20,\ \beta_{10,3} = 1.20,\ \beta_{10,4} = 1.60,\ \beta_{10,5} = 1.51$. The remaining covariates have no impact on the process.
- case 4: $\beta_{14,1} = 1.0,\ \beta_{14,2} = 1.1,\ \beta_{14,3} = 1.9,\ \beta_{14,4} = 1.5,\ \beta_{14,5} = 1.7$. The remaining covariates have no impact on the process.
- case 5: $\beta_{15,1} = 1.0,\ \beta_{15,2} = 1.1,\ \beta_{15,3} = 1.9,\ \beta_{15,4} = 1.5,\ \beta_{15,5} = 1.7$. The remaining covariates have no impact on the process.

Then we use function 'rThomas' in R package 'spatstat' to generate Thomas process under each setting. Some other parameters are $\kappa = 5 \times 10^{-5}$, $w = 30$ and $m = 10$. The estimation follows the method described before. For simplicity, we set all $\gamma_{ij} = 1$. Also, since all groups have the same group size, we don't have to add additional weights to get rid of the group size effect. Moreover, we fit the model with a path of tuning parameter γ and summarized the results obtained by fixing $\gamma = 0.147$, which is selected by empirical to make a large set of covariates to have coefficients zero. For each setting, the process is repeated 200 times. The results are summarized as histograms of squared ℓ_2 norm of estimated coefficients in each group over 200 times repetition. Some of the plots are ignored due to the space limitation.

- case 1: The histogram for slope centered at 2, which indicates that slope has significant effect. Moreover, since the correlation between slope and elevation, some of the estimates of elevation coefficient are also significantly differ from 0. As for the noise predictor $x_{13}(s)$, its estimated coefficients are located around 1. For all the other covariates, their histograms show that the magnitude of them are ignorable compared to $x_{13}(s), x_{14}(s), x_{15}(s)$, thus we didn't include their plots in this section.
- case 2: The characteristics of plots are similar to case 1.
- case 3: In this setting, two noise predictors $x_1(s)$ and $x_{10}(s)$ are assigned non-zero coefficients and all others are zero. For all the histograms, there is no strong signal among them, including $x_1(s)$ and $x_{10}(s)$. That is because the only two covariates related to non-zero coefficients are randomly generated from standard normal distribution, thus in any place of the whole domain the expected distribution intensity is the same, which leads to a homogeneous point process in the whole space. So, the proposed method cannot capture any siginificant factor under this situation.





- case 4: Here we let elevation be the only influential factor among all others. The histogram of elevation indicates that the magnitude of signal from elevation is the strongest. Also, due to correlation induced by the correlation matrix, estimates related to slope and $x_{13}(s)$ are significantly depart from 0. All the others have ignorable signal strength.
- case 5: Here we let slope be the only influential factor among all others. Similar to the analysis in case 4, slope and elevation both have large enough magnitude to be selected by our method, and $x_{13}(s)$ has signal strength slightly larger than other noise covariates due to its correlation with slope and elevation.

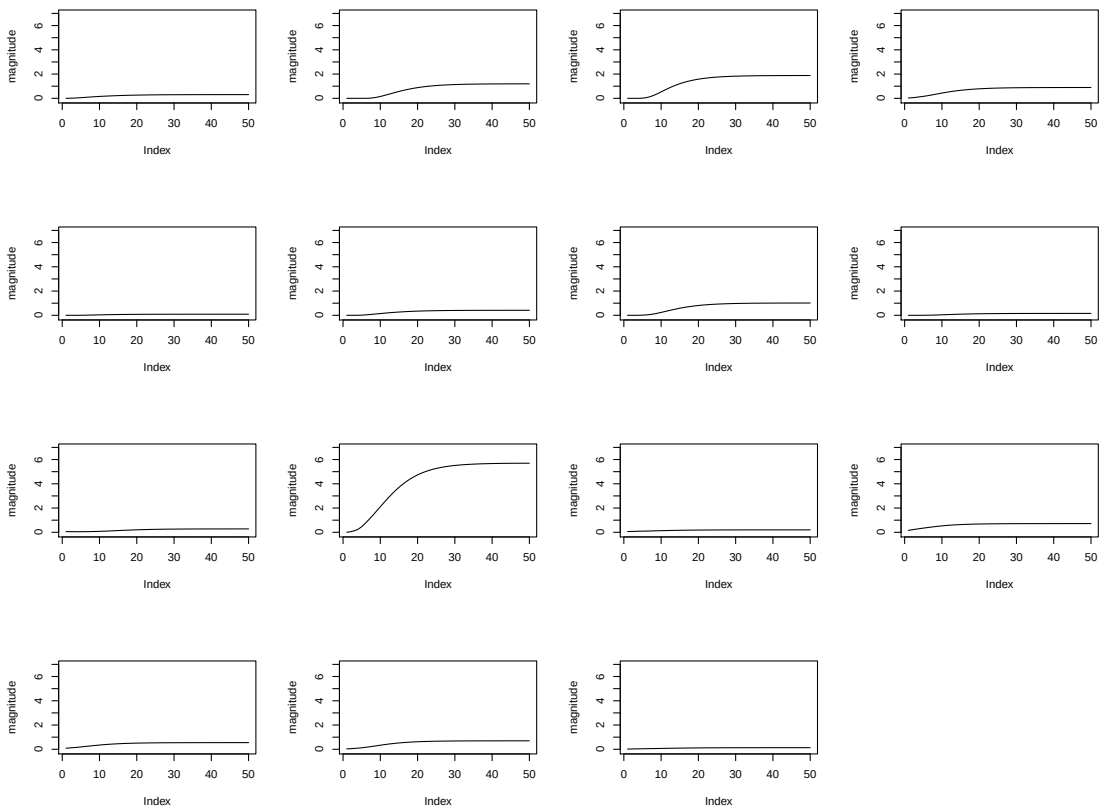
Application

In this section, we apply our model to the Barro Colorado Island data. The domain of interest is a rectangular space $[0, 1000] \times [0, 500]$ with unit meter. Seven censuses of all the trees in this region has been conducted in the past years, the information obtained including the location, height of the trees, also the species, genus or family that a tree belongs to. The considered genus of trees are (The format is like: Genus, species (mnemonic, number of trees)):

- Inga: acuminata(ingas1, 334), goldmanii(ingago, 154), marginata(ingama, 409), nobilis(ingaqu, 303), sapindoides(ingasa, 113), thibaudiana(ingath, 148), umbellifera(ingaum, 403).
- Ocotea: cernua(ocotce, 230), oblonga(ocotob, 144), puberula(ocotpu, 108), whitei(ocotwh, 203).
- Eugenia: coloradoensis(eugeco, 318), galalonensis(eugega, 1020), nesiotica(eugene, 306), oerstediana(eugeoe, 969).

Then, for each genus we apply our method to find out the factors that are important to the growth of this kind of trees. Finally, we compare the located factors of different genus of trees to obtain the information that if different genus of trees have different flavor to its surrounding enviromental

factors. Take Inga as an example, the plots for the squared ℓ_2 norm for each group estimates along the solution path are



Thus according to the strength observed from the plots, for Inga, the selected set including: 2(B),3(Ca),4(Cu),7(Mg), 10(Zn), and elevation. Similarly, for the other two genus, we have that, for Eugenia, the selected set including: 3(Ca),6(K),7(Mg) and elevation; for Ocotea, the selected set including: 2(B),3(Ca),6(K),7(Mg), elevation and slope. There is different between sets selected by different genus of trees.

Discussion

In this project, we simply generalize the method introduced in Thurman (2015) into the one that can handle several clustered point processes simulataneously and select variables by using group Lasso. This model can be used to specify the set of varibles that are influential to all the processes in a group by combining information from each one of them. One drawback of our model is the assumption that given covariates all the processes considered are independent to each other. In some situations, there are also some impacts from one process to another. For example, two species of trees in a neighborhood may compete for limited resources. Thus, it is necessary to consider models that can take into consideration of dependence between several processes.

Reference

Condit, R., Lao, S., Pérez, R., Dolins, S.B., Foster, R.B. Hubbell, S.P. 2012. Barro Colorado Forest Census Plot Data, 2012 Version. DOI http://dx.doi.org/10.5479/data.bci.20130603

Baddeley, Adrian, and Rolf Turner. 2000. "Practical Maximum Pseudolikelihood for Spatial Point Patterns." *Australian & New Zealand Journal of Statistics* 42 (3): 283–322. https://doi.org/10.1111/1467-842X.00128.

Berman, Mark, and T. Rolf Turner. 1992. "Approximating Point Process Likelihoods with Glim." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41 (1). [Wiley, Royal Statistical Society]: 31–38. http://www.jstor.org/stable/2347614.

Gaetan, Carlo, and Xavier Guyon. 2010. *Spatial Statistics and Modeling*. Springer Series in Statistics. New York, NY: Springer Science+Business Media, LLC.

Guan, Yongtao, and Ye Shen. 2010. "A Weighted Estimating Equation Approach for Inhomogeneous Spatial Point Processes." *Biometrika* 97 (4). [Oxford University Press, Biometrika Trust]: 867–80. http://www.jstor.org/stable/29777142.

Møller, Jesper, and Rasmus Plenge Waagepetersen. 2004. *Statistical Inference and Simulation for Spatial Point Processes*. Book; Book/Illustrated. Boca Raton, Fla. : Chapman & Hall/CRC.

Waagepetersen, Rasmus Plenge. 2007. "An Estimating Function Approach to Inference for Inhomogeneous Neyman-Scott Processes." *Biometrics* 63 (1). [Wiley, International Biometric Society]: 252–58. http://www.jstor.org/stable/4541321.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476): 1418–29. https://doi.org/10.1198/016214506000000735.