Regularization techniques for inhomogeneous Gibbs point process models with a diverging number of covariates

Ismaila Ba

2019-05-21

Gibbs Point Process

- ▶ Gibbs Point Processes (GPP) constitute a large class of point processes with interaction between the points.
- ▶ The interaction between the points can be repulsive or attractive. This means that GPP can model both clustering or inhibition.
- ▶ In statistical physics, (infinite) GPP are standard models for modelling systems with a large number of interacting particles.

How a realization of GPP looks like?

swedishpines
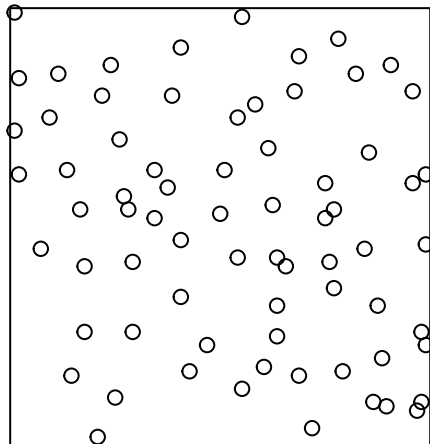
Characterization of GPP

Let $\boldsymbol{X}$ be a GPP defined on $W \subseteq \mathbb{R}^d$ (usually d=2,3). We denote by $\boldsymbol{x}$ a realization of the GPP $\boldsymbol{X}$.

▶ Probability density (exponential family):

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = c_{\boldsymbol{\theta}} \exp(\beta^\top Z(\boldsymbol{x}) + \psi^\top S(\boldsymbol{x})) \qquad (1)$$

where $\boldsymbol{\theta} = (\psi^\top, \beta^\top)^\top \in \mathbb{R}^p$ is a parameter vector to be estimated. We assume that ψ and β are of dimension respectively p_1 and p_2; and $p = p_1 + p_2$.

$c_{\boldsymbol{\theta}}$: the normalizing constant;

$\beta^\top Z(\boldsymbol{x})$: trend term describing spatial inhomogeneity through spatial covariates;

$\psi^\top S(u, \boldsymbol{x})$: interaction term describing spatial dependence among points.

Characterization of GPP

▶ Papangelou conditional intensity
For any location $u \in W$, we define the conditional intensity as follows:

$$\lambda_{\boldsymbol{\theta}}(u, \boldsymbol{x}) = \begin{cases} f(\boldsymbol{x} \cup u; \boldsymbol{\theta})/f(\boldsymbol{x}; \boldsymbol{\theta}) & \text{for} \quad u \notin \boldsymbol{x} \\ f(\boldsymbol{x}; \boldsymbol{\theta})/f(\boldsymbol{x} \setminus u; \boldsymbol{\theta}) & \text{for} \quad u \in \boldsymbol{x} \end{cases} \tag{2}$$

with $a/0 := 0$ for $a \geq 0$. For exponential family of GPP, (2) becomes

$$\lambda_{\boldsymbol{\theta}}(u, \boldsymbol{x}) = \exp(\beta^\top Z(u) + \psi^\top S(u, \boldsymbol{x}))$$

More on the conditional intensity

- $\lambda_\theta(u, \boldsymbol{x})\mathrm{d}u$ may be interpreted as the probability that has the process $\boldsymbol{X}$ to insert a point in a region $\mathrm{d}u$ around u that also respects the already existing configuration outside this region.

- At points u close to the edge of W, $\lambda_\theta(u, \boldsymbol{x})$ may depend on unobserved points of $\boldsymbol{x}$. This is called the boundary effect problem and to handle this, we use the minus sampling $D := W \ominus R$, where $R < \infty$ is the range of $\boldsymbol{X}$.

- $\boldsymbol{X}$ has the finite range property if

$$\lambda_\theta(u, \boldsymbol{x}) = \lambda_\theta(u, \boldsymbol{x} \cap B(u, R))$$

where $B(u, R)$ is the ball centered at u and radius $0 < R < \infty$. This means that the process does not allow the points to interact from each other except they are at a distance less or equal than R apart.

Likelihood Inference on GPP

- The normalizing constant c_θ is generally intractable for GPP.
- The latter point makes difficult to fit Gibbs models via maximum likelihood.
- A reasonable alternative is to maximize the pseudolikehood function which does not involve the normalizing constant.

The log-pseudolikelihood function is defined in D by

$$\mathbf{LPL}(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{u \in \boldsymbol{x} \cap D} \log \lambda_{\boldsymbol{\theta}}(u, \boldsymbol{x}) - \int_D \lambda_{\boldsymbol{\theta}}(u, \boldsymbol{x}) \mathrm{d}u. \qquad (3)$$

Regularization techniques

The penalized log-Pseudolikelihood function:

$$Q(\boldsymbol{x}; \boldsymbol{\theta}) = \textbf{LPL}(\boldsymbol{x}; \boldsymbol{\theta}) - |D| \sum_{j=1}^{p} p_{\lambda_j}(|\theta_j|). \qquad (4)$$

- ▶ We look to maximize the function Q.
- ▶ We do not make penalizations on the parameter ψ. The second term in (4) becomes

$$|D| \sum_{j=p_1+1}^{p} p_{\lambda_j}(|\theta_j|).$$

Penalty functions

- ℓ_1 norm: $p_\lambda(\theta) = \lambda\theta$,
- ℓ_2 norm: $p_\lambda(\theta) = \frac{1}{2}\lambda\theta^2$,
- Elastic net: for $0 < \gamma < 1$, $p_\lambda(\theta) = \lambda\left\{\gamma\theta + \frac{1}{2}(1-\gamma)\theta^2\right\}$,
- SCAD: for any $\gamma > 2$,

$$p_\lambda(\theta) = \begin{cases} \lambda\theta & \text{if} \quad \theta \leq \lambda \\ \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1} & \text{if} \quad \lambda \leq \theta \leq \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if} \quad \theta \geq \gamma\lambda, \end{cases}$$

- MC+: for any $\gamma > 1$, $p_\lambda(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{if} \quad \theta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if} \quad \lambda \leq \theta \leq \gamma\lambda. \end{cases}$

We also consider adaptive version of the convex penalty functions, i.e. adaptive lasso and adaptive elastic net.

Numerical method

From Baddeley and Turner (2000), we have the following finite sum approximation:

$$\int_D \lambda_\theta(u, \boldsymbol{x}) \mathrm{d}u \approx \sum_{j=1}^{n+m} w_j \lambda_\theta(u_j, \boldsymbol{x})$$

$$\textbf{LPL}(\boldsymbol{x}; \boldsymbol{\theta}) \approx \sum_{j=1}^{n+m} w_j (y_j \log \lambda_\theta(u_j, \boldsymbol{x}) - \lambda_\theta(u_j, \boldsymbol{x})) \tag{5}$$

where $y_j = \frac{1}{w_j}$ for $u_j \in \boldsymbol{x}$ and $y_j = 0$ otherwise. The quadrature weights w_j sum to the volume of D i.e. $\sum_{j=1}^{n+m} w_j = |D|$.

(5) corresponds to a quasi Poisson log-likelihood function.

This approximation often requires large m to perform well.

Numerical method

$$Q(\boldsymbol{x}; \boldsymbol{\theta}) \approx \underbrace{\sum_{j=1}^{n+m} w_j(y_j \log \lambda_\theta(u_j, \boldsymbol{x}) - \lambda_\theta(u_j, \boldsymbol{x}))}_{\text{ppm function}} - \underbrace{|D| \sum_{j=1}^{p} p_{\lambda_j}(|\theta_j|)}_{+ \text{ (glmnet or ncvreg) function}}$$

For the implementation of the method in R, we need three functions or R packages:

▶ ppm function in the spatstat package;

▶ glmnet function for convex penalty functions;

▶ ncvreg function for non convex penalty functions. Note that a modification of the ncvreg function is required to include options for weights.

Main results

- We present the asymptotic properties of the regularized pseudolikelihood estimator.

- The eroded observation domain D expands to $\mathbb{R}^d$, i.e. $D = D_n$, $n = 1, 2, \ldots$

- Let $\lambda = \lambda_{n,j}$, $\mathbf{LPL} = \mathbf{LPL}_n$ and $Q = Q_n$ be indexed by n.

- The dimension of the parameter θ is now of dimension p_n which diverges as $n \to \infty$.

- Let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^\top, \boldsymbol{\theta}_{02}^\top)^\top = (\boldsymbol{\theta}_{01}^\top, \mathbf{0}^\top)^\top$ be the p_n-dimensional vector of true coefficients, where $\boldsymbol{\theta}_{01}$ is of dimension $p_1 + s$ and $\boldsymbol{\theta}_{02}$ of dimension $p_n - p_1 - s$.

- $\boldsymbol{\theta}_{01}$ and $\boldsymbol{\theta}_{02}$ are respectively the vector of nonzero coefficients and that of zero coefficients.

Main results

Theorem

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$$

where $\hat{\boldsymbol{\theta}}$ is a local maximizer of $Q_n(\boldsymbol{X}; \theta)$. This means that the regularized pseudolikelihood estimator is root-$(|D_n|/p_n)$ consistent.

Theorem

(i) *Sparsity:* $\mathrm{P}(\hat{\boldsymbol{\theta}}_2 = 0) \to 1$ *as* $n \to \infty$,

(ii) *Asymptotic Normality:*
   $$|D_n|^{1/2}\boldsymbol{\Sigma}_n(\boldsymbol{X}; \boldsymbol{\theta}_0)^{-1/2}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}) \xrightarrow{d} \mathcal{N}(0, \mathsf{I}_m),$$

where $m = p_1 + s$.

Some comments on the results

- The above results are also valid for the unregularized pseudolikelihood function.
- Necessary conditions to satisfy for the two main results do not hold for the Lasso, Rigde and Elastic Net regularization methods. This prevents us for applying the two mentionned theorems for these methods.

Simulation study

- We conduct a simulation study with three differents scenarios.
- We consider the inhomogeneous Strauss model with $\gamma = .5$.
- To investigate the performance of the estimation, we compute the empirical bias, variance and mean squarred error (MSE).
- We also investigate the performance of the selection through the empirical true positive rate (TPR) and false positive rate (FPR). TPR corresponds to the ratio of the selected true covariates over the number of the true covariates, while FPR corresponds to the ratio of the selected noisy covariates over the number of noisy covariates.
- For the selection of the tuning parameter, we consider two criteria: the composite bayesian information criterion (cBIC) and the composite extented regularized information criteria (cERIC).

Simulation set-up

- ▶ The setting is simalar to that of Choiruddin et al. (2018).
- ▶ The spatial domain is $D = [0, 1000] \times [0, 500]$. The 201×101 pixel images of elevation (z_1) and gradient of elevation (z_2) contained in the **bei** datasets of **spatstat** library in **R** are first centered and scaled; and used as true two covariates.
- ▶ We set the true conditional intensity to be $\lambda_\theta(u; \boldsymbol{x}) = \exp\{\beta_0 + \beta_1 z_1(u) + \beta_2 z_2(u) + \psi S(u; \boldsymbol{x})\}$ where $\beta_1 = 2$, $\beta_2 = 0.75$ and $\psi = \log(0.5)$. β_0 is chosen such that each realization has in average 4000 points with the inhomogeneous Poisson model i.e when $\psi = 0$.
- ▶ In addition to the two covariates, we create three different scenarios to define extra covariates.

Simulation set-up

▶ Scenario 1: We generate 48 201 $\times$ 101 pixel images of covariates as standard Gaussian white noise and denote them by $x_3(u), \ldots, x_{50}(u)$. We define $\boldsymbol{z}(u) = \{1, x_1(u), \ldots, x_{50}\}$ as the covariates vector. The regression coefficients for $z_3, \ldots, z_{50}$ are set to zero.

▶ Scenario 2: We first generate 48 201 $\times$ 101 pixel images of covariates as in Scenario 1. Second, we transform them, together with x_1 and x_2, to have multicolinearity. In particular, we define $\boldsymbol{z}(u) = \boldsymbol{V}^{\top}\boldsymbol{x}(u)$, where $\boldsymbol{x}(u) = \{x_1(u), \ldots, x_{50}\}^{\top}$. More precisely, $\boldsymbol{V}$ is such that $\boldsymbol{\Omega} = \boldsymbol{V}^{\top}\boldsymbol{V}$, and $(\Omega)_{ij} = (\Omega)_{ji} = 0.7^{|i-j|}$ for $i, j = 1, \ldots, 50$ except $(\Omega)_{12} = (\Omega)_{21} = 0$, to preserve the correlation between x_1 and x_2. The regression coefficients for $z_3, \ldots, z_{50}$ are set to zero.

Simulation set-up

- ▶ Scenario 3: We center and scale the 13 50×25 pixel images of soil nutrients covariates obtained from the study in tropical forest of Barro Colorado Island (BCI) in central Panama, convert them to be 201×101 pixel images as x_1 and x_2, and use them as the extra covariates. Together with x_1 and x_2, we keep the structure of the covariance matrix to preserve the complexity of the situation. In this setting, we have $z(u) = \{1, x_1(u), \ldots, x_{15}\}^\top$. The regression coefficents for $z_3, \ldots, z_{15}$ are set to zero.

Simulation results (Scenario 1, BIC)

Table 1: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 1 with BIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.41	0.01	0.18	1.96	100
Ridge	0.56	0.04	0.36	100	100
Enet	0.61	0.02	0.39	7.79	100
ALasso	0.09	0.01	0.02	0.04	100
AEnet	0.12	0.01	0.03	0	100

Simulation Results (Scenario 1, ERIC)

Table 2: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 1 with ERIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.41	0.01	0.18	1.81	100
Ridge	0.56	0.04	0.35	100	100
Enet	0.61	0.01	0.38	9.08	100
ALasso	0.18	0.01	0.04	0	100
AEnet	0.18	0.01	0.04	0	100

Simulation results (Scenario 2, BIC)

Table 3: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 2 with BIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.4	0.32	0.48	38.08	100
Ridge	1.39	0.08	2.03	100	100
Enet	1.45	0.02	2.12	7.46	100
ALasso	0.1	0.06	0.07	20.83	100
AEnet	0.1	0.06	0.07	21.58	100

Simulation results (Scenario 2, ERIC)

Table 4: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 2 with ERIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.58	0.11	0.45	11.69	100
Ridge	1.43	0.08	2.11	100	100
Enet	1.39	0.02	1.96	8.75	100
ALasso	0.37	0.02	0.16	1.15	100
AEnet	0.36	0.03	0.17	2.17	100

Table 5: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 3 with BIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.38	0.06	0.21	14.31	100
Ridge	0.74	0.04	0.59	100	100
Enet	0.27	0.15	0.23	59.31	100
ALasso	0.09	0.01	0.02	0	100
AEnet	0.16	0.02	0.04	0.08	100

Simulation results (Scenario 3, ERIC)

Table 6: Empirical prediction properties (Bias, Var and MSE) and empirical selection properties (TPR, FPR in %) based on 100 replications of inhomogeneous Strauss process on the domain D for Scenario 3 with ERIC criterion.

	Biais	Var	MSE	FPR	TPR
Lasso	0.45	0.03	0.24	6.69	100
Ridge	0.75	0.04	0.61	100	100
Enet	0.47	0.1	0.32	36	100
ALasso	0.23	0.01	0.06	0	100
AEnet	0.24	0.01	0.07	0.08	100

References

Baddeley, Adrian, and R. Turner. 2000. "Practical Maximum Pseudolikelihood for Spatial Point Patterns." *Aust. N. Z. J. Stat* 42 (3): 283–322.

Choiruddin, Achmad, Jean-François Coeurjolly, Frédérique Letue, and others. 2018. "Convex and Non-Convex Regularization Methods for Spatial Point Processes Intensity Estimation." *Electronic Journal of Statistics* 12 (1): 1210–55.