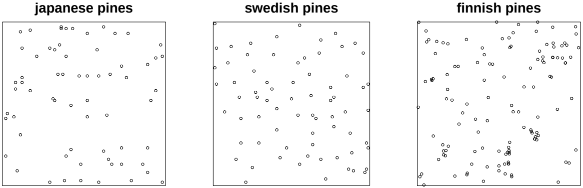
Regularization techniques for inhomogeneous Gibbs point process models with a diverging number of covariates

Ismaila Ba (with Jean François Coeurjolly)

2019-05-29

Gibbs Point Process

- Gibbs Point Processes (GPP) constitute a large class of point processes with interaction between the points.
  $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_i \in W \subseteq \mathbb{R}^d$, (usually d=2,3), n random.
- The interaction between the points can be repulsive or attractive. This means that GPP can model both clustering or inhibition.



Characterization of GPP

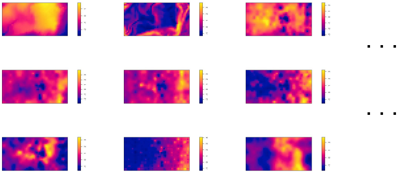
- Density with respect to a Poisson Process, say $f(\boldsymbol{x};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter vector to estimate;
- Papangelou conditional intensity $\lambda_{\boldsymbol{\theta}}(u,\boldsymbol{x})$ defined for any location $u \in W$ as follows:

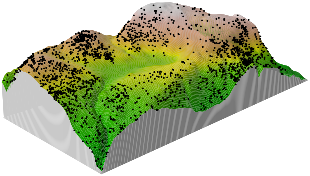
$$\lambda_{\boldsymbol{\theta}}(u,\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x} \cup u;\boldsymbol{\theta})/f(\boldsymbol{x};\boldsymbol{\theta}) & \text{for} \quad u \notin \boldsymbol{x} \\ f(\boldsymbol{x};\boldsymbol{\theta})/f(\boldsymbol{x} \setminus u;\boldsymbol{\theta}) & \text{for} \quad u \in \boldsymbol{x} \end{cases} \tag{1}$$

with $a/0 := 0$ for $a \geq 0$.

Data: Barro Colorado Island (Hubell et al., 1999, 2005)

- $D = [0, 1000m] \times [0, 500m]$
- $> 300{,}000$ locations of trees
- ≈ 100 spatial covariates observed at fine scale (altitude, nature of soils,...)





Density of GPP [Daniel et al., 2018]:

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

$Z(\boldsymbol{x}) = \sum_{v \in \boldsymbol{x}} Z(v)$, $\boldsymbol{\psi} \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_2}$,
$\boldsymbol{\theta} = (\psi^\top, \beta^\top)^\top \in \mathbb{R}^p$,
$Z(v) = (Z_1(v), \dots, Z_{p_2}(v))^\top$ and
$S(\boldsymbol{x})$, the interaction function.
Problem: p_2 large, covariates very correlated, $c_{\boldsymbol{\theta}}$ intractable.

Estimation of $\boldsymbol{\theta}$ when p_2 is large and $c_{\boldsymbol{\theta}}$ intractable

- The conditional intensity:

$$\lambda_{\boldsymbol{\theta}}(u, \boldsymbol{x}) = \exp(\beta^\top Z(u) + \psi^\top S(u, \boldsymbol{x})) \tag{2}$$

- The log-Pseudolikehood function:

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

- The penalized log-Pseudolikelihood function:

$$Q(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbf{LPL}(\boldsymbol{x}; \boldsymbol{\theta}) - |D| \sum_{j=p_1+1}^{p} p_{\lambda_j}(|\theta_j|). \tag{4}$$

$$\boxed{\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\ Q(\boldsymbol{x}; \boldsymbol{\theta})}$$

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

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

Numerical method

From Baddeley, Rubak, and Turner (2015), we have the following finite sum approximation:

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

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

$$Q(\boldsymbol{x}; \boldsymbol{\theta}) \approx \underbrace{\sum_{j=1}^{n+m} w_j (y_j \log \lambda_{\boldsymbol{\theta}}(u_j, \boldsymbol{x}) - \lambda_{\boldsymbol{\theta}}(u_j, \boldsymbol{x}))}_{\textbf{spatstat}} - \underbrace{|D| \sum_{j=p_1+1}^{p} p_{\lambda_j}(|\theta_j|)}_{\textbf{glmnet} \text{ or } \textbf{ncvreg}}$$

where $y_j = \frac{1}{w_j}$ for $u_j \in \boldsymbol{x}$ and $y_j = 0$ otherwise.
This approximation often requires large m to perform well.

Main results

- D expands to $\mathbb{R}^d$, i.e. $D = D_n$, $n = 1, 2, \ldots$
- $\lambda = \lambda_{n,j}$, $\mathbf{LPL} = \mathbf{LPL}_n$ and $Q = Q_n$.
- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top = (\boldsymbol{\theta}_1^\top, \mathbf{0}^\top)^\top$, $\boldsymbol{\theta}_1 \in \mathbb{R}^{p_1+s}$, $\boldsymbol{\theta}_2 \in \mathbb{R}^{p_n-p_1-s}$.
- Define the sequences

$$
\begin{aligned}
a_n &= \max_{j=1,\ldots,s} |p'_{\lambda_{n,j+p_1}}(|\beta_{0j}|)|, \\
b_n &= \inf_{j=p_1+s+1,\ldots,p_n} \inf_{\substack{|\theta| \le \epsilon_n \\ \theta \neq 0}} p'_{\lambda_{n,j}}(\theta), \text{ for } \epsilon_n = K_1 \sqrt{\frac{p_n}{|D_n|}}, \\
c_n &= \max_{j=1,\ldots,s} |p''_{\lambda_{n,j+p_1}}(|\beta_{0j}|)|
\end{aligned}
$$

where K_1 is any positive constant.

Main results

Theorem 1 [BaCoeurjolly19+]

- Under some assumptions such that it works ...
- $a_n = O(|D_n|^{-1/2})$, $c_n = o(1)$.

Then there exists $\hat{\boldsymbol{\theta}}$ such that

$$\boxed{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_{\mathrm{P}}(\sqrt{p_n}(|D_n|^{-1/2} + a_n))}$$

Theorem 2 [BaCoeurjolly19+]

- Under some assumptions such that it works ...
- $p_n^3/|D_n| \to 0$, $a_n\sqrt{|D_n|} \to 0$, $b_n\sqrt{|D_n|/p_n^2} \to \infty$.

Then, as $n \to \infty$

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

(ii) Asymptotic Normality:

$$|D_n|^{1/2}\boldsymbol{\Sigma}_n(\boldsymbol{X};\boldsymbol{\theta})^{-1/2}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_m),$$

where $m = p_1 + s$.

Values of a_n, b_n and c_n for some given regularization methods

Possible? $\Leftrightarrow a_n\sqrt{|D_n|} \to 0$ and $b_n\sqrt{|D_n|/p_n^2} \to \infty$

- Lasso:

$$a_n = b_n = \lambda_n, \quad c_n = 0 \quad \text{and} \quad \textbf{Possible?} = \textbf{NO}.$$

- Ridge:

$$a_n = \lambda_n \max_{j=1,\dots,s}\{|\beta_{0j}|\}, \quad b_n = 0, \quad c_n = \lambda_n \quad \text{and} \quad \textbf{Possible?} = \textbf{NO}$$

- Adaptive Lasso:

$$a_n = \max_{j=1,\dots,s}\{\lambda_{n,j+p_1}\}, \quad b_n = \inf_{j=p_1+s+1,\dots,p_n}\{\lambda_{n,j}\}, \quad c_n = 0$$

and **Possible?**=**YES**.

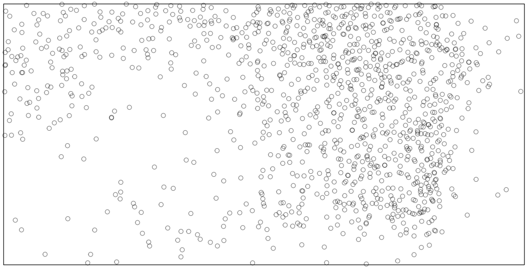
Example of a GPP

- $\boldsymbol{X}$: Inhomogeneous strauss point process:

$$\lambda(u; \boldsymbol{x}) = \beta(u)\gamma^{s_R(u;\boldsymbol{x})} \quad \text{where} \quad s_R(u; \boldsymbol{x}) = \sum_{v \in \boldsymbol{x}} 1(\|u - v\| \leq R).$$

- In log-linear form, $\beta(u) = \exp(\beta^\top Z(u))$ and $\gamma = \exp(\psi)$.
- $\boldsymbol{x}$ in $D = [0, 1000] \times [0, 500]$ with $Z(u) = (\underbrace{Z_1(u), Z_2(u)}_{\text{BCI cov.}})^\top$,

$\beta_1 = 2$, $\beta_2 = 0.75$, $\gamma = 0.5$ and $R = 12$.



Simulation study (similar to Choiruddin et al. (2018))

- $D = [0, 1000] \times [0, 500]$; $\boldsymbol{X}$: Strauss model; $m = 500$ replications
- $\lambda_{n,j}$ is chosen using **BIC**-type criterion for composite likelihood

$$cBIC(\lambda) = -2\mathbf{LPL}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_\lambda) + \log(n) tr(\hat{J}_\lambda \hat{H}_\lambda) \quad \text{where}$$

$$\hat{J}_\lambda = \text{var}(\mathbf{LPL}^{(1)}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_\lambda)) \quad \text{and} \quad \hat{H}_\lambda = \mathbb{E}(-\mathbf{LPL}^{(2)}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_\lambda)).$$

- $Z(u) = (\underbrace{Z_1(u), Z_2(u)}_{\text{true BCI cov.}}, \underbrace{Z_3(u), \dots, Z_{50}}_{\text{noisy correlated cov.}})^\top$
- Z_i's are then standardized, $\beta_1 = 2$, $\beta_2 = .75$ and $\psi = \log(.5)$

1170 points in average			
	TPR (%)	FPR (%)	MSE
Lasso	100	34	0.4
A. Lasso	100	22	0.05
A. Enet	100	23	0.07

References

Ba, Ismaila, and Jean François Coeurjolly, "Regularization techniques for inhomogeneous Gibbs models with a diverging number of covariates", in preparation.

Daniel, Jeffrey, Julie Horrocks, and Gary J Umphrey. 2018."Penalized Composite Likelihoods for Inhomogeneous Gibbs PointProcess Models."Computational Statistics & Data Analysis124:104–16.

Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2015. *Spatial Point Patterns: Methodology and Applications with R*.

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.