

Regularization techniques for spatial point processes intensity estimation

Jean-François Coeurjolly

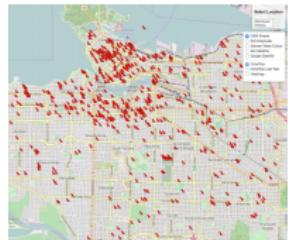
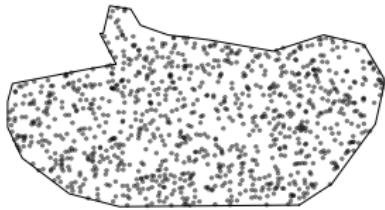
(joint work with A. Choiruddin, F. Letu )



UQAM | **D partement de math matiques**
FACULT  DES SCIENCES
Universit  du Qu bec   Montr al

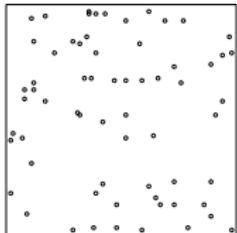
Point processes / Point patterns

- Spatial point pattern : (locally finite random measure)
 $\mathbf{x} = \{x_1, \dots, x_n\}, x_i \in W \subset \mathbb{R}^d$ (e.g. $d = 2, 3$), $n = \text{random}$.
- Patterns¹ can be : homogeneous or inhomogeneous

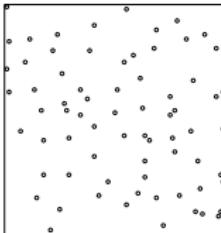


- ...can exhibit independence, neg. or pos. dependence

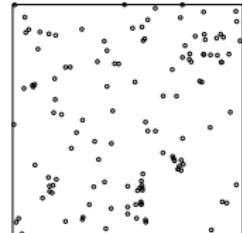
japanese pines



swedish pines



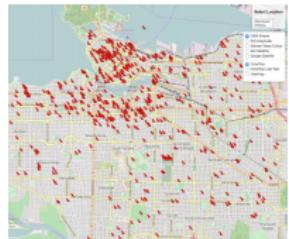
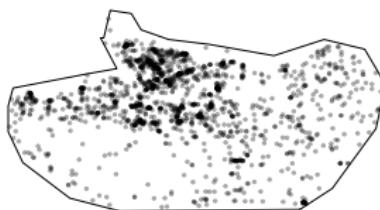
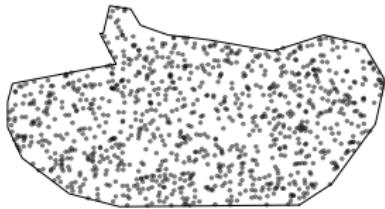
finnish pines



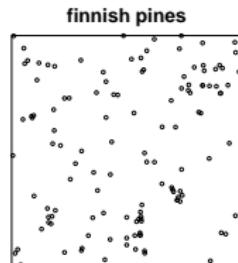
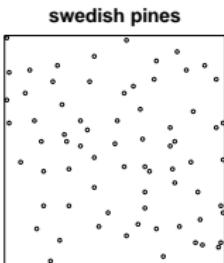
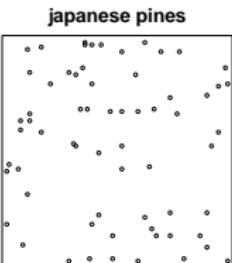
1. courtesy Jim O'Leary, crows attacks in Vancouver in 2017

Point processes / Point patterns

- Spatial point pattern : (locally finite random measure)
 $\mathbf{x} = \{x_1, \dots, x_n\}$, $x_i \in W \subset \mathbb{R}^d$ (e.g. $d = 2, 3$), $n = \text{random}$.
- Patterns¹ can be : homogeneous or **inhomogeneous**



- ...can exhibit **independence, neg. or pos. dependence**



1. courtesy Jim O'Leary, crows attacks in Vancouver in 2017

How to characterize or summarize a point process ?

- finite-dimensional distributions of counting variables ;
- void probabilities : $P(N(B) = 0)$, for any compact set B ;
- via a density (with respect to a Poisson process) ;
- intensity functions : $\boxed{\rho}$, $\rho^{(2)}$, $\rho^{(3)}$, ...
- summary statistics : Ripley's K function, functions F, G, J, L, \dots

Intensity function : $\rho : W \rightarrow \mathbb{R}^+$

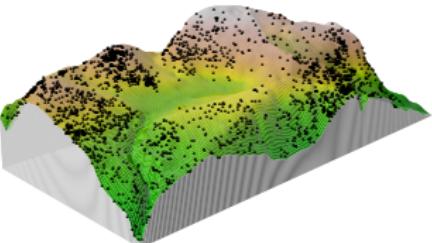
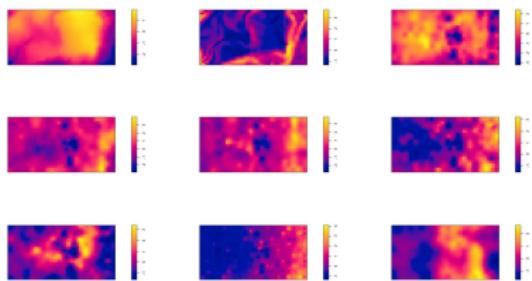
$\rho(u) \approx P(\mathbf{X} \text{ has a point in } B(u, du))$

$$E \sum_{u \in \mathbf{X}} h(u) = \int h(u) \rho(u) du, \quad [\text{Campbell Theorem}]$$

Objective : parametric modeling ρ

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

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



Modeling ρ : for one species e.g.

$$\log \rho(u; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)),$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{z}(u) = (z_1(u), \dots, z_p(u))^\top$$

Problem : p large, covariates very correlated.

Estimation of β when p is moderate (1)

$$\hat{\beta} = \operatorname{argmax}_{\beta} \ell(\beta), \quad \ell(\beta) = \sum_{u \in \mathbf{X} \cap W} \underbrace{\beta^\top \mathbf{z}(u)}_{\log \rho(u; \beta)} - \int_W \underbrace{\exp(\beta^\top \mathbf{z}(u))}_{\rho(u; \beta)} du$$

- ℓ = log-likelihood if \mathbf{X} is a Poisson point process ;
- but $\ell^{(1)}(\beta)$ remains an *unbiased estimating equation* for general point processes. Indeed,

$$\ell^{(1)}(\beta) = \sum_{u \in \mathbf{X}} \mathbf{z}(u) - \int_W \mathbf{z}(u) \rho(u; \beta) du$$

so $E\ell^{(1)}(\beta) = 0$ by Campbell Theorem (under the true model).



Estimation of β when p is moderate (1)

$$\hat{\beta} = \operatorname{argmax}_{\beta} \ell(\beta), \quad \ell(\beta) = \sum_{u \in \mathbf{X} \cap W} w(u) \underbrace{\beta^\top \mathbf{z}(u)}_{\log \rho(u; \beta)} - \int_W w(u) \underbrace{\exp(\beta^\top \mathbf{z}(u))}_{\rho(u; \beta)} du$$

- ℓ = log-likelihood if \mathbf{X} is a Poisson point process ;
- but $\ell^{(1)}(\beta)$ remains an *unbiased estimating equation* for general point processes. Indeed,

$$\ell^{(1)}(\beta) = \sum_{u \in \mathbf{X}} w(u) \mathbf{z}(u) - \int_W w(u) \mathbf{z}(u) \rho(u; \beta) du$$

so $E\ell^{(1)}(\beta) = 0$ by Campbell Theorem (under the true model).

- Nothing changes if we add a weight surface $w(\cdot)!!$

Estimation of β when p is moderate (2)

- $\hat{\beta} \rightarrow \beta$ and is asympt. normal ($W = W_n \rightarrow \mathbb{R}^d$) for \dots ;
- asympt. cov. matrix can be optimized in terms of $w(\cdot)$.
- Bermann-Turner approximation : discretize

$$\int_W w(u) \rho(u; \beta) du \approx \sum_{i=1}^{n+m} q_i \rho(u_i; \beta)$$

m : # dummy points ; n : # data points ; q_i quadrature weights. Then, with $y_i = q_i^{-1} \mathbf{1}(u_i \in \mathbf{X})$

$$\ell(\beta) \approx \sum_{i=1}^{n+m} q_i \left\{ y_i \log \rho(u_i; \beta) - \rho(u_i; \beta) \right\} \stackrel{\text{R}}{=} \underbrace{\text{glm}(\dots, \text{family}=quasipoisson)}_{\text{spatstat package}}$$

- BT approx. can be avoided (but $\text{Var}(\hat{\beta}') > \text{Var}(\hat{\beta})$)

$$\ell'(\beta) = \sum_{u \in \mathbf{X} \cap W} \log \frac{\rho(u; \beta)}{\delta + \rho(u; \beta)} - \delta \int_W \log \frac{\rho(u; \beta) + \delta}{\delta} du \approx \dots \stackrel{R}{=} \text{glm}(\dots, \text{binomial})$$

When p is large (even $p = p_n$, $W = W_n \rightarrow \mathbb{R}^d$)

- Penalization techniques : $\hat{\beta} = \operatorname{argmax}_{\beta} Q(\beta)$ where

$$Q(\beta) = \ell(\beta) - |W_n| \sum_{j=1}^{p_n} \pi_{\lambda_{n,j}}(|\beta_j|)$$

- ↪ $\lambda_{n,j} \geq 0$ are regularization parameters
- ↪ $\pi_\lambda(\cdot)$: penalty function convex or non-convex
 $\|\cdot\|_1, \|\cdot\|_2, \dots$ SCAD, MC+, ...
- ↪ Example : adaptive Lasso $\pi_{\lambda_{n,j}} = \lambda_{n,j} |\beta_j|$.

When p is large (even $p = p_n$, $W = W_n \rightarrow \mathbb{R}^d$)

- Penalization techniques : $\hat{\beta} = \operatorname{argmax}_{\beta} Q(\beta)$ where

$$Q(\beta) = \ell(\beta) - |W_n| \sum_{j=1}^{p_n} \pi_{\lambda_{n,j}}(|\beta_j|)$$

- ↪ $\lambda_{n,j} \geq 0$ are regularization parameters
- ↪ $\pi_\lambda(\cdot)$: penalty function convex or non-convex
 $\|\cdot\|_1, \|\cdot\|_2, \dots$ SCAD, MC+, ...
- ↪ Example : adaptive Lasso $\pi_{\lambda_{n,j}} = \lambda_{n,j} |\beta_j|$.

- Computational point of view : thanks to the BT approximation

$$\begin{aligned}\min(-Q(\beta)) &= \min(-\ell(\beta) + \text{penalty}) \\ &= \text{convex} + \text{convex/non-convex} \\ &\stackrel{\mathbb{R}}{=} \text{spatstat} + \text{glmnet / ncvreg}\end{aligned}$$

What can we prove? (well expected results!)

- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, $\boldsymbol{\beta}_1 \in \mathbb{R}^s$, $\boldsymbol{\beta}_2 = 0 \in \mathbb{R}^{p_n - s}$.
- for the *adaptive lasso* : let

$$a_n = \max_{j=1,\dots,s} \lambda_{n,j}, \quad b_n = \min_{j=s+1,\dots,p_n} \lambda_{n,j}.$$

Theorem ($|W_n| \rightarrow \infty$, $p_n \rightarrow \infty$) [Choiruddin, C., Letué'18]

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

Then, as $n \rightarrow \infty$

$$\boxed{P(\hat{\boldsymbol{\beta}}_2 = 0) \rightarrow 1 \quad \text{and} \quad \hat{\Sigma}_n^{-1/2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} N(0, I_s)}$$

Technical difficulties : (compared to standard regr. type problems)

- asymptotic is different.
- most importantly : points are **dependent**.

For other penalties

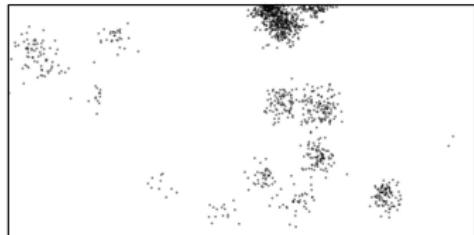
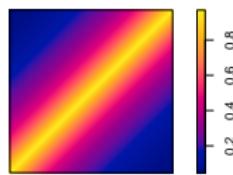
Possible ? $\Leftrightarrow a_n \sqrt{|W_n|} \rightarrow 0$ and $b_n \sqrt{|W_n|/p_n^2} \rightarrow \infty$

Method	a_n	b_n	Possible ?
Ridge	$\lambda_n \max_{j=1,\dots,s} \{ \beta_{0j} \}$	0	✗
Lasso	λ_n	λ_n	✗
Enet	$\lambda_n [(1 - \alpha) \max_{j=1,\dots,s} \{ \beta_{0j} \} + \alpha]$	$\lambda_n \alpha$	✗
ALasso	$\max_{j=1,\dots,s} \{\lambda_{n,j}\}$	$\inf_{j=s+1,\dots,p} \{\lambda_{n,j}\}$	✓
Aenet	$\max_{j=1,\dots,s} \{\lambda_{n,j}((1 - \alpha) \beta_{0j} + \alpha)\}$	$\alpha \inf_{j=s+1,\dots,p} \{\lambda_{n,j}\}$	✓
SCAD	0^*	λ_n^*	✓
MC+	0^*	$\lambda_n - \frac{K}{\gamma \sqrt{ D_n }}^*$	✓

* if $\lambda_n \rightarrow 0$ and $\lambda_n \sqrt{|W_n|/p_n^2} \rightarrow \infty$.

Short simulation

- $W = [0, 1000] \times [0, 500]$; \mathbf{X} : Thomas process (generates clustered patterns); $m = 5,000$ replications.
- $\lambda_{n,j} = \lambda / |\tilde{\beta}_j|$, $\tilde{\beta}$ is the non-regularized estimator (or ridge); λ is chosen using BIC-type criterion (CV is dangerous in this setting).
- $\mathbf{z}(u) = (\underbrace{z_1(u), z_2(u)}, \underbrace{z_3(u), \dots, z_{100}(u)}_{\text{true BCI cov.}})^T$ noisy correlated cov.
- the z_i 's are then centered and reduced;
 $\beta_1 = 2, \beta_2 = .75$.



1600 points in average

	FPR (%)	FNR (%)	RMSE
Lasso	97	23	0.73
A. Lasso	95	3	0.63
SCAD	96	7	0.65

Additional remarks and perspectives

- Applied to the **bei** dataset - selection of 8 (relevant) covariates (among 93) ; higher AUC.
- current work : compare AL with **Dantzig type selector**

$$\text{Minimize} \quad \sum_j \lambda_{n,j} |\beta_j| \quad \text{subject to} \quad \|\ell^{(1)}(\boldsymbol{\beta})\| \leq \lambda_{n,j}$$

- 300 species of trees ? **high-dimensional multivariate LGCP** (much more challenging) : for $i = 1, \dots, 300$

$$\Lambda_i(u) = \rho(u; \boldsymbol{\beta}_i) \exp Z_i(u) \quad Z_i(u) = \sum_{l=1}^q \alpha_{il} E_l(u) + U_i(u)$$

- ↪ E_l, U_i are indep. stationary Gaussian random fields ;
- ↪ typically approx. 10,000 interaction parameters