Regularization techniques and point processes

Jean-François Coeurjolly

(joint works with I Ba, A Choiruddin, T Espinasse, AL Fougères, F Lavancier, F Letué, F Cuevas-Pacheco, MH Descary, J Møller and R Waagepetersen)

UQÀM

Université Grenoble Alpes

GDR GeoSto

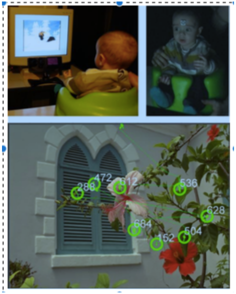
Online Seminar on Spatial and spatio-temporal Point processes and beyond (OSSP)

Journées MAS, Rouen 2022

1. Examples of (high-dimensional) spatial point patterns
2. Standard models and methods
3. Regularization techniques
4. Other approaches/problems
5. Conclusion/perspectives

Table of Contents

Eye-movement data (1)



Eye-movement (on an image or video) is composed of

- sacades : exploratory step, local, very quick 120ms.
- fixations ($< 1°$ of oscillation) ; analysing fixations allows to understand how a subject explores an image ; locations of fixations as well as their number are random.

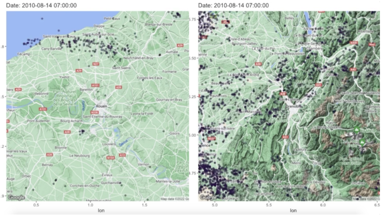
Oculo-nimbus project (Univ. Grenoble) : aim to understand mechanisms of newborns vision

- Dozens of images
- Newborns of 3-, 6-, 9- and 12-month + adults control group
- $\simeq$ 40 subjects per age group
- $\simeq$ 15 – 20 fixations by subject


adults
12 months
9 months
6 months
3 months

Lightning strikes in France (2)

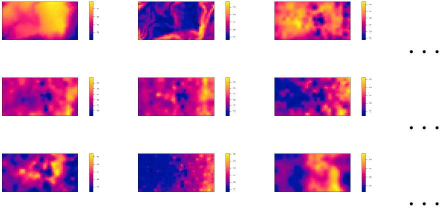
- Spatio-temporal point process : the time event as well as the location are random.

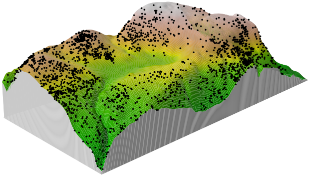


- Observed with plenty of spatial and spatio-temporal covariates : topography, wind direction/speed, population density type covariates, urbanization, . . .
- Highly challenging in particular since the number of points is very large (more than 2 millions).

Data : Barro Colorado Island (Hubell et al., 1999, 2005) (3)

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





- Even for one species of trees : how to relate locations of trees to $\mathbf{z}_1, \dots, \mathbf{z}_p$?
- <u>Problem :</u> p large, covariates very correlated.

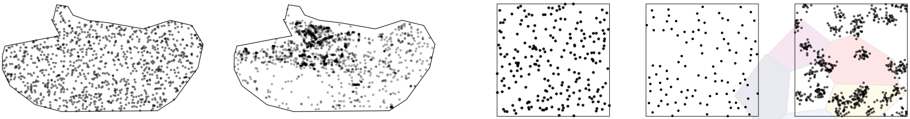
Table of Contents

Intensity and conditional intensity functions

- Let $\mathbf{X}$ be an SPP on $S \subseteq \mathbb{R}^d$; view $\mathbf{X}$ as a locally finite random measure ;
- A realization is of the form :

$$\mathbf{x} = \{x_1, \dots, x_n\}, x_i \in W \subset \mathbb{R}^d (\text{e.g. } d = 2, 3)$$

  where x_i and n are random ; W domain of observation with volume $|W|$ (note that S can be $\neq, =$ W)
- Observed patterns can be homogeneous/inhomogeneous **and/or** exhibit independence between points or dependence (clustering **and/or** repulsion)



Intensity and conditional intensity functions

Or I should say Campbell vs Georgii-Nguyen-Zessin theorems

1. Let $h : S \to \mathbb{R}$ (s.t. ...)

$$\mathrm{E} \sum_{u \in \mathbf{X}} h(u) = \int h(u)\rho(u)\mathrm{d}u$$

2. Let $h : S \times N_{lf} \to \mathbb{R}$ (s.t. ...)

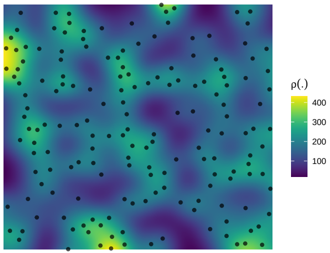
$$\mathrm{E} \sum_{u \in \mathbf{X}} h(u, \mathbf{X} \setminus u) = \mathrm{E} \int h(u, \mathbf{X})\lambda(u, \mathbf{X})\mathrm{d}u$$

- ρ and λ are respectively the first-order intensity function and the (first-order) Papangelou conditional intensity function.

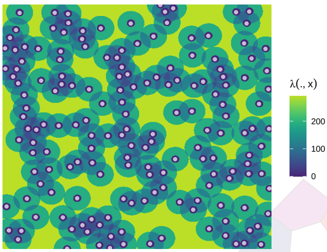
Interpretation

- Taking h as indicator functions we may interprete
  1. $\rho(u)\mathrm{d}u \approx$ Probability to observe a point in the vicinity of u.
  2. $\lambda(u, \mathbf{x})\mathrm{d}u \approx$ Probability to observe a point in the vicinity of u given the rest of the configuration is x.

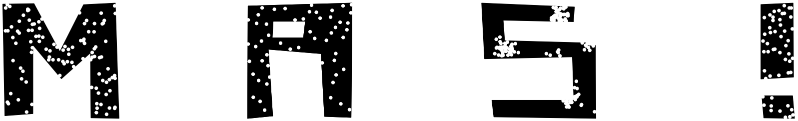
Intensity $\rho(u)$



Conditional intensity $\lambda(u, \mathbf{x})$



Are ρ and/or λ explicit for standard models ?



Model	Type of interaction	Is $\rho(\cdot)$ explicit ?	Is $\lambda(u, \mathbf{x})$ explicit ?
Poisson	**no interaction**	$\surd$	$\surd$
Gibbs	**attraction/repulsion**	$\times$	$\surd$
Cox	**attraction**	$\surd$	$\times$
DPP	**repulsion**	$\surd$	$\surd$ $\times$

(1st-order inhomogeneous) parametric models

Standard models = exponential family models

- Intensity function : $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{z}(u) = (z_1(u), \dots, z_p(u))^\top$, $z_i : S \to \mathbb{R}$

$$\rho(u) = \exp\left(\boldsymbol{\beta}^\top \mathbf{z}(u)\right)$$

- Papangelou conditional intensity function : $\boldsymbol{\psi} \in \mathbb{R}^l$, $\boldsymbol{\beta} \in \mathbb{R}^p$; $\mathbf{S}(u, \mathbf{x}) = (s_1(u, \mathbf{x}), \cdots, s_l(u, \mathbf{x}))^\top$=interaction terms.

$$\lambda(u, \mathbf{x}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u) + \boldsymbol{\psi}^\top \mathbf{S}(u, \mathbf{x}))$$

Examples

- $s_1(u, \mathbf{x}) = \sum_{v \in \mathbf{x}} g(\|v - u\|)$ PIPP ; $g(r) = \mathbf{1}(r \in (0, R))$=Strauss.
- $s_1(u, \mathbf{x}) = |A(\mathbf{x} \cup u)| - |A(\mathbf{x})|$ where $A(\mathbf{x}) = \cup_{v \in \mathbf{x}} B(v, R) =$ area-interaction model
- Geyer saturation model, piecewise Srauss models,...

Aside . . .existence point process models defined on $S \subseteq \mathbb{R}^d$ with **prescribed ρ or λ ?**

- Obvious for ρ : Poisson, LGCP, Neymann-Scott point processes, DPP,. . .
- for $\lambda \Leftrightarrow$ existence of inhomogeneous GPP on the infinite volume ; challenging probabilistic question even when $p = 1, z_1(u) = 1$!

Aside . . . existence point process models defined on $S \subseteq \mathbb{R}^d$ with **prescribed ρ or λ ?**

- Obvious for ρ : Poisson, LGCP, Neymann-Scott point processes, DPP,. . .
- for $\lambda \Leftrightarrow$ existence of inhomogeneous GPP on the infinite volume ; challenging probabilistic question even when $p = 1, z_1(u) = 1$!

Proposition (C., Dereudre, Vasseur('21))

Let $\lambda : \mathbb{R}^d \times N_{lf} \to \mathbb{R}^+$, finite-range (FR) and local stability (LS) assumptions, then there exists at least one infinite volume Gibbs measure, i.e. Gibbs model $\mathbf{X}$, with Papangelou conditional intensity λ.

- FR : $\lambda(u, \mathbf{x}) = \lambda(u, \mathbf{x} \cap B(u, R))$ for some $R < \infty$
- LS : $\lambda(u, \mathbf{x}) \leq \bar{\lambda}$ uniformly
- Very simple to check : ok for Strauss, area-interaction, Geyer,. . .

Standard parametric methodology

- Assume we observe a single realization $\mathbf{x}$ of $\mathbf{X}$ on W.
- To estimate ρ : Poisson likelihood (composite likelihood)

$$\mathrm{PL}_\rho = \sum_{u \in \mathbf{x} \cap W} \log \rho(u) - \int_W \rho(u) \mathrm{d}u$$

- To estimate λ : Pseudolikelihood

$$\mathrm{PL}_\lambda = \sum_{u \in \mathbf{x} \cap W} \log \lambda(u, \mathbf{x} \setminus u) - \int_W \lambda(u, \mathbf{x}) \mathrm{d}u$$

Remarks

- PL_ρ is the likelihood under the Poisson case, but $\mathrm{PL}_\rho^{(1)}$ remains an estimating equation for general PP.
- [Jensen and Møller'92] PL_λ is the limit of a product of conditional densities ;

Comments :

- For ρ and λ (when $p = 1$, i.e. stationary case)
  - Asymptotic results well-established as $|W_n| \to \mathbb{R}^d$ [Guan and Loh'07, Guan and Waagepetersen'09] [Billot, C. and Drouilhet'08].
  - Weighted versions : [Guan and Shen'14] [C., Guan, Khanmohammadi and Waagepetersen'16]
  - Not restricted to exponential family models [Guan and Waagepeterser'09], [Coeurjolly and Drouilhet'10], and for GPP to non hereditary models [Dereudre and Lavancier'09]
- Specifically for ρ : optimal estimation (quasilikelihood) [Guan, Jalilian and Waagepetersen'15], mispecified models and infill asymptotics [Choiruddin, C. and Waagepetersen'20]
- Specifically for λ : results valid for $p > 1$ [Ba and Coeurjolly'20]

Other alternatives : (incomplete)

- Palm likelihood (for ρ) [Prokešová and Jensen'13], variational approach [Baddeley and Dereudre'13] [C. and Møller'14], logisitc regression likelihoods [Waagepetersen'07] [Baddeley, C., Rubak and Waageepetersen'14], . . .

Computational aspects

- To evaluate PL_ρ (in terms of $\boldsymbol{\beta}$) or PL_λ (in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$) we have to discretize $\int_W \rho(u)\mathrm{d}u$ or $\int_W \lambda(u,\mathbf{x})\mathrm{d}u$
- Bermann-Turner approximation : [Baddeley and Turner'00]

$$\int_W \rho(u)\mathrm{d}u \ \text{ or } \ \int_W \lambda(u,\mathbf{x})\mathrm{d}u \approx \sum_{i=1}^{n+m} q_i \mu(u_i)$$

  where
  - $\mu(u_i) = \rho(u_i)$ or $\lambda(u_i,\mathbf{x})$
  - n = # data points ; q_i quadrature weights ;
  - m=dummy points ; $m >> n$
- Then, with $y_i = q_i^{-1}\mathbf{1}(u_i \in \mathbf{X})$

$$\mathrm{PL}_\rho \text{ or } \mathrm{PL}_\lambda \approx \sum_{i=1}^{n+m} q_i \Big\{ y_i \log \mu(u_i) - \mu(u_i) \Big\} \stackrel{\mathtt{R}}{=} \underbrace{\mathtt{glm(\ldots,family=quasipoisson)}}_{\mathtt{spatstat\ package}}$$

Logistic regression as a computational alternative

Definition of the contrast [Waagepetersen'07][Baddeley et al'14]

$$\mathrm{LR}_\bullet = \sum_{u \in \mathbf{x} \cap W} \log\left(\frac{\mu(u)}{\nu + \mu(u)}\right) - \int_W \nu \log\left(\frac{\nu}{\nu + \mu(u)}\right) \mathrm{d}u$$

where $\mu(u) = \rho(u)$ or $\lambda(u, \mathbf{x})$ when $\bullet = \rho$ or λ.

- When ν is large, $\mathrm{LR}_\bullet \approx \mathrm{PL}_\bullet$
- Interest : if we discretize the integral using only dummy points s.t. $m = \nu|W|$; with u_i = data point $(i = 1, \dots, n)$ or dummy point $i = n+1, \dots, n+m$

$$\mathrm{LR}_\rho \text{ or } \mathrm{LR}_\lambda \approx \sum_{i=1}^{n} \log\left(\frac{\mu(u_i)}{\nu + \mu(u_i)}\right) - \sum_{j=1}^{m} \log\left(\frac{\nu}{\nu + \mu(u_{j+n})}\right)$$

$$\overset{\mathtt{R}}{=} \underbrace{\texttt{glm(...,family=binomial,offset=-log(nu))}}_{\texttt{spatstat package, ppm(...,method='logi')}}$$

Selection criteria : How to select among models ?

$$\mathcal{M}_l = \left\{\rho(\cdot;\boldsymbol{\beta}_l) \text{ or } \lambda(\cdot,\mathbf{x};\boldsymbol{\beta}_l) \mid \boldsymbol{\beta}_l = \left\{\beta_0, (\beta_j)_{j\in I_l}\right\} \in \mathbb{R}^{p_l}\right\}, \quad l = 1,\dots,2^p.$$

where $\boldsymbol{\beta}_l$ (and eventually $\boldsymbol{\psi}$) is estimated using $\mathrm{PL}_{\rho,l}$ or $\mathrm{PL}_{\lambda,l}$

Criteria [Choiruddin, C. and Waagepetersen'20]

- Composite Akaike's information type criterion

$$\mathrm{CIC}_{\bullet,l} = -2\,\widehat{\mathrm{PL}_{\bullet,l}} + 2\widehat{p_l^*}$$

  where $\widehat{p_l^*}$ estimates $p_l^* = \mathrm{Tr}\left(\mathbf{S}^{-1}\boldsymbol{\Sigma}\right)$, $\mathbf{S} = -\mathrm{E}\left(\mathrm{PL}_{\bullet,l}^{(2)}\right)$ $\boldsymbol{\Sigma} = \mathrm{Var}\left(\mathrm{PL}_{\bullet,l}^{(1)}\right)$

- Composite Bayesian information criterion

$$\mathrm{CBIC}_{\bullet,l} = -2\,\widehat{\mathrm{PL}_{\bullet,l}} + \widehat{p_l^*}\log(n)$$

  where n is the observed number of points.

Note that under the Poisson case, $p_l^* = p_l$.

On the BCI dataset ...

Estimation of ρ (using PL_ρ or LR_ρ) :

- `bei` dataset : $W = [0, 1000] \times [0, 500]$; $n \simeq 3000$ locations of trees ;
- 93 spatial covariates (single and interaction terms) : some of them are highly correlated and/or are little informative ;
- Standard method combined with a naive selection procedure (e.g. stepwise based on some criterion) :
  - very expensive from a computational point of view (more than 10 hours using a forward/backward stepwise procedure using a CBIC type criterion ;
  - some of the investigated models produced numerical errors ;
  - of course : all coefficients are $\neq 0$! Signs are incoherent with expertise, ...

$\Rightarrow$ make sense to investigate a simultaneous selection/estimation procedure, especially if we assume **sparsity**.

Table of Contents

References (Absolutely not exhaustive)

$d = 1$:

- Lasso Poisson : [Reynaud-Bouret'03] [Ivanoff, Picard and Rivoirard] (lasso and group lasso - methodology and concentration inequalities)
- Multivariate Hawkes point processes : [Hansen, Reynaud-Bouret and Rivoirard'15] (concentration inequalities)

$d > 1$:

- Methodology for ρ : [Thurman, Fu, Guan and Zhu'15] (adaptive lasso , PL_ρ)
- Methodology for λ : [Yue and Loh'15], [Daniel, Horrocks and Umphrey'18] (adaptive lasso, enet for PL_λ and LR_λ)
- For multivariate point patterns : GPP [Rajala, Murrell and Olhede'17], LGCP [Choiruddin, Cueva-Pacheco, C. and Waagepetersen'20]

Contributions : Methodology and theoretical results for

- several contrasts (including PL or LR), large class of PP, convex/non-convex penalty functions, Dantzig selector, ...

Context and penalized criteria

- Single observation of an SPP on $(W_n)_{n\geq 1}$, $W_n \to \mathbb{R}^d$ as $n \to \infty$
- Sparse model with a diverging number of parameters : we assume $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top\right)^\top = \left(\boldsymbol{\beta}_1^\top, \mathbf{0}^\top\right)^\top$ with $\boldsymbol{\beta}_1 \in \mathbb{R}^s$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_n - s}$ and where $p_n \to \infty$.
- We define : $\hat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}} Q_\rho$, or $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) = \mathrm{argmax}_{\boldsymbol{\beta},\boldsymbol{\psi}} Q_\lambda$ where

$$\boxed{Q_\bullet = \mathrm{PL}_\bullet - |W_n| \sum_{j=1}^{p_n} \pi_{\lambda_{n,j}}(|\beta_j|) \quad \text{or} \quad Q_\bullet = \mathrm{LR}_\bullet - |W_n| \sum_{j=1}^{p_n} \pi_{\lambda_{n,j}}(|\beta_j|)}$$

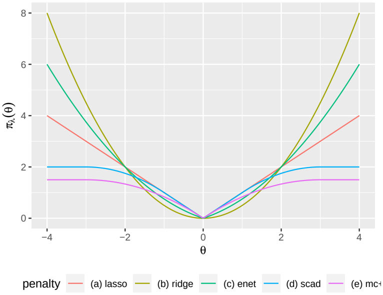
(for GPP we do not penalize $\boldsymbol{\psi}$)

$\hookrightarrow$ $\lambda_{n,j} \geq 0$ are regularization parameters

$\hookrightarrow$ $\pi_\lambda(\cdot)$: penalty function

Examples of convex and non-convex penalties

- lasso or ridge : $\lambda_{n,j} = \lambda_n$, $\pi_\lambda(\theta) = \lambda|\theta|$ or $\lambda\theta^2/2$
- elastic net : $\lambda_{n,j} = \lambda_n$, $\pi_\lambda(\theta) = \lambda(\alpha|\theta| + (1-\alpha)\frac{1}{2}\theta^2)$
- adaptive lasso $\pi_{\lambda_{n,j}}(\theta) = \lambda_{n,j}|\theta|$



- SCAD penalty : $\gamma > 2$, $\pi_\lambda(\theta) = \begin{cases} \lambda\theta & \text{if} \quad \theta \le \lambda \\ \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1} & \text{if} \quad \lambda \le \theta \le \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if} \quad \theta \ge \gamma\lambda, \end{cases}$

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

Computational aspects

- For exponential familty models : PL_ρ, PL_λ, LR_ρ, LR_λ are convex functions of $\boldsymbol{\beta}$ or $(\boldsymbol{\beta},\boldsymbol{\psi})$
- Hence, thanks to Bermann-Turner approximation, we can take advantage of GLMs adapted procedures !

$$
\begin{aligned}
\min(-Q_\bullet) &= \min\left(\underbrace{-\mathrm{PL}_\bullet \text{ or } -\mathrm{LR}_\bullet} + \text{penalty} \right) \\
&= \text{convex} \quad + \quad \text{convex/non-convex} \\
&\overset{\mathtt{R}}{=} \mathtt{spatstat} + \mathtt{glmnet} \ / \ \mathtt{ncvreg}
\end{aligned}
$$

How to tune the $\lambda_{n,j}$?

- Standard procedure ([Zou et al] : $\lambda_{n,j} = \dfrac{\lambda}{|\hat\beta_j|^\gamma}$ where $\hat\beta_j$ is the PL$_\bullet$ or LR$_\bullet$ estimate, γ = extra parameter often set to 1.
- So the question is how to tune λ ?
  - Ideas from Bootstrapping/resampling techniques (see OSSP talk by O. Cronnie)
  - Extend standard criterions : let CL$_\bullet$ = PL$_\bullet$ or LR$_\bullet$. Select λ minimizing a criterions such as

    1. $\mathrm{CIC}(\lambda) = -2\widehat{\mathrm{CL}}(\lambda) + 2\hat d(\lambda)$
    2. $\mathrm{CBIC}(\lambda) = -2\widehat{\mathrm{CL}}(\lambda) + \hat d(\lambda)\log(n)$
    3. $\mathrm{CERIC}(\lambda) = -2\widehat{\mathrm{CL}}(\lambda) + \hat d(\lambda)\log\left(\frac{n}{|W_n|\lambda}\right)$ (Bayesian prior)

    where $\hat d(\lambda)$ is an estimate of $d(\lambda) = \mathrm{Tr}\left(\mathbf{S}^{-1}(\lambda)\boldsymbol{\Sigma}(\lambda)\right)$, $\mathbf{S}(\lambda) = -\mathrm{E}\left(\mathrm{PL}_{\bullet,l}(\lambda)^{(2)}\right)$ $\boldsymbol{\Sigma}(\lambda) = \mathrm{Var}\left(\mathrm{PL}^{(1)}_{\bullet,l}(\lambda)\right)$.
  - Under the Poisson case, $d(\lambda)$ = # non-zero coefficients.

What can we prove ? (well expected results !)

- Let $\mu_n = \mathrm{E}N(W_n) = \begin{cases} \int_{W_n} \rho(u)\mathrm{d}u \\ \int_{W_n} \mathrm{E}(\lambda(u,\mathbf{X}))\mathrm{d}u \end{cases}$
- Asymptotic framework : $s = s_n$, $p = p_n$, $\mu_n \to \infty$ **(includes infill and increasing domain asymptotics)**
- For simplicity, we focus on the *adaptive lasso* here ; let

$$a_n = \max_{j=1,\ldots,s_n} \lambda_{n,j}, \qquad b_n = \min_{j=s_n+1,\ldots,p_n} \lambda_{n,j}.$$

Theorem [Choiruddin, C. and Letué'18,'22] [Ba and C.'22]

- Under some assumptions (such that it works ...)
- $\max\left(\frac{p_n^4}{\mu_n}, \frac{s_n^2 p_n^3}{\mu_n}\right) \to 0,\ a_n\sqrt{s_n\mu_n} \to 0,\ b_n / \sqrt{\frac{\mu_n}{p_n^2}} \to \infty.$

Then, as $n \to \infty$ and $\forall \phi \in \mathbb{R}^{s_n} \setminus \{0\}$ s.t. $\|\phi\| < \infty$

$$\boxed{\mathrm{P}\left(\hat{\boldsymbol{\beta}}_2 = 0\right) \to 1 \quad \text{and} \quad \sigma_\phi^{-1}\phi^\top S_{11}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} N(0, I_s)}$$

where $\sigma_\phi^2 = \phi^\top V_{11}\phi$ and where S_{11} and V_{11} are the sensitivity and variance of the score ...

For other penalties

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

Method	a_n	b_n	Possible ?
ridge	$\lambda_n \max_{j=1,\dots,s_n} \{\lvert\beta_{0j}\rvert\}$	0	✗
lasso	λ_n	λ_n	✗
enet	$\lambda_n \left[(1-\alpha) \max_{j=1,\dots,s_n} \{\lvert\beta_{0j}\rvert\} + \alpha\right]$	$\lambda_n \alpha$	✗
aLasso	$\max_{j=1,\dots,s_n} \{\lambda_{n,j}\}$	$\inf_{j=s_n+1,\dots,p_n} \{\lambda_{n,j}\}$	✓
aenet	$\max_{j=1,\dots,s_n} \{\lambda_{n,j}((1-\alpha)\lvert\beta_{0j}\rvert + \alpha)\}$	$\alpha \inf_{j=s_n+1,\dots,p_n} \{\lambda_{n,j}\}$	✓
SCAD	0^*	$\lambda_n{}^*$	✓
MC+	0^*	$\lambda_n - \frac{K}{\gamma \sqrt{\mu_n}}{}^*$	✓

* if $\lambda_n \to 0$ and $\lambda_n \sqrt{\mu_n / p_n^2} \to \infty$.

(Too) brief illustration for Inhom Strauss PP ($\gamma = .5$)

- PL_λ with adaptive lasso with CERIC(λ)
- $W_1 = [0, 250] \times [0, 125]$, $W_2 = 2\,W_1$ and $W_3 = 4\,W_1$
- $p_1 = 39$, $p_2 = 56$ and $p_3 = 79$ covariates
- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \mathbf{0})^\top$, $\boldsymbol{\beta}_1 = b\mathbf{1} \in \mathbb{R}^s$ with $s = 2$ (first row), $s = 5$ (second row) [the higher b the stronger the signal !]
- $\mathbf{z}1, \mathbf{z}_2$ are generated using BCI covariates
- All FPRs are $< 4\%$

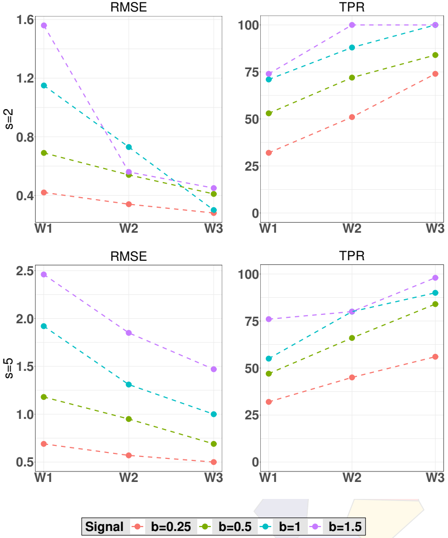


Table of Contents

Dantzig selector

- Lasso for linear models :

$$\text{Minimizing} \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

$$\iff \text{Minimizing } \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \text{ subj. to } \|\boldsymbol{\beta}\|_1 \leq \nu.$$

- Another point of view by [Candès and Tao'04] :

$$\iff \text{Minimizing } \|\boldsymbol{\beta}\|_1 \text{ subj. to } \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_\infty \leq \nu.$$

  - as Lasso, performs features selection ; can be efficiently implemented using linear programming ;
  - [Candès and Tao'04] provided some oracle inequalities (in particular) for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$;
  - then compared to Lasso by e.g. [Bickel et al.'09], extended to GLM by [James and Radchenko'09], [Dicker'10]

Dantzig selector for ρ (2) [Choiruddin, C. and Letué'20]

1. First substitute residuals (for a standard LM) by $\mathrm{PL}_\rho^{(1)}$

$$\text{Minimizing } \sum_{j=1}^{p_n} |\beta_j| \text{ subject to } \|\mathrm{PL}_\rho^{(1)}\|_\infty \le \lambda_n$$

2. + Adaptive version : let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_{n,j}, j = 1, \dots, p_n)$

$$\text{Minimizing } \|\mathbf{\Lambda}\boldsymbol{\beta}\|_1 \text{ subject to } \|\mathbf{\Lambda}^{-1}\mathrm{PL}^{(1)}\|_\infty \le 1.$$

3. + Linearization of the constraint (to use of linear programming)

$$\text{Minimizing } \|\mathbf{\Lambda}\boldsymbol{\beta}\|_1 \text{ subject to } \left\|\mathbf{\Lambda}^{-1}\left(\mathrm{PL}_\rho^{(1)}(\widetilde{\boldsymbol{\beta}}) + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathrm{PL}_\rho^{(2)}(\widetilde{\boldsymbol{\beta}})\right)\right\|_\infty \le 1.$$

[Choiruddin, C. and Letué'20] for details on the methodology, results, . . . : lots of similarities with adaptive lasso.

Log-convolution model for ρ

- In the context of eye-movement data, [Cuevas-Pacheco,C. and Descary'20] proposed the model

$$\log \rho(u) = \beta * z(u)$$

where $\beta : W\mathbb{R}^d$, $z(u)$ is a saliency map (deterministic prediction map), $*=$ convolution product

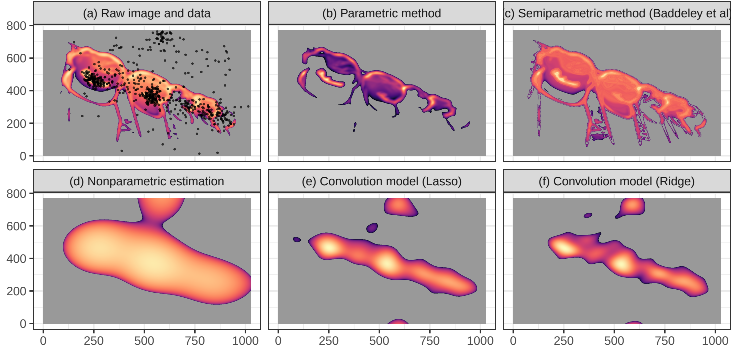
- Taking advantage of the Fourier basis $\phi_{\boldsymbol{\kappa}}$, $\boldsymbol{\kappa} \in \mathbb{Z}^d$

$$\log \rho(u) \approx \beta_{\boldsymbol{\kappa}_0} Z_{\boldsymbol{\kappa}_0} + \sum_{i=1}^{K} \Big\{ 2\beta^R_{\boldsymbol{\kappa}_i} \mathcal{R}[Z_{\boldsymbol{\kappa}_i} \phi_{\boldsymbol{\kappa}_i}(\boldsymbol{s})] - 2\beta^I_{\boldsymbol{\kappa}_i} \mathcal{I}[Z_{\boldsymbol{\kappa}_i} \phi_{\boldsymbol{\kappa}_i}(\boldsymbol{s})] \Big\}$$

$Z^{R,I}_{\boldsymbol{\kappa}}$ $\beta^{R,I}_{\boldsymbol{\kappa}}$ real or imaginary Fourier coefficient of $z(u)$ and $\beta(u)$.

- Close to a log-linear model in the spectral domain ; since K can be large $\Rightarrow$ regularization must be investigated.

Illustration



Method/model	AUC
Parametric estimate : Log-linear model $\rho(u) = \beta \times z(u)$	0.785
Semiparam. est. $\rho(u) = f(z(u))$ [Baddeley, Chang, Song and Turner'12]	0.774
Nonparametric estimate (kernel density estimate)	0.869
Log-convolution model (adaptive lasso)	0.900
Log-convolution model (adaptive ridge)	0.918

Table of Contents

Brief conclusion

- Regularization techniques for SPP is now a mature topic
- main methodologies ensue from links between PL/LR with GLMs ; treatment is now quite common to estimate either $\rho(u)$ or $\lambda(u, \mathbf{x})$

A few perspectives

- Finite-sample size results : requires concentration inequalities (quite complex)
- Understand more criteria to tune the regularization parameters (CERIC(λ),...)
- Extension to spatio-temporal PP (with an adapted penalty)
- squared-root lasso, group lasso, fused lasso
- Distribution of estimates ; post-selection inference

Thank you for your attention

Main references

- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. JASA.
- Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. The AOS.
- Candes, E., & Tao, T. (2007). The Dantzig selector : Statistical estimation when p is much larger than n. AOS.
- Dicker, L., & Lin, X. (2013). Parallelism, uniqueness, and large-sample asymptotics for the Dantzig selector. Canadian Journal of Statistics, 41(1), 23-35.
- James, G. M., & Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. Biometrika.

- I Ba and JF Coeurjolly. Inference for high-dimensional parametric inhomogeneous Gibbs point processes, in revision, 2020
- A Choiruddin, JF C. and F Letué. Convex and non-convex regularization methods for spatial point processes intensity estimation, Electronic Journal of Statistics, 12(1):1210-1255, 2018
- A Choiruddin, JF C. and F Letué. Dantzig selector and Adaptive Lasso methods for spatial point processes intensity estimation with a diverging number of parameters, submitted 2020.
- A Choiruddin, JF Coeurjolly and R Waagepetersen. Information critera for inhomogeneous spatial point processes, submitted, 2020.
- A Choiruddin, F Cuevas-Pacheco, JF Coeurjolly and R Waagepetersen. Regularized estimation for highly multivariate log Gaussian Cox processes, Statistics and Computing, 30 :649-662, 2020.
- F Cuevas-Pacheco, JF Coeurjolly, MH Descary. Fast estimation of a convolution type model for the intensity of spatial point processes, submitted, 2020.