Estimation Lasso et interactions poissoniennes sur le cercle

Laure SANSONNET

Université catholique de Louvain, Belgique
J-3 : AgroParisTech

en collaboration avec :
Patricia Reynaud-Bouret *(Université Nice Sophia-Antipolis)*,
Vincent Rivoirard *(Université Paris Dauphine)* et
Rebecca Willett *(University of Wisconsin-Madison)*.

Vendredi 29 août 2014

Journées MAS2014

Motivation

- Estimation in Hawkes model (Reynaud-Bouret and Schbath 2010, Hansen *et al.* 2012, . . .) and for Poissonian interactions (Sansonnet 2014) needs the knowledge of an a priori bound on the support of interaction fonctions.

Motivation

- Estimation in Hawkes model (Reynaud-Bouret and Schbath 2010, Hansen *et al.* 2012, ...) and for Poissonian interactions (Sansonnet 2014) needs the knowledge of an a priori bound on the support of interaction fonctions.
- Aim: Overcome knowledge of this bound in a high-dimensional discrete setting.
    - A discrete version of the Poissonian interactions model that is in the heart of my PhD thesis.
    - A circular model: to avoid boundary effects and also to reflect a certain biological reality.

Poissonian discrete model

- Parents: n i.i.d. uniform random variables $U_1, \ldots, U_n$ on the set $\{0, \ldots, p-1\} \rightarrow$ points on a circle (we work modulus p).

Poissonian discrete model

- Parents: n i.i.d. uniform random variables $U_1, \ldots, U_n$ on the set $\{0, \ldots, p-1\} \to$ points on a circle (we work modulus p).
- Children:
  - Each U_i gives birth independently to some Poisson variables.
  - If $x^* = (x_0^*, \ldots, x_{p-1}^*)^{\mathsf{H}} \in \mathbb{R}_+^p$, then $N^i_{U_i+j} \sim \mathcal{P}(x_j^*)$ independent of anything else.
  - The variable $N^i_{U_i+j}$ represents the number of children that a certain individual U_i has at distance j.
  - We set $Y_k = \sum_{i=1}^n N_k^i$ the total number of children at position k whose distribution conditioned on the U_i's is given by

$$Y_k \sim \mathcal{P}\left(\sum_{i=1}^n x^*_{k-U_i}\right).$$

Poissonian discrete model

- Parents: n i.i.d. uniform random variables $U_1, \ldots, U_n$ on the set $\{0, \ldots, p-1\} \to$ points on a circle (we work modulus p).
- Children:
  - Each U_i gives birth independently to some Poisson variables.
  - If $x^* = (x_0^*, \ldots, x_{p-1}^*)^{\mathsf{H}} \in \mathbb{R}_+^p$, then $N^i_{U_i+j} \sim \mathcal{P}(x_j^*)$ independent of anything else.
  - The variable $N^i_{U_i+j}$ represents the number of children that a certain individual U_i has at distance j.
  - We set $Y_k = \sum_{i=1}^n N_k^i$ the total number of children at position k whose distribution conditioned on the U_i's is given by

$$Y_k \sim \mathcal{P}\left(\sum_{i=1}^n x^*_{k-U_i}\right).$$

Aim: Estimate x^* with the observation of the U_i's and the Y_k's.

Another translation

$$Y = (Y_0, \ldots, Y_{p-1})^{\mathsf{H}} \sim \mathcal{P}(Ax^*),$$

Another translation

$$Y = (Y_0, \dots, Y_{p-1})^{\mathsf{H}} \sim \mathcal{P}(Ax^*),$$

where

- $$A = \begin{pmatrix} \mathbb{N}(0) & \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) \\ \mathbb{N}(1) & \mathbb{N}(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{N}(p-1) \\ \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) & \mathbb{N}(0) \end{pmatrix}$$
  is a $p \times p$ circulant matrix, with
- $\mathbb{N}(k) = \text{card}\,\{i : \ U_i = k[p]\}$
  representing the number of parents at position k on the circle.

Another translation

$$Y = (Y_0, \dots, Y_{p-1})^{\mathsf{H}} \sim \mathcal{P}(Ax^*),$$

where

- $$A = \begin{pmatrix} \mathbb{N}(0) & \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) \\ \mathbb{N}(1) & \mathbb{N}(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{N}(p-1) \\ \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) & \mathbb{N}(0) \end{pmatrix}$$
  is a $p \times p$ circulant matrix, with
- $\mathbb{N}(k) = \mathrm{card}\,\{i : \; U_i = k[p]\}$
  representing the number of parents at position k on the circle.

Aim: Recover $x^* \Leftrightarrow$ solve an inverse problem (potentially ill-posed) where the operator A is random and depends on the U_i's.

RIP property I

- The eigenvalues of A are given by $\sigma_k = \sum_{i=1}^{n} e^{-2\pi \hat{\imath} k U_i / p}$, for k in $\{0, \ldots, p-1\}$. In particular,

$$\mathbb{E}(\sigma_k) = \begin{cases} n & \text{if } k = 0[p] \\ 0 & \text{if } k \neq 0[p] \end{cases} .$$

RIP property I

- The eigenvalues of A are given by $\sigma_k = \sum_{i=1}^{n} e^{-2\pi \hat{\imath} k U_i / p}$, for k in $\{0, ..., p-1\}$. In particular,

$$\mathbb{E}(\sigma_k) = \begin{cases} n & \text{if } k = 0[p] \\ 0 & \text{if } k \neq 0[p] \end{cases} .$$

- So, with high probability, many eigenvalues of A may be close to zero. For these reasons, our problem is potentially ill-posed, which justifies the use of nonparametric procedures, such as the Lasso, even if p is not large with respect to n.

RIP property I

- The eigenvalues of A are given by $\sigma_k = \sum_{i=1}^{n} e^{-2\pi \hat{\imath} k U_i / p}$, for k in $\{0, \dots, p-1\}$. In particular,

$$\mathbb{E}(\sigma_k) = \begin{cases} n & \text{if } k = 0[p] \\ 0 & \text{if } k \neq 0[p] \end{cases} .$$

- So, with high probability, many eigenvalues of A may be close to zero. For these reasons, our problem is potentially ill-posed, which justifies the use of nonparametric procedures, such as the Lasso, even if p is not large with respect to n.
- We first focus on proving a Restricted Isometry Property (Candès and Tao, 2005): there exist positive constants r and R such that with high probability, for any K-sparse vector x (i.e. with at the most K non-zero coordinates)

$$r\|x\|^2 \leqslant \|Ax\|^2 \leqslant R\|x\|^2,$$

with R as close to r as possible.

RIP property II

- Under conditions, a RIP is satisfied by $\widetilde{A} = A - \frac{n-\sqrt{n}}{p}\mathbb{1}\mathbb{1}^{\mathsf{H}}$:

$$\frac{n}{2}\|x\|_2^2 \leqslant \|\widetilde{A}x\|_2^2 \leqslant \frac{3n}{2}\|x\|_2^2.$$

RIP property II

- Under conditions, a RIP is satisfied by $\widetilde{A} = A - \frac{n-\sqrt{n}}{p}\mathbb{1}\mathbb{1}^{\mathsf{H}}$:

$$\frac{n}{2}\|x\|_2^2 \leqslant \|\widetilde{A}x\|_2^2 \leqslant \frac{3n}{2}\|x\|_2^2.$$

- We obtain also a classical Restricted Eigenvalue (RE) type condition (see Bickel 2009) on an event of probability larger than $1 - 5.54\ pe^{-\theta}$ s.t. for all $d \in \{0, \ldots, p-1\}$,

$$|\mathbb{U}(d)| \leqslant \kappa\left(\frac{n}{\sqrt{p}}\theta + \theta^2\right) =: n\xi(\theta),$$

with $\mathbb{U}(d) = \sum_{u=0}^{p-1}\sum_{i=1}^{n}\sum_{j\neq i, j=1}^{n}\left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{U_j=u+d[p]} - \frac{1}{p}\right)$ and for p and n be fixed integers larger than 1 and for all $\theta > 1$.

Lasso estimator by using a weighted penalty I

- $\widetilde{A} = A - \frac{n-\sqrt{n}}{p} \mathbb{1}\mathbb{1}^{\mathsf{H}}$ and $\widetilde{Y}_k = Y_k - \frac{n-\sqrt{n}}{p} \overline{Y}$, with $\overline{Y} = \frac{1}{n} \sum_{k=0}^{p-1} Y_k$.
- $\mathbb{E}(\overline{Y}) = \|x^*\|_1$ and conditionally on the U_i's, $\widetilde{Y}$ is an unbiased estimate of $\widetilde{A}x^*$:
$$\mathbb{E}(\widetilde{Y} | U_1, \dots, U_n) = \widetilde{A}x^*.$$

Lasso estimator by using a weighted penalty I

- $\widetilde{A} = A - \frac{n-\sqrt{n}}{p}\mathbb{1}\mathbb{1}^{\mathrm{H}}$ and $\widetilde{Y}_k = Y_k - \frac{n-\sqrt{n}}{p}\overline{Y}$, with $\overline{Y} = \frac{1}{n}\sum_{k=0}^{p-1} Y_k$.
- $\mathbb{E}(\overline{Y}) = \|x^*\|_1$ and conditionally on the U_i's, $\widetilde{Y}$ is an unbiased estimate of $\widetilde{A}x^*$:

$$\mathbb{E}(\widetilde{Y}|U_1, \ldots, U_n) = \widetilde{A}x^*.$$

We first introduce the following Lasso estimate for some $\gamma > 0$:

$$\widehat{x} := \underset{x\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=0}^{p-1} d_k|x_k| \right\},$$

which is based on random and data-dependent weights d_k.

Lasso estimator by using a weighted penalty II

The Lasso estimate satisfies

$$\begin{cases} (\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x}))_k & = & \frac{\gamma \boldsymbol{d_k}}{2} \operatorname{sign}(\widehat{x}_k) & \text{for } k \text{ s.t. } \widehat{x}_k \neq 0 \\ |\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x})|_k & \leq & \frac{\gamma \boldsymbol{d_k}}{2} & \text{for } k \text{ s.t. } \widehat{x}_k = 0 \end{cases},$$

and in particular, for any k,

$$|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x})|_k \leq \frac{\gamma d_k}{2}.$$

Lasso estimator by using a weighted penalty II

The Lasso estimate satisfies

$$\begin{cases} (\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x}))_k & = & \frac{\gamma d_k}{2}\,\mathrm{sign}(\widehat{x}_k) & \text{for } k \text{ s.t. } \widehat{x}_k \neq 0 \\ |\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x})|_k & \leq & \frac{\gamma d_k}{2} & \text{for } k \text{ s.t. } \widehat{x}_k = 0 \end{cases},$$

and in particular, for any k,

$$|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x})|_k \leq \frac{\gamma d_k}{2}.$$

- Our approach where weights are random is similar to Zou 2006, Bertin *et al.* 2011, Hansen *et al.* 2012, ...: in some sense, the weights play the same role as the thresholds in the estimation procedure proposed in Donoho and Johnstone 1994, Reynaud-Bouret and Rivoirard 2010, Sansonnet 2014, ...

Lasso estimator by using a weighted penalty III

The double role of the weights:

- They have to control the random fluctuations of $\widetilde{A}^{\mathrm{H}}\widetilde{Y}$ around its mean conditionally on the U_i's due to the Poisson setting:

Lasso estimator by using a weighted penalty III

The double role of the weights:

- They have to control the random fluctuations of $\widetilde{A}^{\mathsf{H}}\widetilde{Y}$ around its mean conditionally on the U_i's due to the Poisson setting:

$$|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}x^*)|_k \leqslant d_k.$$

Remark: If $\gamma \geqslant 2$, x^* will also satisfy $|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x}^*)|_k \leq \frac{\gamma d_k}{2}$. This is a key technical point to prove optimality of our approach.

Lasso estimator by using a weighted penalty III

The double role of the weights:

- They have to control the random fluctuations of $\widetilde{A}^{\mathsf{H}}\widetilde{Y}$ around its mean conditionally on the U_i's due to the Poisson setting:

$$|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}x^*)|_k \leqslant d_k.$$

  Remark: If $\gamma \geqslant 2$, x^* *will also satisfy* $|\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}\widehat{x}^*)|_k \leq \frac{\gamma d_k}{2}$. *This is a key technical point to prove optimality of our approach.*

- They need to be high enough, so that they work even if A is not invertible enough.

How to choose the d_k's?

Proposition 1

For any $\theta > 0$, there exists an event $\Omega_v(\theta)$ of probability larger than $1 - 2pe^{-\theta}$ on which, for all k in $\{0, \ldots, p-1\}$,

$$|\widetilde{A}^{\mathrm{H}}(\widetilde{Y} - \widetilde{A}x^*)|_k \leqslant \sqrt{2v_k\theta} + \frac{B\theta}{3},$$

where

$$B = \max_{u \in \{0,\ldots,p-1\}} \left| \mathbb{N}(u) - \frac{n-1}{p} \right|$$

and

$$v_k = \sum_{u=0}^{p-1} w(k-u) x_u^*,$$

with $w(d) = \sum_{u=0}^{p-1} \left(\mathbb{N}(u) - \frac{n-1}{p} \right)^2 \mathbb{N}(u+d)$ for all d in $\{0, \ldots, p-1\}$.

Derivation of constant weights

$$v_k \leqslant W\|x^*\|_1,$$

where

- $W := \max_d w(d) = \max_{u\in\{0,\dots,p-1\}} \sum_{\ell=0}^{p-1} \left(\mathbb{N}(\ell+u) - \frac{n-1}{p}\right)^2 \mathbb{N}(\ell)$ is observable, and
- $\|x^*\|_1$ is unbiasedly estimated by $\bar{Y}$.

Derivation of constant weights

$$v_k \leqslant W\|x^*\|_1,$$

where

- $W := \max_d w(d) = \max_{u \in \{0,\dots,p-1\}} \sum_{\ell=0}^{p-1} \left(\mathbb{N}(\ell+u) - \frac{n-1}{p} \right)^2 \mathbb{N}(\ell)$ is observable, and
- $\|x^*\|_1$ is unbiasedly estimated by $\bar{Y}$.

Constant weights

$$d := \sqrt{2W\theta} \left[\sqrt{\bar{Y} + \frac{5\theta}{6n}} + \sqrt{\frac{\theta}{2n}} \right] + \frac{B\theta}{3}$$

satisfies $|\widetilde{A}^{\mathrm{H}}(\widetilde{Y} - \widetilde{A}x^*)|_k \leqslant d_k = d$ on an event of probability larger than $1 - (2p+1)e^{-\theta}$.

Derivation of non constant weights

We reestimate the v_k's:

$$\hat{v}_k = \sum_{\ell=0}^{p-1} \left(\mathbb{N}(\ell - k) - \frac{n-1}{p} \right)^2 Y_\ell.$$

Derivation of non constant weights

We reestimate the v_k's:

$$\hat{v}_k = \sum_{\ell=0}^{p-1} \left(\mathbb{N}(\ell - k) - \frac{n-1}{p} \right)^2 Y_\ell .$$

Non constant weights

$$\tilde{d}_k = \sqrt{2\theta} \left[\sqrt{\hat{v}_k + \frac{5\theta B^2}{6}} + \sqrt{\frac{\theta B^2}{2}} \right] + \frac{B\theta}{3}$$

satisfies $|\widetilde{A}^{\mathrm{H}}(\widetilde{Y} - \widetilde{A}x^*)|_k \leqslant d_k = \tilde{d}_k$ on an event of probability larger than $1 - 3pe^{-\theta}$.

Control of the weights I

We consider the following special relationships between n, p and θ. For any integers n, p and for any $\theta > 1$,

$$\kappa' \sqrt{p}\theta \leqslant n \leqslant \kappa'' p\theta^{-1},$$

where $\kappa' := \max(2\kappa, 1)$ and κ'' is an absolute constant small enough.

- In particular, if we choose θ proportional to $\log p$ (which is natural to have results on events with large probability), then the regime becomes

$$\sqrt{p}\log(p) << n << p\log^{-1}(p).$$

Control of the weights II

Theorem

- On a event of probability larger than $1 - (2 + 8.54p)e^{-\theta}$

$$\bar{c}\left(n\theta\|x^*\|_{\ell_1} + \theta^2\right) \leqslant d^2 \leqslant c\left(n\theta\|x^*\|_{\ell_1} + \theta^4\right).$$

- On a event of probability larger than $1 - 10.54pe^{-\theta}$, for any $k \in \{1, \ldots, p\}$,

$$c''\left(n\theta x_k^* + n^2\theta p^{-1}\sum_{u\neq k} x_u^* + \theta^2\right)$$

$$\leqslant \tilde{d}_k^2 \leqslant c'\left(n\theta x_k^* + n^2\theta^2 p^{-1}\sum_{u\neq k} x_u^* + \theta^4\right).$$

Control of the weights II

For instance, in the asymptotic regime

$$n\theta = o(p) \quad \text{and} \quad \theta^2 = o(n^2\|x^*\|_1/p),$$

then if $x_k^* = 0$,

$$\tilde{d}_k^2 = O(n^2\theta^2 p^{-1}\|x^*\|_1) = o(n\theta\|x^*\|_1).$$

Therefore

$$\tilde{d}_k^2 = o(d^2),$$

i.e. $\tilde{d}_k$ can be much more smaller than d.

Oracle inequalities for the lasso estimate with d

Theorem

Let $\gamma > 2$ and $0 < \varepsilon < 1$. Let s a positive integer satisfying

$$\frac{3\gamma + 2}{\gamma - 2} s\xi(\theta) \leqslant 1 - \varepsilon.$$

Then, there exists a constant $C_{\gamma,\varepsilon}$ depending on γ and ε such that on an event of probability larger than $1 - (1 + 7.54p)e^{-\theta}$, the lasso estimate $\widehat{x}$ satisfies

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leqslant C_{\gamma,\varepsilon} \inf_{x:\, |\mathrm{supp}(x)| \leqslant s} \left\{ \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \frac{sd(\theta)^2}{n} \right\}.$$

And if x^* is s-sparse, under the same assumptions, we get that

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leqslant \frac{C_{\gamma,\varepsilon} sd(\theta)^2}{n}.$$

Remark : We also obtain oracle inequalities for the ℓ_1 and ℓ_∞-losses.

Oracle inequalities for the lasso estimate with $\tilde{d}_k$

Theorem

Let $\gamma > 2$ and $0 < \varepsilon < 1$. Let s be a positive integer and assume that we are on an event s.t.

$$\frac{\max_{1 \leqslant k \leqslant p} \tilde{d}_k(\theta)}{\min_{1 \leqslant k \leqslant p} \tilde{d}_k(\theta)} \leqslant \frac{\left((1-\varepsilon)s^{-1}\xi(\theta)^{-1} - 1\right)(\gamma - 2)}{2(\gamma + 2)}.$$

Then, there exists a constant $C_{\gamma,\varepsilon}$ depending on γ and ε such that on an event of probability larger than $1 - 8.54pe^{-\theta}$, the lasso estimate $\widehat{x}$ satisfies

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leqslant C_{\gamma,\varepsilon} \inf_{x:\, |\mathrm{supp}(x)| \leqslant s} \left\{ \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \frac{1}{n} \sum_{k \in \mathrm{supp}(x)} \tilde{d}_k^2(\theta) \right\}.$$

And if x^* is s-sparse, under the same assumptions, we get that

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leqslant \frac{C_{\gamma,\varepsilon}}{n} \sum_{k \in \mathrm{supp}(x^*)} \tilde{d}_k^2(\theta).$$

And what happens next?

- We will consider a second alternative consisting in assuming that x^* is supported by S^* with $|S^*| = o(p)$. We then introduce the pseudo-estimate $\widehat{x}^{(S^*)}$ defined by

$$\widehat{x}^{(S^*)} \in \underset{x \in \mathbb{R}^p : \mathrm{supp}(x) \subseteq S^*}{\mathrm{argmin}} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=0}^{p-1} d_k |x_k| \right\}.$$

  In this case, we shall see that under some conditions the support of $\hat{x}$ is included into S^* (support property), enabling us to derive oracle inequalities quite easily.

And what happens next?

- We will consider a second alternative consisting in assuming that x^* is supported by S^* with $|S^*| = o(p)$. We then introduce the pseudo-estimate $\widehat{x}^{(S^*)}$ defined by

$$\widehat{x}^{(S^*)} \in \underset{x \in \mathbb{R}^p : \mathrm{supp}(x) \subseteq S^*}{\operatorname{argmin}} \left\{ \| \widetilde{Y} - \widetilde{A}x \|_2^2 + \gamma \sum_{k=0}^{p-1} d_k |x_k| \right\}.$$

  In this case, we shall see that under some conditions the support of $\hat{x}$ is included into S^* (support property), enabling us to derive oracle inequalities quite easily.

- Simulations...

And what happens next?

- We will consider a second alternative consisting in assuming that x^* is supported by S^* with $|S^*| = o(p)$. We then introduce the pseudo-estimate $\widehat{x}^{(S^*)}$ defined by

$$\widehat{x}^{(S^*)} \in \underset{x \in \mathbb{R}^p : \mathrm{supp}(x) \subseteq S^*}{\operatorname{argmin}} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=0}^{p-1} d_k |x_k| \right\}.$$

  In this case, we shall see that under some conditions the support of $\hat{x}$ is included into S^* (support property), enabling us to derive oracle inequalities quite easily.
- Simulations...

Merci pour votre attention !