Hawkes process as models for some genomic data

P. Reynaud-Bouret

CNRS - LJAD University of Nice

January 27th, Institut Curie

Contents

Contents

Contents

Contents

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Motifs

= words in the DNA-alphabet $\{\texttt{actg}\}$.

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Motifs

= words in the DNA-alphabet $\{\mathtt{actg}\}$.

How can statistician suggest functional motifs based on the statistical properties of their occurrences ?

- Unexpected frequency → Markov models (see for a review Reinert, Schbath, Waterman (2000))

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Motifs

= words in the DNA-alphabet $\{\texttt{actg}\}$.

How can statistician suggest functional motifs based on the statistical properties of their occurrences ?

- Unexpected frequency → Markov models (see for a review Reinert, Schbath, Waterman (2000))
- Poor or rich regions → scan statistics (see, for instance, Robin Daudin (1999) or Stefanov (2003))

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Motifs

= words in the DNA-alphabet {actg}.

How can statistician suggest functional motifs based on the statistical properties of their occurrences ?

- Unexpected frequency → Markov models (see for a review Reinert, Schbath, Waterman (2000))
- Poor or rich regions → scan statistics (see, for instance, Robin Daudin (1999) or Stefanov (2003))
- If two motifs are part of a common biological process, the space between their occurrences (not necessarily consecutive) should be somehow fixed → favored or avoided distances (Gusto, Schbath (2005))

Several examples

There are several "events" of different types on the DNA that may "work" together in synergy.

Motifs

= words in the DNA-alphabet {`actg`}.

How can statistician suggest functional motifs based on the statistical properties of their occurrences ?

- Unexpected frequency → Markov models (see for a review Reinert, Schbath, Waterman (2000))
- Poor or rich regions → scan statistics (see, for instance, Robin Daudin (1999) or Stefanov (2003))
- If two motifs are part of a common biological process, the space between their occurrences (not necessarily consecutive) should be somehow fixed → favored or avoided distances (Gusto, Schbath (2005)) pairwise study.

Several examples

TRE

Transcription Regulatory Elements = "everything" that may enhance or repress gene expression

Several examples

TRE

Transcription Regulatory Elements = "everything" that may enhance or repress gene expression

- promoter, enhancer, silencer, histone modifications on the DNA.... They should interact but how ? Can we have a statistical guess ?

Several examples

TRE

Transcription Regulatory Elements = "everything" that may enhance or repress gene expression

- promoter, enhancer, silencer, histone modifications on the DNA.... They should interact but how ? Can we have a statistical guess ?
- There are methods (ChIP-chip experiments, ChIP-seq experiments) where after preprocessing the data one has access to the (almost exact) positions of several type of TREs at one time, and this under different experimental conditions. (ENCODE)

Several examples

TRE

Transcription Regulatory Elements = "everything" that may enhance or repress gene expression

- promoter, enhancer, silencer, histone modifications on the DNA.... They should interact but how ? Can we have a statistical guess ?
- There are methods (ChIP-chip experiments, ChIP-seq experiments) where after preprocessing the data one has access to the (almost exact) positions of several type of TREs at one time, and this under different experimental conditions. (ENCODE)
- → analysis of dependance between the different positions (see Carstensen, Sandelin, Winther, Hansen (2010)), based on favored or avoided distances.

Several examples

TRE

Transcription Regulatory Elements = "everything" that may enhance or repress gene expression

- promoter, enhancer, silencer, histone modifications on the DNA.... They should interact but how ? Can we have a statistical guess ?
- There are methods (ChIP-chip experiments, ChIP-seq experiments) where after preprocessing the data one has access to the (almost exact) positions of several type of TREs at one time, and this under different experimental conditions. (ENCODE)
- → analysis of dependance between the different positions (see Carstensen, Sandelin, Winther, Hansen (2010)), based on favored or avoided distances.

Why just DNA ? RNA etc ...

A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition

A simplified/informal vision

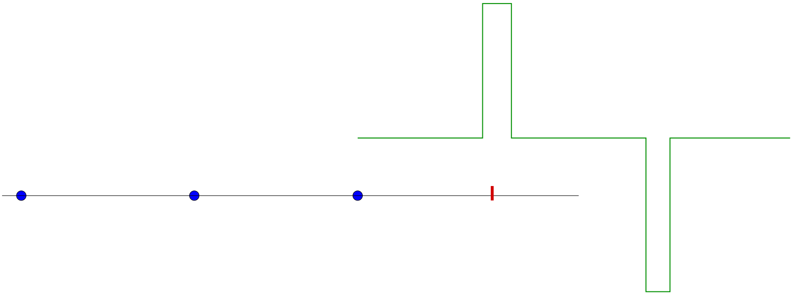
How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction

A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

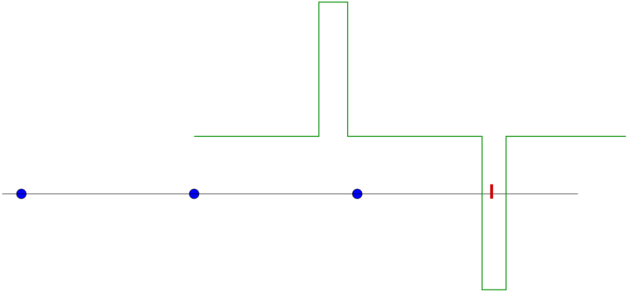
- Spontaneous apparition
- Self-interaction



A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

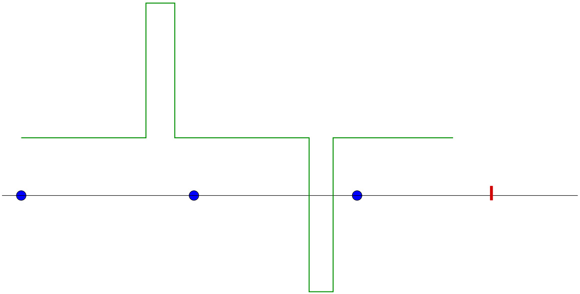
- Spontaneous apparition
- Self-interaction



A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction



A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction

A simplified/informal vision

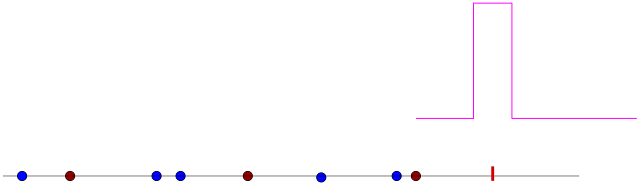
How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction
- Interaction with another event

A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction
- Interaction with another event



A simplified/informal vision

How can we explain the positions of the occurrences of one event (motifs, TRE, etc) wrt the others and itself ?

- Spontaneous apparition
- Self-interaction
- Interaction with another event

Maximal sensible interaction distance: 5 000 - 10 000 bases because of the 3d DNA structure

Point process on the real line

Point process

$N=$ random countable set of points of $\mathbb{R}$ (here).

Point process on the real line

Point process

$N=$ random countable set of points of $\mathbb{R}$ (here).

N_A number of points of N in A, $N_t = N_{[0,t]}$,
$dN_t = \sum_{T \text{ point de } N} \delta_T$. $\int f(t) dN_t = \sum_{T \in N} f(T)$

Point process on the real line

Point process

N= random countable set of points of $\mathbb{R}$ (here).

N_A number of points of N in A, $N_t = N_{[0,t]}$,
$dN_t = \sum_{T \text{ point de } N} \delta_T$. Usually $\mathbb{R}$ is thought as time

Intensity

$t \to \lambda(t)$ represents the probability to have a point at time t conditionnally to the past before t $(s < t)$

Point process on the real line

Point process

N= random countable set of points of $\mathbb{R}$ (here).

N_A number of points of N in A, $N_t = N_{[0,t]}$,
$dN_t = \sum_{T \text{ point de } N} \delta_T$. Usually $\mathbb{R}$ is thought as time

Intensity

$t \to \lambda(t)$ represents the probability to have a point at time t conditionnally to the past before t $(s < t)$

"Past" contains in particular the previous occurrences of points.

Point process on the real line

Point process

N= random countable set of points of $\mathbb{R}$ (here).

N_A number of points of N in A, $N_t = N_{[0,t]}$,
$dN_t = \sum_{T \text{ point de } N} \delta_T$. Usually $\mathbb{R}$ is thought as time

Intensity

$t \to \lambda(t)$ represents the probability to have a point at time t conditionnally to the past before t $(s < t)$

"Past" contains in particular the previous occurrences of points.
NB : here $\mathbb{R}$ is the DNA strand. The "past" may be interpreted as what has already been read in a prescribed direction (e.g. 5'-3' or 3'-5').

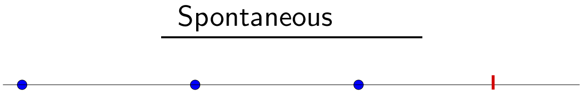
The simple Hawkes process

The intensity $\lambda(t)$ is given by

The simple Hawkes process

The intensity $\lambda(t)$ is given by

$$\nu$$

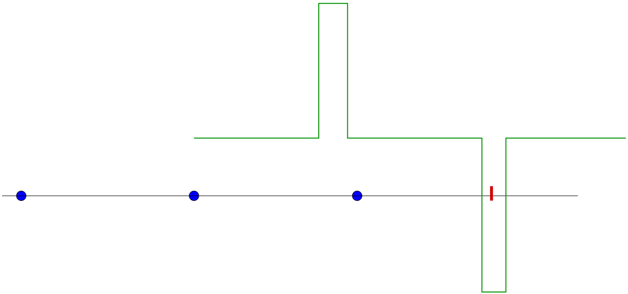


The simple Hawkes process

The intensity $\lambda(t)$ is given by

$$\nu \qquad + \quad \sum_{T\in N} h(t-T)$$

Spontaneous Self-exciting



The simple Hawkes process

The intensity $\lambda(t)$ is given by

$$\underbrace{\nu}_{\text{Spontaneous}} \quad + \quad \underbrace{\sum_{T \in N} h(t - T)}_{\text{Self-exciting}}$$

The simple Hawkes process

The intensity $\lambda(t)$ is given by

$$\nu \qquad + \quad \sum_{T \in N} h(t - T)$$

Spontaneous	Self-exciting

The most classical case corresponds to $h > 0$ (see Hawkes (1971)) and a stationary version exists if $\int h < 1$. There is also in this specific case a branching / cluster process representation (see Hawkes and Oakes (1974)).

The simple Hawkes process

The intensity $\lambda(t)$ is given by

$$\left(\nu \qquad + \quad \sum_{T \in N} h(t-T) \right)_+$$

Spontaneous Self-exciting

The most classical case corresponds to $h > 0$ (see Hawkes (1971)) and a stationary version exists if $\int h < 1$. There is also in this specific case a branching / cluster process representation (see Hawkes and Oakes (1974)).

One can actually consider any 1-Lipschitz modification, there is still a stationary version of it if $\int |h| < 1$ (Brémaud and Massoulié (1996)).

The Hawkes process interaction with itself + an additional interaction

$$\lambda(t) =$$

The Hawkes process interaction with itself + an additional interaction

$\lambda(t) =$

ν

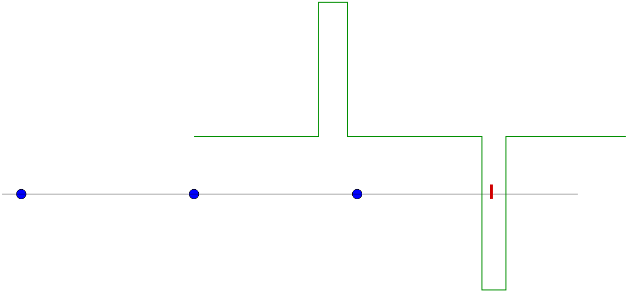
Spontaneous

The Hawkes process interaction with itself + an additional interaction

$\lambda(t) =$

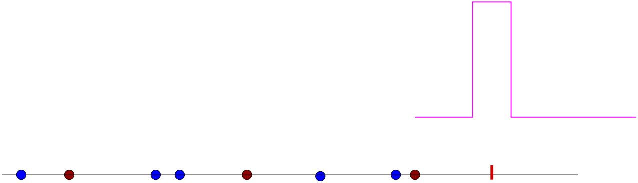
$$\nu \quad + \quad \sum_{T \in N} h(t - T)$$

Spontaneous Self-interaction



The Hawkes process interaction with itself + an additional interaction

$$\lambda(t) = \underbrace{\nu}_{\text{Spontaneous}} + \underbrace{\sum_{T \in N} h(t - T)}_{\text{Self-interaction}} + \underbrace{\sum_{X \in N_2} h_2(t - X)}_{\text{Interaction with other type}}$$



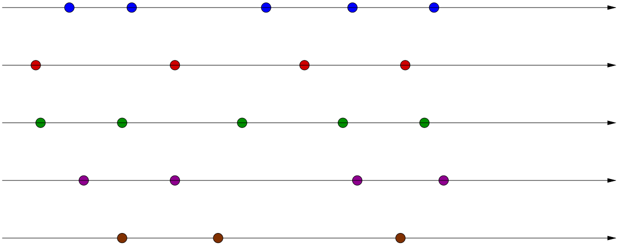
The Hawkes process interaction with itself + an additional interaction

$$\lambda(t) = \left(\underbrace{\nu}_{\text{Spontaneous}} + \underbrace{\sum_{T \in N} h(t-T)}_{\text{Self-interaction}} + \underbrace{\sum_{X \in N_2} h_2(t-X)}_{\text{Interaction with other type}} \right)_+$$

If h is null and if N_2 is fixed (no reciprocal interaction), then N is a Poisson process given N_2.

The multivariate Hawkes process

One observes $N^{(1)}, \ldots, N^{(r)}, \ldots, N^{(M)}$ processes such that



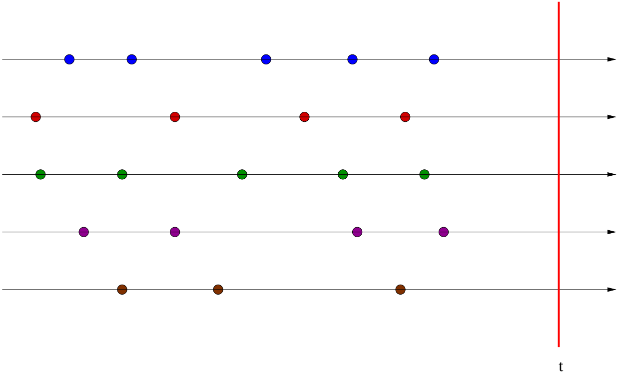
The multivariate Hawkes process

One observes $N^{(1)}, \ldots, N^{(r)}, \ldots, N^{(M)}$ processes such that

$\lambda^{(1)}(t) =$

$\lambda^{(2)}(t) =$

$\lambda^{(r)}(t) =$



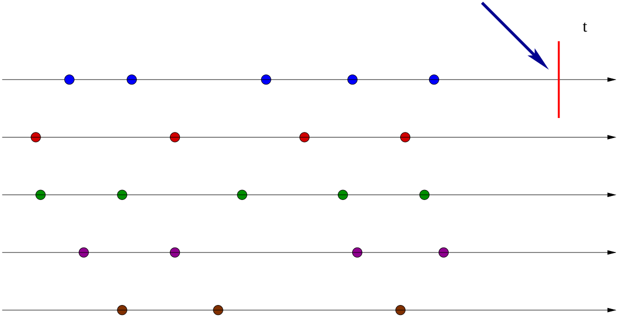
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$\lambda^{(1)}(t) = \quad \nu_1$

$\lambda^{(2)}(t) =$

$\lambda^{(r)}(t) =$



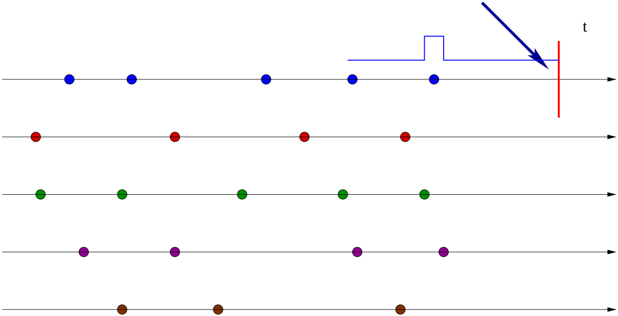
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \quad \nu_1 \quad + \sum_{T \in N^{(1)}} h_1^{(1)}(t - T)$$

$$\lambda^{(2)}(t) =$$

$$\lambda^{(r)}(t) =$$



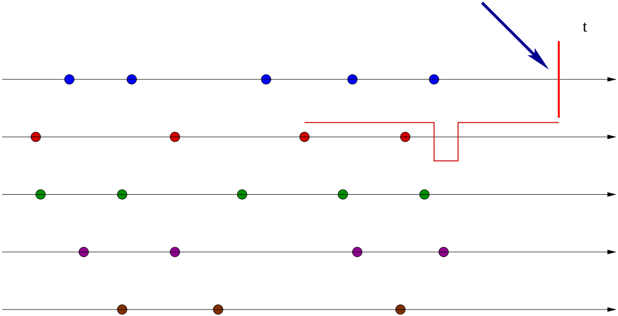
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \quad \nu_1 \quad + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T)$$

$$\lambda^{(2)}(t) =$$

$$\lambda^{(r)}(t) =$$



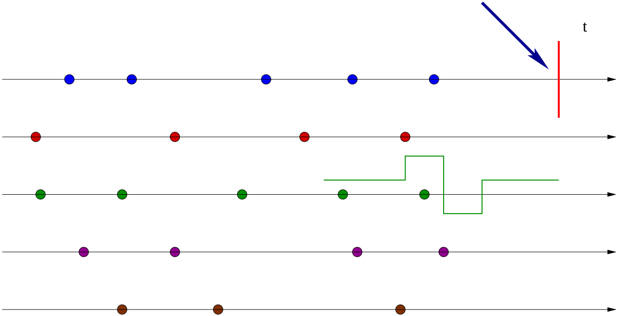
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \quad \nu_1 \quad + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T)$$

$$\lambda^{(2)}(t) =$$

$$\lambda^{(r)}(t) =$$



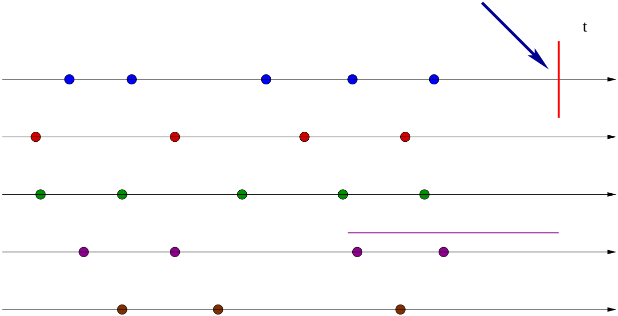
The multivariate Hawkes process

One observes $N^{(1)}, \dots, N^{(r)}, \dots, N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \quad \nu_1 \quad + \sum_{T \in N^{(1)}} h_1^{(1)}(t - T)$$

$$\lambda^{(2)}(t) =$$

$$\lambda^{(r)}(t) =$$



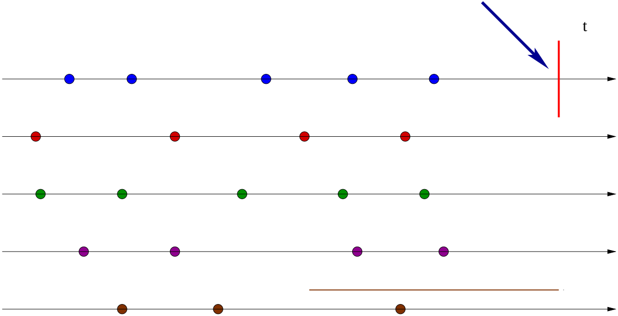
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T\in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell\neq 1} \sum_{T\in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) =$$

$$\lambda^{(r)}(t) =$$



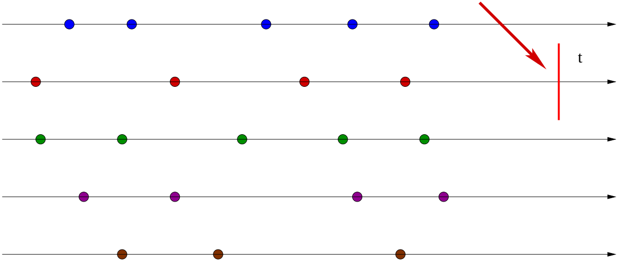
The multivariate Hawkes process

One observes $N^{(1)}, \ldots, N^{(r)}, \ldots, N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell \neq 1} \sum_{T \in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2$$

$$\lambda^{(r)}(t) =$$



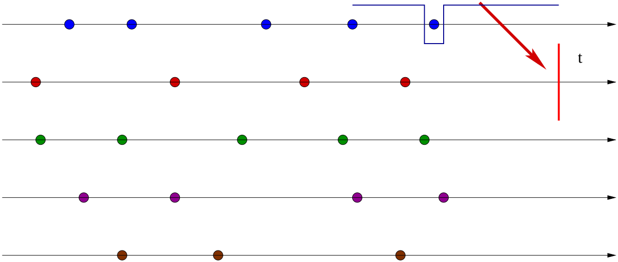
The multivariate Hawkes process

One observes $N^{(1)}, \dots, N^{(r)}, \dots, N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell \neq 1} \sum_{T \in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2$$

$$\lambda^{(r)}(t) =$$



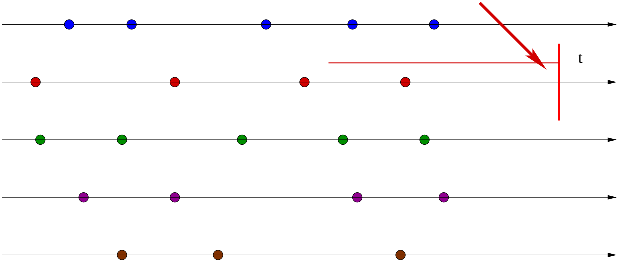
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T\in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell\neq 1}\sum_{T\in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2 + \sum_{T\in N^{(2)}} h_2^{(2)}(t-T)$$

$$\lambda^{(r)}(t) =$$



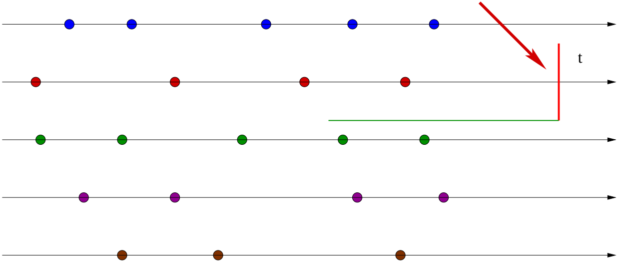
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell \neq 1} \sum_{T \in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2 + \sum_{T \in N^{(2)}} h_2^{(2)}(t-T)$$

$$\lambda^{(r)}(t) =$$



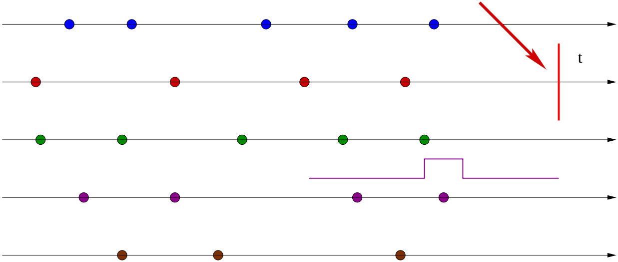
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell \neq 1} \sum_{T \in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2 + \sum_{T \in N^{(2)}} h_2^{(2)}(t-T)$$

$$\lambda^{(r)}(t) =$$



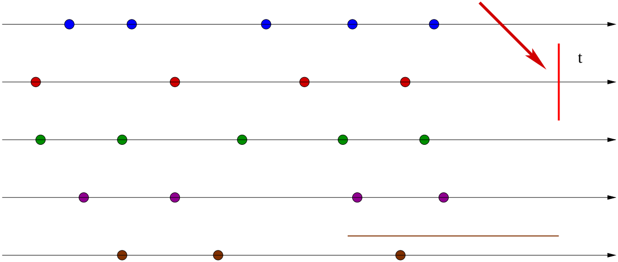
The multivariate Hawkes process

One observes $N^{(1)}, ..., N^{(r)}, ..., N^{(M)}$ processes such that

$$\lambda^{(1)}(t) = \nu_1 + \sum_{T \in N^{(1)}} h_1^{(1)}(t-T) + \sum_{\ell \neq 1} \sum_{T \in N^{(\ell)}} h_\ell^{(1)}(t-T)$$

$$\lambda^{(2)}(t) = \nu_2 + \sum_{T \in N^{(2)}} h_2^{(2)}(t-T) + \sum_{\ell \neq 2} \sum_{T \in N^{(\ell)}} h_\ell^{(2)}(t-T)$$

$$\lambda^{(r)}(t) =$$

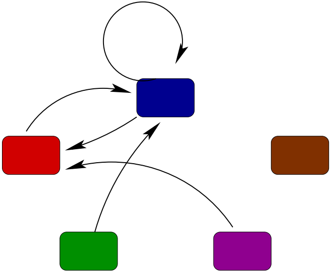


The multivariate Hawkes process(2)

Link with graphical model of local independence (see Didelez (2008))

The multivariate Hawkes process(2)

Link with graphical model of local independence (see Didelez (2008))



The multivariate Hawkes process(3)

- Only excitation (all the $h_\ell^{(r)}$ are positive): for all r,

$$\lambda^{(r)}(t) = \nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u)dN_u^{(\ell)}.$$

Branching / Cluster representation, stationary process if the spectral radius of $\left(\int h_\ell^{(r)}(t)dt\right)$ is < 1.

The multivariate Hawkes process(3)

- Only excitation (all the $h_\ell^{(r)}$ are positive): for all r,

$$\lambda^{(r)}(t) = \nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u) dN_u^{(\ell)}.$$

  Branching / Cluster representation, stationary process if the spectral radius of $\left(\int h_\ell^{(r)}(t)dt\right)$ is < 1.

- Interaction, for instance

$$\lambda^{(r)}(t) = \left(\nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u) dN_u^{(\ell)} \right)_+ .$$

The multivariate Hawkes process(3)

- Only excitation (all the $h_\ell^{(r)}$ are positive): for all r,

$$\lambda^{(r)}(t) = \nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u)dN_u^{(\ell)}.$$

  Branching / Cluster representation, stationary process if the spectral radius of $\left(\int h_\ell^{(r)}(t)dt\right)$ is < 1.

- Interaction, for instance

$$\lambda^{(r)}(t) = \left(\nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u)dN_u^{(\ell)}\right)_+ .$$

- Exponential

$$\lambda^{(r)}(t) = \exp\left(\nu_r + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(r)}(t-u)dN_u^{(\ell)}\right).$$

  Multiplicative shape but no guarantee of a stationary version

Maximum likelihood estimation

- Only the shape of the intensity is necessary.

Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,

$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$

Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,

$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$

- maximizing gives $\hat{\theta}$.

Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,

$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$

- maximizing gives $\hat{\theta}$.
- If several processes or n sample → sum

Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,

$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$

- maximizing gives $\hat{\theta}$.
- If several processes or n sample → sum
- not necessary to have a stationary version, it depends on the considered asymptotics.

Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,
$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$
- maximizing gives $\hat{\theta}$.
- If several processes or n sample $\longrightarrow$ sum
- not necessary to have a stationary version, it depends on the considered asymptotics.
- Consistence, asymptotic normality under smooth conditions, see Ogata and Akaike (1982), Ozaki (1979) or Andersen et al (1993).

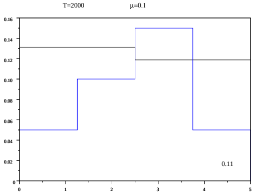
Maximum likelihood estimation

- Only the shape of the intensity is necessary.
- Given the "past" before 0, for one process observed until time T,

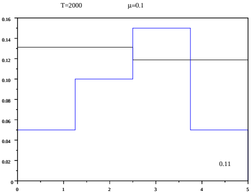
$$L_T(\theta) = \int_0^T \ln(\lambda_\theta(t))dN_t - \int_0^T \lambda_\theta(t)dt.$$

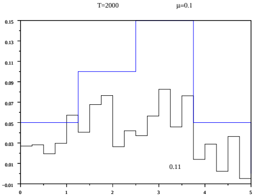
- maximizing gives $\hat{\theta}$.
- If several processes or n sample → sum
- not necessary to have a stationary version, it depends on the considered asymptotics.
- Consistence, asymptotic normality under smooth conditions, see Ogata and Akaike (1982), Ozaki (1979) or Andersen et al (1993).
- Even tests of the nullity of $h_\ell^{(r)}$ and access to a graphical model (see Carstensen et al (2010)) .

Model choice

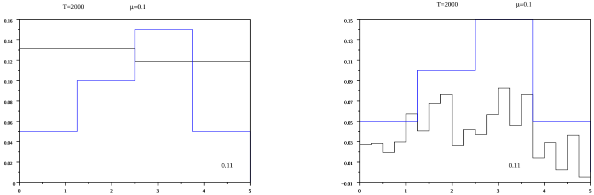


Model choice



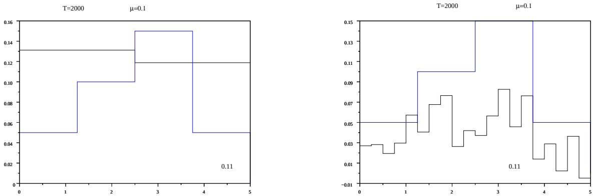


Model choice

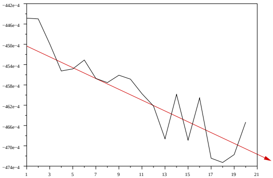


The biggest model is not the correct one, it overfits but ...

Model choice



The biggest model is not the correct one, it overfits but ...



AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$

AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$
- one can always select one model by minimizing

$$-L_T(\hat{\theta}_m) + D_m$$

AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$
- one can always select one model by minimizing

$$-L_T(\hat{\theta}_m) + D_m$$

- If the family of models is fixed and if there is a true model, it should select the correct one asymptotically (see for instance Vere-Jones and Ozaki (1982), Gusto and Schbath (2005)).

AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$
- one can always select one model by minimizing

$$-L_T(\hat{\theta}_m) + D_m$$

- If the family of models is fixed and if there is a true model, it should select the correct one asymptotically (see for instance Vere-Jones and Ozaki (1982), Gusto and Schbath (2005)).
- What if no true model ? What if family too big with respect to T ?
- Gusto PhD thesis numerical study shows that AIC does not select a sparse model if spline with not equally spaced knots.

AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$
- one can always select one model by minimizing

$$-L_T(\hat{\theta}_m) + D_m$$

- If the family of models is fixed and if there is a true model, it should select the correct one asymptotically (see for instance Vere-Jones and Ozaki (1982), Gusto and Schbath (2005)).
- What if no true model ? What if family too big with respect to T ?
- Gusto PhD thesis numerical study shows that AIC does not select a sparse model if spline with not equally spaced knots.
- Gusto and Schbath program, FADO, only works for equally spaced knots.

AIC criterion

If several parametric models m with different dimensions D_m,

- for each model one has an MLE $\hat{\theta}_m$
- one can always select one model by minimizing

$$-L_T(\hat{\theta}_m) + D_m$$

- If the family of models is fixed and if there is a true model, it should select the correct one asymptotically (see for instance Vere-Jones and Ozaki (1982), Gusto and Schbath (2005)).
- What if no true model ? What if family too big with respect to T ?
- Gusto PhD thesis numerical study shows that AIC does not select a sparse model if spline with not equally spaced knots.
- Gusto and Schbath program, FADO, only works for equally spaced knots.
- Hence no clear access to favoured or avoided distance

Aim

- We want to perform an adaptive estimation of the functions h!

Aim

- We want to perform an adaptive estimation of the functions h!
- ie select a model $m \longrightarrow \hat{h}$ sparse and picky : one can read on the selected estimate favored or avoided distances if any.

Aim

- We want to perform an adaptive estimation of the functions h!
- ie select a model $m \longrightarrow \hat{h}$ sparse and picky : one can read on the selected estimate favored or avoided distances if any.
- More likely that no true model and that the family of models grows with T.

Aim

- We want to perform an adaptive estimation of the functions h!
- ie select a model $m \longrightarrow \hat{h}$ sparse and picky : one can read on the selected estimate favored or avoided distances if any.
- More likely that no true model and that the family of models grows with T.
- → penalized model selection (Birgé and Massart, see Massart course in St Flour (2007))

Aim

- We want to perform an adaptive estimation of the functions h!
- ie select a model $m \longrightarrow \hat{h}$ sparse and picky : one can read on the selected estimate favored or avoided distances if any.
- More likely that no true model and that the family of models grows with T.
- $\longrightarrow$ penalized model selection (Birgé and Massart, see Massart course in St Flour (2007))
- but also, thresholding or Lasso methods.

Least-square contrast for one process

- The meaningful quantity is $\lambda(t)$.

Least-square contrast for one process

- The meaningful quantity is $\lambda(t)$.
- It may depends on the process itself.

Least-square contrast for one process

- The meaningful quantity is $\lambda(t)$.
- It may depends on the process itself.
- It is the "mean" version of the point measure $\delta_{T=t}$.

Least-square contrast for one process

- The meaningful quantity is $\lambda(t)$.
- It may depends on the process itself.
- It is the "mean" version of dN_t.

Least-square contrast for one process

- The meaningful quantity is $\lambda(t)$.
- It may depends on the process itself.
- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.

Least-square contrast for one process

- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.
- Let η be an intensity candidate.

Least-square contrast

$$\gamma_T(\eta) = -\frac{2}{T}\int_0^T \eta(t)dN_t + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

Least-square contrast for one process

- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.
- Let η be an intensity candidate.

Least-square contrast

$$\gamma_T(\eta) = -\frac{2}{T}\int_0^T \eta(t)dN_t + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

- η is better if $\gamma_T(\eta)$ is small.

Least-square contrast for one process

- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.
- Let η be an intensity candidate.

Least-square contrast

$$\gamma_T(\eta) = -\frac{2}{T}\int_0^T \eta(t)dN_t + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

- η is better if $\gamma_T(\eta)$ is small.

$$\gamma_T(\eta) \simeq -\frac{2}{T}\int_0^T \eta(t)\lambda(t)dt + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

Least-square contrast for one process

- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.
- Let η be an intensity candidate.

Least-square contrast

$$\gamma_T(\eta) = -\frac{2}{T}\int_0^T \eta(t)dN_t + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

- η is better if $\gamma_T(\eta)$ is small.

$$\gamma_T(\eta) \simeq -\frac{2}{T}\int_0^T \eta(t)\lambda(t)dt + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

$$\simeq \frac{1}{T}\int_0^T [\eta(t)-\lambda(t)]^2 dt - \frac{1}{T}\int_0^T \lambda(t)^2 dt$$

Least-square contrast for one process

- It is the "mean" version of dN_t.
- A good estimate of the parameters should correspond to an intensity candidate close to the true one.
- Let η be an intensity candidate.

Least-square contrast

$$\gamma_T(\eta) = -\frac{2}{T}\int_0^T \eta(t)dN_t + \frac{1}{T}\int_0^T \eta(t)^2 dt$$

- η is better if $\gamma_T(\eta)$ is small.

$$\begin{aligned}\gamma_T(\eta) &\simeq -\frac{2}{T}\int_0^T \eta(t)\lambda(t)dt + \frac{1}{T}\int_0^T \eta(t)^2 dt \\ &\simeq \frac{1}{T}\int_0^T [\eta(t)-\lambda(t)]^2 dt - \frac{1}{T}\int_0^T \lambda(t)^2 dt\end{aligned}$$

- We will minimize γ_T for precise shape of intensity candidates.

Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t - T \leq b} = \alpha N_{[t-b, t-a]}$$

There is only one parameter α → minimizing γ_T

Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t-T \leq b} = \alpha N_{[t-b,t-a]}$$

There is only one parameter $\alpha \to$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N^2_{[t-b,t-a]} dt}$$

Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t - T \leq b} = \alpha N_{[t-b, t-a]}$$

There is only one parameter $\alpha \rightarrow$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N_{[t-b,t-a]}^2 dt}$$

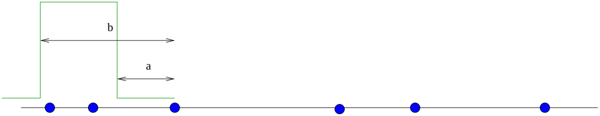
Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t-T \leq b} = \alpha N_{[t-b,t-a]}$$

There is only one parameter $\alpha \rightarrow$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N^2_{[t-b,t-a]} dt}$$



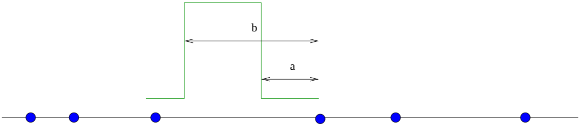
Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t-T \leq b} = \alpha N_{[t-b,t-a]}$$

There is only one parameter $\alpha \rightarrow$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N^2_{[t-b,t-a]} dt}$$



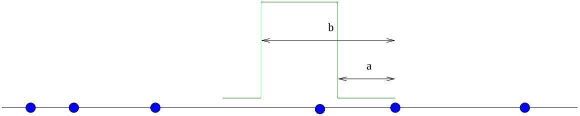
Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t-T \leq b} = \alpha N_{[t-b,t-a]}$$

There is only one parameter $\alpha \to$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N^2_{[t-b,t-a]} dt}$$



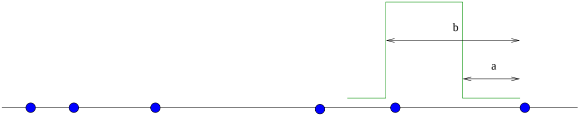
Least-square contrast on a toy model

Let $[a, b]$ an interval of $\mathbb{R}^+$.

$$\eta(t) = \sum_{T \in N} \alpha \mathbf{1}_{a \leq t-T \leq b} = \alpha N_{[t-b,t-a]}$$

There is only one parameter $\alpha \rightarrow$ minimizing γ_T

$$\hat{\alpha} = \frac{\int_0^T N_{[t-b,t-a]} dN_t}{\int_0^T N^2_{[t-b,t-a]} dt}$$



For the simple Hawkes process

- $\mathbb{L}_2 = \left\{ f = (\mu, g) \ : \ \text{Supp}(g) = [0, A], \|f\|^2 = \mu^2 + \int_0^A g^2(x)dx < +\infty \right\}$
- $s = (\nu, h)$ to estimate

For the simple Hawkes process

- $\mathbb{L}_2 = \left\{ f = (\mu, g) \ : \ \mathrm{Supp}(g) = [0, A], \|f\|^2 = \mu^2 + \int_0^A g^2(x)dx < +\infty \right\}$
- $s = (\nu, h)$ to estimate
- Intensity candidate : $\eta(t) = \Psi_f(t) = \mu + \int_{t-A}^{t-} g(t-u)dN_u$

For the simple Hawkes process

- $\mathbb{L}_2 = \left\{ f = (\mu, g) \ : \ \mathrm{Supp}(g) = [0, A], \|f\|^2 = \mu^2 + \int_0^A g^2(x)dx < +\infty \right\}$
- $s = (\nu, h)$ to estimate
- Intensity candidate : $\eta(t) = \Psi_f(t) = \mu + \int_{t-A}^{t-} g(t-u)dN_u$
- Least-square contrast :

$$\gamma_T(f) = -\frac{2}{T}\int_0^T \Psi_f(t)dN_t + \frac{1}{T}\int_0^T \Psi_f(t)^2 dt$$

For the simple Hawkes process

- $\mathbb{L}_2 = \left\{ f = (\mu, g) \ : \ \mathrm{Supp}(g) = [0, A], \|f\|^2 = \mu^2 + \int_0^A g^2(x)dx < +\infty \right\}$
- $s = (\nu, h)$ to estimate
- Intensity candidate : $\eta(t) = \Psi_f(t) = \mu + \int_{t-A}^{t-} g(t-u)dN_u$
- Least-square contrast :

$$\gamma_T(f) = -\frac{2}{T}\int_0^T \Psi_f(t)dN_t + \frac{1}{T}\int_0^T \Psi_f(t)^2 dt$$

- $\hat{s}_m = \arg\min_{f \in S_m} \gamma_T(f)$. no close formula in general.

Risk study

Only for simple Hawkes process

Risk

The risk of $\hat{s}$ is $\mathbb{E}(\|s - \hat{s}\|^2)$.

Risk study

Only for simple Hawkes process

Risk

The risk of $\hat{s}$ is $\mathbb{E}(\|s - \hat{s}\|^2)$.

One can show, under technical assumptions,

$$\mathbb{E}(\|\hat{s}_m - s\|^2) \leq \square \left[d(s, S_m)^2 + D_m \frac{\log T}{T} \right]$$

Risk study

Only for simple Hawkes process

Risk

The risk of $\hat{s}$ is $\mathbb{E}(\|s-\hat{s}\|^2)$.

One can show, under technical assumptions,

$$\mathbb{E}(\|\hat{s}_m - s\|^2) \leq \square \left[d(s,S_m)^2 + D_m \frac{\log T}{T} \right]$$

- The bigger is S_m the smaller is the bias, $d(s,S_m)^2$ but D_m is big.

Risk study

Only for simple Hawkes process

Risk

The risk of $\hat{s}$ is $\mathbb{E}(\|s - \hat{s}\|^2)$.

One can show, under technical assumptions,

$$\mathbb{E}(\|\hat{s}_m - s\|^2) \leq \square \left[d(s, S_m)^2 + D_m \frac{\log T}{T} \right]$$

- The bigger is S_m the smaller is the bias, $d(s, S_m)^2$ but D_m is big.
- → compromise: the best (or oracle) in a family is the one that minimizes the sum.

Model selection with dimension-based penalty

Joint work with S. Schbath (2010). Only for simple Hawkes process

Model selection with dimension-based penalty

Joint work with S. Schbath (2010). Only for simple Hawkes process

Models

$S_m = \left\{ f = (\mu, g), g \text{ piecewise constant on } m \right\}$ where m set of disjoint intervals in $[0, A]$ =model

Model selection with dimension-based penalty

Joint work with S. Schbath (2010). Only for simple Hawkes process

Models

$S_m = \left\{f = (\mu, g), g \text{ piecewise constant on } m\right\}$ where m set of disjoint intervals in $[0, A]$ =model

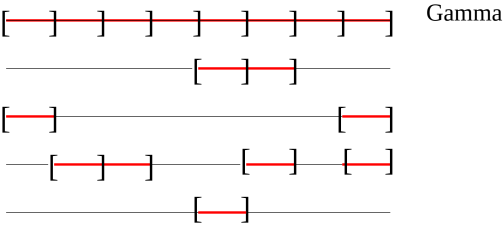
- NESTED Regular dyadic partitions constructed on Γ.
- REGULAR Regular partitions until a certain prescribed bandwidth
- IRREGULAR All irregular partitions contructed on Γ.

Model selection with dimension-based penalty

Joint work with S. Schbath (2010). Only for simple Hawkes process

Models

$S_m = \{f = (\mu, g), g$ piecewise constant on $m\}$ where m set of disjoint intervals in $[0, A]$ =model



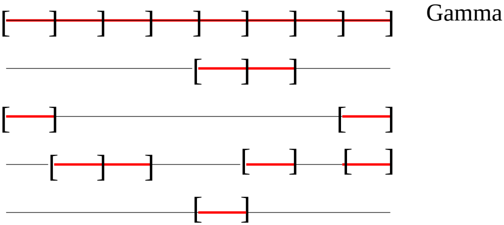
- ISLANDS

Model selection with dimension-based penalty

Joint work with S. Schbath (2010). Only for simple Hawkes process

Models

$S_m = \{f = (\mu, g), g$ piecewise constant on $m\}$ where m set of disjoint intervals in $[0, A]$ =model



- ISLANDS
  All models included in Γ. The most adapted to the biological question.

Model selection with dimension-based penality(2)

Only for simple Hawkes process
Let $\mathcal{M}_T$ be a family of models ($m \in \mathcal{M}_T$).

Model selection with dimension-based penality(2)

Only for simple Hawkes process

Let $\mathcal{M}_T$ be a family of models ($m \in \mathcal{M}_T$).

Let $\mathrm{pen} : \mathcal{M}_T \to \mathbb{R}^+$ and

Model selection with dimension-based penality(2)

Only for simple Hawkes process

Let $\mathcal{M}_T$ be a family of models ($m \in \mathcal{M}_T$).

Let $\mathrm{pen} : \mathcal{M}_T \to \mathbb{R}^+$ and

Penalized Criterion

$$\hat{m} = \mathrm{argmin}_{m \in \mathcal{M}_T} \left[\gamma_T(\hat{s}_m) + \mathrm{pen}(m) \right].$$

Model selection with dimension-based penality(2)

Only for simple Hawkes process

Let $\mathcal{M}_T$ be a family of models ($m \in \mathcal{M}_T$).

Let $\mathrm{pen} : \mathcal{M}_T \to \mathbb{R}^+$ and

Penalized Criterion

$$\hat{m} = \mathrm{argmin}_{m\in\mathcal{M}_T}\left[\gamma_T(\hat{s}_m) + \mathrm{pen}(m)\right].$$

Selected estimator

$$\tilde{s} = (\tilde{\nu}, \tilde{h}) = \hat{s}_{\hat{m}}.$$

Model selection with dimension-based penality(3)

Only for simple POSITIVE Hawkes process

Under technical assumptions, if

$$\mathrm{pen}(m) = \kappa Q(|m| + 1)\frac{\log(T)^2}{T},$$

with κ depending on unobserved parameters

Model selection with dimension-based penality(3)

Only for simple POSITIVE Hawkes process

Under technical assumptions, if

$$\mathrm{pen}(m) = \kappa Q(|m|+1)\frac{\log(T)^2}{T},$$

with κ depending on unobserved parameters then

$$\mathbb{E}(\|\tilde{s}-s\|)^2 \leq \square \inf_{m\in\mathcal{M}_T}\left[d(s,S_m)^2 + (|m|+1)\frac{\log(T)^2}{T}\right] + \square\frac{\#\{\mathcal{M}_T\}}{T^Q}.$$

Model selection with dimension-based penality(3)

Only for simple POSITIVE Hawkes process

Under technical assumptions, if

$$\mathrm{pen}(m) = \kappa Q(|m|+1)\frac{\log(T)^2}{T},$$

with κ depending on unobserved parameters then

$$\mathbb{E}(\|\tilde{s}-s\|)^2 \leq \square \inf_{m\in\mathcal{M}_T}\left[d(s,S_m)^2 + (|m|+1)\frac{\log(T)^2}{T}\right] + \square\frac{\#\{\mathcal{M}_T\}}{T^Q}.$$

- It is an oracle inequality!

Model selection with dimension-based penality(3)

Only for simple POSITIVE Hawkes process

Under technical assumptions, if

$$\mathrm{pen}(m) = \kappa Q(|m| + 1)\frac{\log(T)^2}{T},$$

with κ depending on unobserved parameters then

$$\mathbb{E}(\|\tilde{s} - s\|)^2 \leq \square \inf_{m \in \mathcal{M}_T} \left[d(s, S_m)^2 + (|m| + 1)\frac{\log(T)^2}{T} \right] + \square \frac{\#\{\mathcal{M}_T\}}{T^Q}.$$

- It is an oracle inequality!
- family grows with T at a moderate rate, especially for Islands

Model selection with dimension-based penality(3)

Only for simple POSITIVE Hawkes process

Under technical assumptions, if

$$\mathrm{pen}(m) = \kappa Q(|m|+1)\frac{\log(T)^2}{T},$$

with κ depending on unobserved parameters then

$$\mathbb{E}(\|\tilde{s}-s\|)^2 \leq \square \inf_{m\in\mathcal{M}_T}\left[d(s,S_m)^2 + (|m|+1)\frac{\log(T)^2}{T}\right] + \square\frac{\#\{\mathcal{M}_T\}}{T^Q}.$$

- It is an oracle inequality!
- family grows with T at a moderate rate, especially for Islands
- also adaptive minimax results for Hölder functions.

Model selection with dimension-based penality(4)

Only for simple Hawkes process

- In practice, one can use least-square estimators even for possibly negative h.

Model selection with dimension-based penality(4)

Only for simple Hawkes process

- In practice, one can use least-square estimators even for possibly negative h.
- Theoretical penalty not observed, Slope heuristics does not work especially for Islands

Model selection with dimension-based penality(4)

Only for simple Hawkes process

- In practice, one can use least-square estimators even for possibly negative h.
- Theoretical penalty not observed, Slope heuristics does not work especially for Islands
- An angle is perfectly clear on the simulations at the correct dimension:

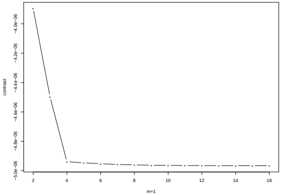
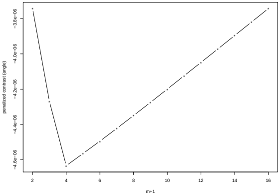
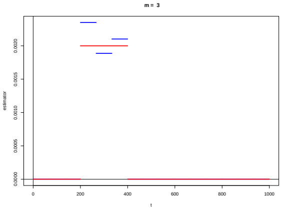
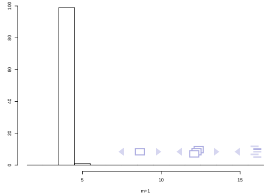
$$-\bar{k} = \frac{\gamma_T(\hat{s}_\Gamma) - \gamma_T(\hat{s}_1)}{|\Gamma| - 1}$$

and

$$\hat{m} = \mathrm{argmin}_{m \in \mathcal{M}_T} \gamma_T(\hat{s}_m) + \bar{k}(|m| + 1).$$

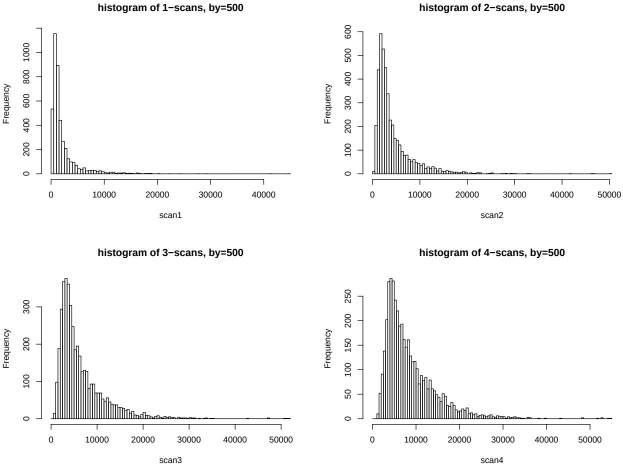
Illustration

Only for simple Hawkes process

Contrast	Penalized contrast
	
Chosen estimator	Selected dimension
	

On real data

Only for simple Hawkes process

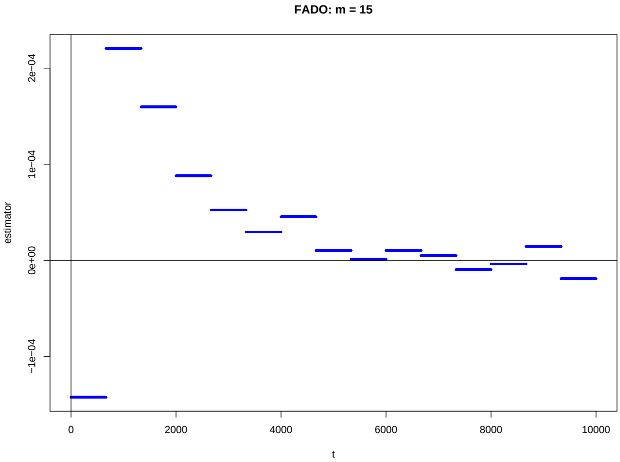


Analysis of the positions of the 4290 genes of E. coli

($T = 9288442$, $A = 10000$) (r-scans)

On real data

Only for simple Hawkes process

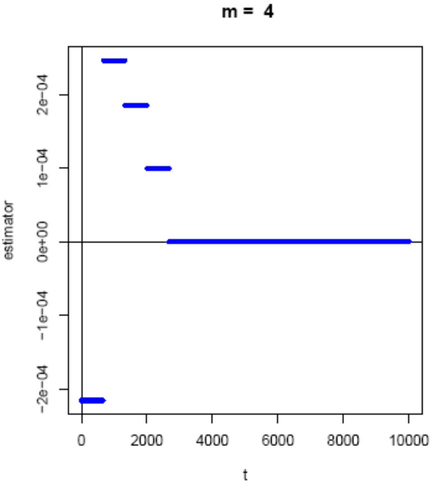


Analysis of the positions of the 4290 genes of E. coli

($T = 9288442$, $A = 10000$) (FADO)

On real data

Only for simple Hawkes process

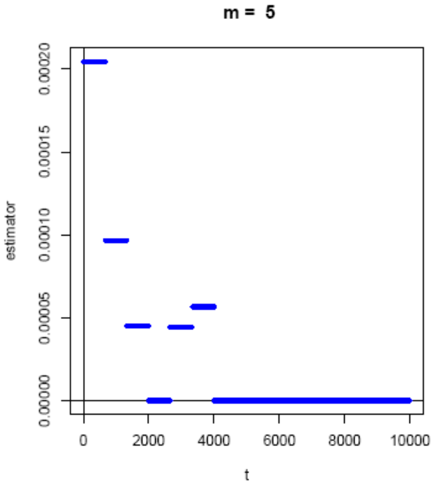


Analysis of the positions of the 4290 genes of E. coli

($T = 9288442,\ A = 10000$) (Islands)

On real data

Only for simple Hawkes process



Analysis of the 1036 occurrences of `tataat` for E. coli.

$(T = 9288442,\ A = 10000)$. (Islands)

Disadvantages of the dimension-based penalty method

- only for simple Hawkes process! (not more than grid with 15 - 26 bins)

Disadvantages of the dimension-based penalty method

- only for simple Hawkes process! (not more than grid with 15 - 26 bins)
- We miss part of the understanding because we do not take external information into account (here interaction `tataat` - genes).

Poissonian Interaction

Work in progress of Laure Sansonnet.

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, ..., U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^{n} h(t - U_i)$.

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, ..., U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^n h(t - U_i)$. (eventually, in practice $(\sum_{i=1}^n h(t - U_i))_+$)

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, \ldots, U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^{n} h(t - U_i)$. (eventually, in practice $\left(\sum_{i=1}^{n} h(t - U_i)\right)_+$)

Pros :

- A pairwise study of two processes is possible (for instance `tataat` and genes). If h has support in $\mathbb{R}_+$ = very special case of multivariate Hawkes process.

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, ..., U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^{n} h(t - U_i)$. (eventually, in practice $(\sum_{i=1}^{n} h(t - U_i))_+$)

Pros :

- A pairwise study of two processes is possible (for instance `tataat` and genes). If h has support in $\mathbb{R}_+$ = very special case of multivariate Hawkes process.
- h with support in $[-A, A]$ possible: we look at both direction at the same time.
- unbiased estimator of $\int h(t)\varphi(t)dt$.

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, ..., U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^{n} h(t - U_i)$. (eventually, in practice $(\sum_{i=1}^{n} h(t - U_i))_+$)

Pros :

- A pairwise study of two processes is possible (for instance `tataat` and genes). If h has support in $\mathbb{R}_+$ = very special case of multivariate Hawkes process.
- h with support in $[-A, A]$ possible: we look at both direction at the same time.
- unbiased estimator of $\int h(t)\varphi(t)dt$.
- Thresholding estimator of h exists. Threshold data-driven !

Poissonian Interaction

Work in progress of Laure Sansonnet.We observe two processes:

1. Parents : $U_1, \ldots, U_n$, n iid uniform random variables on $[0, T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^n h(t - U_i)$. (eventually, in practice $\left(\sum_{i=1}^n h(t - U_i)\right)_+$)

Pros :

- A pairwise study of two processes is possible (for instance `tataat` and genes). If h has support in $\mathbb{R}_+$ = very special case of multivariate Hawkes process.
- h with support in $[-A, A]$ possible: we look at both direction at the same time.
- unbiased estimator of $\int h(t)\varphi(t)dt$.
- Thresholding estimator of h exists. Threshold data-driven !
- Computation quite fast. Precision $>>$ Islands

Poissonian Interaction

Work in progress of Laure Sansonnet. We observe two processes:

1. Parents : $U_1, \ldots, U_n$, n iid uniform random variables on $[0.T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^n h(t-U_i)$. (eventually, in practice $(\sum_{i=1}^n h(t-U_i))_+$)

Cons:

- The presence of Parents will not be explained and is not linked to Children.

Poissonian Interaction

Work in progress of Laure Sansonnet. We observe two processes:

1. Parents : $U_1, \ldots, U_n$, n iid uniform random variables on $[0.T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^n h(t - U_i)$. (eventually, in practice $(\sum_{i=1}^n h(t - U_i))_+$)

Cons:

- The presence of Parents will not be explained and is not linked to Children.
- Crucial to say who are the parents/ the children (it is not symetric!).

Poissonian Interaction

Work in progress of Laure Sansonnet. We observe two processes:

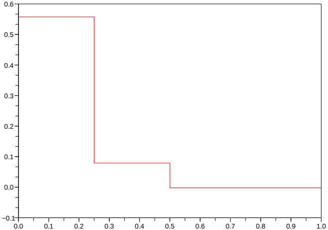
1. Parents : $U_1, \dots, U_n$, n iid uniform random variables on $[0.T]$.
2. Children : Poisson process with intensity $\sum_{i=1}^{n} h(t - U_i)$. (eventually, in practice $(\sum_{i=1}^{n} h(t - U_i))_+$)

Cons:

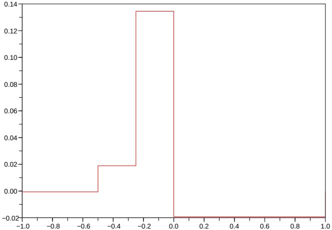
- The presence of Parents will not be explained and is not linked to Children.
- Crucial to say who are the parents/ the children (it is not symetric!).
- For the moment, one cannot explain the Children position, even with an extra spontaneous apparition.

Poissonian Interaction (2)

Interaction `tataat` - genes.



Parents = `tataat`



Parents = Genes

Full Multivariate Hawkes processes and lasso procedure

Joint Work with N.R. Hansen (Copenhagen) and V. Rivoirard (Dauphine), in progress.

Full Multivariate Hawkes processes and lasso procedure

Joint Work with N.R. Hansen (Copenhagen) and V. Rivoirard (Dauphine), in progress.

We want to estimate $s = \left((\nu_r, (h_\ell^{(r)})_{\ell=1,\dots,M})_{r=1,\dots,M} \right)$ in

$$\mathbb{L}_2 = \Big\{ f = \left((\mu_r, (g_\ell^{(r)})_{\ell=1,\dots,M})_{r=1,\dots,M} \right) \ \ / \ \ g_\ell^{(r)} \text{ with support in } (0,A] \text{ and } \|f\|^2 = \sum_r (\mu_r)^2 + \sum_r \sum_\ell \int_0^A (g_\ell^{(r)})^2(x)dx < \infty \Big\}.$$

Full Multivariate Hawkes processes and lasso procedure

Joint Work with N.R. Hansen (Copenhagen) and V. Rivoirard (Dauphine), in progress.

We want to estimate $s = \left((\nu_r, (h_\ell^{(r)})_{\ell=1,\dots,M})_{r=1,\dots,M}\right)$ in

$$\mathbb{L}_2 = \left\{ f = \left((\mu_r, (g_\ell^{(r)})_{\ell=1,\dots,M})_{r=1,\dots,M}\right) \quad / \quad g_\ell^{(r)} \text{ with support in } (0, A] \text{ and } \|f\|^2 = \sum_r (\mu_r)^2 + \sum_r \sum_\ell \int_0^A (g_\ell^{(r)})^2(x)dx < \infty \right\}.$$

Intensity candidate per mark

$\psi_f^{(r)}(t) = \mu_r + \sum_\ell \int_{-\infty}^t g_\ell^{(r)}(t-u)dN_u^{(\ell)}$.

Full Multivariate Hawkes processes and lasso procedure(2)

Least-square contrast (full form)

$\gamma_T(f) = \sum_r \gamma_T^{(r)}(f)$ where

$$\gamma_T^{(r)}(f) = -\frac{2}{T}\int_0^T \psi_f^{(r)}(t)dN_t^{(m)} + \frac{1}{T}\int_0^T \psi_f^{(r)}(t)^2 dt.$$

Full Multivariate Hawkes processes and lasso procedure(2)

Least-square contrast (full form)

$\gamma_T(f) = \sum_r \gamma_T^{(r)}(f)$ where

$$\gamma_T^{(r)}(f) = -\frac{2}{T}\int_0^T \psi_f^{(r)}(t)dN_t^{(m)} + \frac{1}{T}\int_0^T \psi_f^{(r)}(t)^2 dt.$$

If each $g_\ell^{(r)}$'s in f are decomposed on a finite orthonormal family of functions with cardinal K, then

- $f \rightarrow (\mathbf{a}_r)_{r\leq M}$ Each $\mathbf{a}_r$ of size $MK+1$.

Full Multivariate Hawkes processes and lasso procedure(2)

Least-square contrast (full form)

$\gamma_T(f) = \sum_r \gamma_T^{(r)}(f)$ where

$$\gamma_T^{(r)}(f) = -\frac{2}{T}\int_0^T \psi_f^{(r)}(t)dN_t^{(m)} + \frac{1}{T}\int_0^T \psi_f^{(r)}(t)^2 dt.$$

If each $g_\ell^{(r)}$'s in f are decomposed on a finite orthonormal family of functions with cardinal K, then

- $f \to (\mathbf{a}_r)_{r\leq M}$ Each $\mathbf{a}_r$ of size $MK+1$.
- $$\gamma_T^{(r)}(f) = -2\mathbf{a}_r^*\mathbf{b}_r + \mathbf{a}_r^*\mathbf{G}\mathbf{a}_r,$$

  where

  - **G** is a random observable matrix, independent of the mark r
  - $\mathbf{b}_r$ is also a random observable vector.

Full Multivariate Hawkes processes and lasso procedure(3)

The Lasso criterion can be expressed independently for each mark.

Lasso criterion

$$\hat{\mathbf{a}}_r = \mathrm{argmin}_{\mathbf{a}_r \in \mathbb{R}^{MK+1}} \{-2\mathbf{a}_r^*\mathbf{b}_r + \mathbf{a}_r^*\mathbf{G}\mathbf{a}_r + 2\mathbf{d}_r^*|\mathbf{a}_r|\}$$

- The vector $\mathbf{d}_r$ is not constant: it is random and depends on the index

Full Multivariate Hawkes processes and lasso procedure(3)

The Lasso criterion can be expressed independently for each mark.

Lasso criterion

$$\hat{\mathbf{a}}_r = \operatorname{argmin}_{\mathbf{a}_r \in \mathbb{R}^{MK+1}} \{-2\mathbf{a}_r^* \mathbf{b}_r + \mathbf{a}_r^* \mathbf{G} \mathbf{a}_r + 2\mathbf{d}_r^* |\mathbf{a}_r|\}$$

- The vector $\mathbf{d}_r$ is not constant: it is random and depends on the index
- → data-driven penalty (see also Bertin, Le Pennec, Rivoirard (2011) in the density setting)

Full Multivariate Hawkes processes and lasso procedure(3)

The Lasso criterion can be expressed independently for each mark.

Lasso criterion

$$\hat{\mathbf{a}}_r = \mathrm{argmin}_{\mathbf{a}_r \in \mathbb{R}^{MK+1}} \{ -2\mathbf{a}_r^* \mathbf{b}_r + \mathbf{a}_r^* \mathbf{G} \mathbf{a}_r + 2\mathbf{d}_r^* |\mathbf{a}_r| \}$$

- The vector $\mathbf{d}_r$ is not constant: it is random and depends on the index
- → data-driven penalty (see also Bertin, Le Pennec, Rivoirard (2011) in the density setting)
- Oracle inequality with "high" probability possible....