Some Lasso procedure for multivariate counting processes and its particular link with some exponential inequalities for martingales

N.R. Hansen, P. Reynaud-Bouret, V. Rivoirard

Copenhagen, CNRS - LJAD University of Nice, Dauphine

Banff, October 13th 2011

Counting processes

Point process

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

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population,

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

- N_A number of points of N in A,

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

- N_A number of points of N in A,
- $N_t = N_{[0,t]}$ counts the number of points between 0 and $t =$ counting process

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

- N_A number of points of N in A,
- $N_t = N_{[0,t]}$ counts the number of points between 0 and t = counting process
- $dN_t = \sum_{T \text{ point of } N} \delta_T$ = point measure

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

- N_A number of points of N in A,
- $N_t = N_{[0,t]}$ counts the number of points between 0 and t = counting process
- $dN_t = \sum_{T \text{ point of } N} \delta_T$ = point measure

Usually $\mathbb{R}$ is thought as time, but also the DNA strand (point= position of transcription regulatory elements).

Counting processes

Point process

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

Examples : breakdowns, earthquakes, lifetimes (or death) in a certain population, action potentials (detected by an electrode on a particular place of a neuron)

- N_A number of points of N in A,
- $N_t = N_{[0,t]}$ counts the number of points between 0 and $t =$ counting process
- $dN_t = \sum_{T \text{ point of } N} \delta_T =$ point measure

Usually $\mathbb{R}$ is thought as time, but also the DNA strand (point= position of transcription regulatory elements).Sometimes it's marked (or multivariate), ie $(N_t^{(m)})_{m=1,\dots,M}$.

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

An intensity is a predictable process wrt a filtration that defines "past". If it exists, $\int_0^t \lambda(x)dx$ is the compensator of N_t, ie

$$M_t = N_t - \int_0^t \lambda(x)dx$$

is a (local) martingale.

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

Predictable linear transformation

- For any parameter $f \in \mathcal{H}$, $f \mapsto \Psi_f^{(m)}$ is a known predictable linear transformation

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

Predictable linear transformation

- For any parameter $f \in \mathcal{H}$, $f \mapsto \Psi_f^{(m)}$ is a known predictable linear transformation
- (statistical model) $\lambda^{(m)}(t) = \Psi_s^{(m)}$ for some unknown parameter s.

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

Predictable linear transformation

- For any parameter $f \in \mathcal{H}$, $f \mapsto \Psi_f^{(m)}$ is a known predictable linear transformation
- (statistical model) $\lambda^{(m)}(t) = \Psi_s^{(m)}$ for some unknown parameter s.

Examples:

- $m = 1$, $\Psi_f = f$ with $f \in \mathbb{L}^2(\mathbb{R})$ = Poisson

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t ($x < t$).

Predictable linear transformation

- For any parameter $f \in \mathcal{H}$, $f \mapsto \Psi_f^{(m)}$ is a known predictable linear transformation
- (statistical model) $\lambda^{(m)}(t) = \Psi_s^{(m)}$ for some unknown parameter s.

Examples:

- $m = 1$, $\Psi_f = f$ with $f \in \mathbb{L}^2(\mathbb{R})$ = Poisson
- $\Psi_f^{(m)} = Y_t f(t, X_m)$ with $f \in \mathbb{L}^2(\mathbb{R} \times \mathcal{X})$ = Aalen multiplicative intensity (right censored survival data, Cox processes etc)

(Conditional) Intensity

(Conditional) Intensity

$t \to \lambda^{(m)}(t)$ where $\lambda^{(m)}(t)dt$ represents the probability to have a point in $N^{(m)}$ at time t conditionally to the past before t $(x < t)$.

Predictable linear transformation

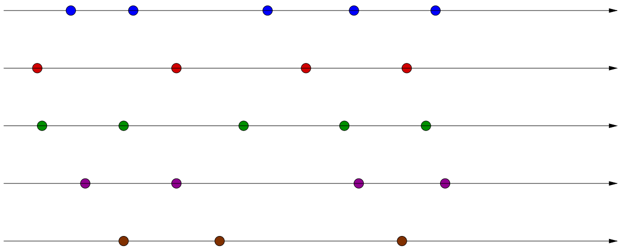
- For any parameter $f \in \mathcal{H}$, $f \mapsto \Psi_f^{(m)}$ is a known predictable linear transformation
- (statistical model) $\lambda^{(m)}(t) = \Psi_s^{(m)}$ for some unknown parameter s.

Examples:

- $m = 1$, $\Psi_f = f$ with $f \in \mathbb{L}^2(\mathbb{R})$ = Poisson
- $\Psi_f^{(m)} = Y_t f(t, X_m)$ with $f \in \mathbb{L}^2(\mathbb{R} \times \mathcal{X})$ = Aalen multiplicative intensity (right censored survival data, Cox processes etc)
- Hawkes ...

Multivariate Hawkes processes

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



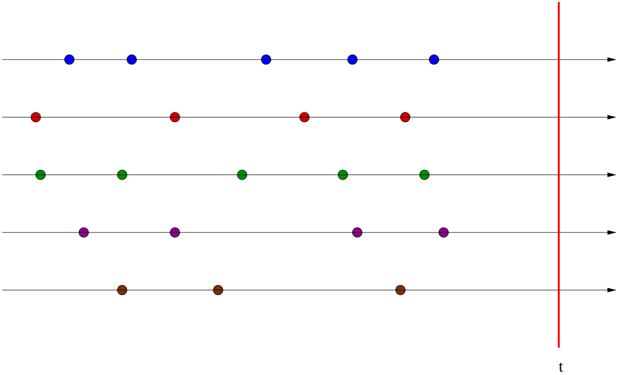
Multivariate Hawkes processes

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

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

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

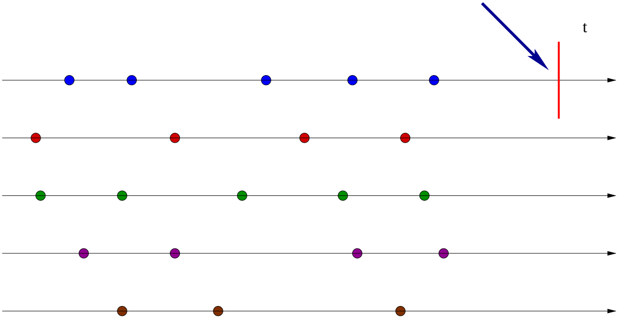
$\lambda^{(r)}(t) =$



Multivariate Hawkes processes

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

$$\lambda^{(1)}(t) = \quad \nu_1$$
$$\lambda^{(2)}(t) =$$
$$\lambda^{(r)}(t) =$$



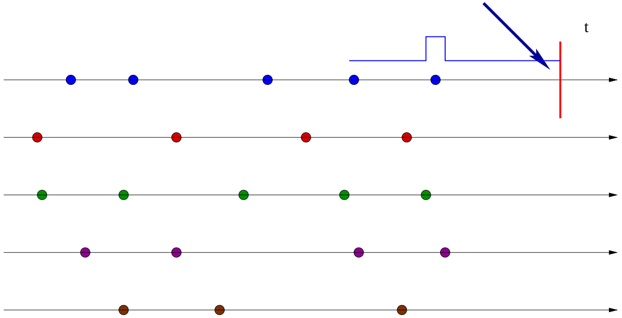
Multivariate Hawkes processes

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

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

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

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



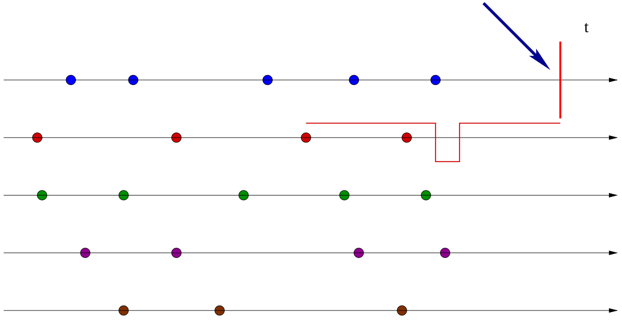
Multivariate Hawkes processes

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

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

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

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



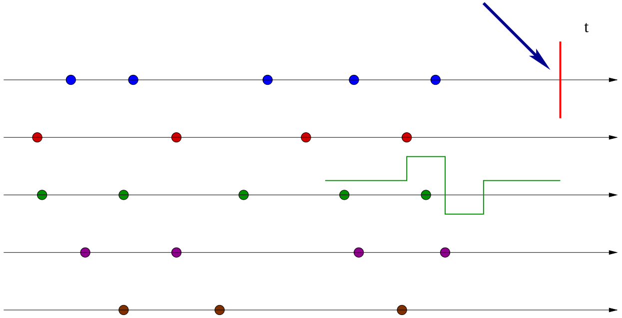
Multivariate Hawkes processes

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

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

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

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



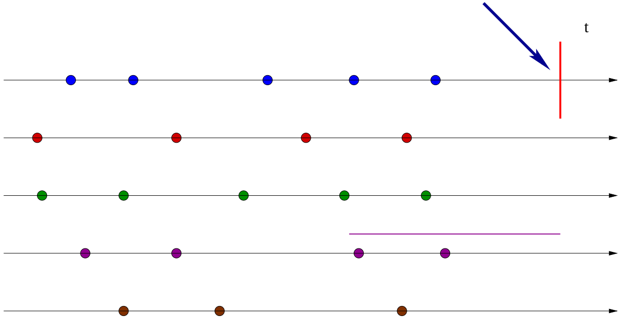
Multivariate Hawkes processes

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

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

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

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



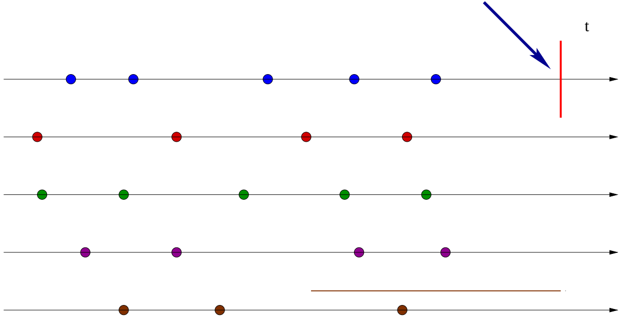
Multivariate Hawkes processes

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

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

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

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



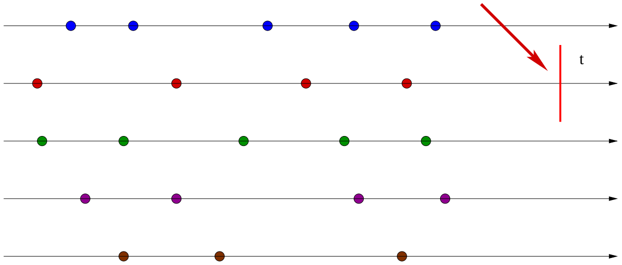
Multivariate Hawkes processes

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

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

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

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



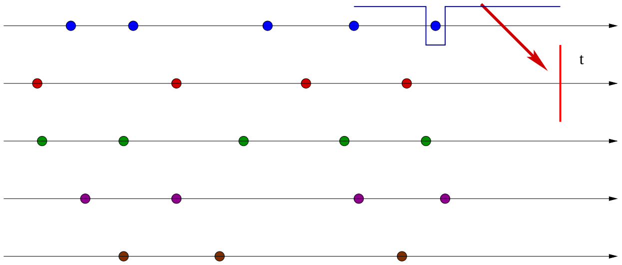
Multivariate Hawkes processes

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

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

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

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



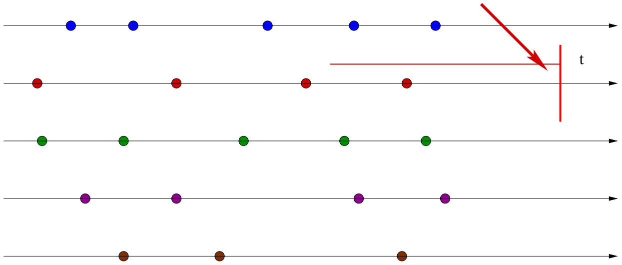
Multivariate Hawkes processes

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

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

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

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



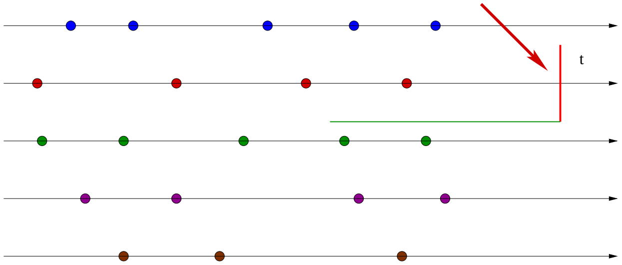
Multivariate Hawkes processes

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

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

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

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



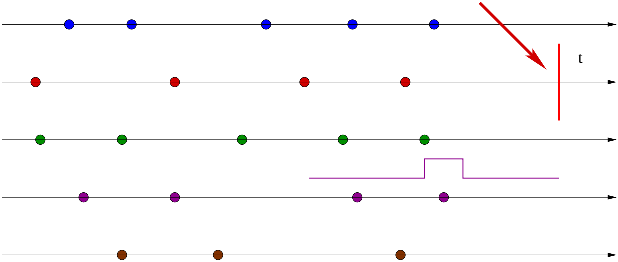
Multivariate Hawkes processes

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

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

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

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



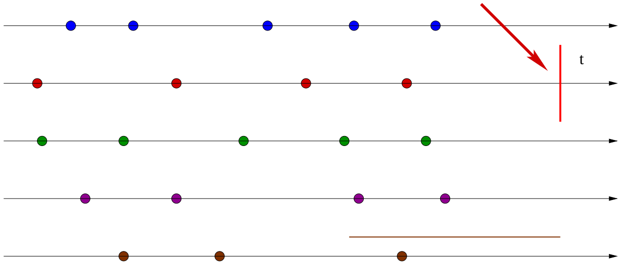
Multivariate Hawkes processes

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

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

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

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

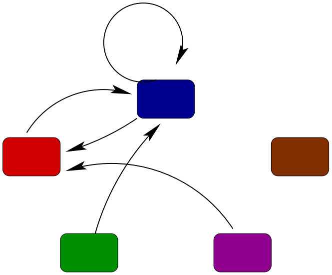


The multivariate Hawkes process(2)

Link with graphical model of local independence (see Didelez (2008)). Estimating the interaction functions and finding out which one is zero gives a picture of the synergy between the different processes (neurons, elements)

The multivariate Hawkes process(2)

Link with graphical model of local independence (see Didelez (2008)). Estimating the interaction functions and finding out which one is zero gives a picture of the synergy between the different processes (neurons, elements)



The multivariate Hawkes process(3)

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

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

The multivariate Hawkes process(3)

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

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

Intensity candidate per mark

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

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda\in\Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda\in\Lambda} a_\lambda\phi_\lambda$. (Hope : decomposition of s sparse)

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda\in\Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda\in\Lambda} a_\lambda \phi_\lambda$. (Hope : decomposition of s sparse)
- $\Psi_f^{(m)} = \sum_{\lambda=0}^{\Lambda} a_\lambda \Psi^{(m,\lambda)}$ and $\Psi^{(m,\lambda)} = \Psi_{\phi_\lambda}^{(m)}$.

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda \in \Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda \in \Lambda} a_\lambda \phi_\lambda$. (Hope : decomposition of s sparse)
- $\Psi_f^{(m)} = \sum_{\lambda=0}^{\Lambda} a_\lambda \Psi^{(m,\lambda)}$ and $\Psi^{(m,\lambda)} = \Psi_{\phi_\lambda}^{(m)}$.

Least-square contrast

$$\gamma(f) = \sum_{m=1}^{M} \left(-2 \int_0^T \Psi_f^{(m)}(t) dN_t^{(m)} + \int_0^T [\Psi_f^{(m)}(t)]^2 dt \right).$$

to minimize in order to find a good estimate.

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda\in\Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda\in\Lambda} a_\lambda \phi_\lambda$. (Hope : decomposition of s sparse)
- $\Psi_f^{(m)} = \sum_{\lambda=0}^{\Lambda} a_\lambda \Psi^{(m,\lambda)}$ and $\Psi^{(m,\lambda)} = \Psi_{\phi_\lambda}^{(m)}$.

Least-square contrast

$$\gamma(f) = \sum_{m=1}^{M} \left(-2\int_0^T \Psi_f^{(m)}(t)dN_t^{(m)} + \int_0^T [\Psi_f^{(m)}(t)]^2 dt \right).$$

to minimize in order to find a good estimate.

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda \in \Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda \in \Lambda} a_\lambda \phi_\lambda$. (Hope : decomposition of s sparse)
- $\Psi_f^{(m)} = \sum_{\lambda=0}^{\Lambda} a_\lambda \Psi^{(m,\lambda)}$ and $\Psi^{(m,\lambda)} = \Psi_{\phi_\lambda}^{(m)}$.

Least-square contrast

$$\gamma(f) = \sum_{m=1}^{M} \left(-2 \int_0^T \Psi_f^{(m)}(t) dN_t^{(m)} + \int_0^T [\Psi_f^{(m)}(t)]^2 dt \right).$$

to minimize in order to find a good estimate.

since $\gamma(f) \simeq -2 \sum_m \int \Psi_f^{(m)}(t) \Psi_s^{(m)}(t) dt + \sum_m \int [\Psi_f^{(m)}(t)]^2 dt$
minimal when $\Psi_f^{(m)} = \Psi_s^{(m)} \rightsquigarrow f = s$

Lasso estimate

- $\Phi = (\phi_\lambda)_{\lambda\in\Lambda}$ = dictionary in $\mathcal{H}$(Orthonormal family ...) and $f = \sum_{\lambda\in\Lambda} a_\lambda \phi_\lambda$. (Hope : decomposition of s sparse)
- $\Psi_f^{(m)} = \sum_{\lambda=0}^{\Lambda} a_\lambda \Psi^{(m,\lambda)}$ and $\Psi^{(m,\lambda)} = \Psi_{\phi_\lambda}^{(m)}$.

Least-square contrast

$$\gamma(f) = \sum_{m=1}^{M} \left(-2 \int_0^T \Psi_f^{(m)}(t) dN_t^{(m)} + \int_0^T [\Psi_f^{(m)}(t)]^2 dt \right).$$

to minimize in order to find a good estimate.

$\gamma(f) = -2a'b + a'Ga$ with

$$b_{\lambda_1} = \sum_{m=1}^{M} \int_0^T \Psi^{(m,\lambda_1)} dN_t^{(m)}, \quad G_{\lambda_1,\lambda_2} = \sum_{m=1}^{M} \int_0^T \Psi^{(m,\lambda_1)} \Psi^{(m,\lambda_2)} dt.$$

Lasso estimate(2)

Lasso estimate

$$\hat{a} \in \mathrm{argmin}_{a \in \mathbb{R}^{|\Lambda|}} \{-2a'b + a'Ga + 2d'|a|\}$$

where d, vector with positive coordinates.

Lasso estimate(2)

Lasso estimate

$$\hat{a} \in \operatorname{argmin}_{a \in \mathbb{R}^{|\Lambda|}} \{-2a'b + a'Ga + 2d'|a|\}$$

where d, vector with positive coordinates.
Because of the ℓ_1 penalty, the resulting estimator $\hat{s} = \sum_\lambda \hat{a}_\lambda \phi_\lambda$ will be sparse (very few non zeros coordinates).

Lasso estimate(2)

Lasso estimate

$$\hat{a} \in \mathrm{argmin}_{a \in \mathbb{R}^{|\Lambda|}} \{-2a'b + a'Ga + 2d'|a|\}$$

where d, vector with positive coordinates.
Because of the ℓ_1 penalty, the resulting estimator $\hat{s} = \sum_\lambda \hat{a}_\lambda \phi_\lambda$ will be sparse (very few non zeros coordinates).
Main point: How to choose d to have a good estimator ?

Lasso estimate(2)

Lasso estimate

$$\hat{a} \in \mathrm{argmin}_{a \in \mathbb{R}^{|\Lambda|}} \{-2a'b + a'Ga + 2d'|a|\}$$

where d, vector with positive coordinates.
Because of the ℓ_1 penalty, the resulting estimator $\hat{s} = \sum_\lambda \hat{a}_\lambda \phi_\lambda$ will be sparse (very few non zeros coordinates).
Main point: How to choose d to have a good estimator ?
Quadratic form (norm ?)

$$\|f\|^2_{T,M} = \sum_{m=1}^{M} \int_0^T [\Psi_f^{(m)}(t)]^2 dt.$$

An analytical result

Theorem

Let $c > 0$. *If*

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$,
2. $\forall \lambda \in \Lambda, \quad |b_\lambda - \bar{b}_\lambda| \leq d_\lambda$, *where* $\bar{b}_\lambda = \sum_{m=1}^{M} \int_0^T \Psi^{(m,\lambda)}(t) \Psi_s^{(m)}(t) dt$,

then, there exists an absolute constant C *such that*

$$\|\hat{s} - s\|_{T,M}^2 \leq C \inf_{a \in \mathbb{R}^{|\Lambda|}} \left\{ \left\| s - \sum_{\lambda \in \Lambda} a_\lambda \phi_\lambda \right\|_{T,M}^2 + c^{-1} \sum_{\lambda \in S(a)} (d_\lambda)^2 \right\},$$

where $S(a)$ *is the support of* a.

An analytical result

Theorem

Let $c > 0$*. If*

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$,
2. $\forall \lambda \in \Lambda, \quad |b_\lambda - \bar{b}_\lambda| \leq d_\lambda$, *where* $\bar{b}_\lambda = \sum_{m=1}^{M} \int_0^T \Psi^{(m,\lambda)}(t) \Psi_s^{(m)}(t) dt$,

then, there exists an absolute constant C *such that*

$$\|\hat{s} - s\|_{T,M}^2 \leq C \inf_{a \in \mathbb{R}^{|\Lambda|}} \left\{ \left\| s - \sum_{\lambda \in \Lambda} a_\lambda \phi_\lambda \right\|_{T,M}^2 + c^{-1} \sum_{\lambda \in S(a)} (d_\lambda)^2 \right\},$$

where $S(a)$ *is the support of* a*.*

Oracle inequality (see also Tsybakov (et al.), Bertin, Le Pennec, Rivoirard (2011))

Two probabilistic keys

One needs to control in probability,

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$.

Two probabilistic keys

One needs to control in probability,

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$.
   In particular this shows that $\|f\|_{T,M}$ is a norm with high probability on the dictionary.

Two probabilistic keys

One needs to control in probability,

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$.
   In particular this shows that $\|f\|_{T,M}$ is a norm with high probability on the dictionary.
   c important for theory, not for practice ....

Two probabilistic keys

One needs to control in probability,

1. $\inf_{x \in \mathbb{R}_*^{|\Lambda|}} \frac{x' G x}{\|x\|_{\ell^2}^2} \geq c$.
   In particular this shows that $\|f\|_{T,M}$ is a norm with high probability on the dictionary.
   c important for theory, not for practice ....
2. $\forall \lambda \in \Lambda, \quad |\sum_{m=1}^{M} \int_0^T \Psi^{(m,\lambda)}(t)(dN_t^{(m)} - \Psi_s^{(m)}(t)dt)| \leq d_\lambda,$
   Choice of d_λ crucial to have a full data-driven procedure

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}.

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket).

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)
   - such a $d = \sqrt{2\gamma\hat{v}x}$ is definitely bad for the estimation procedure when $\gamma < 1$.

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)
    - such a $d = \sqrt{2\gamma\hat{v}x}$ is definitely bad for the estimation procedure when $\gamma < 1$.
    - is good for moderate γ

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)
    - such a $d = \sqrt{2\gamma \hat{v} x}$ is definitely bad for the estimation procedure when $\gamma < 1$.
    - is good for moderate γ
    - becomes bad again if γ too large

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)
    - such a $d = \sqrt{2\gamma\hat{v}x}$ is definitely bad for the estimation procedure when $\gamma < 1$.
    - is good for moderate γ
    - becomes bad again if γ too large
    - optimal on simulations when $\gamma = 1$

Aim

1. One needs to find a data-driven bound $d(x)$ such that if $M_T = \int_0^T H_t(dN_t - \lambda(t)dt)$ and H_s predictable, $\mathbb{P}(M_T \geq d(x))$ exponentially small - of order e^{-x}. Indeed, we will control $|\Lambda| \simeq T^\alpha$ of them, $x \simeq \log(T)$ (not the large deviation regime !)
2. d should be as close as possible to the "CLT" rate ie $d(x) \simeq \sqrt{2vx}$ with v the variance of the process (or bracket). Indeed, (Bertin, Le Pennec, Rivoirard / RB Rivoirard / RB, Rivoirard, Tuleau-Malot) in other settings ($\hat{v}$ unbiased estimate of v)
    - such a $d = \sqrt{2\gamma\hat{v}x}$ is definitely bad for the estimation procedure when $\gamma < 1$.
    - is good for moderate γ
    - becomes bad again if γ too large
    - optimal on simulations when $\gamma = 1$

Existing exponential inequalities

- (classical, van de Geer (1995))
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\rho x} + Bx/3 \text{ and } \int_0^\tau H_t^2 \lambda(t)dt \leq \rho \text{ and } \sup_{t\leq\tau} |H_t| \leq B\right)$
  e^{-x}

Existing exponential inequalities

- (classical, van de Geer (1995))
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\rho x} + Bx/3 \text{ and } \int_0^\tau H_t^2 \lambda(t)dt \leq \rho \text{ and } \sup_{t\leq\tau} |H_t| \leq B\right)$
  e^{-x}
- (Dzhaparidze and van Zanten (2001))
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\theta x} \text{ and } \int_0^\tau H_t^2 \lambda(t)dt + \int_0^\tau H_t^2 dN_t \leq \theta\right) \leq e^{-x}.$

Existing exponential inequalities

- (classical, van de Geer (1995))
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\rho x} + Bx/3 \text{ and } \int_0^\tau H_t^2 \lambda(t)dt \leq \rho \text{ and } \sup_{t\leq\tau} |H_t| \leq B\right)$
  e^{-x}
- (Dzhaparidze and van Zanten (2001))
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\theta x} \text{ and } \int_0^\tau H_t^2 \lambda(t)dt + \int_0^\tau H_t^2 dN_t \leq \theta\right) \leq e^{-x}.$
- (Dzhaparidze and van Zanten (2001), Barlow, Jacka, Yor (1986), de la Peña (1999) and Bercu and Touati (2008)) If symetric (or heavy on the left)
  $\mathbb{P}\left(M_\tau \geq \sqrt{2\xi x} \text{ and } \int_0^\tau H_t^2 dN_t \leq \xi\right) \leq e^{-x},$

One satisfying exponential inequality

Theorem

Let $B > 0$ and $v > w > 0$. For every $x > 0$ and $\mu > 0$ such that $\mu > \phi(\mu)$, define

$$\hat{V}_t^{\mu} = \frac{\mu}{\mu - \phi(\mu)} \int_0^t H_s^2 dN_s + \frac{B^2 x}{\mu - \phi(\mu)},$$

where $\phi(u) = \exp(u) - 1 - u$. Then for any almost surely finite stopping time τ and any $\varepsilon > 0$

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)\hat{V}_\tau^{\mu} x} + \frac{Bx}{3} \text{ and } w \leq \hat{V}_\tau^{\mu} \leq v \text{ and } \sup_{t \in [0,\tau]} |H_t| \leq B \right)$$

$$\leq 2 \frac{\log(v/w)}{\log(1+\varepsilon)} e^{-x}.$$

inspired by Lipster and Spokoiny (2000)