Estimation non-paramétrique dans un modèle d'interactions poissoniennes et application à des données génomiques

Laure SANSONNET - Université Paris-Sud 11

Mardi 17 avril 2012
Colloque des jeunes probabilistes et statisticiens
- CIRM Marseille -

Contents

Biological motivation

- Two given events modeled by point processes $P1$ and $P2$: how does $P1$ influence $P2$?
- Any type of interaction, for example:
  in neurosciences, in economics, in genomics, ...
- "DNA case": study of favored or avoided distances between two given motifs along a genome.
- motif = sequence of letters in the alphabet {a,c,g,t}.
- Genomes are long and motifs of interest are short.
  $\rightarrow$ we work in a continuous framework.
  $\rightarrow$ occurrences of a motif = a point process lying in $[0; T]$, where T is the normalized length of the studied genome.
- To study the influence of $P2$ on $P1$, we just invert their roles in the model.

Poisson process on the real line

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

- N_A number of points of N in A,
- $dN = \sum_{X \in N} \delta_X$.

Poisson process

- N_A obeys a Poisson law $\mathcal{P}(\nu(A))$,
- if $A_1, \ldots, A_\ell$ are disjoint measurable sets, $N_{A_1}, \ldots, N_{A_\ell}$ are independent random variables.

ν is a measure called "mean measure".
Generally, $d\nu(t) = h(t)\, dt$.
If $h = constant$, N is a homogeneous Poisson process.

Our model

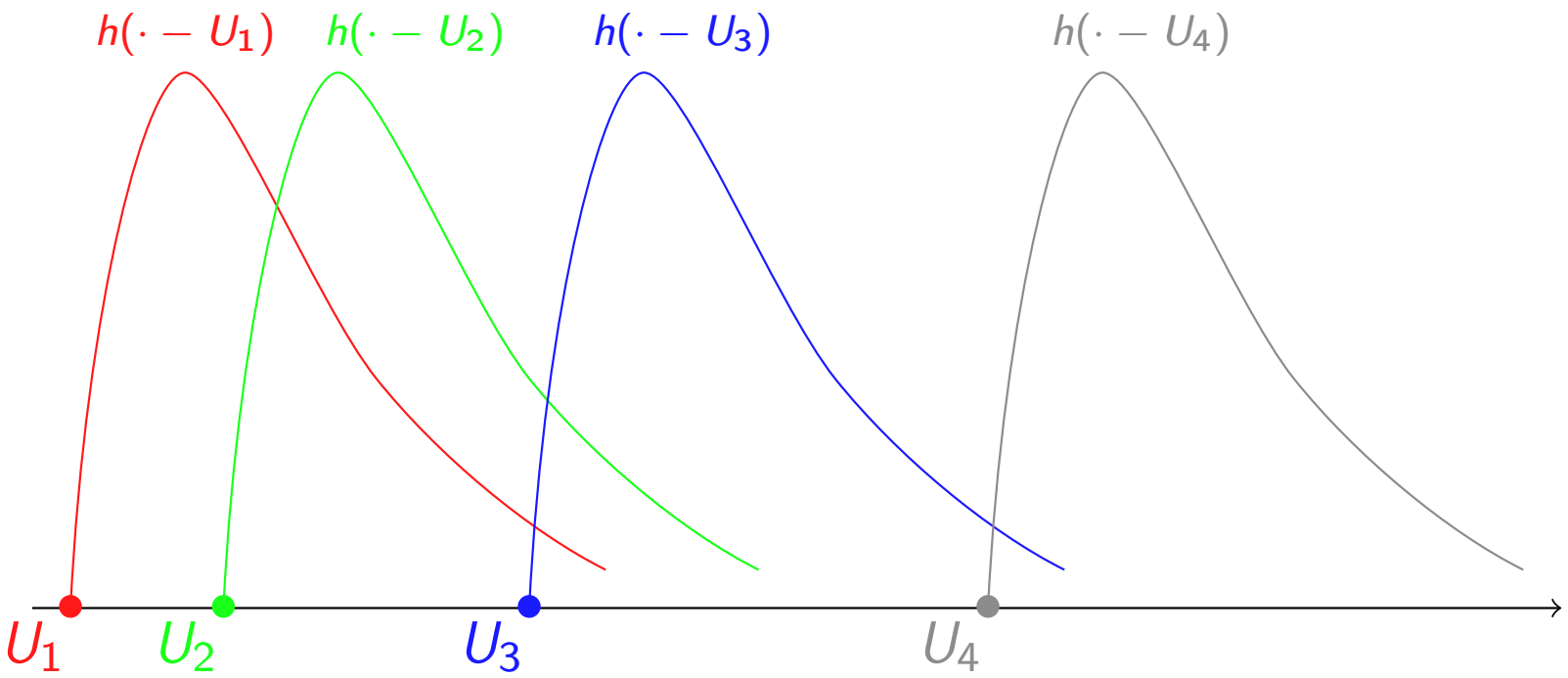
We observe the occurrences of both given motifs:

Our model

We observe the occurrences of both given motifs:

- Parents : $U_1, \ldots, U_n$ i.i.d. uniform random variables on $[0; T]$.

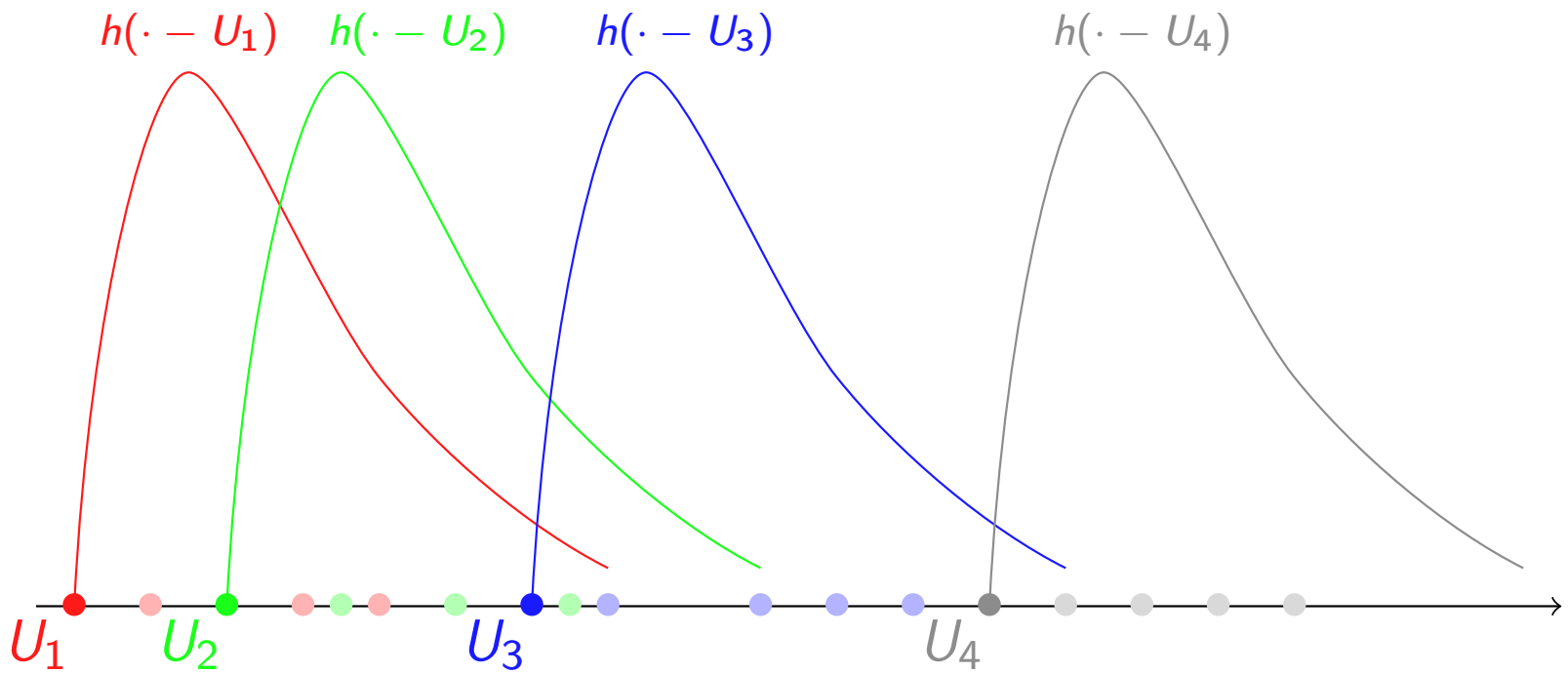
Our model



We observe the occurrences of both given motifs:

- Parents : $U_1, \ldots, U_n$ i.i.d. uniform random variables on $[0; T]$.

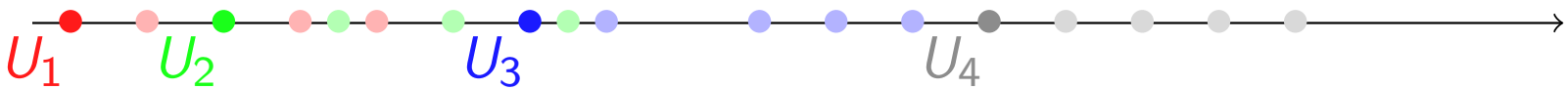
Our model



We observe the occurrences of both given motifs:

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

Our model



We observe the occurrences of both given motifs:

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

Aim: Estimate h.

In genomics:

- The first motif of interest is a rare word and is modeled by a homogeneous Poisson process N^0 on $[0; T]$.
- Conditionally to the event "the number of points falling into $[0; T]$ is n", the points of N^0 (i.e. the parents) obey the same law as a n-sample of uniform random variables on $[0; T]$.
- With very high probability, n is proportional to T.
  $\rightarrow$ the asymptotic considered in genomics: "DNA case".

In genomics:

- The first motif of interest is a rare word and is modeled by a homogeneous Poisson process N^0 on $[0; T]$.
- Conditionally to the event "the number of points falling into $[0; T]$ is n", the points of N^0 (i.e. the parents) obey the same law as a n-sample of uniform random variables on $[0; T]$.
- With very high probability, n is proportional to T.
  $\rightarrow$ the asymptotic considered in genomics: "DNA case".

Hawkes process:

- Gusto and Schbath (2005), Reynaud-Bouret and Schbath (2010), Carstensen *et al.* (2010).

In genomics:

- The first motif of interest is a rare word and is modeled by a homogeneous Poisson process N^0 on $[0; T]$.
- Conditionally to the event "the number of points falling into $[0; T]$ is n", the points of N^0 (i.e. the parents) obey the same law as a n-sample of uniform random variables on $[0; T]$.
- With very high probability, n is proportional to T.
  $\rightarrow$ the asymptotic considered in genomics: "DNA case".

Hawkes process:

- Gusto and Schbath (2005), Reynaud-Bouret and Schbath (2010), Carstensen *et al.* (2010).

Our model:

- no phenomenons of spontaneous apparition and self-excitation,
- but a nonparametric method of estimation, using a wavelet thresholding rule (no sparsity issues) and a double asymptotic.

Framework

- Assumption: $h \in \mathbb{L}_1(\mathbb{R}) \cap \mathbb{L}_\infty(\mathbb{R})$.
- Decomposition of h on the Haar basis (obtained by dilatations and translations of $\phi = \mathbf{1}_{[0;1]}$ and $\psi = \mathbf{1}_{]\frac{1}{2};1]} - \mathbf{1}_{[0;\frac{1}{2}]}$):

$$h = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda \quad \text{with} \quad \beta_\lambda = \int_{\mathbb{R}} h(x)\varphi_\lambda(x)\,dx,$$

where $\Lambda = \{\lambda = (j,k) : j \geqslant -1, k \in \mathbb{Z}\}$ and $\forall x \in \mathbb{R}$,

$$\forall \lambda = (j,k) \in \Lambda, \varphi_\lambda(x) = \begin{cases} \phi(x-k) & \text{if } j = -1 \\ 2^{j/2}\psi(2^j x - k) & \text{otherwise} \end{cases}.$$

- For the theoretical results, we have used the decomposition of h on a particular biorthogonal wavelet basis, built by Cohen *et al.* (1992).

Framework

- Assumption: $h \in \mathbb{L}_1(\mathbb{R}) \cap \mathbb{L}_\infty(\mathbb{R})$.
- Decomposition of h on the Haar basis (obtained by dilatations and translations of $\phi = \mathbf{1}_{[0;1]}$ and $\psi = \mathbf{1}_{]\frac{1}{2};1]} - \mathbf{1}_{[0;\frac{1}{2}]}$):

$$h = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda \quad \text{with} \quad \beta_\lambda = \int_{\mathbb{R}} h(x)\varphi_\lambda(x)\,dx,$$

where $\Lambda = \{\lambda = (j,k) : j \geqslant -1, k \in \mathbb{Z}\}$ and $\forall x \in \mathbb{R}$,

$$\forall \lambda = (j,k) \in \Lambda, \varphi_\lambda(x) = \begin{cases} \phi(x-k) & \text{if } j = -1 \\ 2^{j/2}\psi(2^j x - k) & \text{otherwise} \end{cases}.$$

- For the theoretical results, we have used the decomposition of h on a particular biorthogonal wavelet basis, built by Cohen *et al.* (1992).

Aim: Estimate the β_λ's.

Framework

For all λ in Λ, $\hat{\beta}_\lambda = \dfrac{G(\varphi_\lambda)}{n}$, with

$$G(\varphi_\lambda) = \int_{\mathbb{R}} \sum_{i=1}^{n} \left[\varphi_\lambda(t - U_i) - \frac{n-1}{n} \mathbb{E}_\pi(\varphi_\lambda(t - U)) \right] dN_t.$$

Framework

For all λ in Λ, $\hat{\beta}_\lambda = \dfrac{G(\varphi_\lambda)}{n}$, with

$$G(\varphi_\lambda) = \int_{\mathbb{R}} \sum_{i=1}^{n} \left[\varphi_\lambda(t - U_i) - \frac{n-1}{n} \mathbb{E}_\pi(\varphi_\lambda(t - U)) \right] dN_t.$$

Lemma

For all $\lambda \in \Lambda$, $\mathbb{E}(G(\varphi_\lambda)) = n \int_{\mathbb{R}} \varphi_\lambda(x) h(x)\, dx$, i.e. $\hat{\beta}_\lambda$ is an unbiased estimator for β_λ.
Furthermore, its variance is upper bounded as follows:

$$\mathrm{Var}(\hat{\beta}_\lambda) \leqslant C \left\{ \frac{1}{n} + \frac{n}{T^2} \right\}.$$

Description of our method

- Assumption: h is compactly supported in $[-A; A]$, with $A > 0$ ($A =$ the maximal memory along DNA sequences).

Description of our method

- Assumption: h is compactly supported in $[-A; A]$, with $A > 0$ ($A =$ the maximal memory along DNA sequences).
- $\Gamma = \left\{\lambda = (j, k) \in \Lambda : -1 \leqslant j \leqslant j_0, k \in \mathcal{K}_j\right\}$ a deterministic subset of Λ with $j_0 \in \mathbb{N}^* \rightarrow |\Gamma| \simeq 2^{j_0}$.

Description of our method

- Assumption: h is compactly supported in $[-A; A]$, with $A > 0$ ($A =$ the maximal memory along DNA sequences).
- $\Gamma = \left\{\lambda = (j,k) \in \Lambda : -1 \leqslant j \leqslant j_0, k \in \mathcal{K}_j\right\}$ a deterministic subset of Λ with $j_0 \in \mathbb{N}^* \to |\Gamma| \simeq 2^{j_0}$.
- Given some parameter $\gamma > 0$, for any $\lambda \in \Gamma$, the threshold:

$$\eta_\lambda(\gamma, \Delta) = \sqrt{2\gamma j_0 \widetilde{V}\left(\frac{\varphi_\lambda}{n}\right)} + \frac{\gamma j_0}{3} B\left(\frac{\varphi_\lambda}{n}\right) + \Delta \frac{N_{\mathbb{R}}}{n}$$

Description of our method

- Assumption: h is compactly supported in $[-A; A]$, with $A > 0$ ($A =$ the maximal memory along DNA sequences).
- $\Gamma = \left\{\lambda = (j,k) \in \Lambda : -1 \leqslant j \leqslant j_0, k \in \mathcal{K}_j\right\}$ a deterministic subset of Λ with $j_0 \in \mathbb{N}^* \to |\Gamma| \simeq 2^{j_0}$.
- Given some parameter $\gamma > 0$, for any $\lambda \in \Gamma$, the threshold:

$$\eta_\lambda(\gamma, \Delta) = \sqrt{2\gamma j_0 \widetilde{V}\left(\frac{\varphi_\lambda}{n}\right)} + \frac{\gamma j_0}{3} B\left(\frac{\varphi_\lambda}{n}\right) + \Delta \frac{N_{\mathbb{R}}}{n}$$

  - Δ a positive quantity (of order $\frac{j_0^2 2^{j_0/2}}{n} + \frac{j_0}{\sqrt{T}} + \frac{\sqrt{j_0 n}}{T}$ for theoretical results),
  - $N_{\mathbb{R}} =$ number of points of the process N lying in $\mathbb{R}$,
  - $B\left(\frac{\varphi_\lambda}{n}\right) = \frac{1}{n}\left\|\sum_{i=1}^n \left[\varphi_\lambda(\cdot - U_i) - \frac{n-1}{n}\mathbb{E}_\pi(\varphi_\lambda(\cdot - U))\right]\right\|_\infty$,
  - $\widetilde{V}\left(\frac{\varphi_\lambda}{n}\right) = \frac{1}{n^2}\left(\hat{V}(\varphi_\lambda) + \sqrt{2\gamma j_0 \hat{V}(\varphi_\lambda) B^2(\varphi_\lambda)} + 3\gamma j_0 B^2(\varphi_\lambda)\right)$,
  - $\hat{V}(\varphi_\lambda) = \int_{\mathbb{R}} \left(\sum_{i=1}^n \left[\varphi_\lambda(t - U_i) - \frac{n-1}{n}\mathbb{E}_\pi(\varphi_\lambda(t - U))\right]\right)^2 dN_t$.

Description of our method

$$\eta_\lambda(\gamma,\Delta)=\sqrt{2\gamma j_0\widetilde{V}\left(\frac{\varphi_\lambda}{n}\right)}+\frac{\gamma j_0}{3}B\left(\frac{\varphi_\lambda}{n}\right)+\Delta\frac{N_{\mathbb{R}}}{n}$$

- B, $\hat{V}$ and $\widetilde{V}$ only depend on the observations and can be exactly computed.
- $\tilde{\beta}$ the estimator of $\beta=(\beta_\lambda)_{\lambda\in\Lambda}$ associated with the previous thresholding rule:

$$\tilde{\beta}=\left(\hat{\beta}_\lambda\mathbf{1}_{|\hat{\beta}_\lambda|\geqslant\eta_\lambda(\gamma,\Delta)}\mathbf{1}_{\lambda\in\Gamma}\right)_{\lambda\in\Lambda}.$$

- $\tilde{h}=\sum_{\lambda\in\Lambda}\tilde{\beta}_\lambda\varphi_\lambda$ an estimator of h that only depends on the choice of (γ,Δ) and j_0 fixed later.

Main results

An oracle type inequality.

Theorem

We assume that $n \geqslant 2$, $j_0 \in \mathbb{N}^*$ *such that* $2^{j_0} \leqslant n < 2^{j_0+1}$, $\gamma > 2\log 2$ *and* Δ *is defined in a technical way.*
Then the estimator $\tilde{h}$, *previously defined, satisfies*

$$\mathbb{E}\left(\|\tilde{h} - h\|_2^2\right) \leqslant C_1 \inf_{m \subset \Gamma} \left\{ \sum_{\lambda \notin m} \beta_\lambda^2 + |m| \left[\frac{1}{n} + \frac{n}{T^2}\right] (\log n)^4 \right\} + C_2 \left[\frac{1}{n} + \frac{n}{T^2}\right].$$

Main results

An oracle type inequality.

Theorem

We assume that $n \geqslant 2$, $j_0 \in \mathbb{N}^*$ *such that* $2^{j_0} \leqslant n < 2^{j_0+1}$, $\gamma > 2\log 2$ *and* Δ *is defined in a technical way.*
Then the estimator $\tilde{h}$, *previously defined, satisfies*

$$\mathbb{E}\left(\|\tilde{h} - h\|_2^2\right) \leqslant C_1 \inf_{m \subset \Gamma} \left\{ \sum_{\lambda \notin m} \beta_\lambda^2 + |m| \left[\frac{1}{n} + \frac{n}{T^2}\right] (\log n)^4 \right\} + C_2 \left[\frac{1}{n} + \frac{n}{T^2}\right].$$

"DNA case" (n proportional to T)

$$\mathbb{E}\left(\|\tilde{h} - h\|_2^2\right) \leqslant C_1 \inf_{m \subset \Gamma} \left\{ \sum_{\lambda \notin m} \beta_\lambda^2 + \frac{(\log T)^4}{T} |m| \right\} + \frac{C_2}{T}$$

Main results

A minimax result on Besov balls still with n proportional to T.

$$\mathcal{B}^s_{2,\infty}(R) = \left\{ f = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda, \forall j \geqslant -1, \sum_{k \in \mathcal{K}_j} \beta^2_{(j,k)} \leqslant R^2 2^{-2js} \right\}$$

Main results

A minimax result on Besov balls still with n proportional to T.

$$\mathcal{B}^s_{2,\infty}(R) = \left\{ f = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda, \forall j \geqslant -1, \sum_{k \in \mathcal{K}_j} \beta^2_{(j,k)} \leqslant R^2 2^{-2js} \right\}$$

Corollary ("DNA case")

Let $R > 0$ and $s \in \mathbb{R}$ such that $0 < s < r + 1$. Assume that $h \in \mathcal{B}^s_{2,\infty}(R)$ and n is proportional to T.

Then the estimator $\tilde{h}$ satisfies

$$\mathbb{E}\left(\|\tilde{h} - h\|_2^2 \right) \leqslant C \left(\frac{(\log T)^4}{T} \right)^{\frac{2s}{2s+1}}.$$

Implementation procedure

From now on, we consider "DNA case": n is proportional to T.
Computation of the family of random thresholds $(\eta_\lambda(\gamma,\delta))_{\lambda\in\Gamma}$:

$$\eta_\lambda(\gamma,\delta)=\sqrt{2\gamma j_0\hat{V}\left(\frac{\varphi_\lambda}{n}\right)}+\frac{\gamma j_0}{3}B\left(\frac{\varphi_\lambda}{n}\right)+\frac{\delta}{\sqrt{T}}\frac{N_{\mathbb{R}}}{n},$$

where $\Delta=\frac{\delta}{\sqrt{T}}$ (because n is proportional to T).

Implementation procedure

From now on, we consider "DNA case": n is proportional to T.
Computation of the family of random thresholds $(\eta_\lambda(\gamma,\delta))_{\lambda\in\Gamma}$:

$$\eta_\lambda(\gamma,\delta)=\sqrt{2\gamma j_0\hat{V}\left(\frac{\varphi_\lambda}{n}\right)}+\frac{\gamma j_0}{3}B\left(\frac{\varphi_\lambda}{n}\right)+\frac{\delta}{\sqrt{T}}\frac{N_{\mathbb{R}}}{n},$$

where $\Delta=\frac{\delta}{\sqrt{T}}$ (because n is proportional to T).

- We set $j_0=5$ in the sequel.

Implementation procedure

From now on, we consider "DNA case": n is proportional to T.
Computation of the family of random thresholds $(\eta_\lambda(\gamma,\delta))_{\lambda\in\Gamma}$:

$$\eta_\lambda(\gamma,\delta)=\sqrt{2\gamma j_0\hat{V}\left(\frac{\varphi_\lambda}{n}\right)}+\frac{\gamma j_0}{3}B\left(\frac{\varphi_\lambda}{n}\right)+\frac{\delta}{\sqrt{T}}\frac{N_{\mathbb{R}}}{n},$$

where $\Delta=\frac{\delta}{\sqrt{T}}$ (because n is proportional to T).

- We set $j_0=5$ in the sequel.
- Computation of $\sum_{i=1}^{n}\left[\varphi_\lambda(t-U_i)-\frac{n-1}{n}\mathbb{E}_\pi(\varphi_\lambda(t-U))\right]$, with a cascade algorithm (inspired by Mallat (1989)), in order to compute the coefficients $\hat{\beta}_\lambda$, $\hat{V}$ and B.

Implementation procedure

From now on, we consider "DNA case": n is proportional to T.
Computation of the family of random thresholds $(\eta_\lambda(\gamma,\delta))_{\lambda\in\Gamma}$:

$$\eta_\lambda(\gamma,\delta)=\sqrt{2\gamma j_0\hat{V}\left(\frac{\varphi_\lambda}{n}\right)}+\frac{\gamma j_0}{3}B\left(\frac{\varphi_\lambda}{n}\right)+\frac{\delta}{\sqrt{T}}\frac{N_{\mathbb{R}}}{n},$$

where $\Delta=\frac{\delta}{\sqrt{T}}$ (because n is proportional to T).

- We set $j_0=5$ in the sequel.
- Computation of $\sum_{i=1}^{n}\left[\varphi_\lambda(t-U_i)-\frac{n-1}{n}\mathbb{E}_\pi(\varphi_\lambda(t-U))\right]$, with a cascade algorithm (inspired by Mallat (1989)), in order to compute the coefficients $\hat{\beta}_\lambda$, $\hat{V}$ and B.
- Choice of the parameters γ and δ?
  $\rightarrow$ calibration of parameters from a practical point of view.

Influence promoters/genes in *E. coli*

Data:

- the sequence composed of both strands of *E. coli* genome of length 4 639 221 bases (we took 10 000 bases for the maximal memory)
  $\rightarrow$ a sequence of length 9 288 442 ($= 2 * 4639221 + 10000$),
- locations of 4 290 genes (we took the positions of the first base of coding sequences),
- locations of 1 036 occurrences of the major promoter: `tataat`.

Influence promoters/genes in *E. coli*

Data:

- the sequence composed of both strands of *E. coli* genome of length 4 639 221 bases (we took 10 000 bases for the maximal memory)
  $\rightarrow$ a sequence of length 9 288 442 ($= 2 * 4639221 + 10000$),
- locations of 4 290 genes (we took the positions of the first base of coding sequences),
- locations of 1 036 occurrences of the major promoter: `tataat`.

For convenience, we work on a scale of 1 : 1000 and we set $T = 9289$ and so $A = 10$.

Influence promoters/genes in *E. coli*

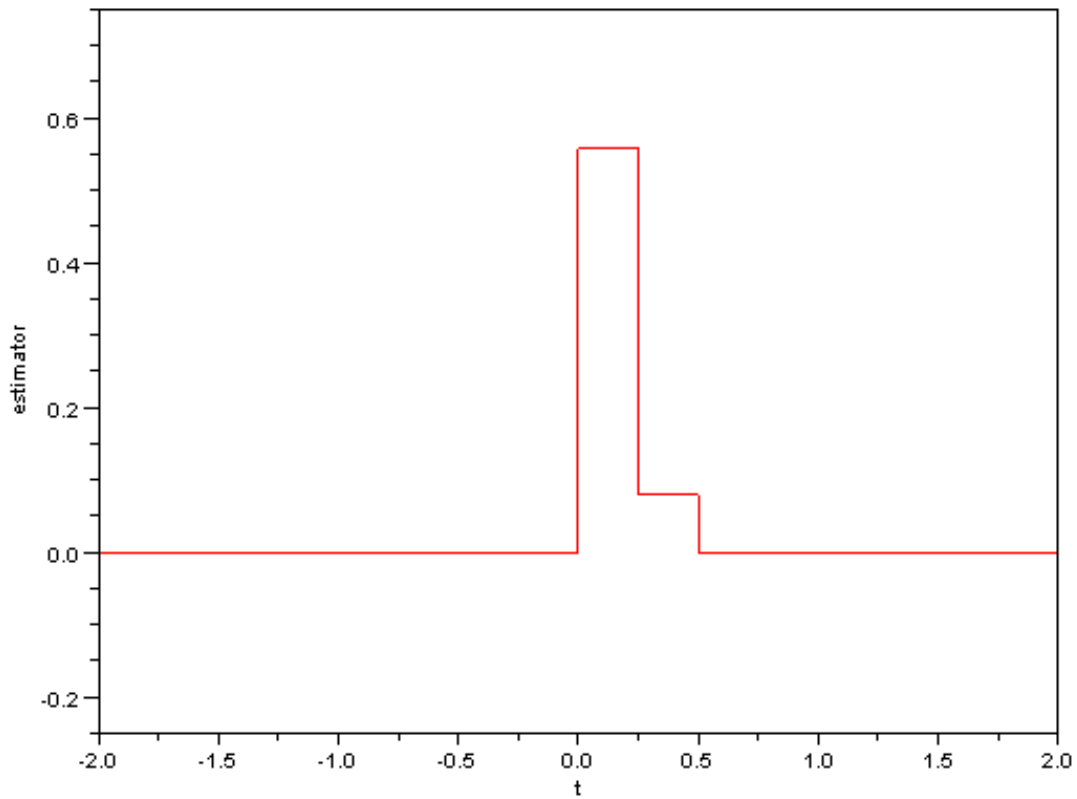
How does the DNA motif `tataat` influence genes?

- parents = `tataat`,
- children = genes.

Influence promoters/genes in *E. coli*

How does the DNA motif `tataat` influence genes?

- parents = `tataat`,
- children = genes.



Influence promoters/genes in *E. coli*

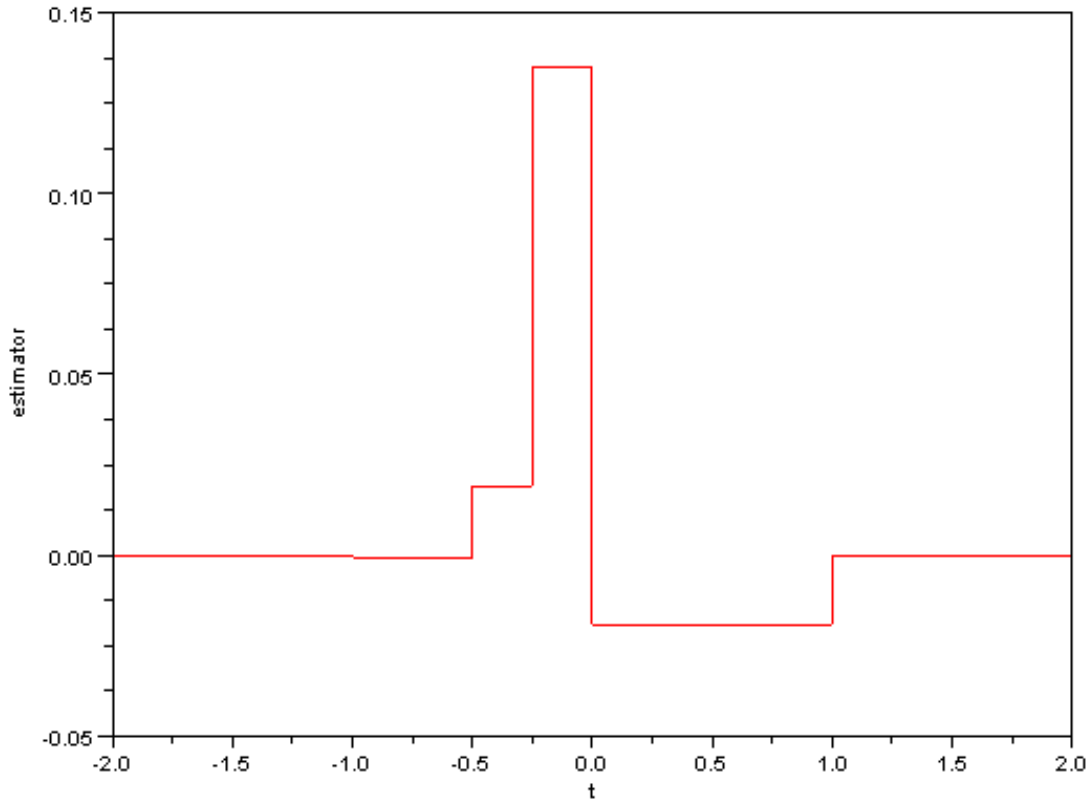
How does genes influence the DNA motif `tataat`?

- parents = genes,
- children = `tataat`.

Influence promoters/genes in *E. coli*

How does genes influence the DNA motif `tataat`?

- parents = genes,
- children = `tataat`.



Conclusion

- Our random thresholding procedure is optimal in the oracle and minimax setting.
- Some simulations illustrate the robustness of our procedure.
- The application to genomic data validates our procedure with a good detection of favored or avoided distances between occurrences of `tataat` and genes along the *E. coli* genome.

Conclusion

- Our random thresholding procedure is optimal in the oracle and minimax setting.
- Some simulations illustrate the robustness of our procedure.
- The application to genomic data validates our procedure with a good detection of favored or avoided distances between occurrences of `tataat` and genes along the *E. coli* genome.

Further possible extensions of our model:

- a more sophisticated model that takes into account the phenomenons of spontaneous apparition and self-excitation,
- an extension of our cascade algorithm to general wavelet bases and not only to Haar bases,
- a study of similar processes in the spatial framework.

Conclusion

- Our random thresholding procedure is optimal in the oracle and minimax setting.
- Some simulations illustrate the robustness of our procedure.
- The application to genomic data validates our procedure with a good detection of favored or avoided distances between occurrences of `tataat` and genes along the *E. coli* genome.

Further possible extensions of our model:

- a more sophisticated model that takes into account the phenomenons of spontaneous apparition and self-excitation,
- an extension of our cascade algorithm to general wavelet bases and not only to Haar bases,
- a study of similar processes in the spatial framework.

Thanks for your attention!