Apprentissage pour l'intensité d'événements avec points de rupture

Avec S. Gaïffas (CMAP - E. Polytechnique), A. Guilloux (LSTA - UPMC)

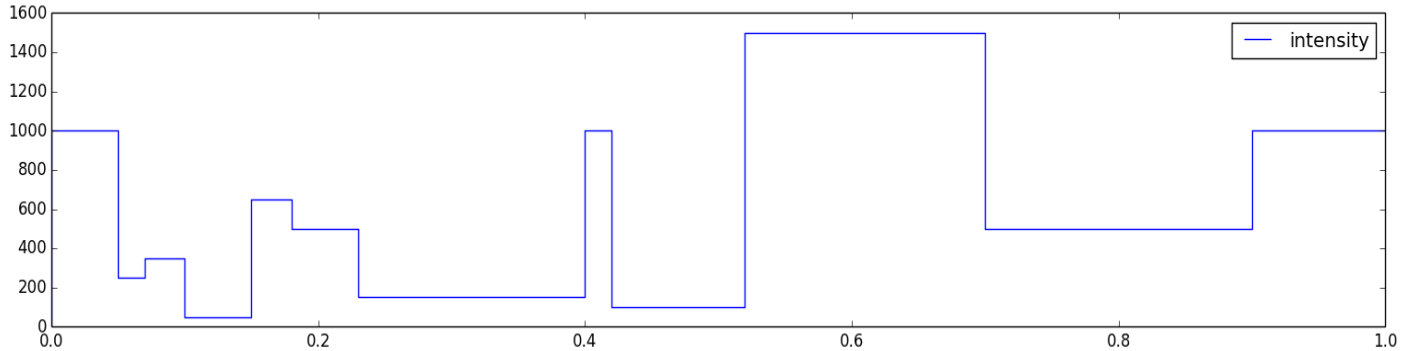
ElMokhtar EzZahdi Alaya

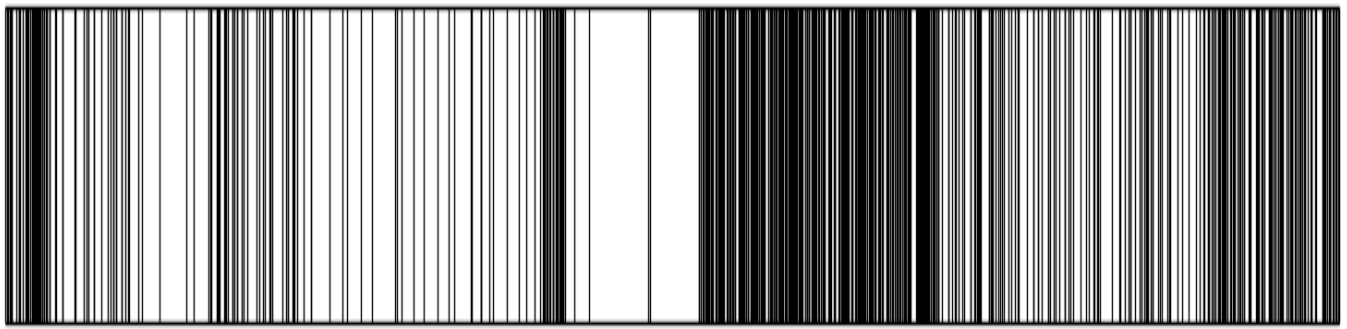
46^{e} Journées de Statistique - Rennes, 2 juin 2014

LSTA
Laboratoire de Statistique
Théorique et Appliquée

Processus de comptage

- $N = \{N(t)\}_{0 \leq t \leq 1}$ un processus de comptage.





- L'intensité de N est défini par :

$$\lambda_0(t)dt = \mathbb{P}[N \text{ a un saut dans } [t, t+dt) | \mathcal{F}(t^-)].$$

Motivations

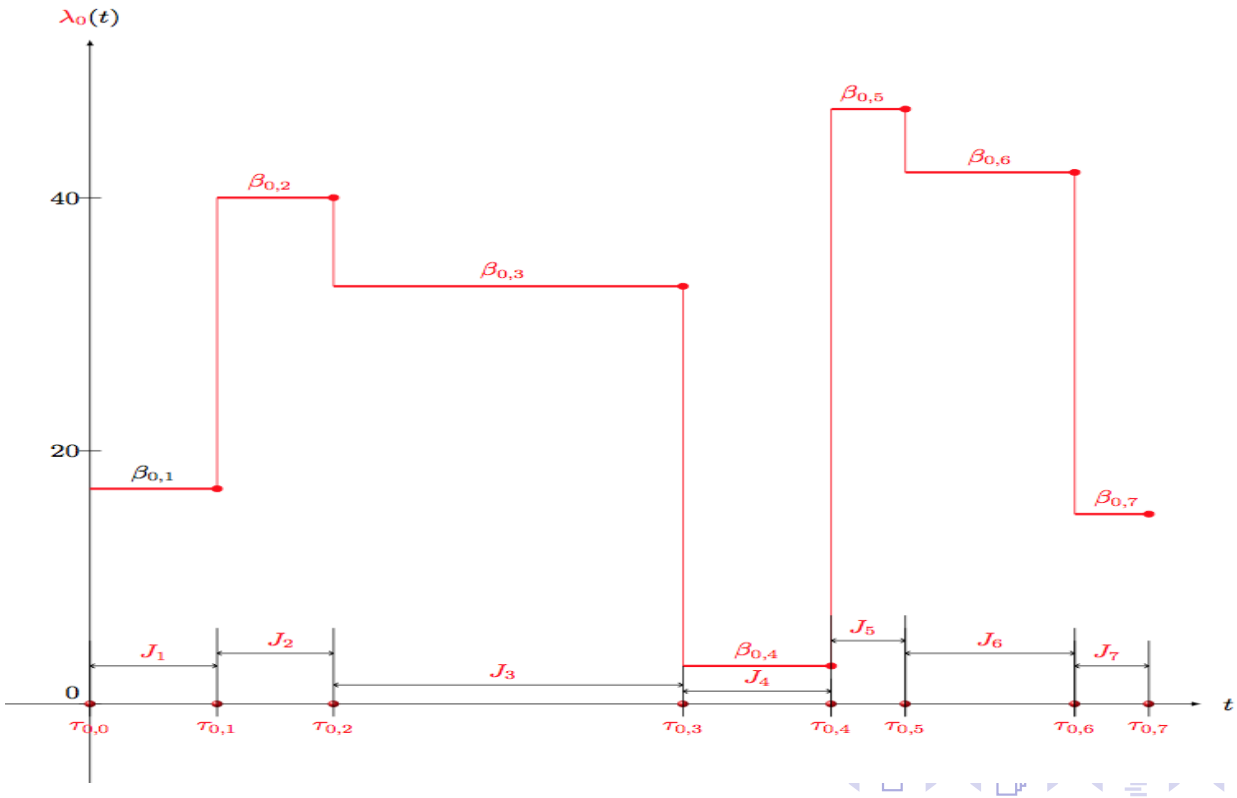
- Une problématique importante
  - segmentation de signaux audio
  - traitement d'image
  - analyse des séries temporelles
  - étude des profils génomiques
- Analyse des données d'expression génétique, données RNA-seq (séquençage de haute fréquence).
- Les données de RNA-seq peuvent être modélisées par des réplications d'un processus de comptage non homogène avec une intensité constante par morceaux (Shen, Zhang (2012)).

LSTA
Laboratoire de Statistique Théorique et Appliquée

Modèle

- Hypothèse de segmentation sparse à priori :

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbf{1}_{J_\ell}(t),\ 0 \leq t \leq 1.$$



LSTA Laboratoire de Statistique Théorique et Appliquée

Modèle

- $\tau_{0,0} = 0 < \tau_{0,1} < \cdots < \tau_{0,L_0-1} < \tau_{0,L_0} = 1.$
- Paramètres à estimer :
  - $(\tau_{0,\ell})$: points de rupture
  - $(\beta_{0,\ell})$: taille des sauts de λ_0
  - L_0 : nombre de points de ruptures.
- Données

On observe n copies i.i.d de N dans $[0,1]$, notées $N_1, \ldots, N_n$.

Équivalent à observer un processus de comptage N avec intensité $n\lambda_0$.

LSTA
Laboratoire de Statistique Théorique et Appliquée

Procédure

- Minimisation du risque empirique

$$R_n(\lambda) = \int_0^1 \lambda(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \int_0^1 \lambda(t) dN_i(t).$$

- Fixons $m = m_n \geq 1$. On approxime λ_0 dans

$$\Lambda_m = \left\{ \lambda_\beta = \sum_{j=1}^{m} \beta_{j,m} \lambda_{j,m} : \beta = [\beta_{j,m}]_{1 \leq j \leq m} \in \mathbb{R}_+^m \right\},$$

avec

$$\lambda_{j,m} = \sqrt{m} \mathbf{1}_{I_{j,m}} \quad \text{et} \quad I_{j,m} = \left(\frac{j-1}{m}, \frac{j}{m} \right].$$

- On considère l'estimateur

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}_+^m} \left\{ R_n(\lambda_\beta) + \|\beta\|_{\text{TV},\hat{w}} \right\}.$$

Procédure : data-driven TV pénalisation

$$\|\beta\|_{\mathrm{TV},\hat{w}} = \sum_{j=2}^{m} \hat{w}_j |\beta_j - \beta_{j-1}|.$$

- Le vecteur de pondérations $[\hat{w}_j]_{1 \leq j \leq m}$ ($\hat{w}_j \geq 0$) contrôle la sparsité dans les différences successives du vecteur β.
- La forme de $\hat{w}_j$ sera précisée dans la suite.
- L'estimateur de λ_0 est donné par

$$\hat{\lambda} = \lambda_{\hat{\beta}}.$$

LSTA
Laboratoire de Statistique Théorique et Appliquée

Inégalité d'oracle

- $\hat{S}$: support du gradient de $\hat{\beta}$,

$$\hat{S} = \{j : \hat{\beta}_{j,m} \neq \hat{\beta}_{j-1,m} \text{ pour } j = 2, \ldots, m\}.$$

- $\hat{L}$: nombre des points de ruptures estimés défini par $\hat{L} = |\hat{S}|$.

Théorème - Alaya et al. (2014)

Fixons $x > 0$. Supposons que $\hat{L} \leq L_{\max}$ *p.s.*, alors

$$\|\hat{\lambda} - \lambda_0\|^2 \leq \inf_{\beta \in \mathbb{R}_+^{\boldsymbol{m}}} \|\lambda_\beta - \lambda_0\|_2^2 + 6(L_{\max} + 2(L_0 - 1)) \max_{j=1,\ldots,m} \hat{w}_j{}^2 + K_1 \frac{\|\lambda_0\|_\infty \big(x + L_{\max}(1 + \log m)\big)}{n} + K_2 \frac{m\big(x + L_{\max}(1 + \log m)\big)^2}{n^2},$$

avec une probabilité supérieure à $1 - L_{\max} e^{-x}$, où $\|\lambda_0\|_\infty = \sup_{t \in [0,1]} \lambda_0(t)$, $K_1 = 1670.89$, et $K_2 = 6683.53$.

Inégalité d'oracle

- Avec

$$\hat{w}_j = 5.66\sqrt{\frac{m(x+\log m+\hat{h}_{n,x,j})\hat{V}_j}{n}}+9.31\frac{\sqrt{m}(x+1+\log m+\hat{h}_{n,x,j})}{n}.$$

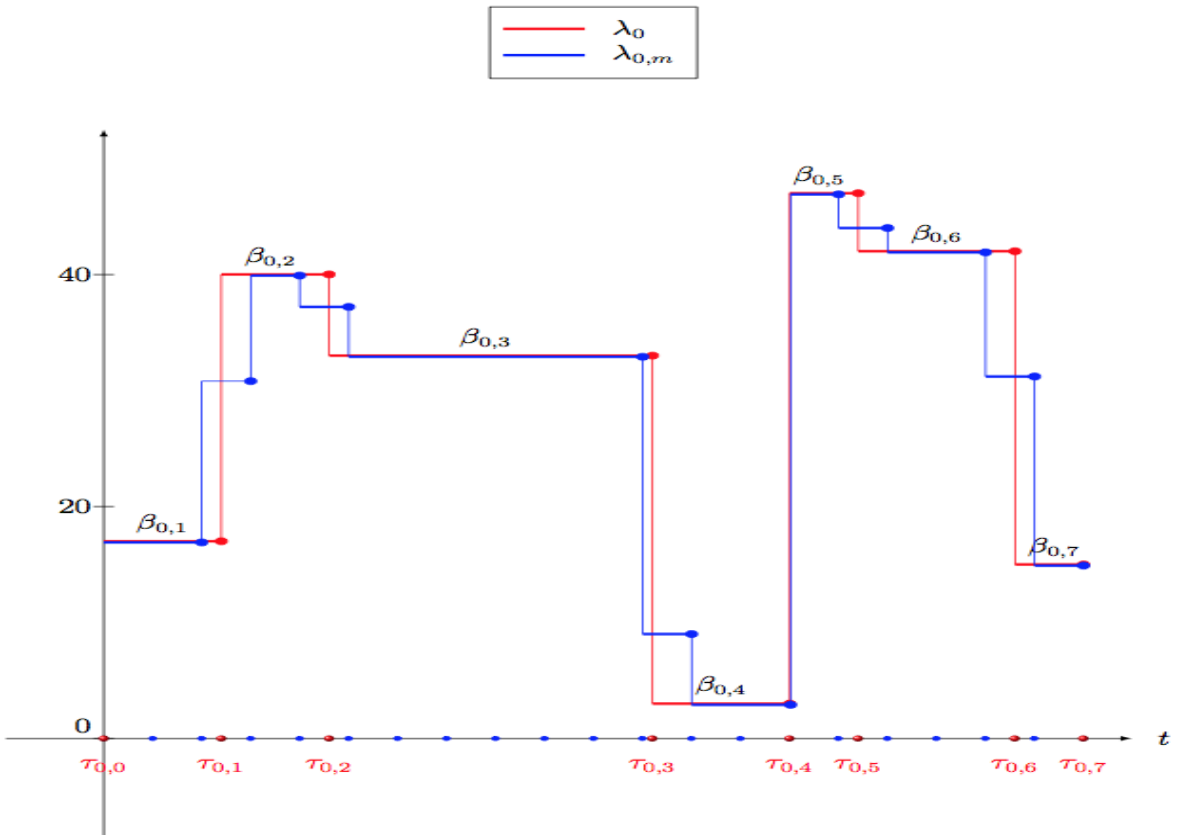
- $\hat{V}_j = \bar{N}_n\left(\left(\frac{j-1}{m},1\right]\right)$ où $\bar{N}_n(I) = \frac{1}{n}\sum_{i=1}^n N_i(I)$, $N_i(I) = \int_I dN_i(t)$.
- $\hat{h}_{n,x,j} = 2\log\log\left(\frac{6en\hat{V}_j+14e(x+\log m)}{28(x+\log m)} \vee e\right)$: terme technique qui vient d'un inégalité de Bernstein pour martingales avec variation optionnelle (Gaïffas, Guilloux (2012)).
- En pratique,

$$\hat{w}_j \approx \sqrt{\frac{m\log m}{n}\hat{V}_j}.$$

LSTA
Laboratoire de Statistique Théorique et Appliquée

Inégalité d'oracle

- Contrôle du terme d'approximation $\inf_{\beta \in \mathbb{R}_+^m} \|\lambda_\beta - \lambda_0\|^2$.
- $\lambda_{0,m}$: la projection orthogonale de λ_0 sur Λ_m.



LSTA
Laboratoire de Statistique Théorique et Appliquée

Inégalité d'oracle

Corollaire

$$\|\hat{\lambda} - \lambda_0\|^2 \leq \frac{2(L_0 - 1)\Delta^2_{\beta,\max}}{m} + 6(L_{\max} + 2(L_0 - 1)) \max_{j=1,\dots,m} \hat{w}_j{}^2 + K_1 \frac{\|\lambda_0\|_\infty \left(x + L_{\max}(1 + \log m)\right)}{n} + K_2 \frac{m\left(x + L_{\max}(1 + \log m)\right)^2}{n^2},$$

avec une probabilité supérieure à $1 - L_{\max} e^{-x}$.

- $\Delta_{\beta,\max} = \max_{1 \leq \ell, \ell' \leq L_0} |\beta_{0,\ell} - \beta_{0,\ell'}|$, le maximun des sauts de λ_0.
- L'équilibre Biais-Variance : $\frac{(L_0 \vee L_{\max}) m \log m}{n}$.
- Si $L_{\max} = O(m) \Rightarrow m \approx n^{1/3}$.
- Si $L_{\max} = O(1) \Rightarrow m \approx n^{1/2}$.

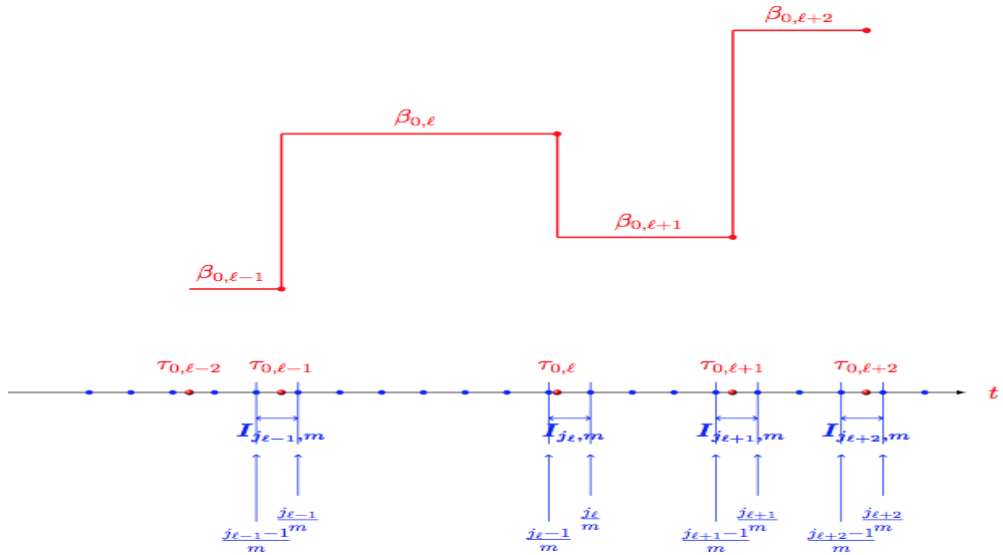
LSTA
Laboratoire de Statistique Théorique et Appliquée

Détection de rupture : consistance

- $(j_\ell)_{\ell=0,\dots,L_0}$ est la suite des *points de rupture approximés* relative à Λ_m définie par la borne droite de l'intervalle $I_{j_\ell,m}$ contenant $\tau_{0,\ell}$, ç.à.d.,

$$\tau_{0,\ell} \in \left(\frac{j_\ell - 1}{m}, \frac{j_\ell}{m}\right],$$

pour $\ell = 1, \dots, L_0 - 1$, tels que $j_0 = 0$ et $j_{L_0} = m$.



Détection de rupture : consistance

- $\hat{S} = \{\hat{j}_1, \ldots, \hat{j}_{\hat{L}}\}$ avec $\hat{j}_1 < \cdots < \hat{j}_{\hat{L}}$. $\hat{j}_0 = 0$ et $\hat{j}_{\hat{L}+1} = m$, on définit

$$\hat{\tau}_\ell = \frac{\hat{j}_\ell}{m}, \text{ pour } \ell = 0, \ldots, \hat{L} + 1.$$

On suppose :

Il existe une constante positive $c \geq 6$ telle que

$$\min_{\ell=1,\ldots,L_0} |\tau_{0,\ell} - \tau_{0,\ell-1}| > \frac{c}{m}.$$

- Les points de rupture dans λ_0 sont suffisamment éloignés. En particulier, chaque intervalle $I_{j,m}$ contient au plus un point de rupture.

LSTA
Laboratoire de Statistique Théorique et Appliquée

Détection de rupture : consistance

- $\Delta_{j,\min} = \min_{1 \le \ell \le L_0 - 1} |j_{\ell+1} - j_\ell|$.
- $\Delta_{\beta,\min} = \min_{1 \le q \le m-1} |\beta_{0,q+1,m} - \beta_{0,q,m}|$.
- $(\varepsilon_n)_{n \ge 1} \downarrow 0$ vérifiant $m\varepsilon_n \ge 6$ pour tout $n \ge 1$.

On suppose

$$\frac{nm\varepsilon_n^2 \Delta_{\beta,\min}^2}{\log m} \to +\infty, \quad \frac{n\Delta_{j,\min}^2 \Delta_{\beta,\min}^2}{m \log m} \to +\infty, \qquad n \to +\infty.$$

- Analyse inspirée de Harchaoui, Lévy-Leduc (2010).

LSTA
Laboratoire de Statistique Théorique et Appliquée

Détection de ruptures : consistance

Théorème - Alaya et al. (2014)

Supposons que $\hat{L} = L_0 - 1$. Alors les points de rupture estimés $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{L}}\}$ satisfont

$$\mathbb{P}\Big[\max_{1 \leq \ell \leq L_0 - 1} |\tau_{0,\ell} - \hat{\tau}_\ell| \leq \varepsilon_n\Big] \to 1, \quad n \to \infty.$$

- Exemples :

	$m = n^{1/3}$	$m = n^{1/2}$
Consistance	$\varepsilon_n = n^{-1/3}$	$\varepsilon_n = n^{-1/2}$
	$\Delta_{\beta,\min} = n^{-1/6}$	$\Delta_{\beta,\min} = n^{-1/6}$

LSTA
Laboratoire de Statistique Théorique et Appliquée

Détection de ruptures : consistance

- Évaluer la distance de Haussdorf $\mathcal{D}(\cdot\|\cdot)$ (non symétrique) entre l'ensemble des points de rupture estimés,

$$\hat{\mathcal{T}} = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{L}}\},$$

et

$$\mathcal{T}_0 = \{\tau_{0,1}, \ldots, \tau_{0,L_0-1}\}.$$

- La distance de Haussdorf entre deux ensembles A et B, est définie par,

$$\mathcal{D}(A\|B) = \sup_{b\in B} \inf_{a\in A} |a-b|.$$

Théorème - Alaya et al. (2014)

Supposons que $\hat{L} \geq L_0 - 1$, on a

$$\mathbb{P}\Big[\mathcal{D}(\hat{\mathcal{T}}\|\mathcal{T}_0) \leq \varepsilon_n\Big] \to 1, \quad n \to +\infty.$$

Aspects numériques

- $R_n(\lambda_\beta) = \sum_{j=1}^m \beta_{j,m}^2 - \frac{2\sqrt{m}}{n} \sum_{j=1}^m \sum_{i=1}^n \beta_{j,m} N_i(I_{j,m})$.
- Réécrivons l'estimateur

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}_+^m} \left\{ \frac{1}{2} \|\mathbf{N} - \beta\|_2^2 + \|\beta\|_{\text{TV},\hat{w}} \right\},$$

avec $\mathbf{N} = [\mathbf{N}_j]_{1 \le j \le m} \in \mathbb{R}_+^m$ est donné par

$$\mathbf{N} = [\sqrt{m}\bar{N}_n(I_{1,m}) \dots \sqrt{m}\bar{N}_n(I_{m,m})].$$

- L'opérateur proximal d'une fonction convexe $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ définit par $\mathbf{prox}_f : \mathbb{R}^p \to \mathbb{R}^p$ tel que

$$\mathbf{prox}_f(y) = \arg\min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - x\|_2^2 + f(x) \right\}.$$

LSTA
Laboratoire de Statistique Théorique et Appliquée

Aspects numériques

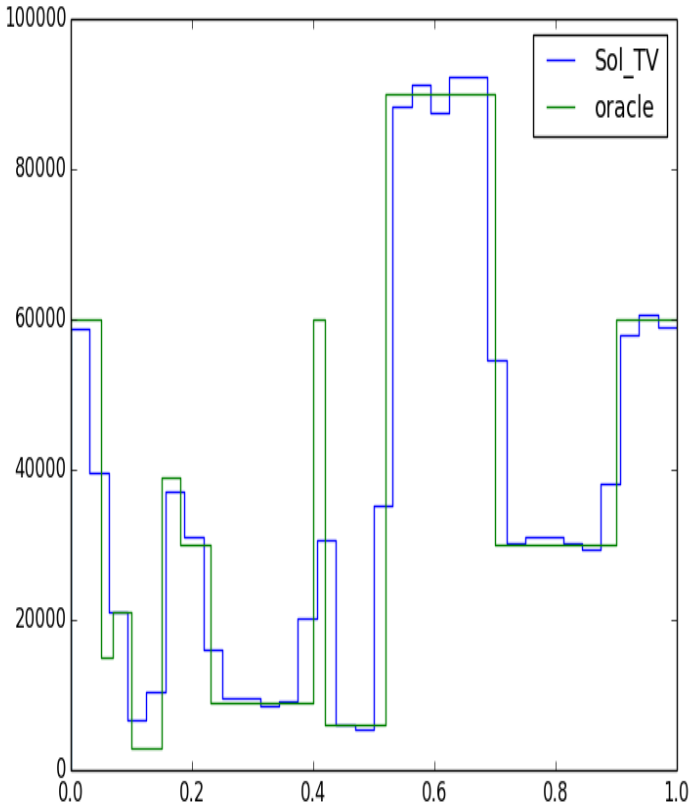
$$\hat{\beta} = \mathbf{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}(\mathbf{N}).$$

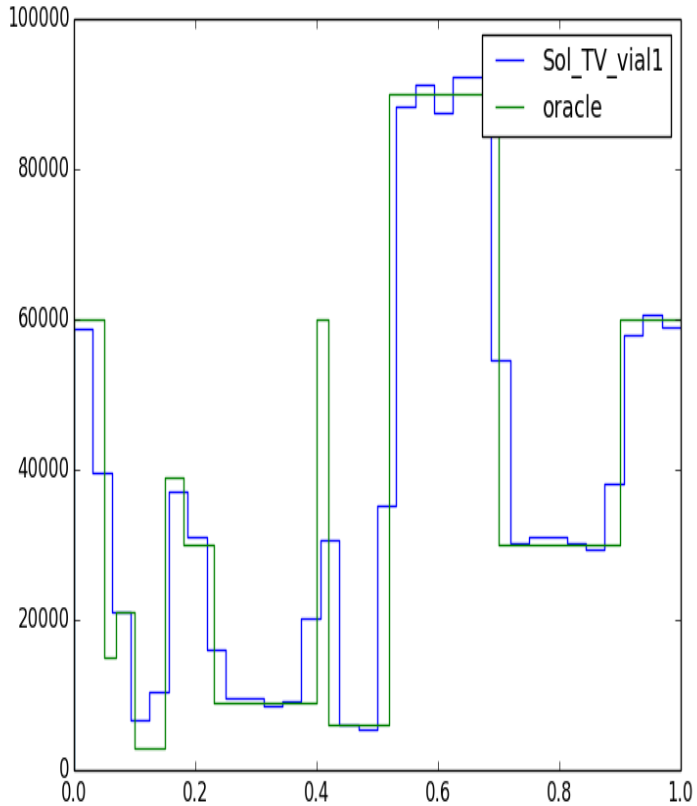
- La pénalisation TV est non séparable $\Rightarrow$ pas de calcul exact de l'opérateur proximal.
- Approche habituelle : changement de variable pour se ramener à une pénalisation ℓ_1.
- Condat (2013) propose un nouvel algorithme pour calculer $\mathbf{prox}_{\|\cdot\|_{\mathrm{TV},1}}$ (TV sans poids), en écrivant le problème dual et en "forçant" les conditions sur le dual
- $\Rightarrow$ Généralisation de l'algorithme de Condat à notre problème pour calculer $\mathbf{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}$ (TV avec poids).

LSTA
Laboratoire de Statistique Théorique et Appliquée

Aspects numériques

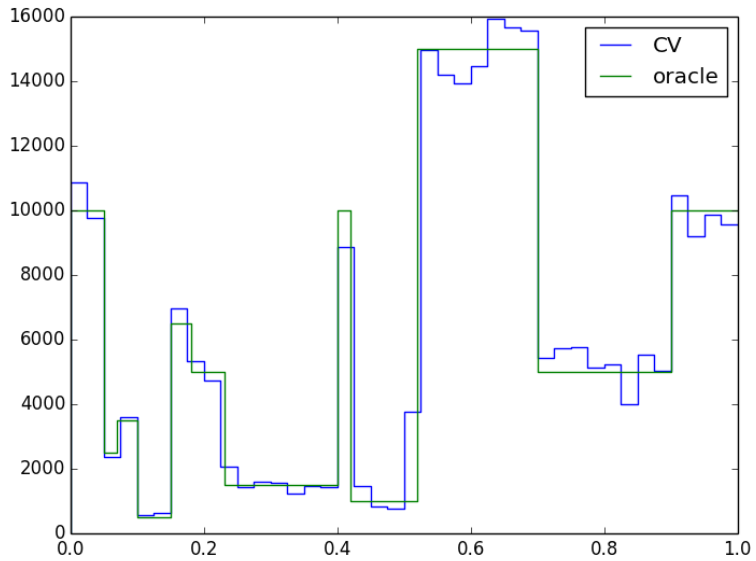
- Temps de calcul : Pour n = 30000
  - FISTA (Beck, Teboulle 2009) : 0.7 secondes (1000 iterations).
  - Proximal TV (Algo. Condat) : 0.009 secondes ($\approx 100\times$ plus rapide).



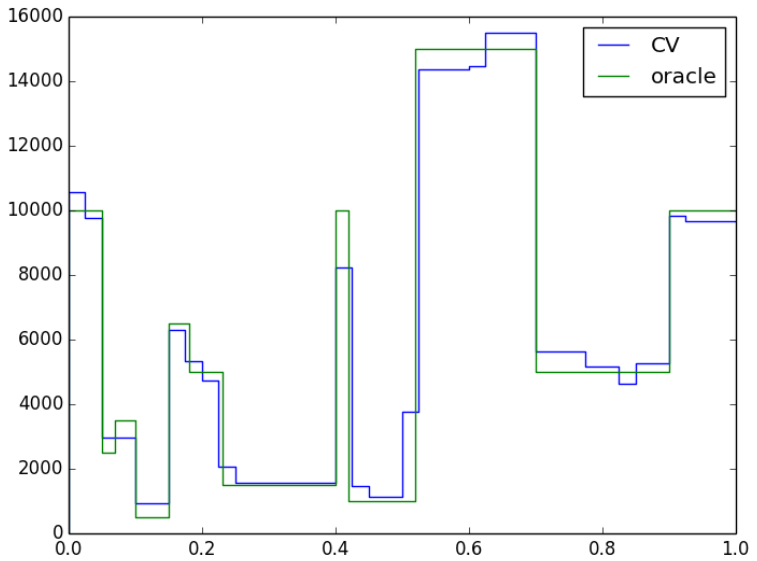


Aspects numériques : sans vs avec pondération data-driven

Comparison penalisation TV avec et sans pondération, pour λ choisi par cross-validation



No weights



With weights

LSTA
Laboratoire de Statistique Théorique et Appliquée

Références

ElMokhtar E. Alaya, S. Gaïffas, et A. Guilloux (2014) :

Learning the intensity of time events with change points,

submitted.

A. Beck, M. Teboulle (2009) :

A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,

SIAM J. IMAGING SCIENCES, 2, 1, 183—202.

L. Condat, (2013) :

A direct algorithm for 1D total variation denoising,

IEEE Signal Proc. Letters, 20, 11, 1054–1057.

S. Gaïffas, A. Guilloux (2012) :

High-dimensional additive hazards models and the Lasso,

Electron. J. Stat., 6, 522–546.

LSTA
Laboratoire de Statistique Théorique et Appliquée

Références

Z. Harchaoui, C. Lévy-Leduc (2010) :

Multiple change-point estimation with a total variation penalty,

J. Amer. Statist. Assoc., 105, 1480–1493.

J. J. Shen and N. R. Zhang, (2012) :

Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing,

Ann. Appl. Stat., 6(2) :476–496.

R. Tibshirani, M. Saunders, S. Rosset, J. Shu, et K. Knight, (2005) :

Sparsity and smoothness via the fused lasso,

J. Amer. Statist. Assoc., 101, 1418–1429.

R. Tibshirani (1996) :

Regression shrinkage and selection via the lasso,

J. Roy. Statist. Soc. Ser. B., 58, 267–288.

LSTA
Laboratoire de Statistique Théorique et Appliquée

Merci !