

Learning the intensity of time events with change-points

joint work with Stéphane Gaïffas² and Agathe Guilloux¹

Mokhtar Zahdi Alaya¹



24 mars 2015

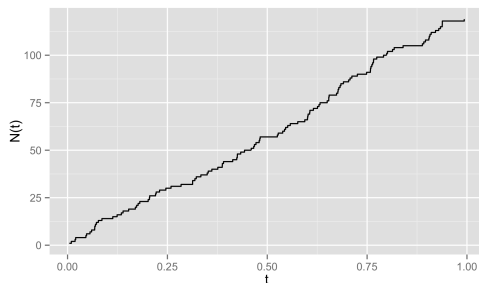
¹LSTA – UPMC

²CMAP – Ecole Polytechnique

Counting process: definitions

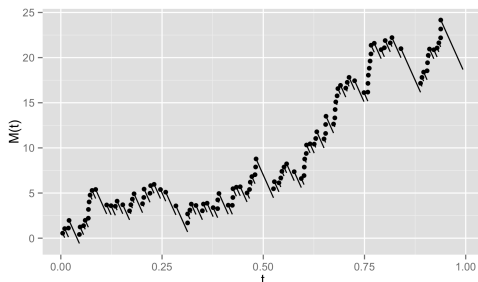
$N = \{N(t)\}_{0 \leq t \leq 1}$ is a counting process if:

- $N(0) = 0$ and $N(t) < \infty$, *a.s.*,
- N is an increasing, right-continuous function *a.s.*,
- $\Delta N(t) = N(t) - N(t^-) \in \{0, 1\}$.



- Doob-Meyer decomposition:

$$N(t) = \underbrace{\Lambda_0(t)}_{\text{compensator}} + \underbrace{M(t)}_{\text{loc. integrable martingale}}, \quad 0 \leq t \leq 1.$$



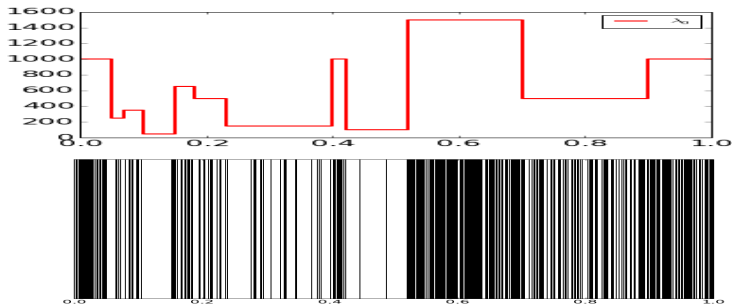
- $\Lambda_0(t) = \mathbb{E}[N(t)] = \int_0^t \lambda_0(s) ds.$
- The intensity of N is defined as follows:

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

Assumption 1

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbf{1}_{(\tau_{0,\ell-1}, \tau_{0,\ell}]}(t), \quad 0 \leq t \leq 1,$$

- Parameters to be estimated:
 - $\{\tau_{0,\ell} : 1 \leq \ell \leq L_0\}$: the set of the true change-points,
 - $\{\beta_{0,\ell} : 1 \leq \ell \leq L_0\}$: the set of the coefficients of the intensity λ_0 ,
 - L_0 : the number of the true the change-points.



Motivations of the sparse segmentation assumption

- Signal processing: Segmentation of the audio signals.
- Time series analysis.
- Study of the genomic profiles: RNA-seq.
- RNA-seq can be modelled mathematically as replications of an inhomogeneous counting process with a piecewise constant intensity (Shen, Zhang (2012)).

- The assumption that the process is in $[0, 1]$ is for the sake of simplicity.

Assumption 2

We observe n i.i.d copies of N on $[0, 1]$, denoted N_1, \dots, N_n .

- We define $\bar{N}_n(I) = \frac{1}{n} \sum_{i=1}^n N_i(I)$, $N_i(I) = \int_I dN_i(t)$, for all subinterval I of $[0, 1]$.
- Assumption 2 is equivalent to observing a single process N with intensity $n\lambda_0$.

A procedure based on total-variation penalization

- We introduce the least-squares functional

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

- Fix $m = m_n \geq 1$, an integer that shall go to infinity as $n \rightarrow \infty$.
- We approximate λ_0 in the set of nonnegative piecewise constant functions on $[0, 1]$ given by

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

where

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

- We consider the estimator

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

- The weighted total-variation penalty is given by:

Data-driven total-variation norm

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

- $[\hat{w}_j]_{1 \leq j \leq m}$, where $\hat{w}_1 = 0$, and $\hat{w}_j \geq 0$, controls the sparsity of the successive difference of the vector β .
- The estimator of λ_0 is defined as follows:

$$\hat{\lambda} = \lambda_{\hat{\beta}} = \sum_{j=1}^m \hat{\beta}_{j,m} \lambda_{j,m}.$$

Fix $x > 0$, and introduce the data-driven weights,

Data-driven weights

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

- $\hat{V}_j = \bar{N}_n\left(\left(\frac{j-1}{m}, 1\right]\right)$.
- $\hat{h}_{n,x,j} = 2 \log \log \left(\frac{6en\hat{V}_j + 14e(x + \log m)}{28(x + \log m)} \vee e \right)$: a technique term given by the Bernstein inequality (Gaïffas, Guilloux (2012)).
- In practical, we consider the dominant term of the data-driven weights

$$\hat{w}_j \approx \sqrt{\frac{m \log m}{n} \bar{N}_n\left(\left(\frac{j-1}{m}, 1\right]\right)}.$$

- The linear space Λ_m is endowed by the norm $\|\lambda\| = (\int_0^1 \lambda^2(t) dt)^{1/2}$.
- We consider the general case of any intensity function of a counting process.
- The estimator $\hat{\lambda}$ satisfies the following:

Theorem 1 [A., Gaïffas, Guillaoux (2014)]

Fix $x > 0$ and let the data-driven weights \hat{w} defined as above. Then, we have

$$\|\hat{\lambda} - \lambda_0\|^2 \leq \inf_{\beta \in \mathbb{R}_+^m} \left(\|\lambda_\beta - \lambda_0\|^2 + 2\|\beta\|_{\text{TV}, \hat{w}} \right)$$

with a probability larger than $1 - 12.85e^{-x}$.

Oracle inequality with fast rate: under Assumption 1

- Let \hat{S} : the support of the discrete gradient of $\hat{\beta}$,

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

- Let \hat{L} : the estimated number of change-points defined by:
 $\hat{L} = |\hat{S}|.$

Theorem 2 [A., Gaïffas, Guilloux (2014)]

Fix $x > 0$, let $\hat{\lambda}$ be the same as in Theorem 1. Assume that \hat{L} satisfies $\hat{L} \leq L_{\max}$. Then, we have

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

with a probability larger than $1 - L_{\max}e^{-x}$, with

$\|\lambda_0\|_\infty = \sup_{t \in [0,1]} \lambda_0(t)$, $K_1 = 1670.89$, and $K_2 = 6683.53$.

Oracle inequality with fast rate: under Assumption 1

- let $\beta_{0,m} = [\beta_{0,j,m}]_{1 \leq j \leq m}$ the coefficients vector of the projection of λ_0 on Λ_m , and $\Delta_{\beta,\max} = \max_{1 \leq \ell, \ell' \leq L_0} |\beta_{0,\ell} - \beta_{0,\ell'}|$.

Lemma: Control of the bias

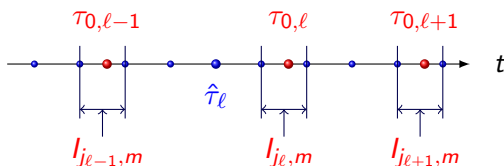
Given Assumption 1, we have

$$\|\lambda_\beta - \lambda_0\|^2 \leq \frac{2(L_0 - 1)\Delta_{\beta,\max}^2}{m}.$$

- Theorem 2 proves that our procedure has a fast rate of convergence of order $\frac{(L_{\max} \vee L_0)m \log m}{n}$.
- A consequence is that an optimal tradeoff between approximation and complexity is given by the choice $m \approx n^{1/2}$.
- If $L_{\max} = O(m) \Rightarrow m \approx n^{1/3}$.
- If $L_{\max} = O(1) \Rightarrow m \approx n^{1/2}$.
- We are able to use the same procedure in Theorems 1 and 2, while it is not the case in the signal + white noise considered (Harchaoui and Levy Leduc (2010)).

Change-point detection: consistency

- The *approximate change-points sequence* $[j_\ell]_{0 \leq \ell \leq L_0}$ is defined as the *right-hand side boundary* of the unique interval $I_{j_\ell, m}$ that contains the change-point $\tau_{0, \ell}$.
- $\tau_{0, \ell} \in \left(\frac{j_{\ell-1}}{m}, \frac{j_\ell}{m}\right]$, for $\ell = 1, \dots, L_0 - 1$, where $j_0 = 0$ and $j_{L_0} = m$ by convention.



- Let $\hat{S} = \{\hat{j}_1, \dots, \hat{j}_{\hat{L}}\}$ with $\hat{j}_1 < \dots < \hat{j}_{\hat{L}}$ of the support of the discrete gradient of $\hat{\beta}$.
- We introduce $\hat{j}_0 = 0$ and $\hat{j}_{\hat{L}+1} = m$, we define simply $\hat{\tau}_\ell = \frac{\hat{j}_\ell}{m}$ for $\ell = 0, \dots, \hat{L} + 1$.

- We will not be able to recover the exact position of two change-points if they lie on the same interval $I_{j,m}$.

Assumption 3

Grant Assumption 1 and assume that there is a positive constant $c \geq 8$ such that

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

- The change-points of λ_0 are sufficiently far apart.
- There cannot be more than one change-point in the “high-resolution” intervals $I_{j,m}$.
- The procedure will be able to recover the (unique) intervals $I_{j_\ell, m}$, for $\ell = 0, \dots, L_0$, where the change-point belongs.

- $\Delta_{j,\min} = \min_{1 \leq \ell \leq L_0 - 1} |j_{\ell+1} - j_\ell|$, the minimum distance between two consecutive terms in the change-points of λ_0 .
- $\Delta_{\beta,\min} = \min_{1 \leq q \leq m-1} |\beta_{0,q+1,m} - \beta_{0,q,m}|$, the smallest jump size of the projection $\lambda_{0,m}$ of λ_0 onto Λ_m .
- $(\varepsilon_n)_{n \geq 1}$, a non-increasing and positive sequence that goes to zero as $n \rightarrow \infty$, and such that $m\varepsilon_n \geq 6$ for any $n \geq 1$.

Assumption 4

We assume that $\Delta_{j,\min}$, $\Delta_{\beta,\min}$ and $(\varepsilon_n)_{n \geq 1}$ satisfy

$$\frac{\sqrt{nm}\varepsilon_n\Delta_{\beta,\min}}{\sqrt{\log m}} \rightarrow \infty \quad \text{and} \quad \frac{\sqrt{n}\Delta_{j,\min}\Delta_{\beta,\min}}{\sqrt{m \log m}} \rightarrow \infty, \quad \text{as } n \rightarrow \infty.$$

- Assumption 4 controls the rate (ε_n) of convergence of $\hat{\tau}_\ell$ towards $\tau_{0,\ell}$.

Theorem 3 [A., Gaïffas, Guillaoux (2014)]

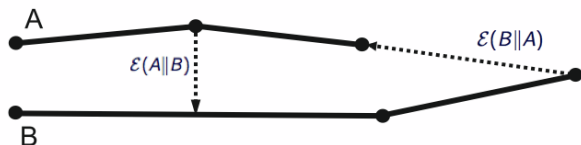
Under Assumptions 3 and 4, and if $\hat{L} = L_0 - 1$, then the change-points estimators $\{\hat{\tau}_1, \dots, \hat{\tau}_{\hat{L}}\}$ satisfy

$$\mathbb{P}\left[\max_{1 \leq \ell \leq L_0 - 1} |\hat{\tau}_\ell - \tau_{0,\ell}| \leq \varepsilon_n\right] \rightarrow 1, \text{ as } n \rightarrow \infty.$$

- If $m = n^{1/3}$, Theorem 3 holds with $\varepsilon_n = n^{-1/3}$, $\Delta_{\beta,\min} = n^{-1/6}$ et $\Delta_{j,\min} \geq 6$.
- $m = n^{1/2}$, Theorem 3 holds with $\varepsilon_n = n^{-1/2}$, $\Delta_{\beta,\min} = n^{-1/6}$ et $\Delta_{j,\min} \geq 6$.

Change-point detection: consistency

- We evaluate a non-symmetrized Hausdorff distance $\mathcal{E}(\hat{\mathcal{T}}\|\mathcal{T}_0)$ between:
 - The set of estimated change-points $\hat{\mathcal{T}} = \{\hat{\tau}_1, \dots, \hat{\tau}_L\}$
 - The set of true change-points $\mathcal{T}_0 = \{\tau_{0,1}, \dots, \tau_{0,L_0-1}\}$,
 - $\mathcal{E}(A\|B) = \sup_{b \in B} \inf_{a \in A} |a - b|$, for two sets A and B .



Theorem 4 [A., Gaïffas, Guilloux (2014)]

Under Assumptions 3 and 4, and if $\hat{L} \geq L_0 - 1$, we have

$$\mathbb{P}\left[\mathcal{E}(\hat{\mathcal{T}}|\mathcal{T}_0) \leq \varepsilon_n\right] \rightarrow 1, \text{ as } n \rightarrow \infty.$$

- Theorem 4 ensures that even when the number of change-points is over-estimated, each true change-point is close to the estimated one.
- We are able to use the same regularization parameters \hat{w} .

Algorithm: Proximal operator of the weighted TV

- The proximal operator prox_f of a proper, lower semi-continuous, convex function $f : \mathbb{R}^m \rightarrow (-\infty, \infty]$, is defined as

$$\text{prox}_f(v) = \underset{x \in \mathbb{R}^m}{\text{argmin}} \left\{ \frac{1}{2} \|v - x\|_2^2 + f(x) \right\}, \text{ for all } v \in \mathbb{R}^m.$$

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

where $\mathbf{N} = [\mathbf{N}_j]_{1 \leq j \leq m} \in \mathbb{R}_+^m$ is given by

$$\mathbf{N} = \begin{bmatrix} \sqrt{m} \bar{N}_n(l_{1,m}) \\ \vdots \\ \sqrt{m} \bar{N}_n(l_{m,m}) \end{bmatrix}$$

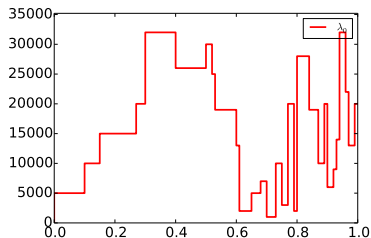
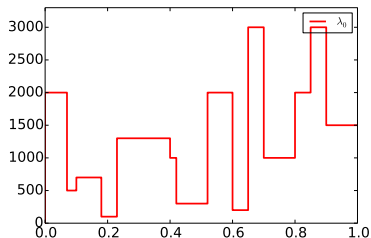
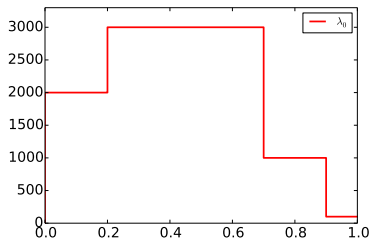
$$\hat{\beta} = \text{prox}_{\|\cdot\|_{\text{TV}, \hat{w}}}(\mathbf{N}).$$

Algorithm: Proximal operator of the weighted TV

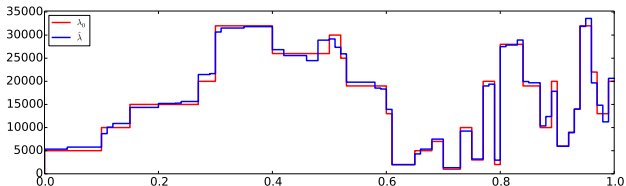
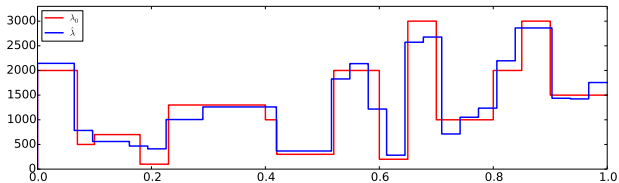
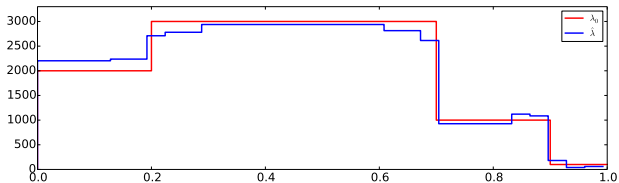
- If we have a feasible dual variable \hat{u} , we can compute the primal solution $\hat{\beta}$, by Fenchel duality.
- The KKT optimality conditions characterize the unique solutions $\hat{\beta}$ and $\hat{\theta}_k := \hat{w}_{k+1} \hat{u}_k$.
- The algorithm consists in running forwardly through the samples $[\mathbf{N}_k]_{1 \leq k \leq m}$.
- Using the KKT, at location k , $\hat{\beta}_k$ stays constant where $|\hat{\theta}_k| < \hat{w}_{k+1}$.
- If this is not possible, it goes back to the last location where a jump can be introduced in $\hat{\beta}$, validates the current segment until this location, starts a new segment, and continues.

Simulated data

- We simulate counting processes with inhomogeneous piecewise intensities λ_0 , with 5, 15 and 30 change points.

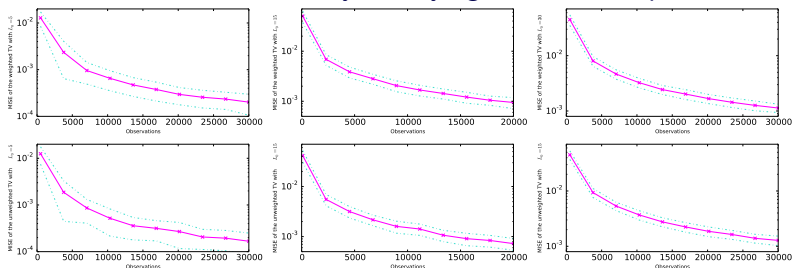


We plot the estimator for the three models

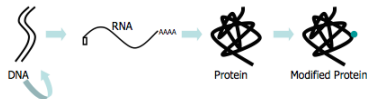


- To evaluate the performance of the total-variation procedure $\hat{\lambda}$, we use a Monte-Carlo averaged mean integrated squared error MISE.
- $\text{MISE}(\hat{\lambda}, \lambda_0) = \mathbb{E} \int_0^1 (\hat{\lambda}(t) - \lambda_0(t))^2 dt.$
- We run 100 Monte-Carlo experiments, for an increasing sample size between $n = 500$ and $n = 30000$, for each 3 examples.

- We plot the MISEs of the weighted and the unweighted total variation, $\hat{w} \equiv 1$, for the three models, as a function of the sample size.
- The estimation error is always decaying with the sample size.

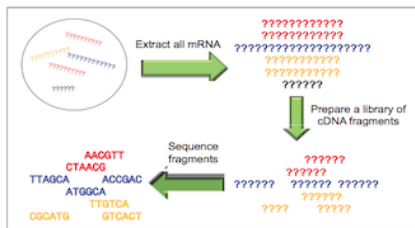


Next generations sequencing (NGS)

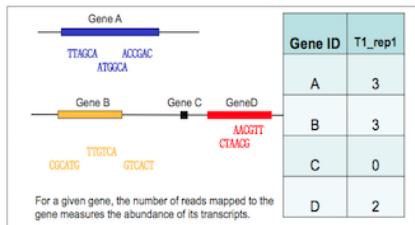


- Complementary base pairing, $A < - > T$ and $C < - > G$
- Genome is a complete set of DNA in an organism.
- Gene is a DNA sequence that encodes a protein or an RNA molecule.
- DNA is transcribed to mRNA, which is translated into protein (central dogma).

Next generations sequencing, RNA-seq

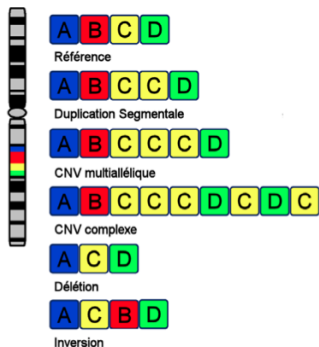


(a) The sequencing step



(b) The mapping step

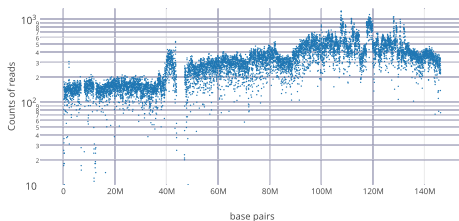
Detection of Copy number variation (CNV)



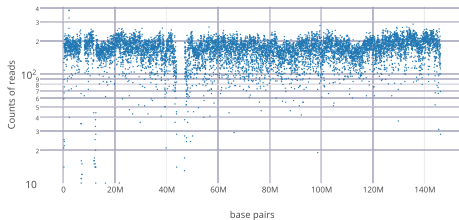
- Copy number variations (CNVs), which are gains or deletions of genomic segments, account for a substantial proportion of human genetic variations.
- CNVs play an important role in the pathogenesis and progression of cancer and confer susceptibility to a variety of human disorders.

- We applied our method to the sequencing data of the breast tumor cell line HCC1954 and its reference cell line BL1954 (Chiang et al. 2009).
- The dataset was produced using the Illumina platform, where the reads are 36bp long.
- There are 7.72 million reads for the tumor (HCC1954) samples.
- There are 6.65 million reads for the normal (BL1954) samples.

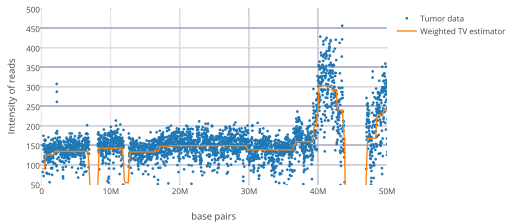
Binned counts of reads on the tumor data



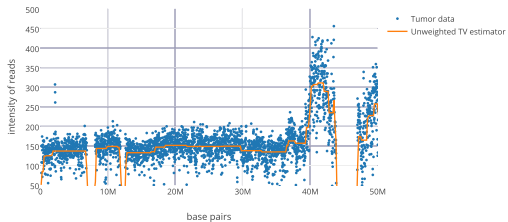
Binned counts of reads on the normal data



Weighted total-variation estimator on the tumor data

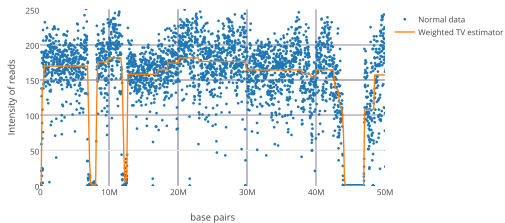


Unweighted total-variation estimator for the tumor data

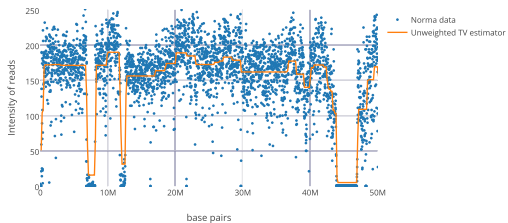


Real data





Weighted total-variation estimator for the normal reads



Unweighted total-variation estimator for the normal data



- We introduce a data-driven weighted total-variation penalization for this problem.
- We prove that convex optimization for the detection of change-points in the intensity of a counting process is a powerful tool.
- We prove two families of theoretical results: oracles inequalities for the prediction error, and consistency in the estimation of change-points.
- The study of maximum likelihood estimation instead of least-squares.
- Multivariate extension of the proposed algorithm.

-  ElMokhtar E. Alaya, S. Gaïffas, et A. Guillaoux (2014) :
Learning the intensity of time events with change-points,
in revision.
-  D.Y Chiang, G. Getz, D. B Jaffe, M.JT O'Kelly, X.vZhao, S. L
Carter, Carsten Russ, C. Nusbaum, M. Meyerson, and E.S Lander.
High-resolution mapping of copy-number alterations with massively
parallel sequencing.
Nature methods, 6(1):99–103, 2009.
-  L. Condat, (2013) :
A direct algorithm for 1D total variation denoising,
IEEE Signal Proc. Letters, 20, 11, 1054–1057.
-  S. Gaïffas, A. Guillaoux (2012) :
High-dimensional additive hazards models and the Lasso,
Electron. J. Stat., 6, 522–546.



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.

Thank you!