Around Supervised Learning with Weighted Total-Variation Penalization

Mokhtar Z. Alaya

Université Paris Nanterre

Seminar of Modal'X, 11th May 2017

Part 0

Supervised Learning in High-Dimensions

→

Setting

- Data $x_i \in \mathcal{X} = \mathbb{R}^p, y_i \in \mathcal{Y}$ for $i = 1, \ldots, n$. The x_i are called **features** and the y_i are called **labels**.
- The labels are scalar numbers. We assume that $\mathcal{Y} \subset \mathbb{R}$.
  $\mathcal{Y} = \{-1, +1\}, \mathcal{Y} = \{0, 1\}$ for binary classification.
  $\mathcal{Y} = \mathbb{R}$ for regression.
- Usually the data $D_n = \{(x_i, y_i) : i = 1, \ldots, n\}$ is supposed to be i.i.d.

Goal

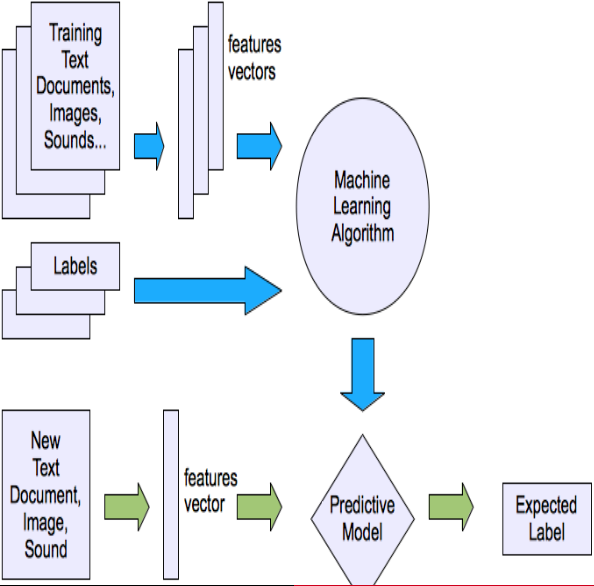
- Based on (x_i, y_i), learn a function that predicts y based on a new x (generalization property).

High-dimension

- p is larger than n.

Work-flow of supervised learning

←



Supervised learning: empirical risk + penalization

Minimize with respect to $f : \mathbb{R}^p \to \mathbb{R}$

$$R_n(f) + \gamma \text{pen}(f)$$

where

- $$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

  is a **goodness-of-fit**, or **empirical risk**, where ℓ is a **loss** function.
- pen is a penalization function, that encodes a prior assumption on f.
- $\gamma > 0$ is a **tuning parameter**, that balances good-of-fitness and penalization.
- **Simplification**: choose a linear function f:

$$f(x) = x^\top \beta = \sum_{j=1}^{p} x_j \beta_j,$$

Supervised learning: empirical risk + penalization

- We end up with:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{R_n(\beta) + \lambda \mathrm{pen}(\beta)\},$$

  where

$$R_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta)$$

  and $\mathrm{pen}(\beta)$ is a penalization on β.

- Choice of penalization !

- L_0-quasi-norm: $\text{pen}(\beta) = \|\beta\|_0 = \#\{j : \beta_j \neq 0\}$.
- Lasso (L_1-norm): $\text{pen}(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ [Tibshirani (1996)].
- Elastic-Net (($L_1 + L_2^2$)-norm): $\text{pen}(\beta) = \|\beta\|_1 + \|\beta\|_2^2$ [Zou and Hastie (2005)].
- Fused Lasso (L_1 + TV): $\text{pen}(\beta) = \|\beta\|_1 + \|\beta\|_{\text{TV}}$ [Tibshirani et al. (2005)] where $\|\cdot\|_{\text{TV}}$ is the total-variation penalization defined as

$$\|\beta\|_{\text{TV}} = \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

Weighted TV

- For a chosen positive vector of weights $\hat{\omega}$, we define the (discrete) weighted total-variation (TV) by

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

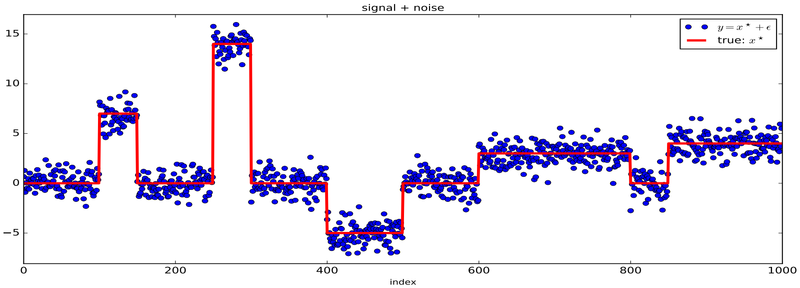
- If $\hat{\omega} \equiv 1$, then we define the unweighted (simple) TV by

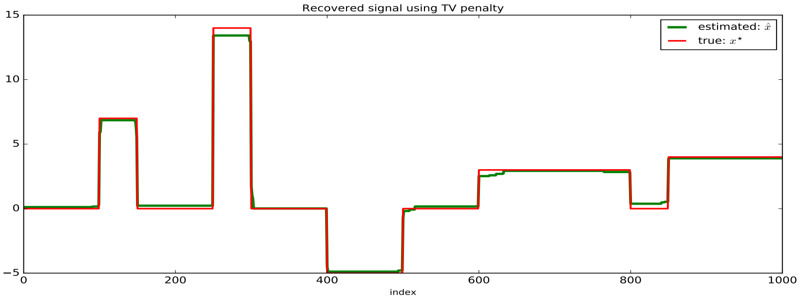
$$\|\beta\|_{\mathrm{TV},1} = \|\beta\|_{\mathrm{TV}} = \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

Motivations for using TV

- Appropriate for multiple change-points estimation.
  $\longrightarrow$ Partitioning a nonstationary signal into several contiguous stationary segments of variable duration [Harchaoui and Lévy-Leduc (2010)].
- Widely used in sparse signal processing and imaging (2D) [Chambolle et al. (2010)].
- Enforces sparsity in the discrete gradient, which is desirable for applications with features ordered in some meaningful way [Tibshirani et al. (2005)].

Toy example: recovery of piecewise constant signal





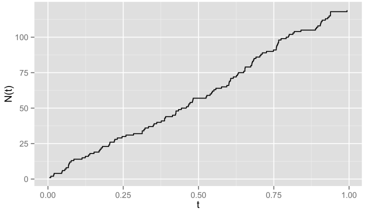
Part I

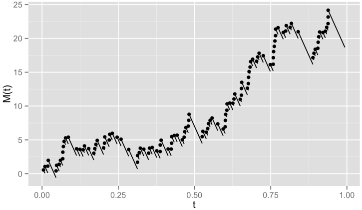
Learning the Intensity of Time Events with Change-Points

[A., Gaïffas, Guilloux (2015), published in IEEE TIT]

Counting process: stochastic setup

- $N = \{N(t)\}_{0 \le t \le 1}$ is a counting process.





- Doob-Meyer decomposition:

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

- The intensity of N is defined by

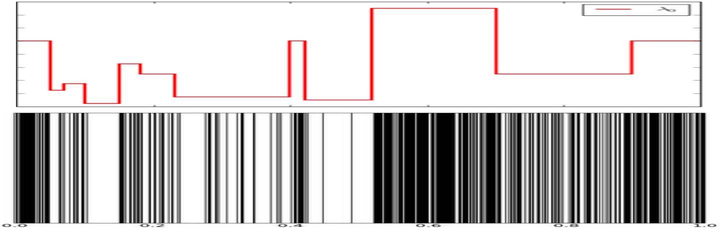
$$\lambda_0(t)dt = d\Lambda_0(t) = \mathbb{P}[N \text{ has a jump in } [t, t + dt) | \mathcal{F}(t)],$$

where $\mathcal{F}(t) = \sigma(N(s), s \le t)$.

- Assume that

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

- $\{\tau_{0,0} = 0 < \tau_{0,1} < \cdots < \tau_{0,L_0-1} < \tau_{0,L_0} = 1\}$: set of true change-points.
- $\{\beta_{0,\ell} : 1 \leq \ell \leq L_0\}$: set of jump sizes of λ_0.
- L_0 : number of true change-points.



Assumption on observations

Data

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 any interval $I \subset [0,1]$.
- This assumption is equivalent to observing a single process N with intensity $n\lambda_0$ (only used to have a notion of growing observations with an increasing n).

A procedure based on weighted TV penalization

- We introduce the least-squares functional

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

[Reynaud-Bouret (2003, 2006), Gaïffas and Guilloux (2012)].

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

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

where

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

A procedure based on weighted TV penalization

- The estimator of λ_0 is defined by

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

where $\hat{\beta}$ is giving by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}_+^m}{\operatorname{argmin}} \Big\{ R_n(\lambda_\beta) + \|\beta\|_{\mathrm{TV},\hat{\omega}} \Big\}.$$

- We consider the dominant term

$$\hat{\omega}_j = \mathcal{O}\left(\sqrt{\frac{m \log m}{n} \bar{N}_n\Big((\frac{j-1}{m}, 1] \Big)} \right).$$

Estimator of change-points number

- The linear space Λ_m is endowed by the norm $\|\lambda\| = \sqrt{\int_0^1 \lambda^2(t)dt}$.
- Let $\hat{S}$ to be the support of the discrete gradient of $\hat{\beta}$,

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

- Let $\hat{L}$ to be the estimated number of change-points defined by:

$$\hat{L} = |\hat{S}|.$$

Oracle inequality with fast rate

The estimator $\hat{\lambda}$ satisfies the following:

Theorem 1

Fix $x > 0$ and let the data-driven weights $\hat{\omega}$ defined as above. Assume that $\hat{L}$ satisfies $\hat{L} \leq L_{\max}$. Then, we have

$$\|\hat{\lambda} - \lambda_0\|^2 \leq \inf_{\beta \in \mathbb{R}_+^m} \big\|\lambda_\beta - \lambda_0\big\|^2 + 6(L_{\max} + 2(L_0 - 1)) \max_{1 \leq j \leq m} \hat{\omega}_j^2$$
$$+ C_1 \frac{\|\lambda_0\|_\infty \big(x + L_{\max}(1 + \log m)\big)}{n}$$
$$+ C_2 \frac{m\big(x + L_{\max}(1 + \log m)\big)^2}{n^2},$$

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

Trade-off bias and variance

- Let $\Delta_{\beta,\max} = \max_{1 \leq \ell, \ell' \leq L_0} |\beta_{0,\ell} - \beta_{0,\ell'}|$, be the maximum of jump size of λ_0.

Corollary

We have

$$\|\lambda_\beta - \lambda_0\|^2 \leq \frac{2 L_0 \Delta^2_{\beta,\max}}{m}.$$

- Our procedure has a fast rate of convergence of order

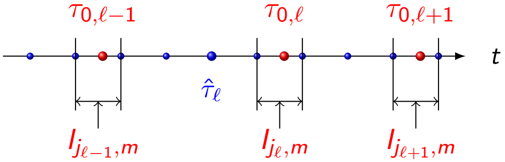
$$\frac{(L_{\max} \vee L_0) m \log m}{n}.$$

- An optimal tradeoff between approximation and complexity is given by the choice:

$$\text{If } L_{\max} = \mathcal{O}(m) \Rightarrow m = \mathcal{O}(n^{1/3}).$$
$$\text{If } L_{\max} = \mathcal{O}(1) \Rightarrow m = \mathcal{O}(n^{1/2}).$$

Consistency of change-points detection

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



- Let $\hat{S} = \{\hat{j}_1, \ldots, \hat{j}_{\hat{L}}\}$ with $\hat{j}_1 < \cdots < \hat{j}_{\hat{L}}$, and $\hat{j}_0 = 0$ and $\hat{j}_{\hat{L}+1} = m$.
- We define simply

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

Consistency of change-points detection

- We can't recover the exact position of two change-points if they lie on the same interval $I_{j,m}$.

Minimal distance between true change-points

Assume that there is a positive constant $c \geq 8$ such that

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

$\longrightarrow$ The change-points of λ_0 are sufficiently far apart.
$\longrightarrow$ 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 $\ell = 0, \ldots, L_0$, where the change-point belongs.

Consistency of change-points detection

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

Technical Assumptions

We assume that $\Delta_{j,\min}$, $\Delta_{\beta,\min}$ and $(\varepsilon_n)_{n\ge1}$ satisfy

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

Consistency of change-points detection

Theorem 2

Under the given Assumptions, and if $\hat{L} = L_0$, then the change-points estimators $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{L}}\}$ satisfy

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

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

Implementation: proximal operator

- We are interested in computing a solution

$$x^\star = \operatorname*{argmin}_{x \in \mathbb{R}^p} \{g(x) + h(x)\},$$

  where g is smooth and h is simple (prox-calculable).

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

$$\text{prox}_h(v) = \operatorname*{argmin}_{x \in \mathbb{R}^m} \Big\{ \frac{1}{2} \|v - x\|_2^2 + h(x) \Big\}, \text{ for all } v \in \mathbb{R}^m.$$

- Proximal gradient descent (PGD) algorithm is based on

$$x^{(k+1)} = \text{prox}_{\varepsilon_k h} \big(x^{(k)} - \varepsilon_k \nabla g(x^{(k)}) \big).$$

  [Daubechies et al. (2004) (ISTA) , Beck and Teboulle (2009) (FISTA)]

Proximal operator of the weighted TV penalization

- We have

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}_+^m} \Big\{ \frac{1}{2} \|\mathbf{N} - \beta\|_2^2 + \|\beta\|_{\mathrm{TV},\hat{\omega}} \Big\},$$

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

$$\mathbf{N} = \Big(\sqrt{m} \bar{N}_n(I_{1,m}), \ldots, \sqrt{m} \bar{N}_n(I_{m,m}) \Big).$$

- Then

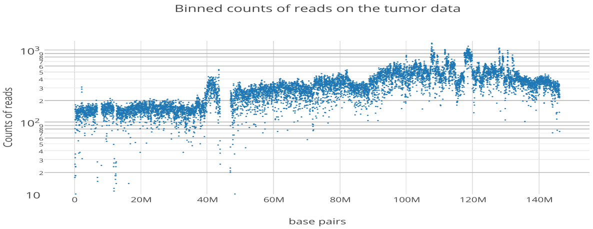
$$\hat{\beta} = \operatorname{prox}_{\|\cdot\|_{\mathrm{TV},\hat{\omega}}}(\mathbf{N}).$$

- Modification of Condat's algorithm [Condat (2013)].
- If we have a feasible dual variable $\hat{u}$, we can compute the primal solution $\hat{\beta}$, by Fenchel duality.
- The Karush-Kuhn-Tucker (KKT) optimality conditions characterize the unique solutions $\hat{\beta}$ and $\hat{u}$.

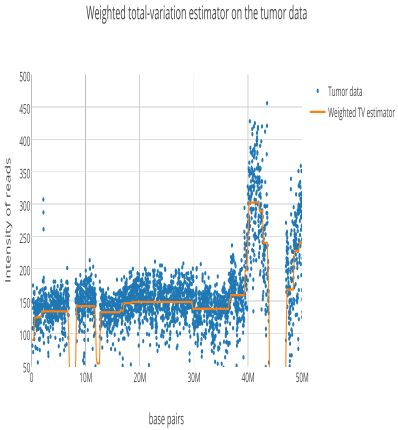
Algorithm 1: $\hat{\beta} = \text{prox}_{\|\cdot\|_{\text{TV},\hat{\omega}}}(\mathbf{N})$

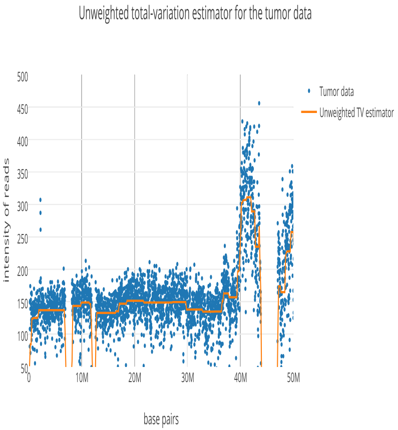
1. **set** $k = k_0 = k_- = k_+ \leftarrow 1$; $\beta_{\min} \leftarrow \mathbf{N}_1 - \hat{\omega}_2$; $\beta_{\max} \leftarrow \mathbf{N}_1 + \hat{\omega}_2$; $\theta_{\min} \leftarrow \hat{\omega}_2$; $\theta_{\max} \leftarrow -\hat{\omega}_2$;
2. **if** $k = m$ **then**
   - $\hat{\beta}_m \leftarrow \beta_{\min} + \theta_{\min}$;
3. **if** $\mathbf{N}_{k+1} + \theta_{\min} < \beta_{\min} - \hat{\omega}_{k+2}$ **then** /* negative jump */
   - $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_-} \leftarrow \beta_{\min}$; $k = k_0 = k_- = k_+ \leftarrow k_- + 1$;
   - $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{\omega}_{k+1} + \hat{\omega}_k$; $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{\omega}_{k+1} + \hat{\omega}_k$; $\theta_{\min} \leftarrow \hat{\omega}_{k+1}$; $\theta_{\max} \leftarrow -\hat{\omega}_{k+1}$;
4. **else if** $\mathbf{N}_{k+1} + \theta_{\max} > \beta_{\max} + \hat{\omega}_{k+2}$ **then** /* positive jump */
   - $\hat{\beta}_{k_0} = \ldots = \hat{\beta}_{k_+} \leftarrow \beta_{\max}$; $k = k_0 = k_- = k_+ \leftarrow k_+ + 1$;
   - $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{\omega}_{k+1} - \hat{\omega}_k$; $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{\omega}_{k+1} - \hat{\omega}_k$; $\theta_{\min} \leftarrow \hat{\omega}_{k+1}$; $\theta_{\max} \leftarrow -\hat{\omega}_{k+1}$;
5. **else** /* no jump */
   - **set** $k \leftarrow k + 1$; $\theta_{\min} \leftarrow \mathbf{N}_k + \hat{\omega}_{k+1} - \beta_{\min}$; $\theta_{\max} \leftarrow \mathbf{N}_k - \hat{\omega}_{k+1} - \beta_{\max}$;
   - **if** $\theta_{\min} \geq \hat{\omega}_{k+1}$ **then**
     - $\beta_{\min} \leftarrow \beta_{\min} + \frac{\theta_{\min} - \hat{\omega}_{k+1}}{k - k_0 + 1}$; $\theta_{\min} \leftarrow \hat{\omega}_{k+1}$; $k_- \leftarrow k$;
   - **if** $\theta_{\max} \leq -\hat{\omega}_{k+1}$ **then**
     - $\beta_{\max} \leftarrow \beta_{\max} + \frac{\theta_{\max} + \hat{\omega}_{k+1}}{k - k_0 + 1}$; $\theta_{\max} \leftarrow -\hat{\omega}_{k+1}$; $k_+ \leftarrow k$;
6. **if** $k < m$ **then**
   - **go to 3.**;
7. **if** $\theta_{\min} < 0$ **then**
   - $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_-} \leftarrow \beta_{\min}$; $k = k_0 = k_- \leftarrow k_- + 1$; $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{\omega}_{k+1} + \hat{\omega}_k$;
   - $\theta_{\min} \leftarrow \hat{\omega}_{k+1}$; $\theta_{\max} \leftarrow \mathbf{N}_k + \hat{\omega}_k - v_{\max}$; **go to 2.**;
8. **else if** $\theta_{\max} > 0$ **then**
   - $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_+} \leftarrow \beta_{\max}$; $k = k_0 = k_+ \leftarrow k_+ + 1$; $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{\omega}_{k+1} - \hat{\omega}_k$;
   - $\theta_{\max} \leftarrow -\hat{\omega}_{k+1}$; $\theta_{\min} \leftarrow \mathbf{N}_k - \hat{\omega}_k - \theta_{\min}$; **go to 2.**;
9. **else**
   - $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_m \leftarrow \beta_{\min} + \frac{\theta_{\min}}{k - k_0 + 1}$;

- RNA-seq can be modelled mathematically as replications of an inhomogeneous counting process with a piecewise constant intensity [Shen and Zhang (2012)].
- We applied our method to the sequencing data of the breast tumor cell line HCC1954 7.72 million reads) and its reference cell line BL1954 (6.65 million reads) [Chiang et al. (2009)].



A zoom into the sequence of reads for tumor data.





Zoom into the weighted (left) and unweighted (right) TV estimators applied to the tumor data.

Part II

Binarsity: a penalization for one-hot encoded features

Features binarization

- Supervised training dataset $(x_i, y_i)_{i=1,\ldots,n}$ containing features $x_i = (x_{i,1}, \ldots, x_{i,p})^\top \in \mathbb{R}^p$ and labels $y_i \in \mathcal{Y} \subset \mathbb{R}$, that are i.i.d.
- We denote $\boldsymbol{X} = [x_{i,j}]_{1 \leq i \leq n; 1 \leq j \leq p}$ the $n \times p$ features matrix.
- Let $\boldsymbol{X}_{\bullet,j}$ be the j-th feature column of $\boldsymbol{X}$.
- The binarized matrix $\boldsymbol{X}^B$ is a matrix with an extended number $d > p$ of columns (only binary).
- The j-th column $\boldsymbol{X}_{\bullet,j}$ is replaced by a number $d_j \geq 2$ of columns $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j}$ containing only zeros and ones.
- The i-th row of $\boldsymbol{X}^B$ is written

$$x_i^B = (x^B_{i,1,1}, \ldots, x^B_{i,1,d_1}, \ldots, x^B_{i,p,1}, \ldots, x^B_{i,p,d_p})^\top \in \mathbb{R}^d.$$

Features binarization

- If $X_{\bullet,j}$ takes values (modalities) in the set $\{1, \ldots, M_j\}$ with cardinality M_j, we take $d_j = M_j$, and use a binary coding of each modality by defining

$$x^B_{i,j,k} = \begin{cases} 1, & \text{if } x_{i,j} = k, \\ 0, & \text{otherwise,} \end{cases}$$

- If $X_{\bullet,j}$ is quantitative, then d_j we consider a partition of intervals $I_{j,1}, \ldots, I_{j,d_j}$ for the range of values of $\boldsymbol{X}_{\bullet,j}$ and define

$$x^B_{i,j,k} = \begin{cases} 1, & \text{if } x_{i,j} \in I_{j,k}, \\ 0, & \text{otherwise,} \end{cases}$$

Features binarization

- A natural choice of intervals is given by the quantiles, namely we can typically choose $I_{j,k} = \big(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\big]$ for $k = 1, \ldots, d_j$.
- To each binarized feature $\boldsymbol{X}^B_{\bullet,j,k}$ corresponds a parameter $\theta_{j,k}$.
- The parameters associated to the binarization of the j-th feature is denoted $\theta_{j,\bullet} = (\theta_{j,1} \cdots \theta_{j,d_j})^\top$.
- The full parameters vector of size $d = \sum_{j=1}^p d_j$, is simply

$$\theta = (\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top)^\top = \big(\theta_{1,1} \cdots \theta_{1,d_1}\, \theta_{2,1} \cdots \theta_{2,d_2} \cdots \theta_{p,1} \cdots \theta_{p,d_p}\big)^\top.$$

Features binarization

- The one-hot-encodings satisfy $\sum_{k=1}^{d_j} \boldsymbol{X}_{i,j,k} = 1$ for all j, meaning that the columns of each block sum to $\mathbf{1}_n$.
  $\rightarrow$ $\boldsymbol{X}^B$ is not of full rank by construction.
- Some of the raw features $\boldsymbol{X}_{\bullet,j}$ might not be relevant for the prediction task, so we want to select raw features from their one-hot encodings.
  $\rightarrow$ **block-sparsity** in θ.

Binarsity

- In our penalization term, we impose $\sum_{k=1}^{d_j} \theta_{j,k} = 0$ for all $j = 1, \ldots, p$ (**sum-to-zero-constraint**).
- We remark that within each block, binary features are ordered.
  $\rightarrow$ We use a within block weighted total-variation penalization

$$\sum_{j=1}^{p} \|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{\omega}_{j,\bullet}}$$

where

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{\omega}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{\omega}_{j,k} |\theta_{j,k} - \theta_{j,k-1}|,$$

Binarsity

- We therefore introduce the following new penalization called *binarsity*

$$\mathrm{bina}(\theta) = \sum_{j=1}^{p} \Big(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| + \delta_1(\theta_{j,\bullet}) \Big),$$

where the indicator function

$$\delta_1(u) = \begin{cases} 0 & \text{if} \quad \mathbf{1}^\top u = 0, \\ \infty & \text{otherwise.} \end{cases}$$

- If a raw feature j is statistically not relevant for predicting the labels, then the full block $\theta_{j,\bullet}$ should be zero.
- If a raw feature j is relevant, then the number of different values for the coefficients of $\theta_{j,\bullet}$ should be kept as small as possible, in order to balance bias and variance.

Weights in Binarsity

We consider the following data-driven weighted version of Binarsity given by

$$\hat{\omega}_{j,k} = \mathcal{O}\left(\sqrt{\frac{\log p}{n}\hat{\pi}_{j,k}}\right),$$

where

$$\hat{\pi}_{j,k} = \frac{\#\left(\left\{i = 1, \ldots, n : x_{i,j} \in \left(q_j\left(\frac{k}{d_j}\right), q_j(1)\right]\right\}\right)}{n}.$$

$\hat{\pi}_{j,k}$ corresponds to the proportion of 1s in the sub-matrix obtained by deleting the first k columns in the j-th binarized block matrix.

Generalized linear models

- The conditional distribution of Y_i given $X_i = x_i$ is assumed to be from one parameter exponential family

$$y|x \mapsto f^0(y|x) = \exp\Big(\frac{y m^0(x) - b(m^0(x))}{\varphi} + c(y)\Big),$$

- The functions $b(\cdot)$ and $c(\cdot)$ are known, while the natural parameter function $m^0(x)$ is *unknown*.
- We have

$$\mathbb{E}[Y_i|X_i = x_i] = b'(m^0(x_i)).$$

- Logistic and probit regression for binary data or multinomial regression for categorical data, Poisson regression for count data, etc ...

Generalized linear models + binarsity

- We consider the empirical risk

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, m_\theta(x_i)\big),$$

where $m_\theta(x_i) = \theta^\top x_i^B$.

- ℓ is the generalized linear model loss function and is given by

$$\ell\big(y, y'\big) = -yy' + b(y').$$

- Our estimator of m^0 is given by $\hat{m} = m_{\hat{\theta}}$, where $\hat{\theta}$ is the solution of the penalized log-likelihood problem

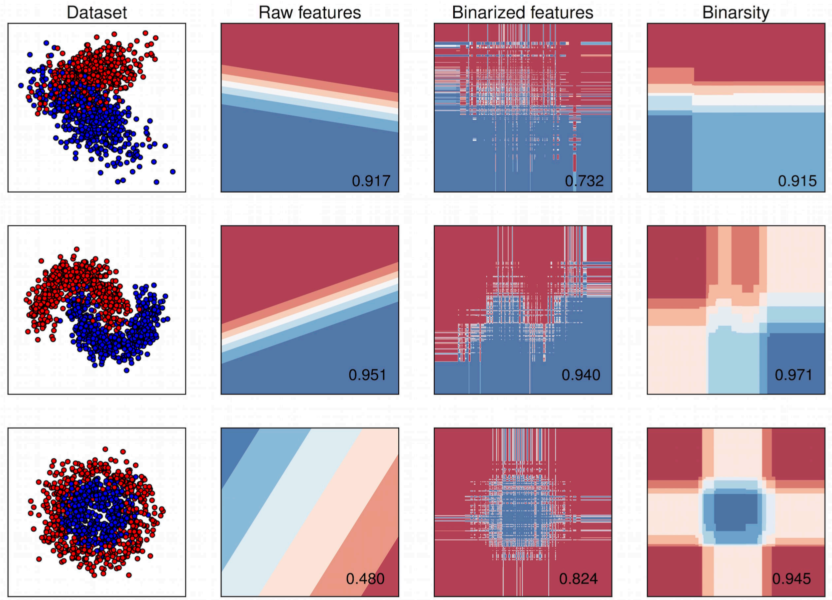
$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \big\{ R_n(\theta) + \text{bina}(\theta) \big\}.$$

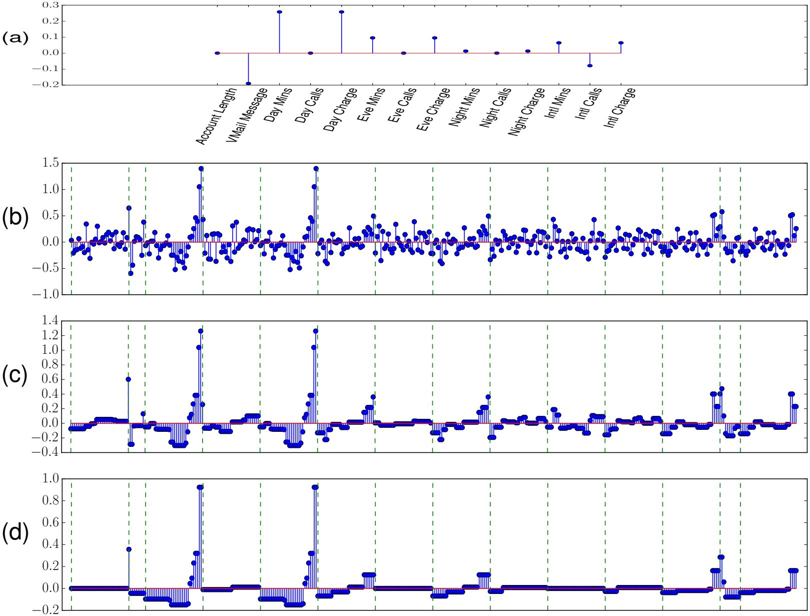
- Since Binarsity is separable by blocks, we have

$$\big(\text{prox}_{\text{bina}_{\hat{\omega}}}(\theta)\big)_{j,\bullet} = \text{prox}_{(\|\cdot\|_{\text{TV},\hat{\omega}_{j,\bullet}} + \delta_{\mathcal{H}_j})}(\theta_{j,\bullet}),$$

  for all $j = 1, \ldots, p$.
- Algorithm 2 expresses $\text{prox}_{\text{bina}_{\hat{\omega}}}$ based on the proximal operator of the weighted TV penalization.

Toy example ($n = 1000$, $p = 2$, $d_1 = d_2 = 100$)




(a)
0.3
0.2
0.1
0.0
−0.1
−0.2
Account Length
VMail Message
Day Mins
Day Calls
Day Charge
Eve Mins
Eve Calls
Eve Charge
Night Mins
Night Calls
Night Charge
Intl Mins
Intl Calls
Intl Charge
(b)
1.5
1.0
0.5
0.0
−0.5
−1.0
(c)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
(d)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Take Home Message

- We introduce a data-driven weighted total-variation penalizations for two problems: change-points detection and generalized linear models with binarized features.
- For each procedure, we give: theoretical guaranties by proving non-asymptotic oracles inequalities for the prediction error and algorithms that efficiently solve the studied convex problems.

Works in Progress

- With S. Bussy and A. Guilloux, we study the estimation problem of high-dimensional Cox model, with covariables having multiple cut-points, using binarsity penalization.
- With T. Allart, we study the complete TV penalty, which is more stable than the simple TV penalization

References

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Chambolle, A., V. Caselles, D. Cremers, M. Novaga, and T. Pock (2010). An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery 9*, 263–340.

Chiang, D. Y., G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods 6*(1), 99–103.

Condat, L. (2013). A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters 20*(11), 1054–1057.

Daubechies, I., M. Defrise, and C. De Mol (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics 57*(11), 1413–1457.

Gaïffas, S. and A. Guilloux (2012). High-dimensional additive hazards models and the lasso. *Electron. J. Statist. 6*, 522–546.

Harchaoui, Z. and C. Lévy-Leduc (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc. 105*(492), 1480–1493.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields 126*(1), 103–153.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli 12*(4), 633–661.

Shen, J. J. 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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*, 301–320.

Thank you!