Lecture 1: Segmentation and classification of genomic profiles

Jean-Philippe Vert

Mines ParisTech / Curie Institute / Inserm
Paris, France

"Optimization, machine learning and bioinformatics" summer school, Erice, Sep 9-16, 2010.

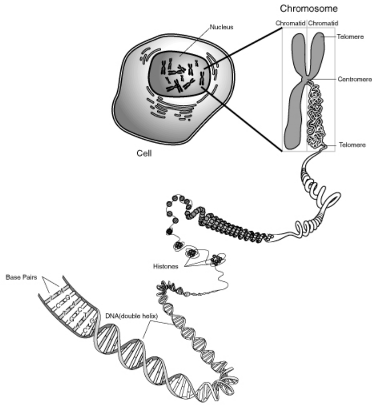
Outline

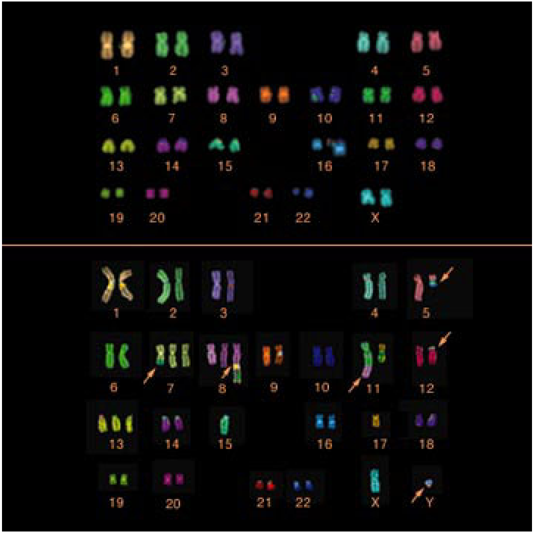
1. Motivation
2. Finding multiple change-points in a single profile
3. Finding multiple change-points shared by many signals
4. Supervised classification of genomic profiles
5. Conclusion

Outline

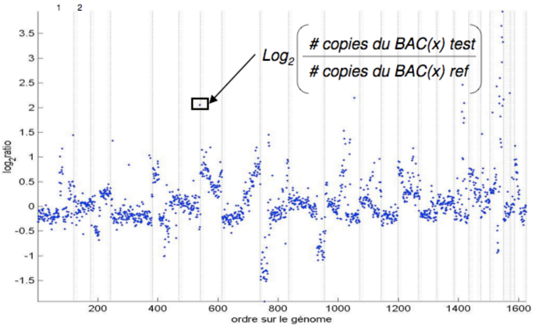
1. Motivation
2. Finding multiple change-points in a single profile
3. Finding multiple change-points shared by many signals
4. Supervised classification of genomic profiles
5. Conclusion

Chromosomic aberrations in cancer



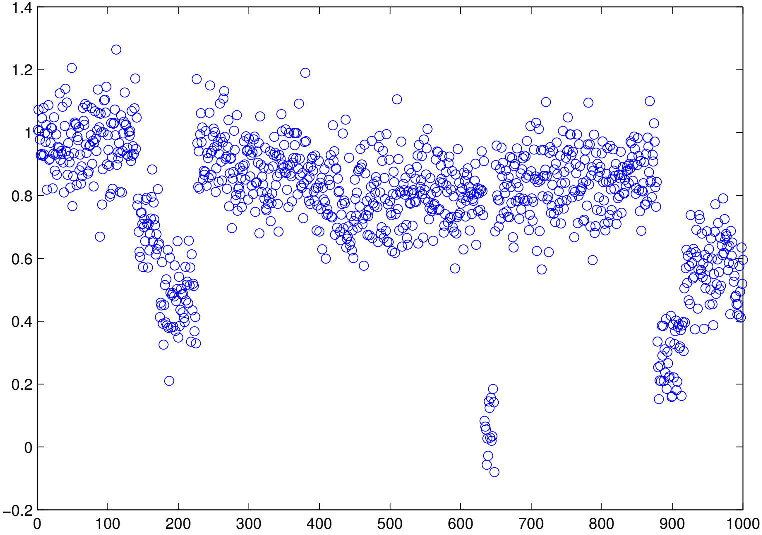


Comparative Genomic Hybridization (CGH)

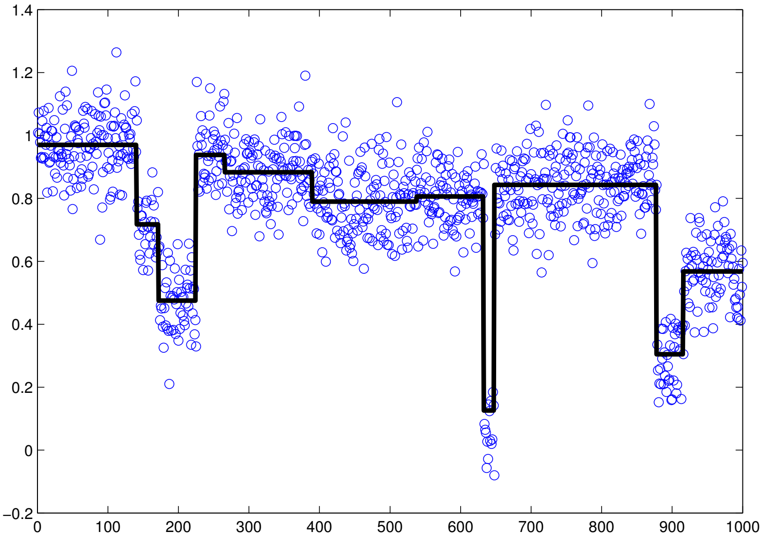


Jain et al. Genome research 2002 12:325-332

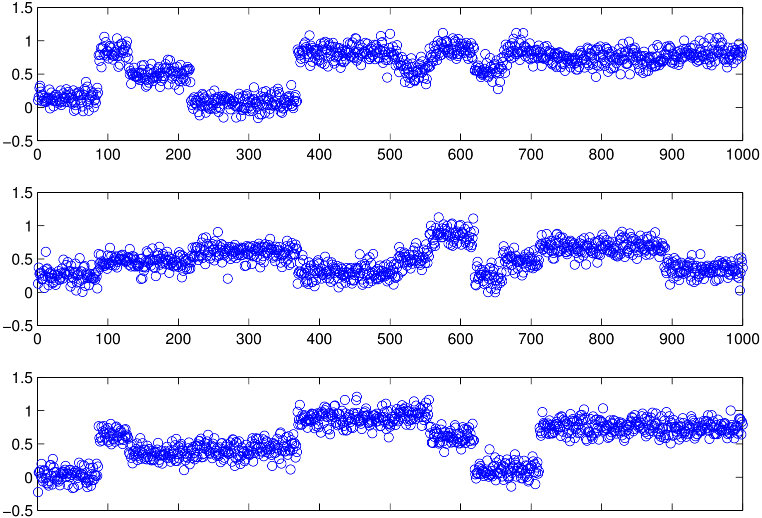
Problem 1: Finding multiple change-points in 1 profile



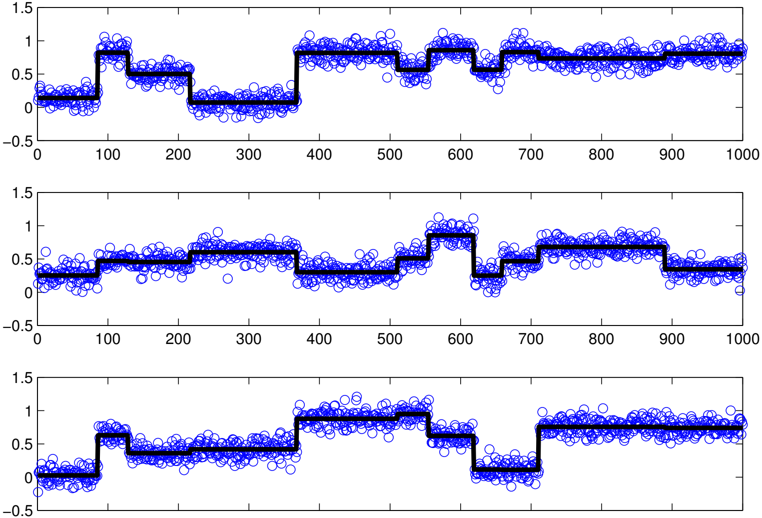
Problem 1: Finding multiple change-points in 1 profile



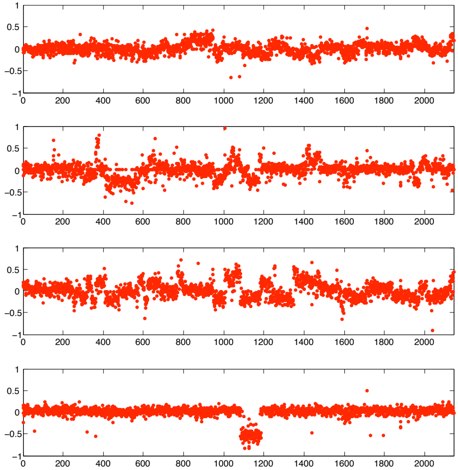
Problem 2: Finding multiple shared change-points in many profiles



Problem 2: Finding multiple shared change-points in many profiles



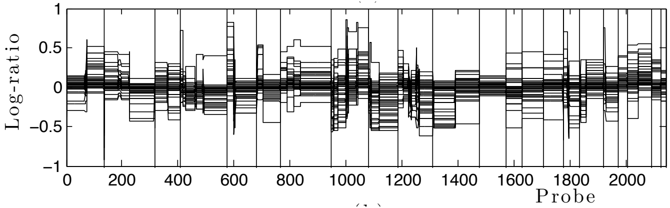
Application: find frequent breakpoints



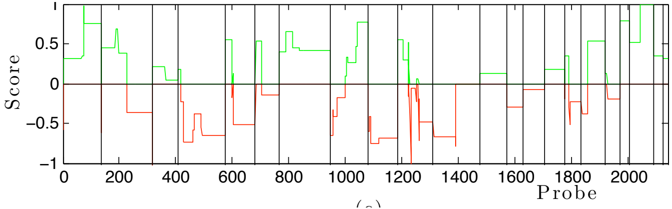
A collection of bladder tumour copy number profiles.

Other applications

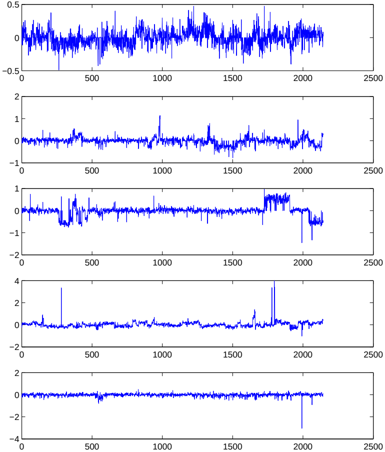
- Low-dimensional summary and visualization of the set of profiles

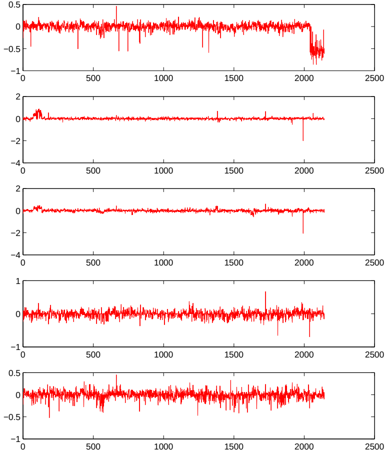


- Detection of frequently altered regions



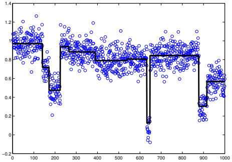
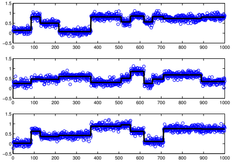
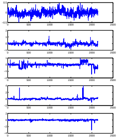
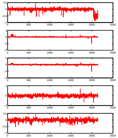
Problem 3: discrimination of genomic profiles





Aggressive (left) vs non-aggressive (right) melanoma.

What I will discuss

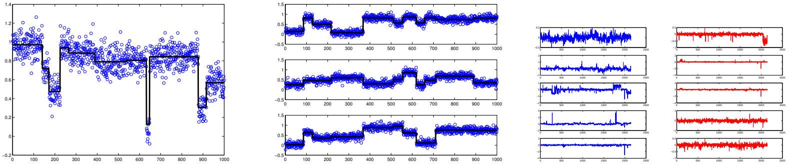





1. A general framework to solve Problems 1, 2 and 3 by rephrasing them as constrainted optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

2. Fast algorithms that scale in time and memory to
   - Profiles length: $p = 10^6 \sim 10^9$
   - Number of profiles (dimension): $n = 10^2 \sim 10^3$
   - Number of change-points: $k = 10^2 \sim 10^3$
3. Analysis of their statistical properties in some situations.

What I will discuss

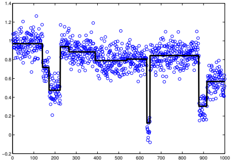


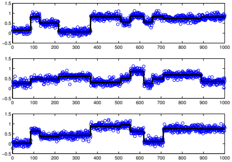
1. A general framework to solve Problems 1, 2 and 3 by rephrasing them as constrainted optimization problems of the form

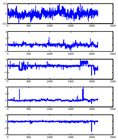
$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C \,.$$

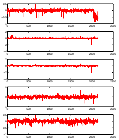
2. Fast algorithms that scale in time and memory to
   - Profiles length: $p = 10^6 \sim 10^9$
   - Number of profiles (dimension): $n = 10^2 \sim 10^3$
   - Number of change-points: $k = 10^2 \sim 10^3$
3. Analysis of their statistical properties in some situations.

What I will discuss







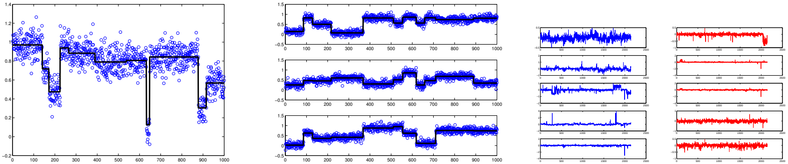


1. A general framework to solve Problems 1, 2 and 3 by rephrasing them as constrainted optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

2. Fast algorithms that scale in time and memory to
    - Profiles length: $p = 10^6 \sim 10^9$
    - Number of profiles (dimension): $n = 10^2 \sim 10^3$
    - Number of change-points: $k = 10^2 \sim 10^3$
3. Analysis of their statistical properties in some situations.

What I will discuss



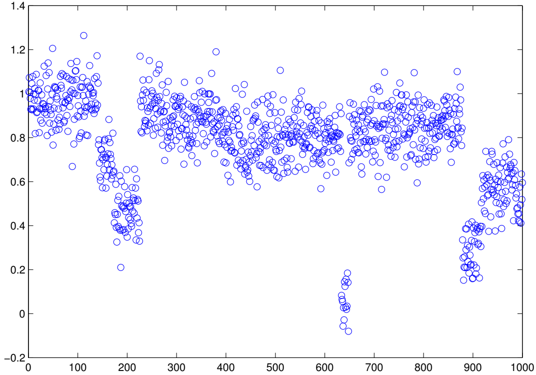
1. A general framework to solve Problems 1, 2 and 3 by rephrasing them as constrainted optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C \,.$$

2. Fast algorithms that scale in time and memory to
   - Profiles length: $p = 10^6 \sim 10^9$
   - Number of profiles (dimension): $n = 10^2 \sim 10^3$
   - Number of change-points: $k = 10^2 \sim 10^3$
3. Analysis of their statistical properties in some situations.

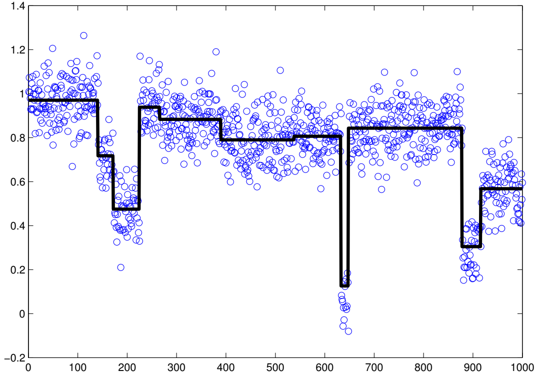
Outline

Reminder: Problem 1



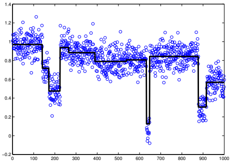
- Let $Y \in \mathbb{R}^p$ the signal
- We want to find a piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ with at most k change-points.

Reminder: Problem 1



- Let $Y \in \mathbb{R}^p$ the signal
- We want to find a piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ with at most k change-points.

An optimal solution?

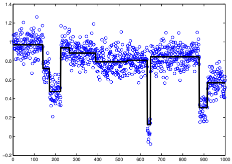


- We can define an "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ as the solution of

$$\min_{U \in \mathbb{R}^p} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1} \neq U_i\right) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

An optimal solution?

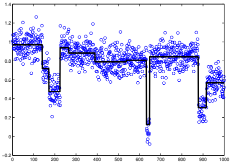


- We can define an "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ as the solution of

$$\min_{U \in \mathbb{R}^p} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}(U_{i+1} \neq U_i) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

An optimal solution?

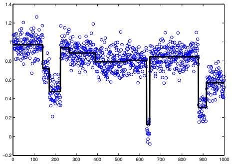


- We can define an "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ as the solution of

$$\min_{U \in \mathbb{R}^p} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1} \neq U_i\right) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

An optimal solution?



- We can define an "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^p$ as the solution of

$$\min_{U \in \mathbb{R}^p} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1} \neq U_i\right) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

Promoting sparsity with the ℓ_1 penalty

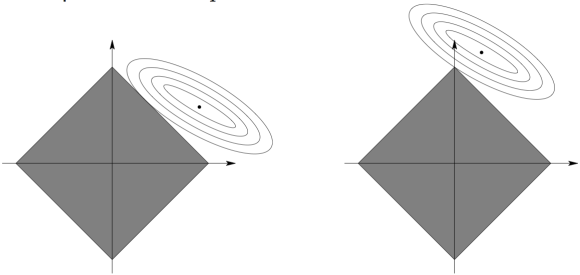
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

is usually sparse.

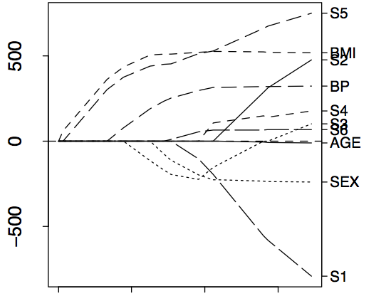
Geometric interpretation with $p = 2$



Efficienty computation of the regularization path

$$\min_{\beta \in \mathbb{R}^p} \| Y - X\beta \|^2 + \lambda \sum_{i=1}^{p} | \beta_i | \qquad (1)$$

- No explicit solution, but this is just a quadratic program.
- LARS (Efron et al., 2004) provides a fast algorithm to compute the solution for all λ's simultaneously (regularization path)



Promoting piecewise constant profiles penalty

The total variation / variable fusion penalty

If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

is usually piecewise constant (Rudin et al., 1992; Land and Friedman, 1996).

Proof:

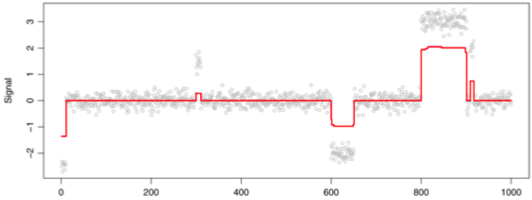
- Change of variable $u_i = \beta_{i+1} - \beta_i$, $u_0 = \beta_1$
- We obtain a Lasso problem in $u \in \mathbb{R}^{p-1}$
- u sparse means β piecewise constant

TV signal approximator

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} | \beta_{i+1} - \beta_i | \leq \mu$$

Adding additional constraints does not change the change-points:

- $\sum_{i=1}^{p} | \beta_i | \leq \nu$ (Tibshirani et al., 2005; Tibshirani and Wang, 2008)
- $\sum_{i=1}^{p} \beta_i^2 \leq \nu$ (Mairal et al. 2010)



Solving TV signal approximator

$$\min_{\beta\in\mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} | \beta_{i+1} - \beta_i | \leq \mu$$

- QP with sparse linear constraints in $O(p^2)$ -> 135 min for $p = 10^5$ (Tibshirani and Wang, 2008)
- Coordinate descent-like method $O(p)$? -> 3s s for $p = 10^5$ (Friedman et al., 2007)
- For all μ with the LARS in $O(pK)$ (Harchaoui and Levy-Leduc, 2008)
- For all μ in $O(p \ln p)$ (Hoefling, 2009)
- For the first K change-points in $O(p \ln K)$ (Bleakley and V., 2010)

Greedy dichotomic segmentation

Require: k number of intervals, $\gamma(I)$ gain function to split an interval I into $I_L(I), I_R(I)$

I_0 represents the interval $[1, p]$

$\mathcal{P} = \{I_0\}$

for $i = 1$ to k **do**

  $I^* \leftarrow \arg\max_{I \in \mathcal{P}} \gamma(I^*)$

  $\mathcal{P} \leftarrow \mathcal{P} \setminus \{I^*\}$

  $\mathcal{P} \leftarrow \mathcal{P} \cup \{I_L(I^*), I_R(I^*)\}$

end for

return $\mathcal{P}$

From greedy segmentation to TV approximator

Theorem

TV approximator is a greedy dichotomic segmentation.

Consequences:

- Fast methods for TV approximator
- Theoretical results for (apparently) greedy segmentation

From greedy segmentation to TV approximator

Theorem

TV approximator is a greedy dichotomic segmentation.

Consequences:

- Fast methods for TV approximator
- Theoretical results for (apparently) greedy segmentation

Technical details

- Represent an interval $[u+1, v]$ by a quadruplet $I = (u, v, \sigma_u, \sigma_v)$ where $\sigma_u, \sigma_v \in \{-1, 0, 1\}$
- Let $F_u = \sum_{i=1}^{u} Y_u$, and for $u < k < v$, $\sigma \in \{-1, 1\}$

$$f_I(k, \sigma) = \begin{cases} \sigma A_k / 2 & \text{if } \sigma_u = \sigma_v \neq 0\,, \\ A_k / \left(\sigma - B_k\right) & \text{otherwise}\ , \end{cases}$$

where

$$A_k = -F_k + \frac{(v-k)\,F_u + (k-u)\,F_v}{v-u}\,,$$

$$B_k = \frac{(v-k)\,\sigma_u + (k-u)\,\sigma_v}{v-u}\,.$$

Technical details (cont.)

Then the functions $\gamma(I)$, $I_L(I)$ and $I_R(I)$ are respectively given by:

$$\gamma(I) = \max_{k\in[u+1,v-1],\sigma\in\{-1,1\}} f_I(k,\sigma)\,,$$

$$(k^*,\sigma^*) = \underset{k\in[u+1,v-1],\sigma\in\{-1,1\}}{\operatorname{argmax}} f_I(k,\sigma)\,,$$

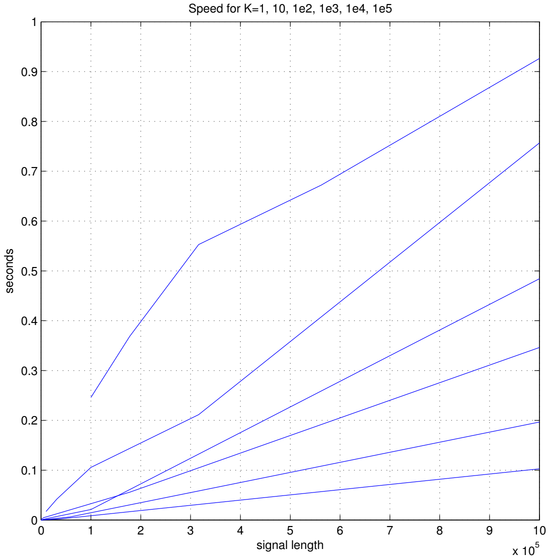
$$I_L(I) = (u, k^*, \sigma_u, \sigma^*)\ ,$$

$$I_R(I) = (k^*, v, \sigma^*, \sigma_v)\ .$$

Proof (sketch)

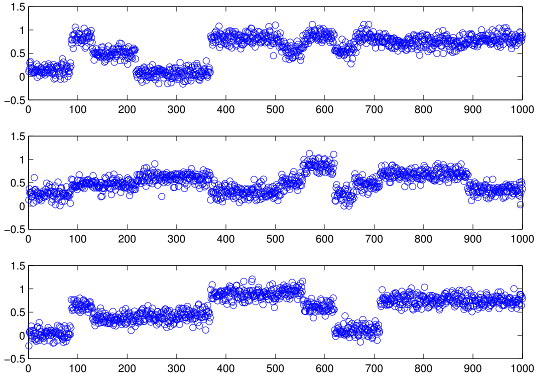
- Homotopy method (LARS)
- Similar to Harchaoui and Levy-Leduc (2008), removing superfluous computations
- The next breakpoint in a segment, and the μ where it appears, is independent of events in other segments

Speed trial : 2 s. for $K = 100$, $p = 10^7$



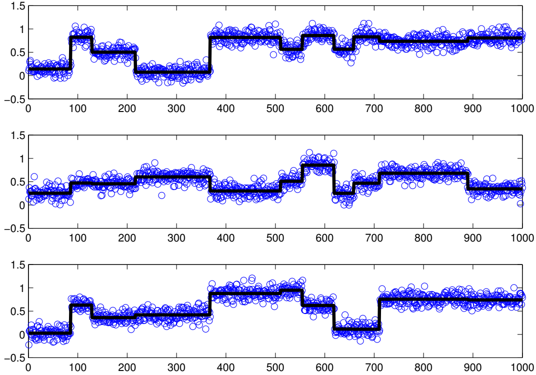
Outline

Reminder: Problem 2



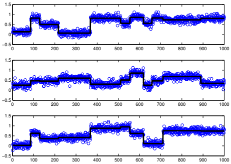
- Let $Y \in \mathbb{R}^{p \times n}$ the n signals of length p
- We want to find a piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ with at most k change-points.

Reminder: Problem 2



- Let $Y \in \mathbb{R}^{p \times n}$ the n signals of length p
- We want to find a piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ with at most k change-points.

"Optimal" segmentation by dynamic programming



- Define the "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ of Y as the solution of

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

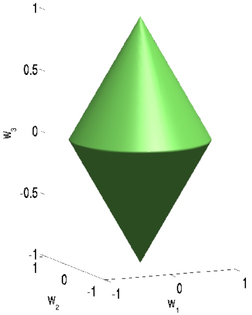
- DP finds the solution in $O(p^2 kn)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

Selecting pre-defined groups of variables

Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the ℓ_1/ℓ_2-norm induces sparse solutions *at the group level*:

$$\Omega_{group}(w) = \sum_g \|w_g\|_2$$



$$\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2 = \sqrt{w_1^2 + w_2^2} + \sqrt{w_3^2}$$

TV approximator for many signals

- Replace

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

by

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

Questions

- Practice: can we solve it efficiently?
- Theory: does it benefit from increasing p (for n fixed)?

TV approximator as a group Lasso problem

- Make the change of variables:

$$\gamma = U_{1,\bullet}\,,$$
$$\beta_{i,\bullet} = w_i \left(U_{i+1,\bullet} - U_{i,\bullet}\right) \quad \text{for } i = 1, \ldots, p-1\,.$$

- TV approximator is then equivalent to the following group Lasso problem (Yuan and Lin, 2006):

$$\min_{\beta \in \mathbb{R}^{(p-1)\times n}} \| \bar{Y} - \bar{X}\beta \|^2 + \lambda \sum_{i=1}^{p-1} \| \beta_{i,\bullet} \|\,,$$

where $\bar{Y}$ is the centered signal matrix and $\bar{X}$ is a particular $(p-1) \times (p-1)$ design matrix.

TV approximator implementation

$$\min_{\beta \in \mathbb{R}^{(p-1)\times n}} \| \bar{Y} - \bar{X}\beta \|^2 + \lambda \sum_{i=1}^{p-1} \| \beta_{i,\bullet} \| \,,$$

Theorem

The TV approximator can be solved efficiently:

- approximately with the group LARS in $O(npk)$ in time and $O(np)$ in memory
- exactly with a block coordinate descent + active set method in $O(np)$ in memory

Proof: computational tricks...

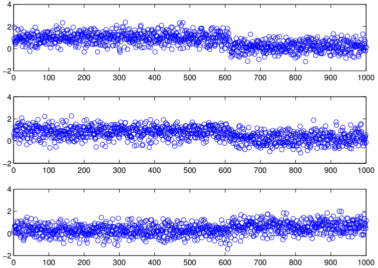
Although $\bar{X}$ is $(p-1) \times (p-1)$:

- For any $R \in \mathbb{R}^{p \times n}$, we can compute $C = \bar{X}^\top R$ in $O(np)$ operations and memory
- For any two subset of indices $A = \left(a_1, \ldots, a_{|A|}\right)$ and $B = \left(b_1, \ldots, b_{|B|}\right)$ in $[1, p-1]$, we can compute $\bar{X}^\top_{\bullet,A}\bar{X}_{\bullet,B}$ in $O\left(|A||B|\right)$ in time and memory
- For any $A = \left(a_1, \ldots, a_{|A|}\right)$, set of distinct indices with $1 \leq a_1 < \ldots < a_{|A|} \leq p-1$, and for any $|A| \times n$ matrix R, we can compute $C = \left(\bar{X}^\top_{\bullet,A}\bar{X}_{\bullet,A}\right)^{-1} R$ in $O(|A|n)$ in time and memory

Consistency for a single change-point

Suppose a single change-point:

- at position $u = \alpha p$
- with increments $(\beta_i)_{i=1,\ldots,n}$ s.t. $\bar{\beta}^2 = \lim_{k\to\infty} \frac{1}{n}\sum_{i=1}^{n} \beta_i^2$
- corrupted by i.i.d. Gaussian noise of variance σ^2



Does the TV approximator correctly estimate the first change-point as p increases?

Consistency of the unweighted TV approximator

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

Theorem

The unweighted TV approximator finds the correct change-point with probability tending to 1 *(resp. 0) as* $n \to +\infty$ *if* $\sigma^2 < \tilde{\sigma}^2_\alpha$ *(resp.* $\sigma^2 > \tilde{\sigma}^2_\alpha$*), where*

$$\tilde{\sigma}^2_\alpha = p\bar{\beta}^2 \frac{(1-\alpha)^2(\alpha - \frac{1}{2p})}{\alpha - \frac{1}{2} - \frac{1}{2p}} .$$

- correct estimation on $[p\epsilon, p(1-\epsilon)]$ with $\epsilon = \sqrt{\frac{\sigma^2}{2p\bar{\beta}^2}} + o(p^{-1/2})$.
- wrong estimation near the boundaries

Consistency of the weighted TV approximator

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

Theorem

The weighted TV approximator with weights

$$\forall i \in [1, p-1] \,, \quad w_i = \sqrt{\frac{i(p-i)}{p}}$$

correctly finds the first change-point with probability tending to 1 *as* $n \to +\infty$.

- we see the benefit of increasing n
- we see the benefit of adding weights to the TV penalty

Proof sketch

- The first change-point $\hat{i}$ found by TV approximator maximizes $F_i = \| \hat{c}_{i,\bullet} \|^2$, where

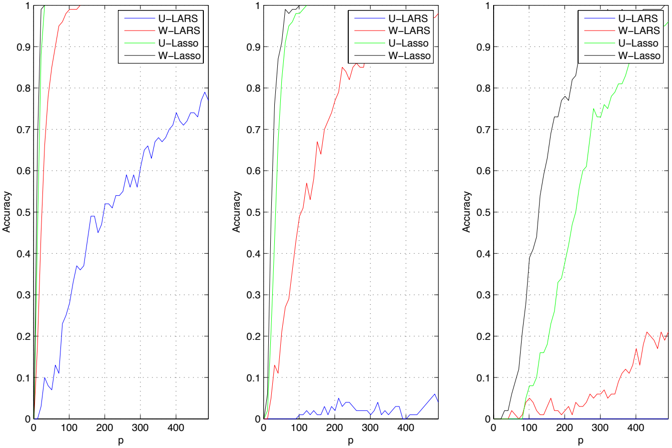
$$\hat{c} = \bar{X}^\top \bar{Y} = \bar{X}^\top \bar{X} \beta^* + \bar{X}^\top W \,.$$

- $\hat{c}$ is Gaussian, and F_i is follows a non-central χ^2 distribution with

$$G_i = \frac{EF_i}{p} = \frac{i(p-i)}{pw_i^2}\sigma^2 + \frac{\bar{\beta}^2}{w_i^2 w_u^2 p^2} \times \begin{cases} i^2 \left(p-u\right)^2 & \text{if } i \leq u \,, \\ u^2 \left(p-i\right)^2 & \text{otherwise.} \end{cases}$$

- We then just check when $G_u = \max_i G_i$

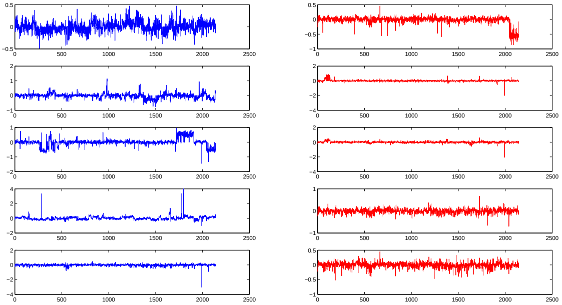
Consistent estimation of more change-points?



$p = 100, k = 10, \bar{\beta}^2 = 1, \sigma^2 \in \{0.05; 0.2; 1\}$

Outline

Reminder: Problem 3



- $x_1, \ldots, x_n \in \mathbb{R}^p$ the n profiles of length p
- $y_1, \ldots, y_n \in [-1, 1]$ the labels
- We want to learn a function $f : \mathbb{R}^p \rightarrow [-1, 1]$

Shrinkage estimators

- Define a large family of "candidate classifiers", e.g., linear predictors $f_\beta(x) = \beta^\top x$ for $\beta \in \mathbb{R}^p$
- For any candidate $\beta \in \mathbb{R}^p$, quantify how "good" f_β is on the training set with some empirical risk, e.g.:

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\beta(x_i), y_i)\,.$$

- Choose β that achieves the minimium empirical risk, subject to some constraint:

$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C\,.$$

Shrinkage estimators

- Define a large family of "candidate classifiers", e.g., linear predictors $f_\beta(x) = \beta^\top x$ for $\beta \in \mathbb{R}^p$
- For any candidate $\beta \in \mathbb{R}^p$, quantify how "good" f_β is on the training set with some empirical risk, e.g.:

$$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) .$$

- Choose β that achieves the minimium empirical risk, subject to some constraint:

$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C .$$

Shrinkage estimators

- Define a large family of "candidate classifiers", e.g., linear predictors $f_\beta(x) = \beta^\top x$ for $\beta \in \mathbb{R}^p$
- For any candidate $\beta \in \mathbb{R}^p$, quantify how "good" f_β is on the training set with some empirical risk, e.g.:

$$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) \,.$$

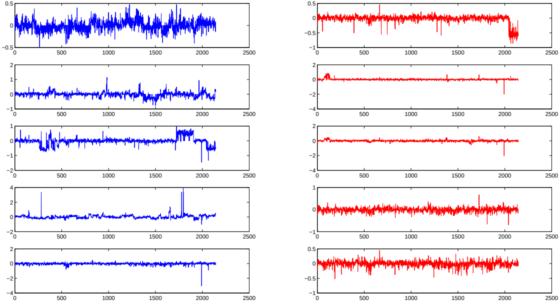
- Choose β that achieves the minimium empirical risk, subject to some constraint:

$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$

Prior knowledge

We expect β to be

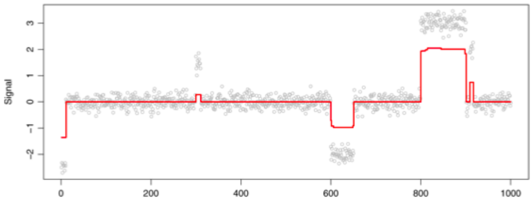
- sparse : not all positions should be discriminative, and we want to identify the predictive region (presence of oncogenes or tumor suppressor genes?)
- piecewise constant : within a selected region, all probes should contribute equally



Fused Lasso signal approximator (Tibshirani et al., 2005)

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{p} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \, .$$

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions



Fused lasso for supervised classification (Rapaport et al., 2008)

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

where ℓ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$.

Implementation

- When ℓ is the hinge loss (fused SVM), this is a linear program -> up to $p = 10^3 \sim 10^4$
- When ℓ is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to $p = 10^8 \sim 10^9$

Fused lasso for supervised classification (Rapaport et al., 2008)

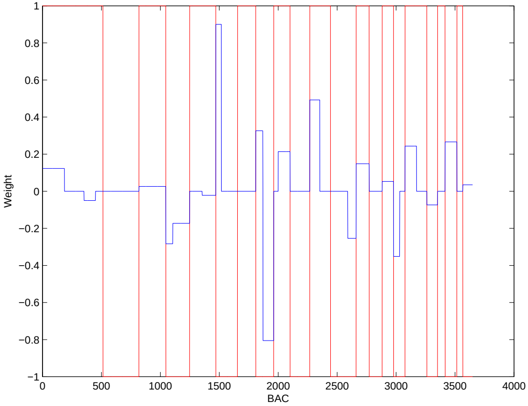
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

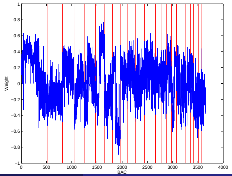
where ℓ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$.

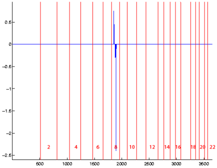
Implementation

- When ℓ is the hinge loss (fused SVM), this is a linear program -> up to $p = 10^3 \sim 10^4$
- When ℓ is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to $p = 10^8 \sim 10^9$

Example: predicting metastasis in melanoma

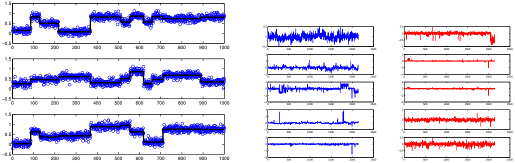






Outline

Conclusion

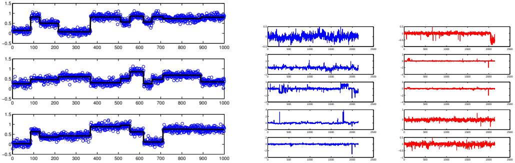


- We formulated 3 related problems as constrained optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C \,.$$

- The risk $R(w)$ depends on the problem we want to solve
- The penalty $\Omega(w)$ depends on the data, here we focused on the total variation and its variants
- Dedicated optimization algorithms lead to fast implementation
- An illustration of a very active and fruitful trend in ML!

Conclusion

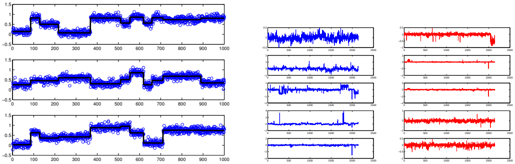


- We formulated 3 related problems as constrained optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

- The risk $R(w)$ depends on the problem we want to solve
- The penalty $\Omega(w)$ depends on the data, here we focused on the total variation and its variants
- Dedicated optimization algorithms lead to fast implementation
- An illustration of a very active and fruitful trend in ML!

Conclusion

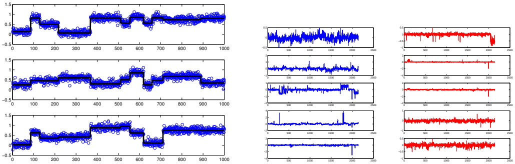


- We formulated 3 related problems as constrained optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

- The risk $R(w)$ depends on the problem we want to solve
- The penalty $\Omega(w)$ depends on the data, here we focused on the total variation and its variants
- Dedicated optimization algorithms lead to fast implementation
- An illustration of a very active and fruitful trend in ML!

Conclusion

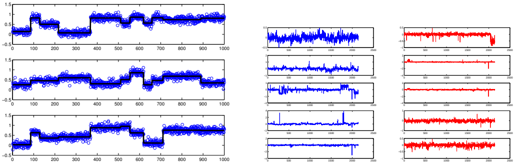


- We formulated 3 related problems as constrained optimization problems of the form

$$\min_{w} R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

- The risk $R(w)$ depends on the problem we want to solve
- The penalty $\Omega(w)$ depends on the data, here we focused on the total variation and its variants
- Dedicated optimization algorithms lead to fast implementation
- An illustration of a very active and fruitful trend in ML!

Conclusion



- We formulated 3 related problems as constrained optimization problems of the form

$$\min_w R(w) \quad \text{s.t.} \quad \Omega(w) \leq C\,.$$

- The risk $R(w)$ depends on the problem we want to solve
- The penalty $\Omega(w)$ depends on the data, here we focused on the total variation and its variants
- Dedicated optimization algorithms lead to fast implementation
- An illustration of a very active and fruitful trend in ML!