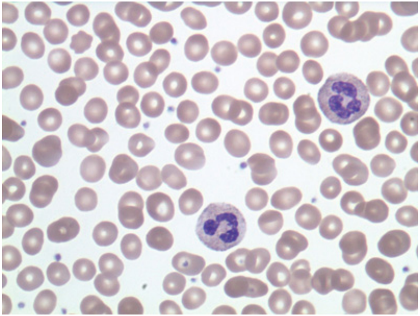
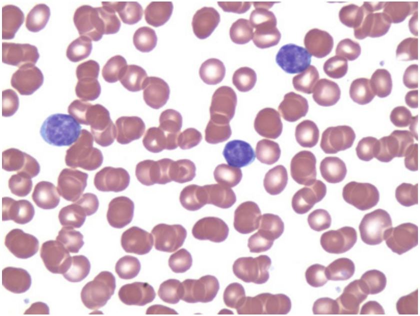
# The group fused Lasso for multiple change-point detection

Jean-Philippe Vert

Inference for Change-Point and Related Processes workshop
Isaac Newton Institute for Mathematical Sciences
Cambridge, UK, January 14, 2014

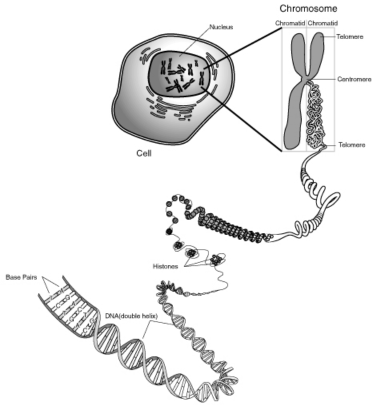
# Normal vs cancer cells

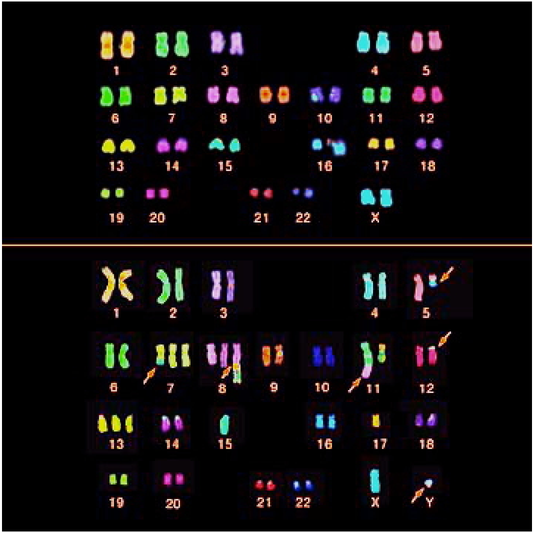




What goes wrong?
How to treat?

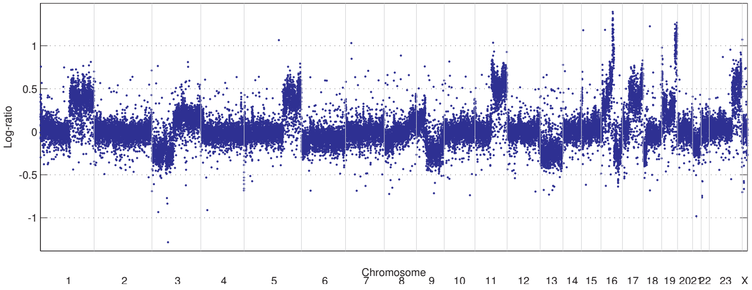
# Chromosomic aberrations in cancer



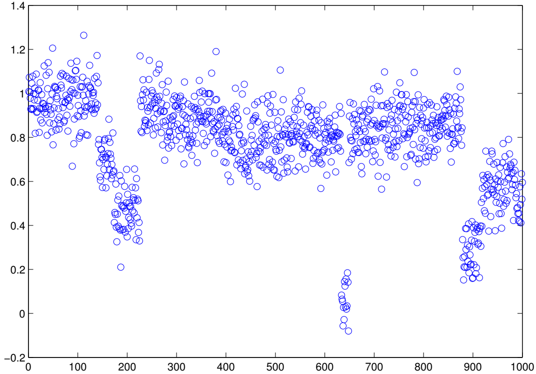



## Motivation

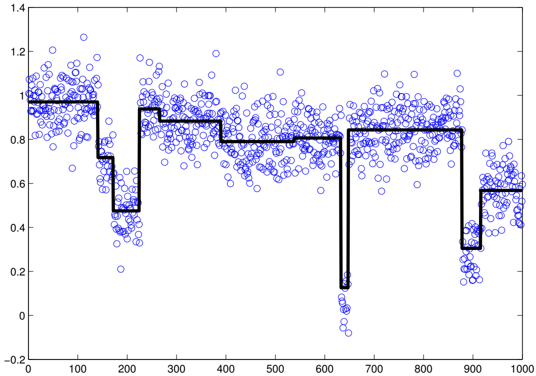
- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research to observe systematically variants in DNA content
- Progressively replaced by high throughput sequencing techniques

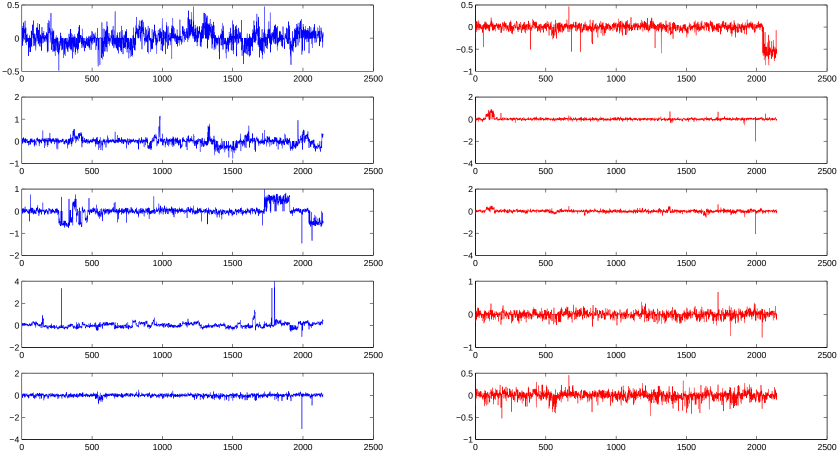
# Problem 1: find change-points in one (long) profile



# Problem 1: find change-points in one (long) profile

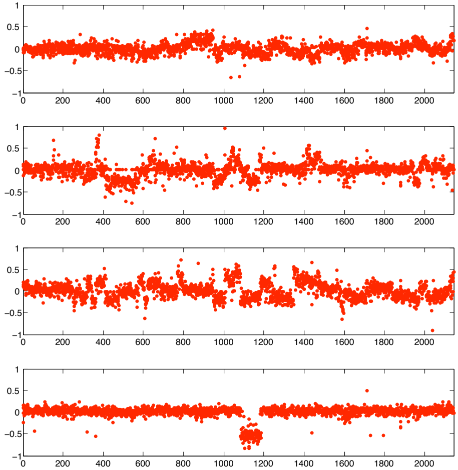


# Problem 2: learn to discriminate profiles



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

# Problem 3: Find frequent breakpoints




*A collection of bladder tumour copy number profiles.*

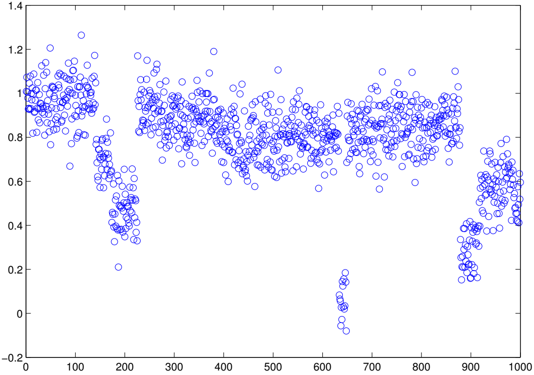
# Outline

1. Fast fused lasso for change-point detection
2. Fused SVM for discrimination of profiles
3. Group fused lasso for multiple frequent change-point detection

# Outline

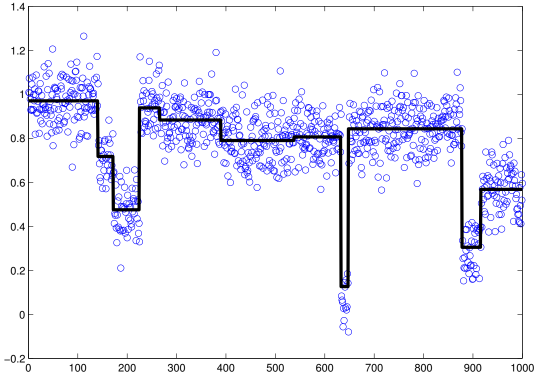
1. Fast fused lasso for change-point detection
2. Fused SVM for discrimination of profiles
3. Group fused lasso for multiple frequent change-point detection

# Can we identify breakpoints and "smooth" each profile?



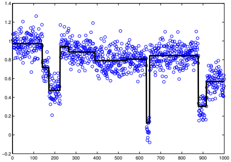

- A classical multiple change-point detection problem
- Should scale to lengths of order $10^6 \sim 10^8$

# Can we identify breakpoints and "smooth" each profile?




- A classical multiple change-point detection problem
- Should scale to lengths of order $10^6 \sim 10^8$

# An optimal solution

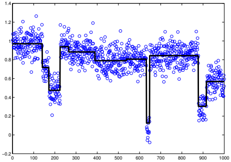


- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

$$\min_{\beta \in \mathbb{R}^p} \parallel Y - \beta \parallel^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(\beta_{i+1} \neq \beta_i\right) \leq k$$

- 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 $p = 10^6 \sim 10^8$...

# An optimal solution

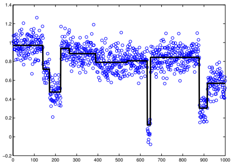



- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

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

- 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 $p = 10^6 \sim 10^8$...

# An optimal solution

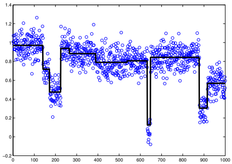



- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

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

- 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 $p = 10^6 \sim 10^8$...

# An optimal solution




- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

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

- 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 $p = 10^6 \sim 10^8$...

# Promoting sparsity with the $\ell_1$ penalty

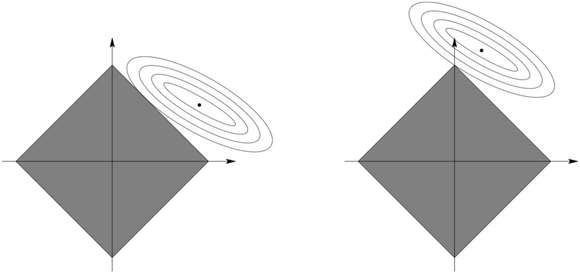
## The $\ell_1$ penalty (Tibshirani, 1996; Chen et al., 1998)

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 $p = 2$



# 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:

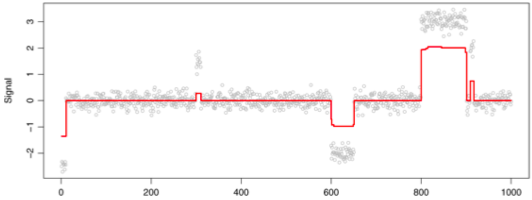
- Change of variable $u_i = \beta_{i+1} - \beta_i$, $u_0 = \beta_1$
- We obtain a Lasso problem in $u \in \mathbb{R}^{p-1}$
- $u$ sparse means $\beta$ piecewise constant

# 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 $\mu$ with the LARS in $O(pK)$ (Harchaoui and Levy-Leduc, 2008)
- For all $\mu$ in $O(p \ln p)$ (Hoefling, 2009)
- For the first $K$ change-points in $O(p \ln K)$ (Bleakley and V., 2010)

# Solving TV signal approximator in $O(p \ln K)$

**Theorem (V. and Bleakley, 2010; see also Hoefling, 2009)**

TV signal approximator is a binary segmentation algorithm

**Algorithm 1** 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 $[1, n]$

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

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

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

$\quad \mathcal{P} \leftarrow \mathcal{P} \backslash \{I^*\}$

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

**end for**

**return** $\mathcal{P}$

Apparently greedy algorithm finds the global optimum!

# Solving TV signal approximator in $O(p \ln K)$

**Theorem (V. and Bleakley, 2010; see also Hoefling, 2009)**

TV signal approximator is a binary segmentation algorithm

Consequences:

- Good news: very fast methods to find the global optimum of TV approximator
- Good news: we can analyze this greedy method by expressing the solution as the global minimum of an objective function
- Bad news: TV approximator is no more than a binary segmentation method...

Extension to *hierarchical clustering*: `ClusterPath` (Hocking et al., ICML 2011)

# 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.)

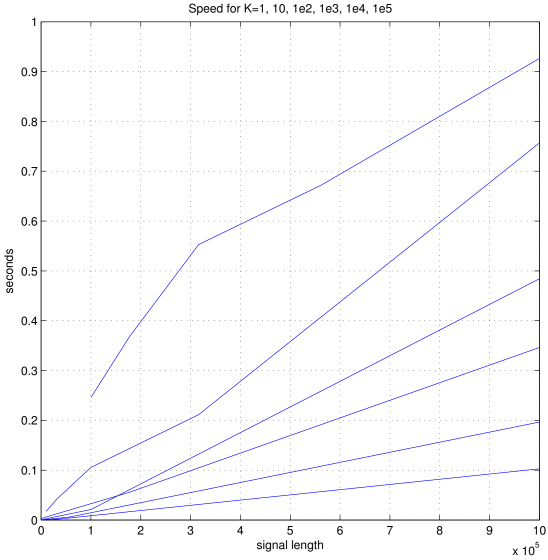
Then the functions $\gamma(I)$, $I_L(I)$ and $I_R(I)$ are respectively given by:

$$
\begin{aligned}
\gamma(I) &= \max_{k\in[u+1,v-1],\sigma\in\{-1,1\}} f_I(k,\sigma)\,,\\
(k^*,\sigma^*) &= \operatorname*{argmax}_{k\in[u+1,v-1],\sigma\in\{-1,1\}} f_I(k,\sigma)\,,\\
I_L(I) &= (u,k^*,\sigma_u,\sigma^*)\ ,\\
I_R(I) &= (k^*,v,\sigma^*,\sigma_v)\ .
\end{aligned}
$$

# Proof (sketch)

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

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



# Application

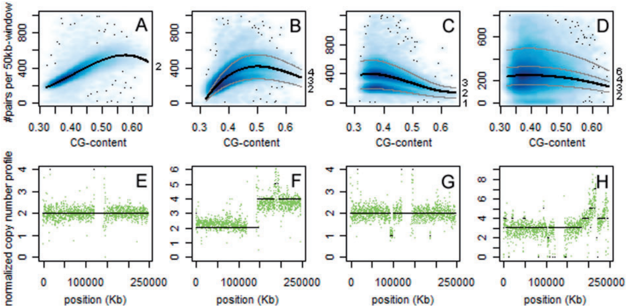
BIOINFORMATICS **APPLICATIONS NOTE**
*Vol. 27 no. 2 2011, pages 268–269*
*doi:10.1093/bioinformatics/btq635*

*Genome analysis*
Advance Access publication November 15, 2010

## Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization

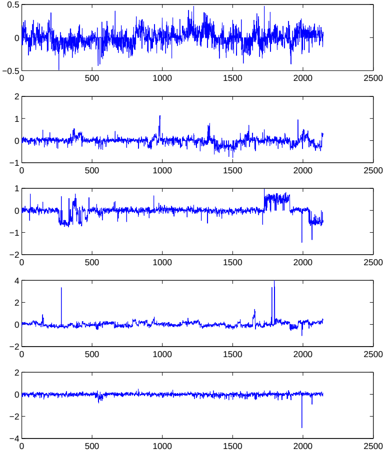
Valentina Boeva[1,2,3,4,*], Andrei Zinovyev[1,2,3], Kevin Bleakley[1,2,3], Jean-Philippe Vert[1,2,3], Isabelle Janoueix-Lerosey[1,4], Olivier Delattre[1,4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, [2]INSERM, U900, Paris, F-75248, [3]Mines ParisTech, Fontainebleau, F-77300 and [4]INSERM, U830, Paris, F-75248 France



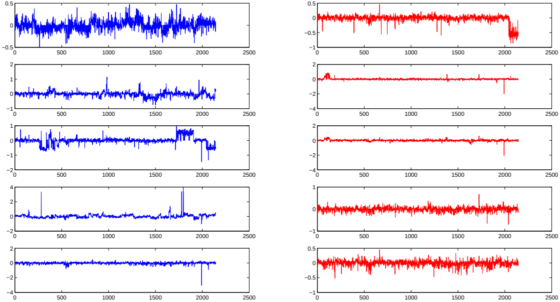
# Outline

# Extension: cancer prognosis




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

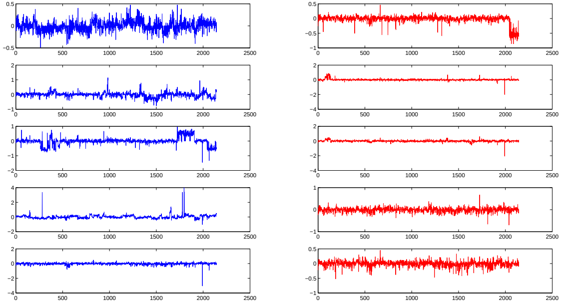
# The problem



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

# Prior knowledge

- Sparsity : 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 for supervised classification (Rapaport et al., 2008)

Find a linear predictor $f(Y) = \beta^\top Y$ that best discriminates the aggressive vs non-aggressive samples, subject to the constraints that it should be sparse and piecewise constant:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

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

## Implementation

- When $\ell$ is the hinge loss (fused SVM), this is a linear program -> up to $p = 10^3 \sim 10^4$
- When $\ell$ 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)

Find a linear predictor $f(Y) = \beta^\top Y$ that best discriminates the aggressive vs non-aggressive samples, subject to the constraints that it should be sparse and piecewise constant:

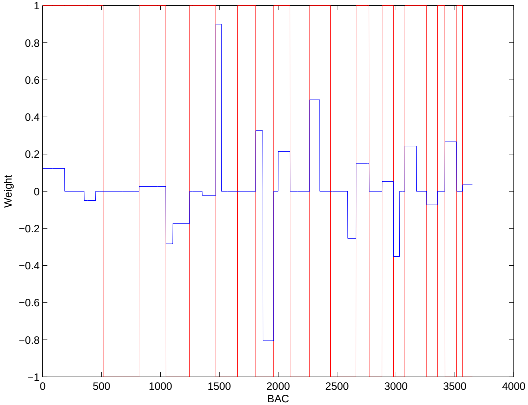
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

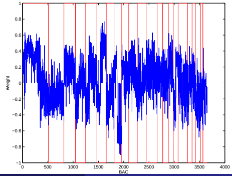
where $\ell$ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$.

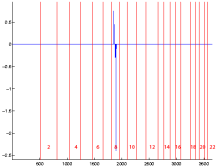
## Implementation

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

# Example: prognosis in melanoma

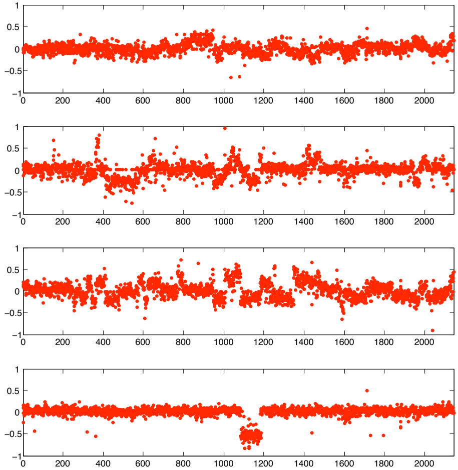


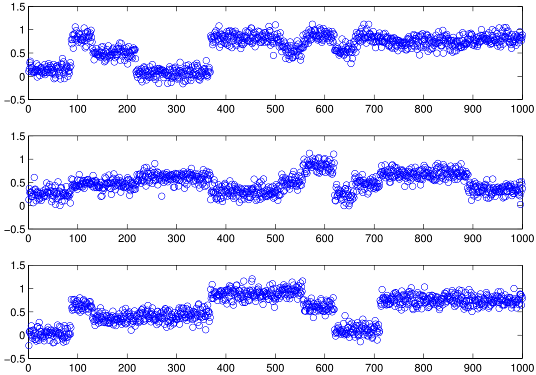
# Outline

# Can we detect frequent breakpoints?

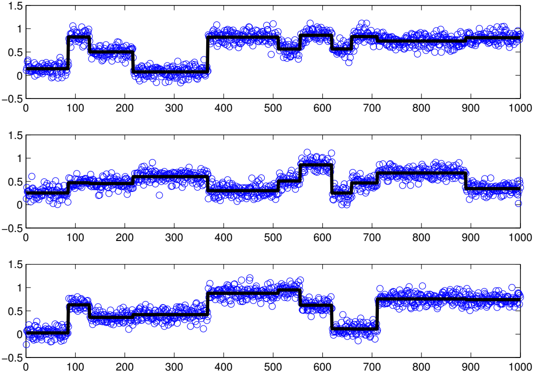


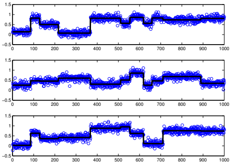

*A collection of bladder tumour copy number profiles.*

# The problem



# The problem





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

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

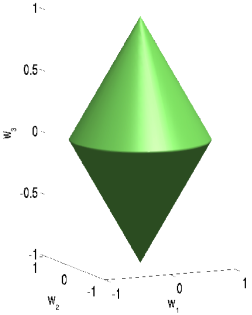
- DP finds the solution in $O(p^2 kn)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^8$...

# Selecting pre-defined groups of variables

## Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the $\ell_1/\ell_2$-norm induces sparse solutions *at the group level*:

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




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

# GFLseg (Bleakley and V., 2011)

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

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

GFLseg = Group Fused Lasso segmentation

## Questions

- Practice: can we solve it efficiently?
- Theory: does it recover the correct segmentation?

# GFLseg (Bleakley and V., 2011)

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

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

GFLseg = Group Fused Lasso segmentation

## Questions

- Practice: can we solve it efficiently?
- Theory: does it recover the correct segmentation?

# GFLseg as a group Lasso problem

- Make the change of variables:

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

- 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... (from Zaid Harchaoui)

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

# Speed trial

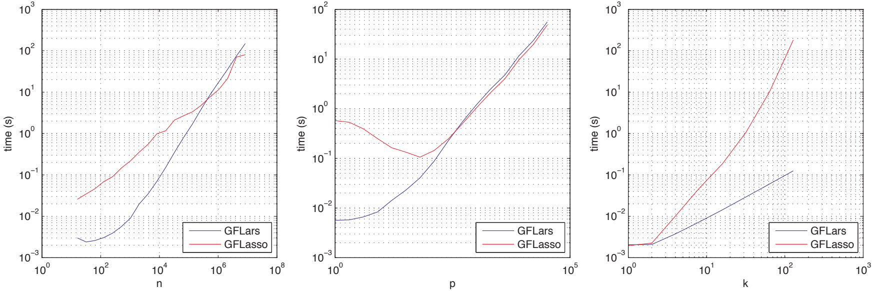
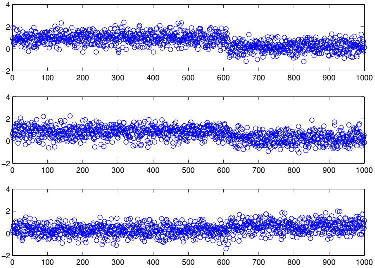



Figure 2: **Speed trials for group fused LARS (top row) and Lasso (bottom row).** *Left column:* varying $n$, with fixed $p = 10$ and $k = 10$; *center column:* varying $p$, with fixed $n = 1000$ and $k = 10$; *right column:* varying $k$, with fixed $n = 1000$ and $p = 10$. Figure axes are log-log. Results are averaged over 100 trials.

# Consistency

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 $\sigma^2$




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* $\sigma^2 < \tilde{\sigma}_\alpha^2$ *(resp.* $\sigma^2 > \tilde{\sigma}_\alpha^2$*), where*

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

- 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 $\chi^2$ 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 $G_u = \max_i G_i$

# Consistency for a single change-point

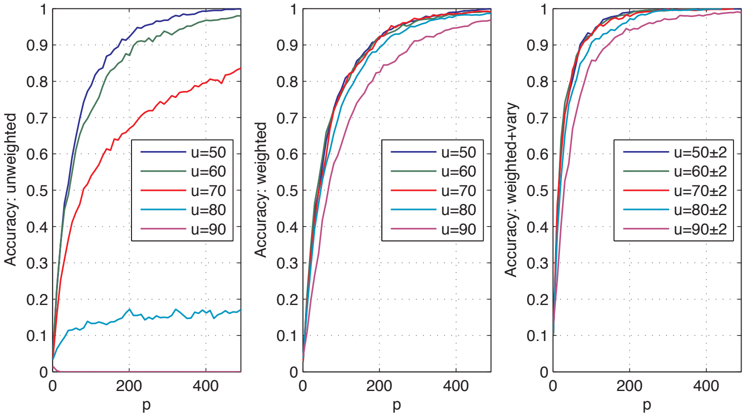



Figure 3: **Single change-point accuracy for the group fused Lasso.** Accuracy as a function of the number of profiles $p$ when the change-point is placed in a variety of positions $u = 50$ to $u = 90$ (left and centre plots, resp. unweighted and weighted group fused Lasso), or: $u = 50 \pm 2$ to $u = 90 \pm 2$ (right plot, weighted with varying change-point location), for a signal of length 100.

# Estimation of several change-points

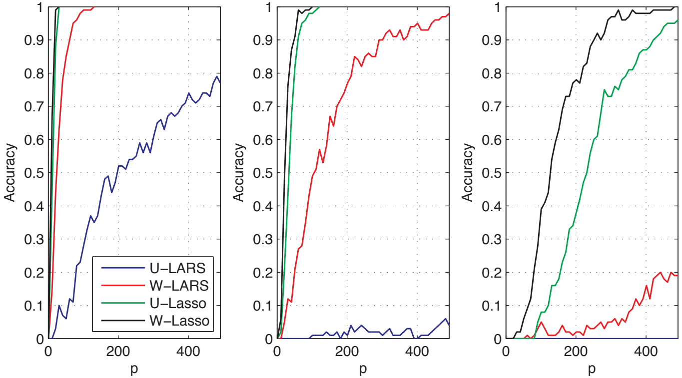
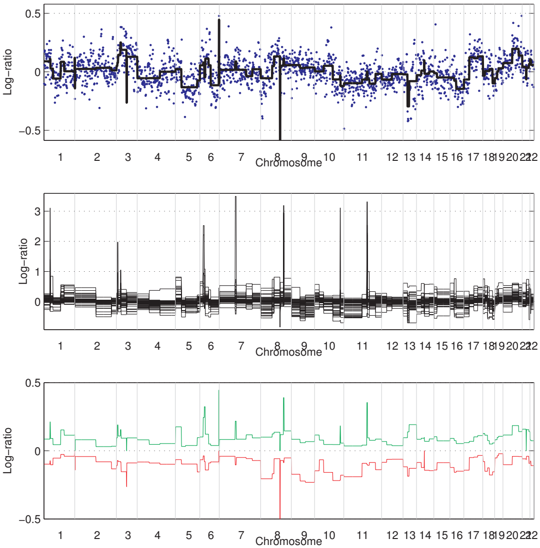



Figure 4: **Multiple change-point accuracy.** Accuracy as a function of the number of profiles $p$ when change-points are placed at the nine positions $\{10, 20, \ldots, 90\}$ and the variance $\sigma^2$ of the centered Gaussian noise is either $0.05$ (left), $0.2$ (center) and $1$ (right). The profile length is $100$.

# Application: detection of frequent abnormalities



# Conclusion

- Convex norms with singularities at piecewise-constant profiles
- Global optimum of fused lasso found by binary segmentation
- Efficient proximal methods for optimization with general loss functions (supervised classification, regression, ...)
- Benefit of increasing the number of profiles

Some questions

- Theoretical results for $K$ change-points in $n$ profiles of length $p$
- What if just a few profiles have a change-point?
- What about time series on a network?
- How to choose the number of change-points?

# Acknowledgements

Kevin Bleakley (INRIA)

European Research Council

AGENCE NATIONALE DE LA RECHERCHE ANR

erc