Structured feature selection for genomic dataa

Jean-Philippe Vert

Mines ParisTech and Curie Institute

Machine Learning and Applications in Biology workshop,
Sapporo, August 6-7, 2012

Outline

Outline

1. Lasso background
2. Frequent breakpoint detection in genomic profiles
3. Gene selection with prior information
4. Conclusion

Feature selection with the lasso

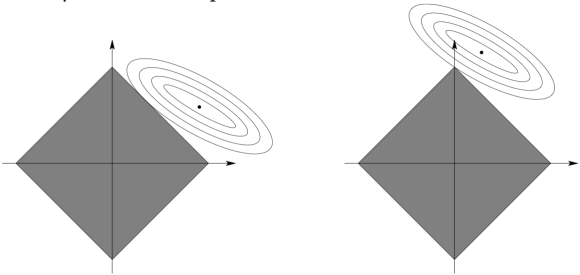
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

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 $p = 2$



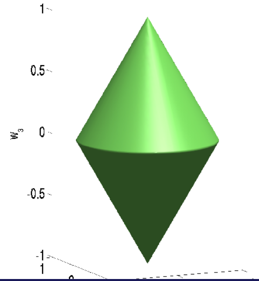
Structured feature selection with the group lasso

The ℓ_1/ℓ_2 penalty (Bach et al., 2004; Yuan & Lin, 2006)

Let $\mathcal{G} = \{g_1, g_2, \ldots\}$ be a partition of $[1, p]$ into disjoint groups. If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{g \in \mathcal{G}} \|\beta_g\|$$

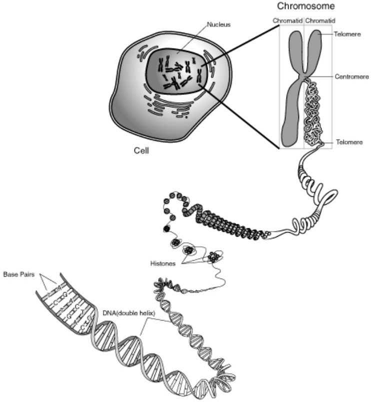
is usually group sparse.

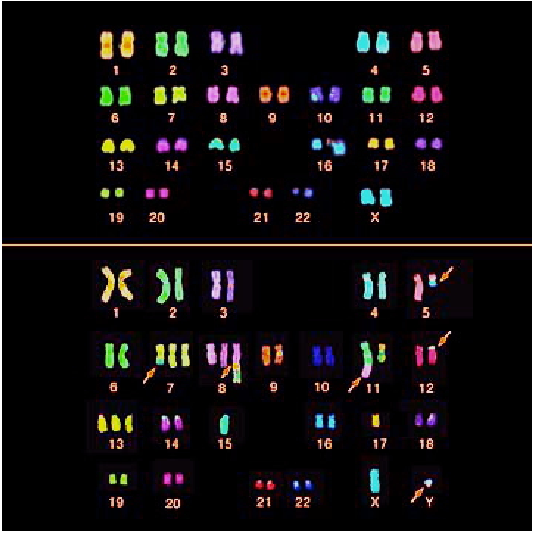


$$\Omega(\beta_1, \beta_2, \beta_3) = \|(\beta_1, \beta_2)\|_2 + \|\beta_3\|_2$$
$$= \sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2}$$

Outline

Chromosomic aberrations in cancer

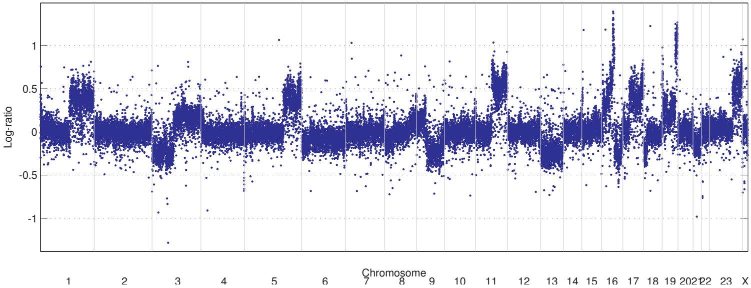




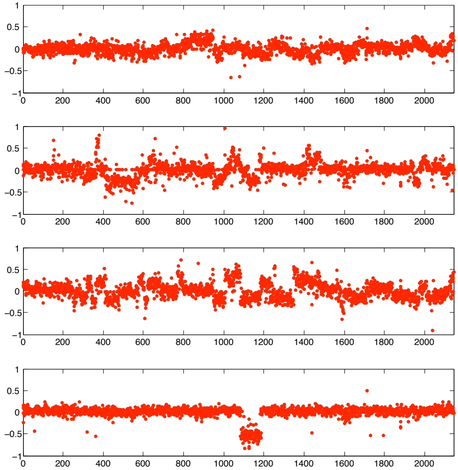
Comparative Genomic Hybridization (CGH)

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

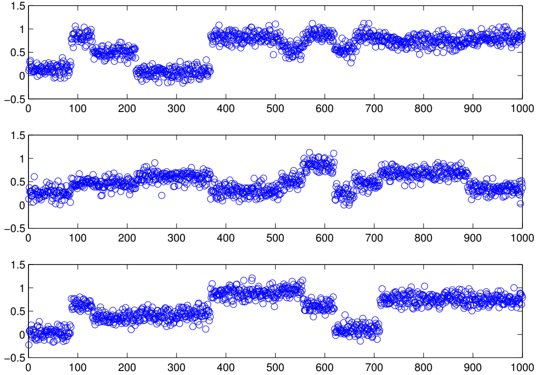


Can we detect frequent breakpoints?

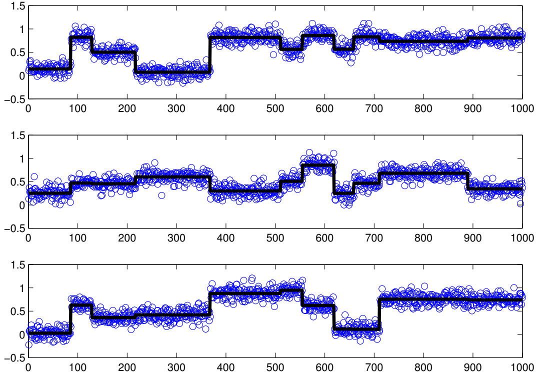


A collection of bladder tumour copy number profiles.

The problem

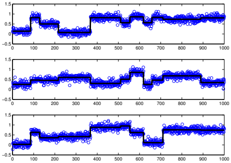


- Let $Y \in \mathbb{R}^{p \times n}$ the n signals of length p
- We want to find a piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ with at most k change-points.



- 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 $p = 10^6 \sim 10^9$...

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

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

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):

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

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 $A = (a_1, \ldots, a_{|A|})$ and $B = (b_1, \ldots, b_{|B|})$ in $[1, p-1]$, we can compute $\bar{X}_{\bullet,A}^\top \bar{X}_{\bullet,B}$ in $O(|A||B|)$ in time and memory
- For any $A = (a_1, \ldots, a_{|A|})$, 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 $C = \left(\bar{X}_{\bullet,A}^\top \bar{X}_{\bullet,A}\right)^{-1} R$ in $O(|A|n)$ in time and memory

Speed trial

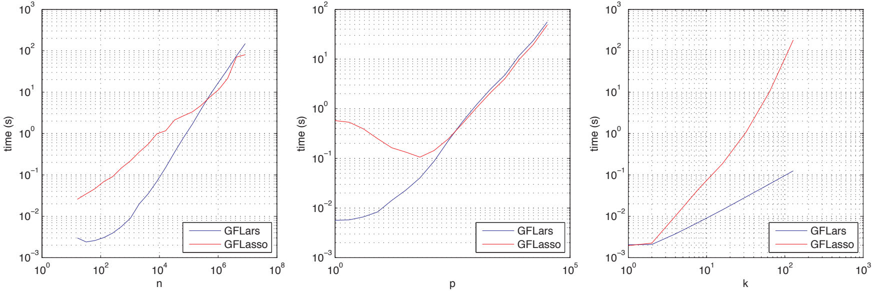
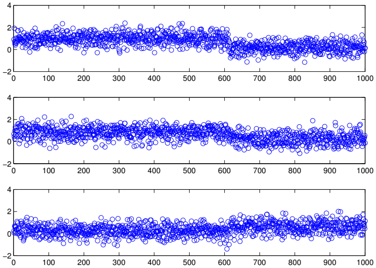


Figure 2: **Speed trials for group fused LARS (top row) and Lasso (bottom row).** *Left column:* varying n, with fixed $p = 10$ and $k = 10$; *center column:* varying p, with fixed $n = 1000$ and $k = 10$; *right column:* varying k, with fixed $n = 1000$ and $p = 10$. Figure axes are log-log. Results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

Consistency for a single change-point

Suppose a single change-point:

- at position $u = \alpha p$
- with increments $(\beta_i)_{i=1,\dots,n}$ 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 $\sigma^2 < \tilde{\sigma}_\alpha^2$ (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

Consistency for a single change-point

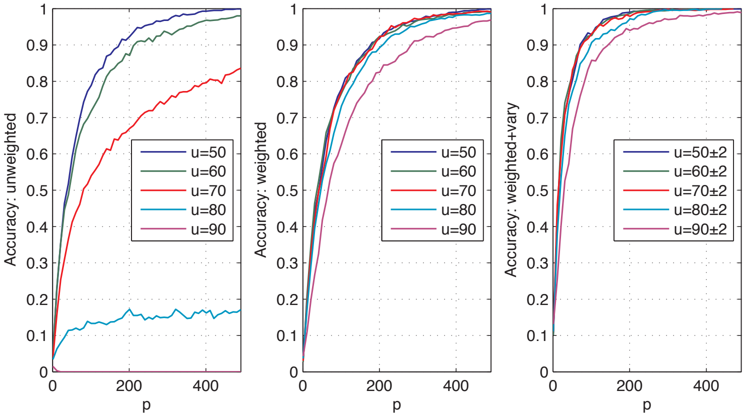


Figure 3: **Single change-point accuracy for the group fused Lasso.** Accuracy as a function of the number of profiles p when the change-point is placed in a variety of positions $u = 50$ to $u = 90$ (left and centre plots, resp. unweighted and weighted group fused Lasso), or: $u = 50 \pm 2$ to $u =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

Estimation of more change-points?

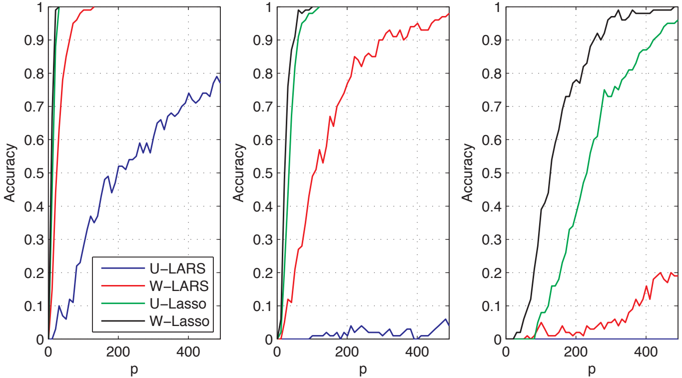
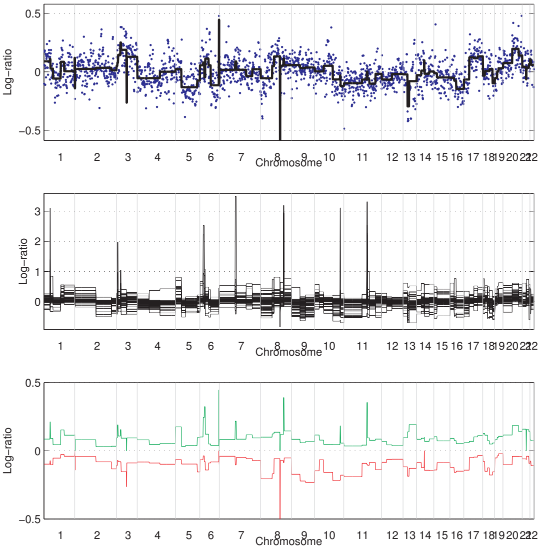


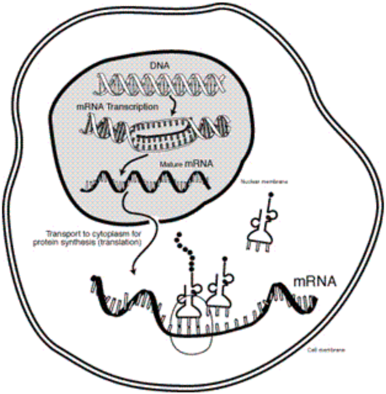
Figure 4: **Multiple change-point accuracy.** Accuracy as a function of the number of profiles p when change-points are placed at the nine positions $\{10, 20, \ldots, 90\}$ and the variance σ^2 of the centered Gaussian noise is either 0.05 (left), 0.2 (center) and 1 (right). The profile length is 100.

Application: detection of frequent abnormalities



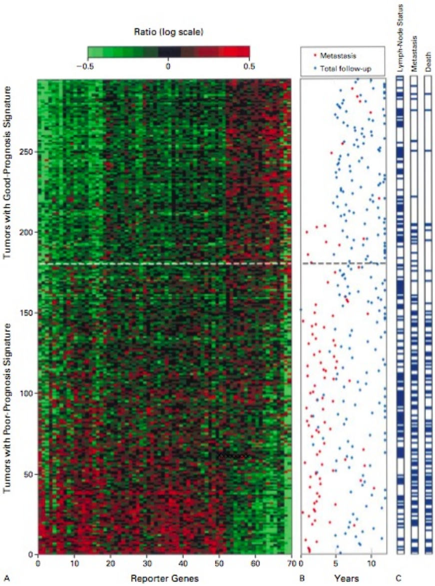
Outline

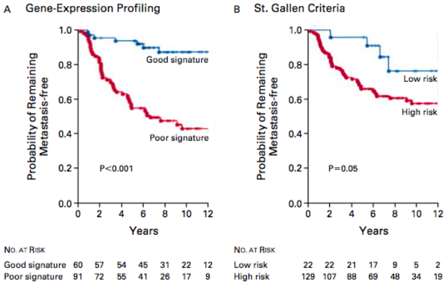
DNA $\rightarrow$ RNA $\rightarrow$ protein



- CGH shows the (static) DNA
- Cancer cells have also abnormal (dynamic) gene expression (= transcription)

Breast cancer prognosis

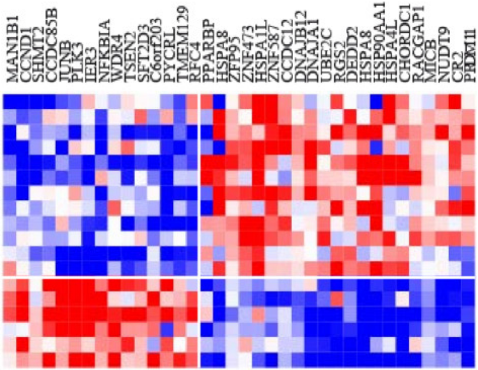




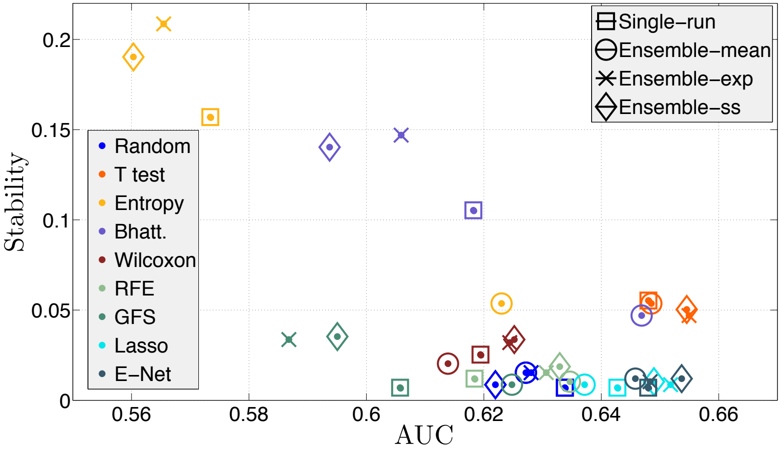
Gene selection, molecular signature

The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology

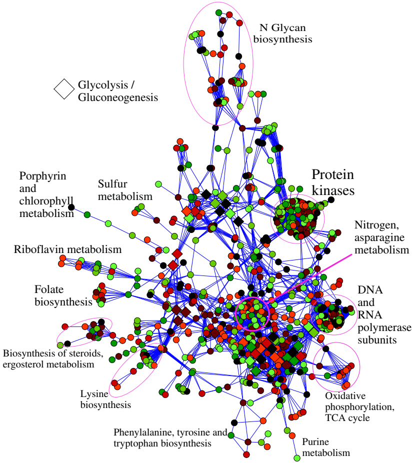


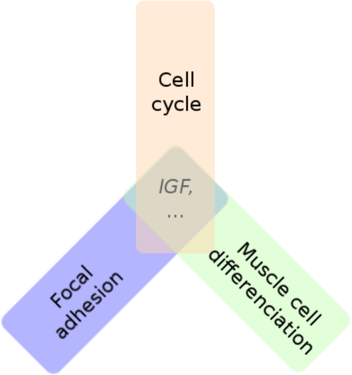
Lack of stability of signatures



Haury et al. (2011)

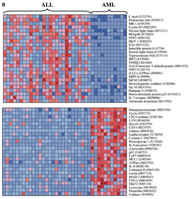
Gene networks, gene groups

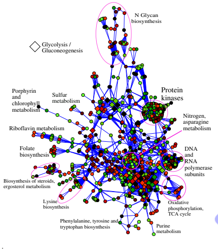


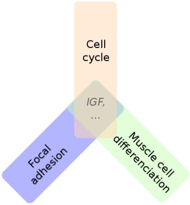


Structured feature selection

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- How to perform structured feature selection, such that selected genes
  - belong to only a few groups?
  - form a small number of connected components on the graph?



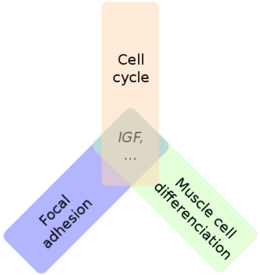




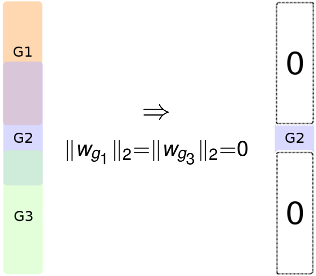
Group lasso with overlapping groups

Idea 1: shrink groups to zero (Jenatton et al., 2009)

- $\Omega_{group}(w) = \sum_g \|w_g\|_2$ sets groups to 0.
- One variable is selected $\Leftrightarrow$ all the groups to which it belongs are selected.



IGF selection $\Rightarrow$ selection of unwanted groups

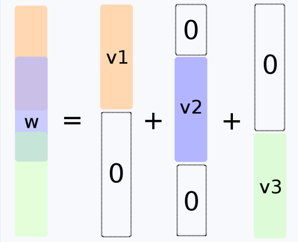


Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

Group lasso with overlapping groups

Idea 2: latent group Lasso (Jacob et al., 2009)

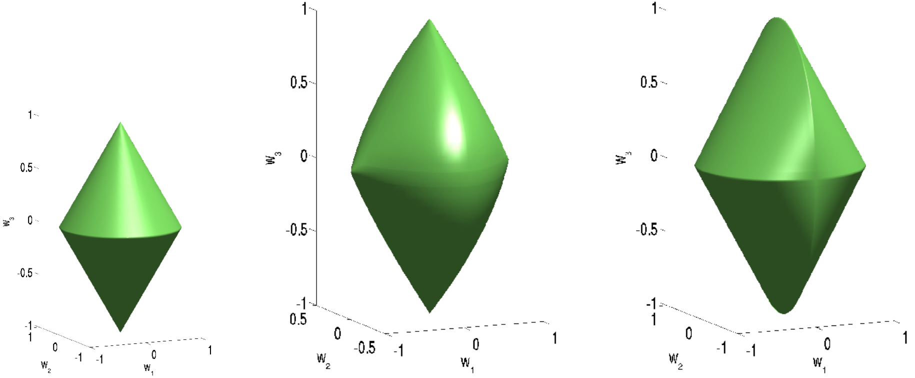
$$\Omega^{\mathcal{G}}_{\text{latent}}(w) \triangleq \begin{cases} \min_{v} \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Equivalent to group lasso when there is no overlap

Overlap and group unity balls



Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega^{\mathcal{G}}_{\text{latent}}(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where w_2 is represented as the vertical coordinate. Left: group-lasso ($\mathcal{G} = \{\{1,2\},\{3\}\}$), for comparison.

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

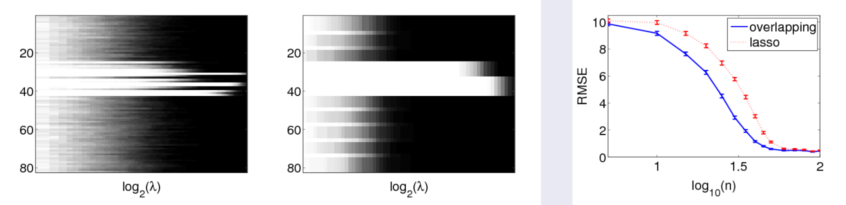
- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g\in\mathcal{G}}$ such that

$$\left\{g \in \mathcal{G} | \hat{v}_g \neq 0\right\} = \left\{g \in \mathcal{G} | \bar{v}_g \neq 0\right\}.$$

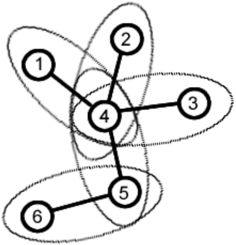
Synthetic data: overlapping groups

- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$.
- Support: union of $4th$ and $5th$ groups.
- Learn from 100 training points.



Frequency of selection of each variable with the lasso (left) and $\Omega^{\mathcal{G}}_{\text{latent}}(.)$ (middle), comparison of the RMSE of both methods (right).

Graph lasso



Two solutions

$$\Omega^{\mathcal{G}}_{\text{group}}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

$$\Omega^{\mathcal{G}}_{\text{latent}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,.$$

Preliminary results

Breast cancer data

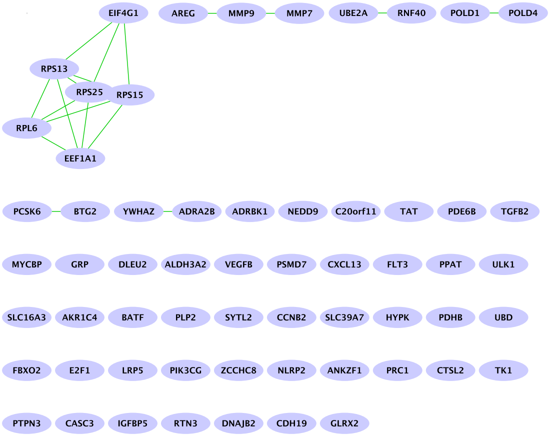
- Gene expression data for $8,141$ genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

METHOD	ℓ_1	$\Omega^{\mathcal{G}}_{\text{LATENT}}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
MEAN $\sharp$ PATH.	130	30

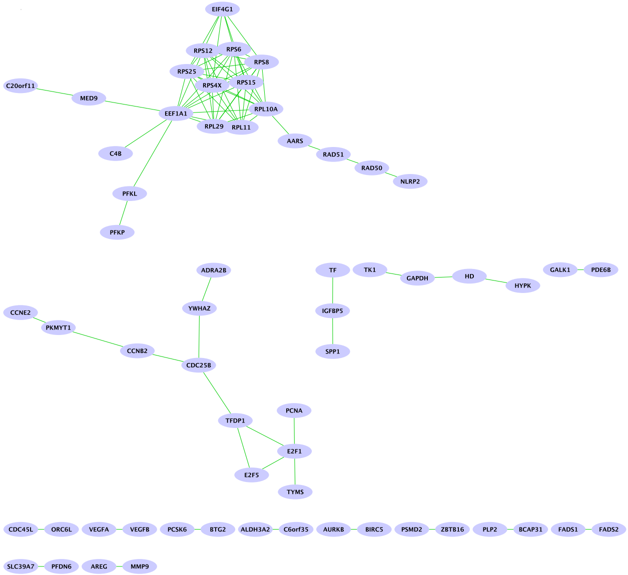
- Graph on the genes.

METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	1.03	1.30

Lasso signature



Graph Lasso signature



Outline

Conclusions

- Convex sparsity-inducing penalties are useful; efficient implementations + consistency results
- Penalty design as a way to incorporate prior knowledge

Kevin Bleakley (INRIA), Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA), Anne-Claire Haury (ParisTech)

AGENCE NATIONALE DE LA RECHERCHE ANR
FRANCE-BERKELEY FUND
FONDS FRANCE-BERKELEY
UNIVERSITY OF CALIFORNIA, BERKELEY | LE MINISTERE DES AFFAIRS ETRANGERES
European Research Council
erc