

Including Prior Knowledge in Machine Learning for Genomic Data

Jean-Philippe Vert

Jean-Philippe.Vert@mines-paristech.fr

Mines ParisTech / Curie Institute / Inserm

10th Annual International Workshop on Bioinformatics and Systems Biology, Kyoto, Japan, July 26-28, 2010.

- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers
- 3 Cancer prognosis from DNA copy number variations
- 4 Diagnosis and prognosis from gene expression data
- 5 Conclusion

- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers
- 3 Cancer prognosis from DNA copy number variations
- 4 Diagnosis and prognosis from gene expression data
- 5 Conclusion

Bienvenue à Paris!

The screenshot shows a web browser window with the title "Bioinformatics and Computational Systems Biology in Paris". The address bar contains the URL "http://sites.google.com/site/sysbioinparis/home". The browser interface includes navigation buttons (back, forward, refresh, home) and a search bar with the Google logo. Below the browser window, the website content is displayed. The main heading is "Bioinformatics and Computational Systems Biology in Paris" in a large, bold, white font on a dark background. To the right of the heading is a search box with the text "Search this site". Below the heading is a horizontal orange bar. On the left side, there is a navigation menu with the following items: "Welcome", "Teams", "Partners", "Events", and "Contact". The main content area features a "Welcome" heading followed by a paragraph: "The 'Bioinformatics and Computational Systems Biology in Paris' initiative gathers several research laboratories in Paris area which participate in an international effort to train and exchange students and young researchers." At the bottom of the page, there is a footer with the text: "Connexion Activités récentes sur le site Conditions Signaler un abus Imprimer la page | Powered by Google Sites". A small "Transl" button is visible in the bottom right corner of the browser window.

Bioinformatics and Computational Systems Biology in Paris

http://sites.google.com/site/sysbioinparis/home

Search this site

Bioinformatics and Computational Systems Biology in Paris

Welcome

- Teams
- Partners
- Events
- Contact

Welcome

The "Bioinformatics and Computational Systems Biology in Paris" initiative gathers several research laboratories in Paris area which participate in an international effort to train and exchange students and young researchers.

Connexion Activités récentes sur le site Conditions Signaler un abus Imprimer la page | Powered by Google Sites

Transl

Transfert des données depuis translate.googleapis.com...

Bioinformatics and Computational Systems Biology in Paris: Teams

http://sites.google.com/site/sysbioinparis/teams

Les plus visités ▾ Débuter avec Firefox À la une ↗

http://sug...0/cfp.txt x http://sug...0/cfp.txt x STAFAV x 2010 5 Day Worksh... x Bioinformatics and ... x

Bioinformatics and Computational Systems Biology in Paris

Search this site

Welcome
Teams
Partners
Events
Contact

Teams

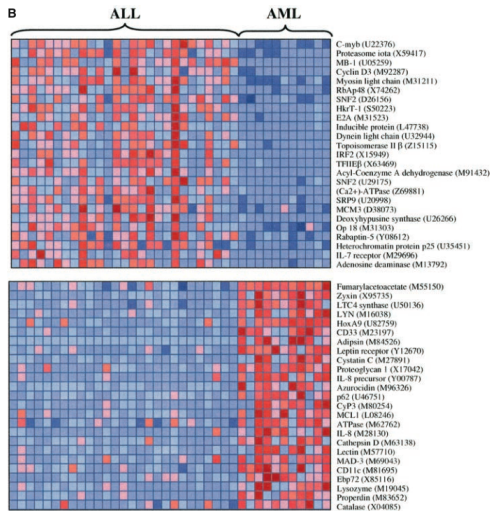
- [Centre for Computational Biology at Mines ParisTech](#) (contact: [Jean-Philippe Vert](#))
 - Keywords: Statistics, machine learning, systems biology, chemoinformatics, cancer genomics
- [Statistics and Genome team at AgroParisTech](#) (contact: [Stéphane Robin](#))
 - Keywords: Statistics, systems biology, genomics
- [Biocomputing and Structure team at Ecole Polytechnique](#) (contact: [Thomas Simonson](#))
 - Keywords: structural biology, computational protein design, molecular dynamics

[Connexion](#) [Activités récentes sur le site](#) [Conditions](#) [Signaler un abus](#) [Imprimer la page](#) | Powered by [Google Sites](#)

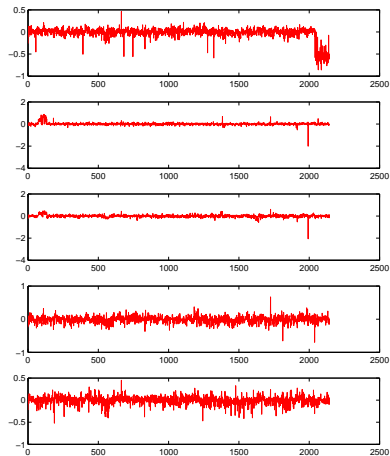
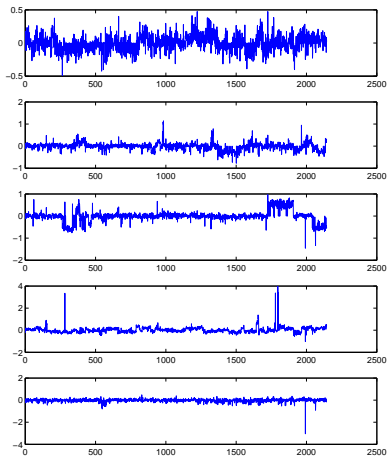
Terminé

- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers**
- 3 Cancer prognosis from DNA copy number variations
- 4 Diagnosis and prognosis from gene expression data
- 5 Conclusion

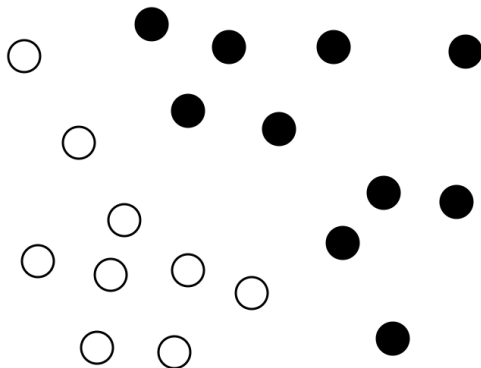
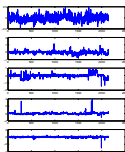
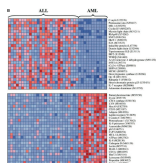
Cancer diagnosis



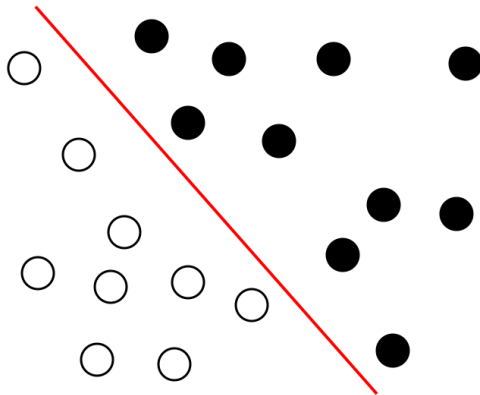
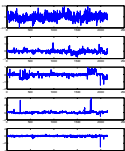
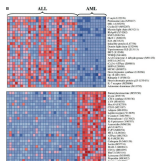
Cancer prognosis



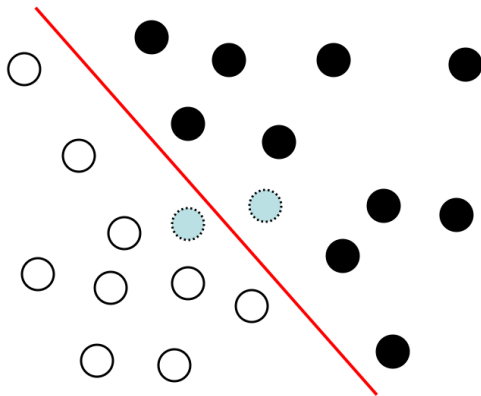
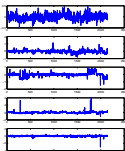
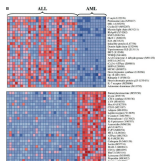
Pattern recognition, *aka* supervised classification



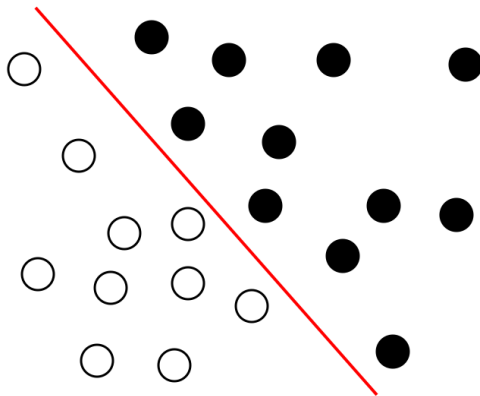
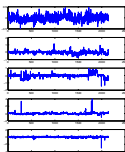
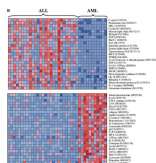
Pattern recognition, *aka* supervised classification

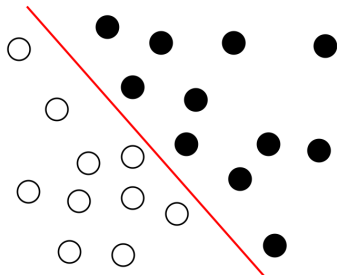


Pattern recognition, *aka* supervised classification



Pattern recognition, *aka* supervised classification





Challenges

- Few samples
- High dimension
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models

Shrinkage estimators

- Define a large family of "candidate classifiers", e.g., linear predictors $f_{\beta}(x) = \beta^{\top} x$
- For any candidate classifier f_{β} , quantify how "good" it 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 minimum 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 any candidate classifier f_{β} , quantify how "good" it 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 minimum 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 any candidate classifier f_{β} , quantify how "good" it is on the training set with some **empirical risk**, e.g.:

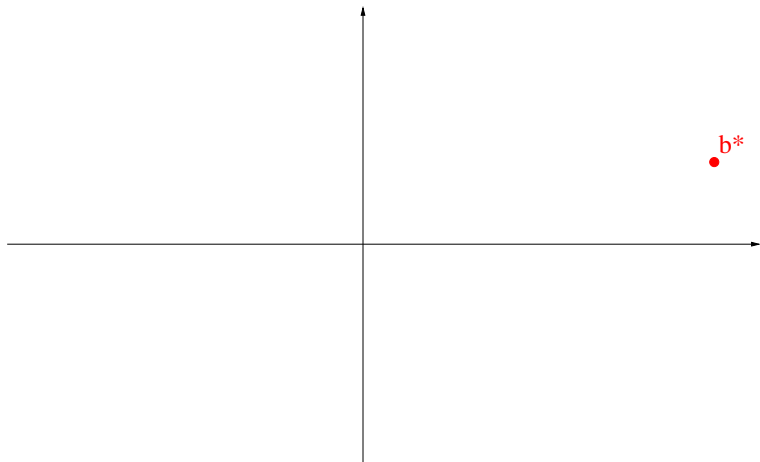
$$R(\beta) = \frac{1}{n} \sum_{i=1}^n l(f_{\beta}(x_i), y_i).$$

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

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

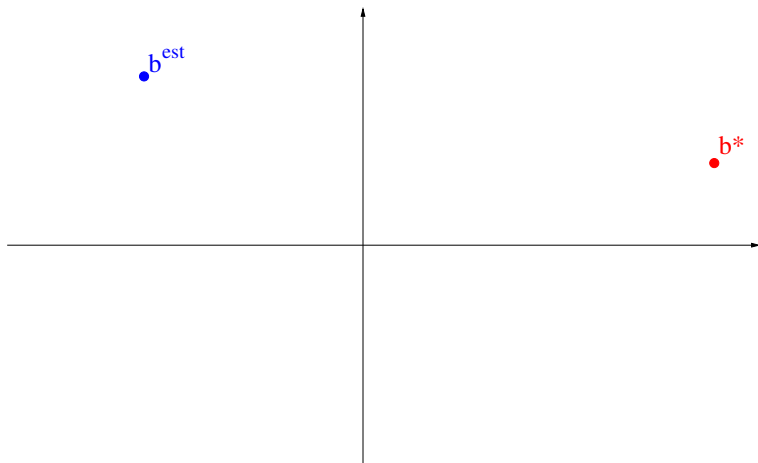
Why shrinkage classifiers?

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



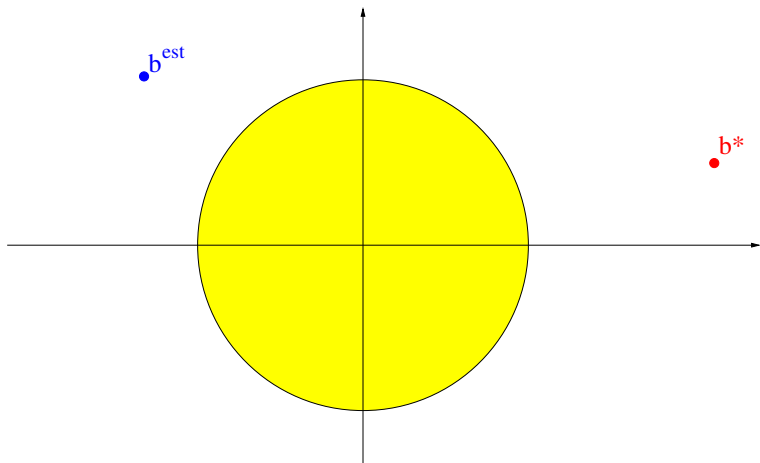
Why shrinkage classifiers?

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



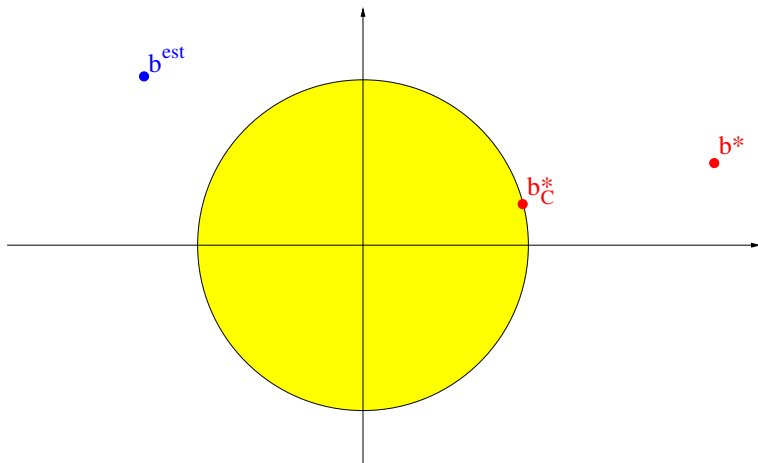
Why shrinkage classifiers?

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



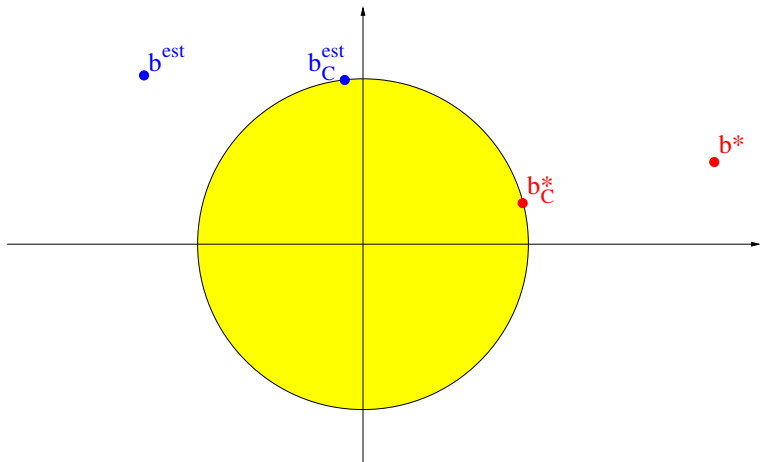
Why shrinkage classifiers?

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



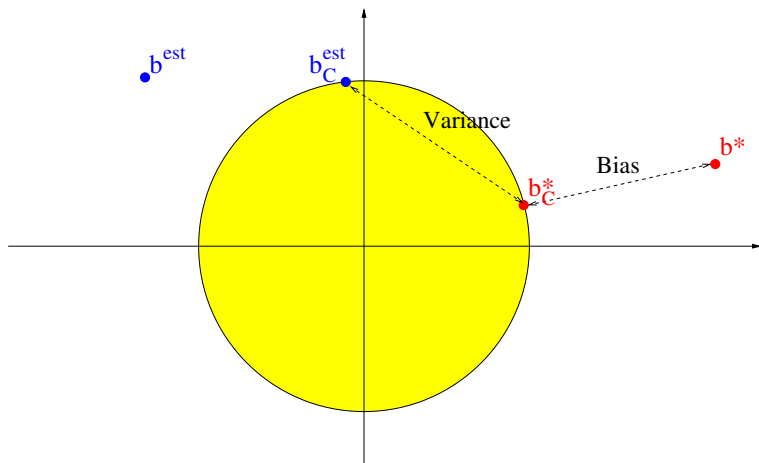
Why shrinkage classifiers?

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

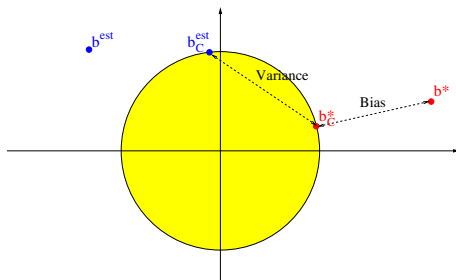


Why shrinkage classifiers?

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



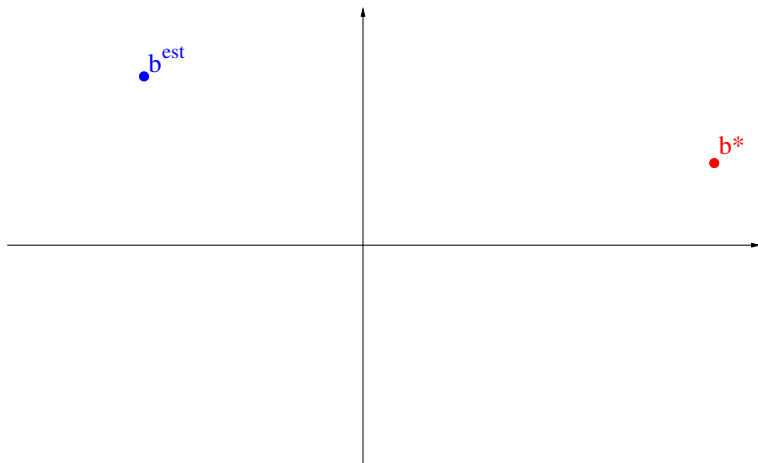
Why shrinkage classifiers?



- "Increases bias and decreases variance"
- Common choices are
 - $\Omega(\beta) = \sum_{i=1}^p \beta_i^2$ (ridge regression, SVM, ...)
 - $\Omega(\beta) = \sum_{i=1}^p |\beta_i|$ (lasso, boosting, ...)

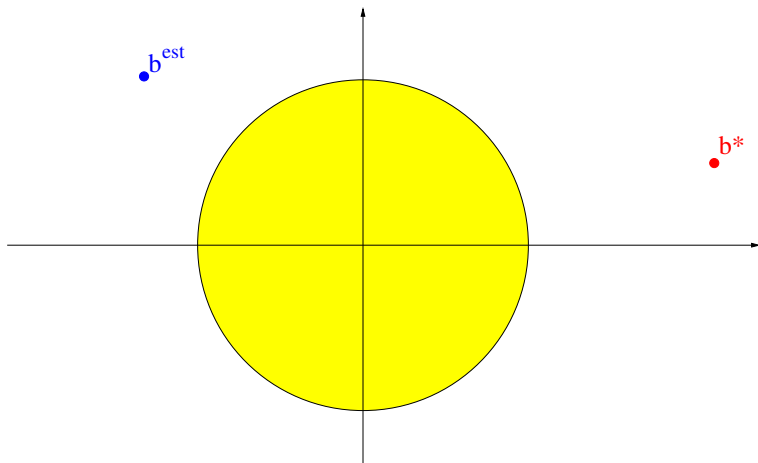
Including prior knowledge in the penalty?

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



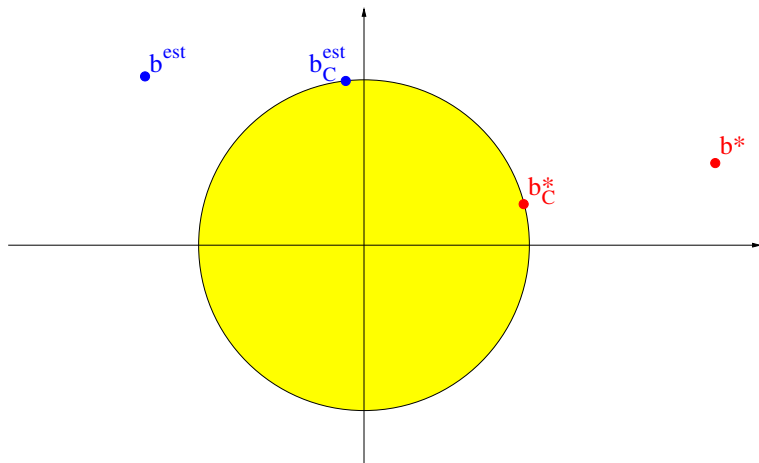
Including prior knowledge in the penalty?

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



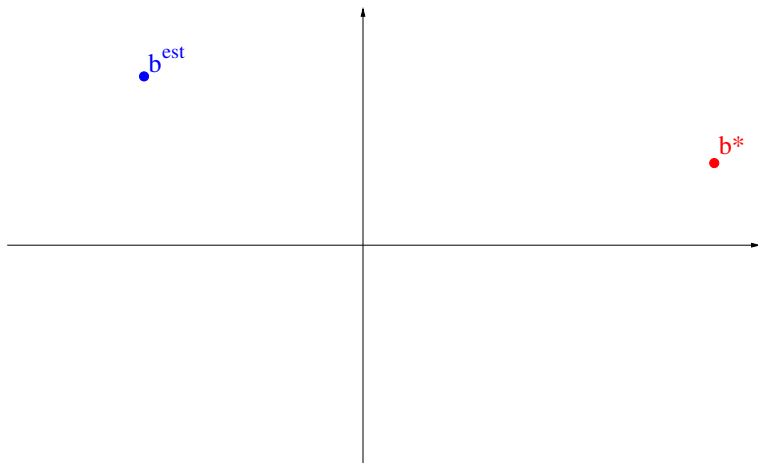
Including prior knowledge in the penalty?

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



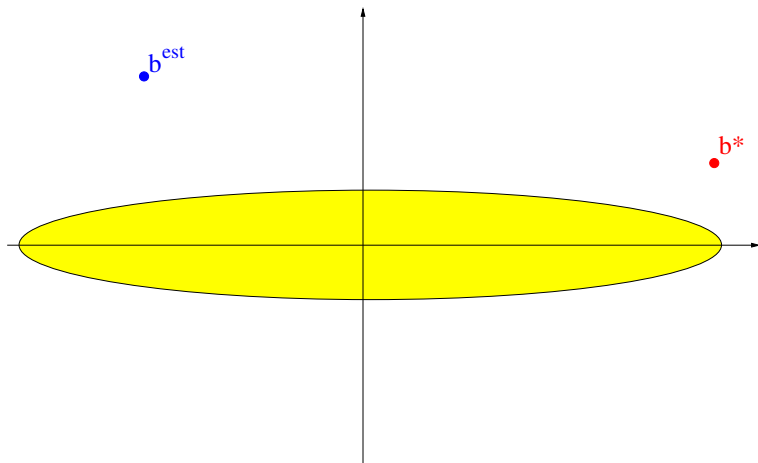
Including prior knowledge in the penalty?

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



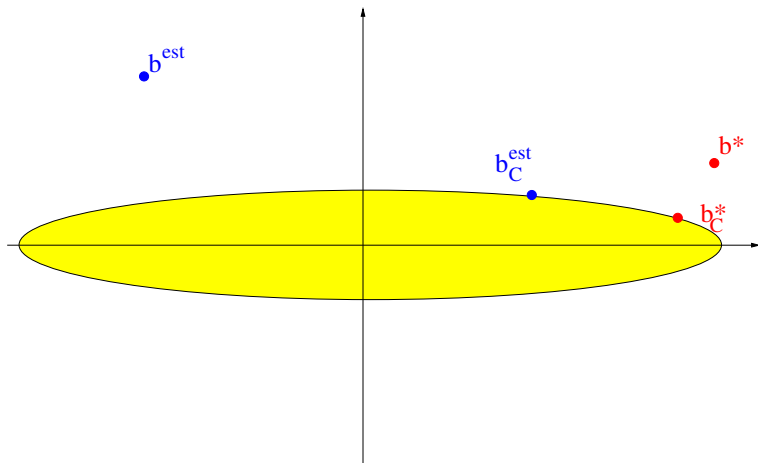
Including prior knowledge in the penalty?

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



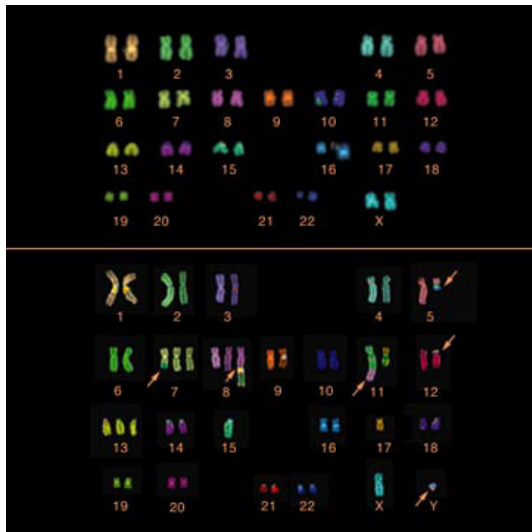
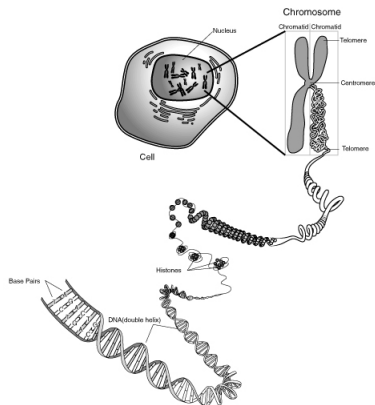
Including prior knowledge in the penalty?

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



- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers
- 3 Cancer prognosis from DNA copy number variations**
- 4 Diagnosis and prognosis from gene expression data
- 5 Conclusion

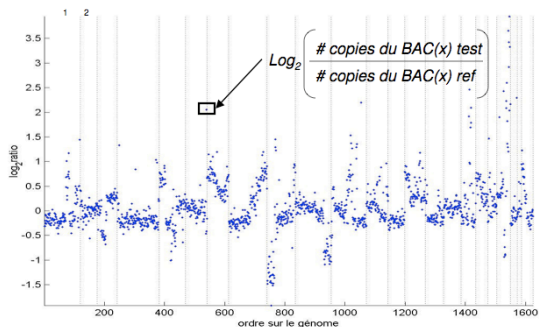
Chromosomal aberrations in cancer



Comparative Genomic Hybridization (CGH)

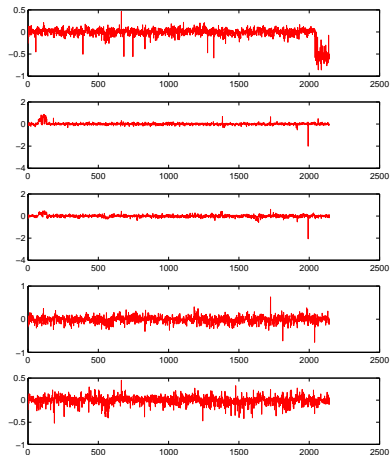
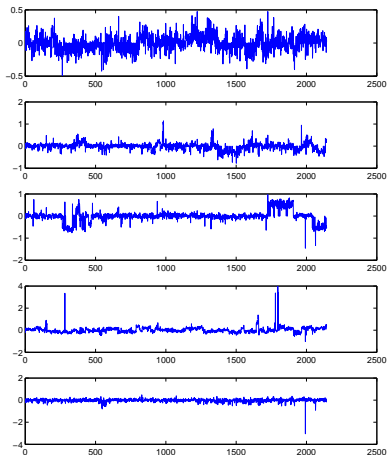
Motivation

- Comparative genomic hybridization (CGH) data measure the **DNA copy number** along the genome
- Very useful, in particular in cancer research
- Can we **classify CGH arrays** for diagnosis or prognosis purpose?



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

Aggressive vs non-aggressive melanoma



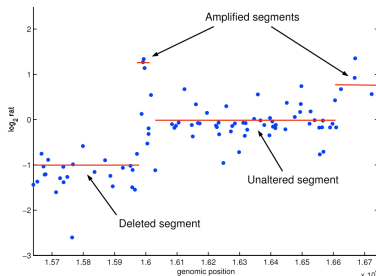
CGH array classification

Prior knowledge

- For a CGH profile $x \in \mathbb{R}^p$, we focus on linear classifiers, i.e., the sign of :

$$f_{\beta}(x) = \beta^{\top} x .$$

- We expect β to be
 - **sparse** : not all positions should be discriminative
 - **piecewise constant** : within a selected region, all probes should contribute equally



Promoting sparsity with the ℓ_1 penalty

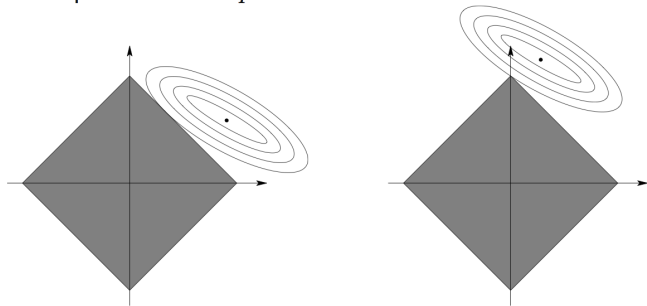
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

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

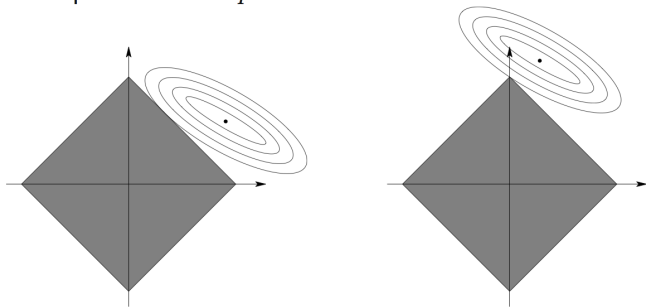
The variable fusion penalty (Land and Friedman, 1996)

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.

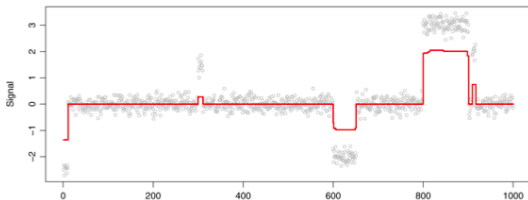
Geometric interpretation with $p = 2$



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)

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

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)

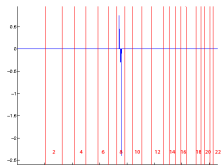
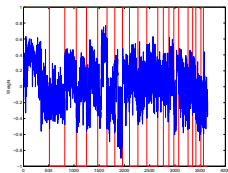
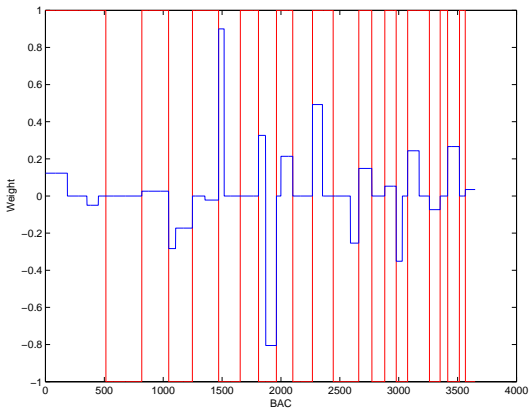
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, \beta^\top x_i) + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|.$$

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

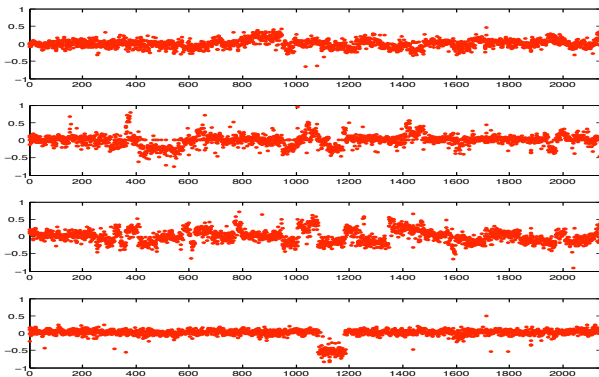
Implementation

- 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

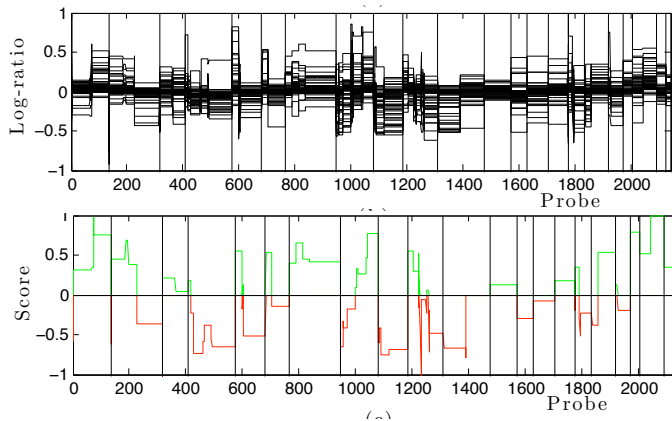


Extension: joint segmentation of many profiles



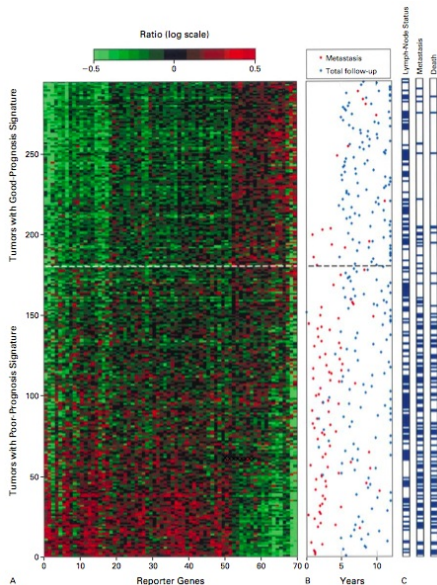
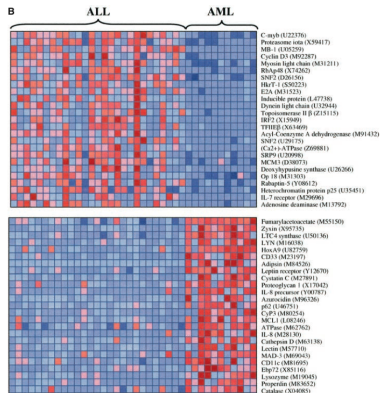
Fused group Lasso signal approximator

$$\min_{\beta \in \mathbb{R}^{n \times p}} \|Y - \beta\|^2 + \lambda \sum_{i=1}^{p-1} \|\beta_{i+1} - \beta_i\|$$



- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers
- 3 Cancer prognosis from DNA copy number variations
- 4 Diagnosis and prognosis from gene expression data**
- 5 Conclusion

Molecular diagnosis / prognosis / theragnosis



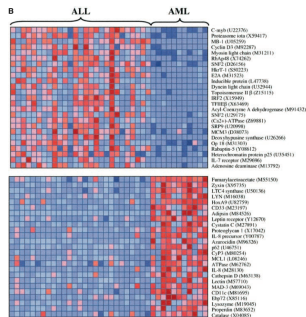
The idea

- We look for a **limited set** of genes that are sufficient for prediction.
- Equivalently, the linear classifier will be **sparse**

Why?

- **Bet on sparsity**: we believe the "true" model is sparse.
- **Interpretation**: we will get a biological interpretation more easily by looking at the selected genes.
- **Statistics**: this is one way to constrain the solution and reduce the complexity to allow learning.

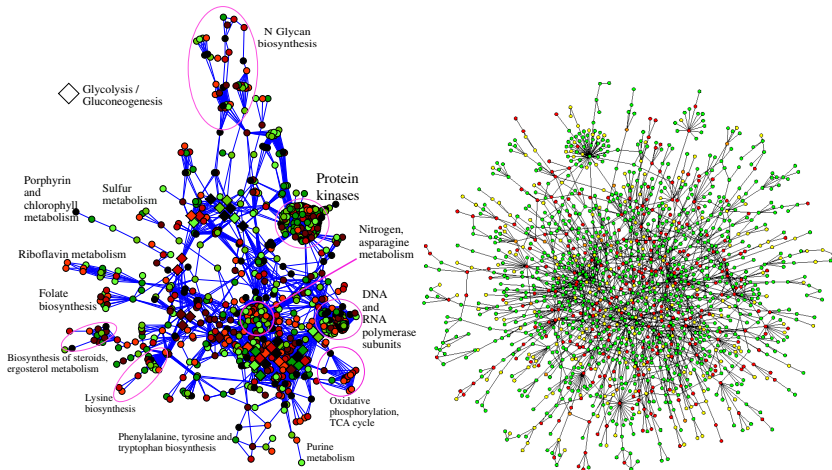
But...



Challenging the idea of gene signature

- We often observe little **stability** in the genes selected...
- Is gene selection the most **biologically relevant** hypothesis?
- What about thinking instead of "**pathways**" or "**modules**" **signatures**?

Gene networks



Prior hypothesis

Genes near each other on the graph should have **similar weights**.

Two solutions (Rapaport et al., 2007, 2008)

$$\Omega_{\text{spectral}}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2,$$

$$\Omega_{\text{graphfusion}}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i|.$$

Prior hypothesis

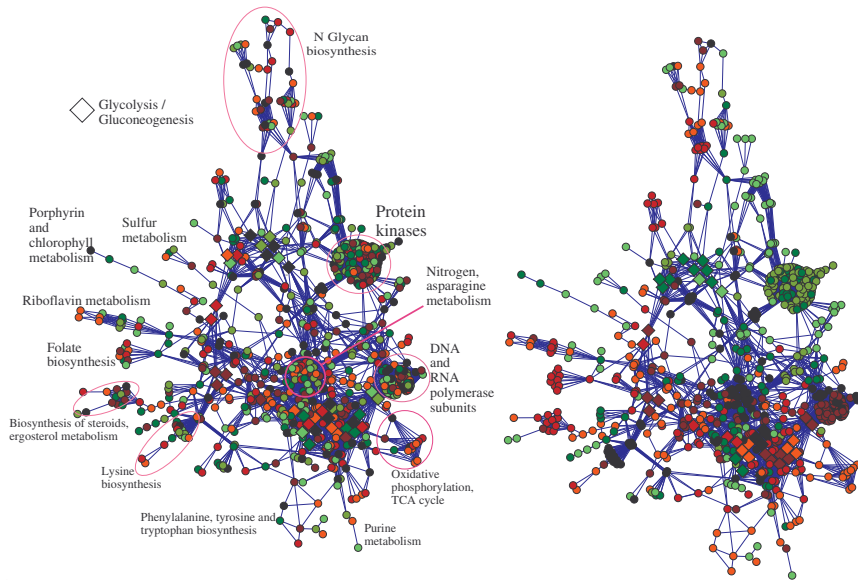
Genes near each other on the graph should have **similar weights**.

Two solutions (Rapaport et al., 2007, 2008)

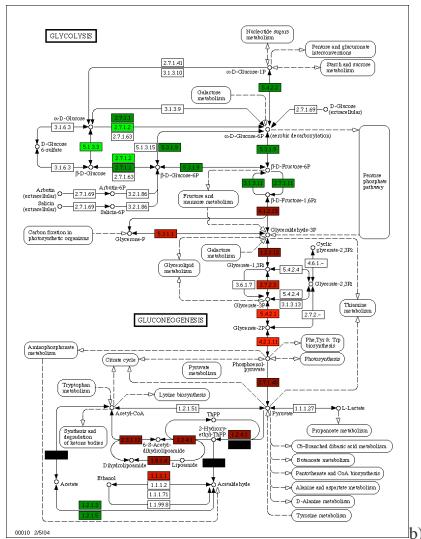
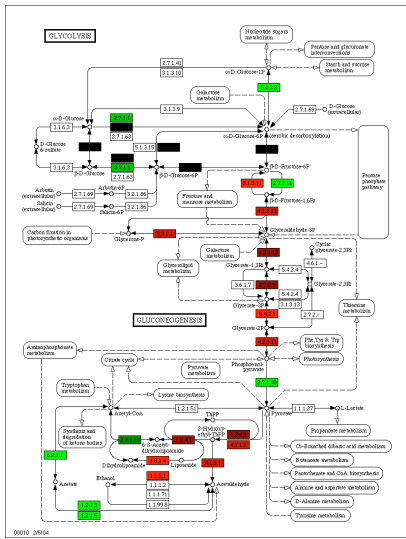
$$\Omega_{\text{spectral}}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2,$$

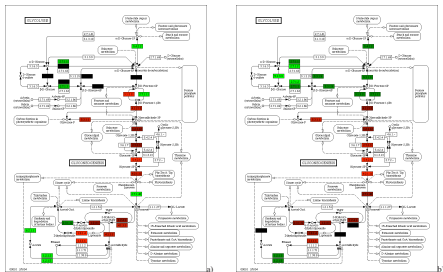
$$\Omega_{\text{graphfusion}}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i|.$$

Classifiers



Classifiers





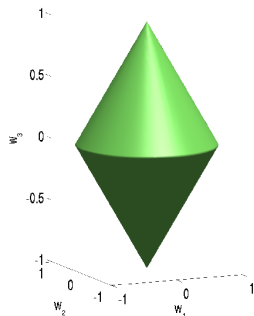
- We are happy to see pathways appear.
- However, in some cases, connected genes should have "opposite" weights (inhibition, pathway branching, etc...)
- **How to capture pathways without constraints on the weight similarities?**

Selecting pre-defined groups of variables

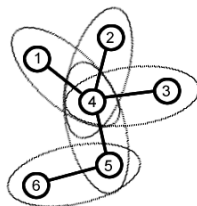
Group lasso (Yuan & Lin, 2006)

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

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



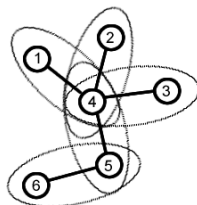
$$\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2$$



- **Hypothesis:** selected genes should form connected components on the graph
- Two solutions (Jacob et al., 2009):

$$\Omega_{group}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2},$$

$$\Omega_{overlap}(\beta) = \sup_{\alpha \in \mathbb{R}^p: \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^T \beta.$$

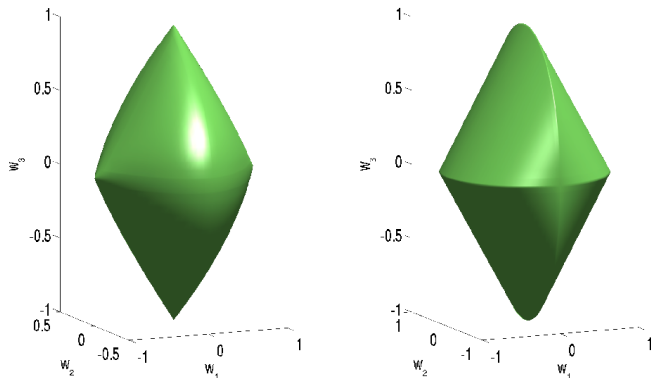


- **Hypothesis:** selected genes should form connected components on the graph
- Two solutions (Jacob et al., 2009):

$$\Omega_{group}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2},$$

$$\Omega_{overlap}(\beta) = \sup_{\alpha \in \mathbb{R}^p: \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta.$$

Overlap and group unity balls



Balls for $\Omega_{\text{group}}^{\mathcal{G}}(\cdot)$ (middle) and $\Omega_{\text{overlap}}^{\mathcal{G}}(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$ where w_2 is represented as the vertical coordinate.

Summary: Graph lasso vs kernel

- Graph lasso:

$$\Omega_{\text{graph lasso}}(\mathbf{w}) = \sum_{i \sim j} \sqrt{w_i^2 + w_j^2}.$$

constrains the **sparsity**, not the values

- Graph kernel

$$\Omega_{\text{graph kernel}}(\mathbf{w}) = \sum_{i \sim j} (w_i - w_j)^2.$$

constrains the values (**smoothness**), not the sparsity

Breast cancer data

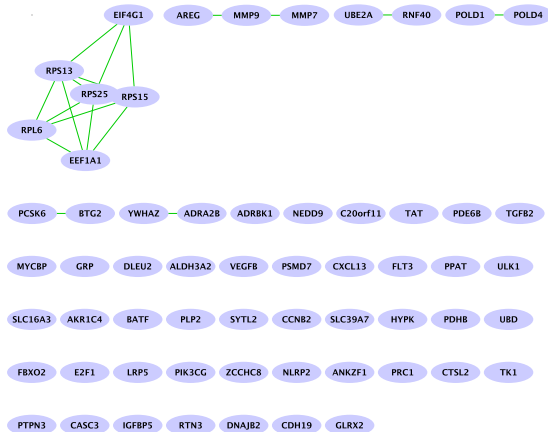
- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

METHOD	l_1	$\Omega_{\text{OVERLAP}}^G(\cdot)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
MEAN \ddagger PATH.	130	30

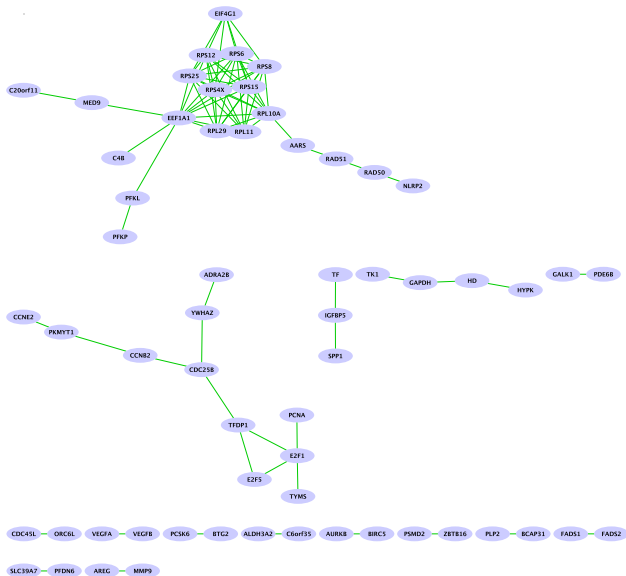
- Graph on the genes.

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

Lasso signature



Graph Lasso signature



- 1 Bioinformatics and Computational Systems Biology in Paris
- 2 Shrinkage classifiers
- 3 Cancer prognosis from DNA copy number variations
- 4 Diagnosis and prognosis from gene expression data
- 5 Conclusion**

- Modern machine learning methods for regression / classification lend themselves well to the **integration of prior knowledge** in the penalization / regularization function.
- Several **computationally efficient** approaches (structured LASSO, kernels...)
- Tight collaborations with domain experts can help develop specific learning machines for specific data
- Natural extensions for **data integration**

People I need to thank



Franck Rapaport (MSKCC), Emmanuel Barillot, Andrei Zynoviev Kevin Bleakley, Anne-Claire Haury (Institut Curie / ParisTech), Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA)