Including prior knowledge in shrinkage classifiers for genomic data

Jean-Philippe Vert
Jean-Philippe.Vert@mines-paristech.fr

Mines ParisTech / Institut Curie / Inserm

Institute for Statistical Mathematics, Tokyo, August 5, 2009.

Outline

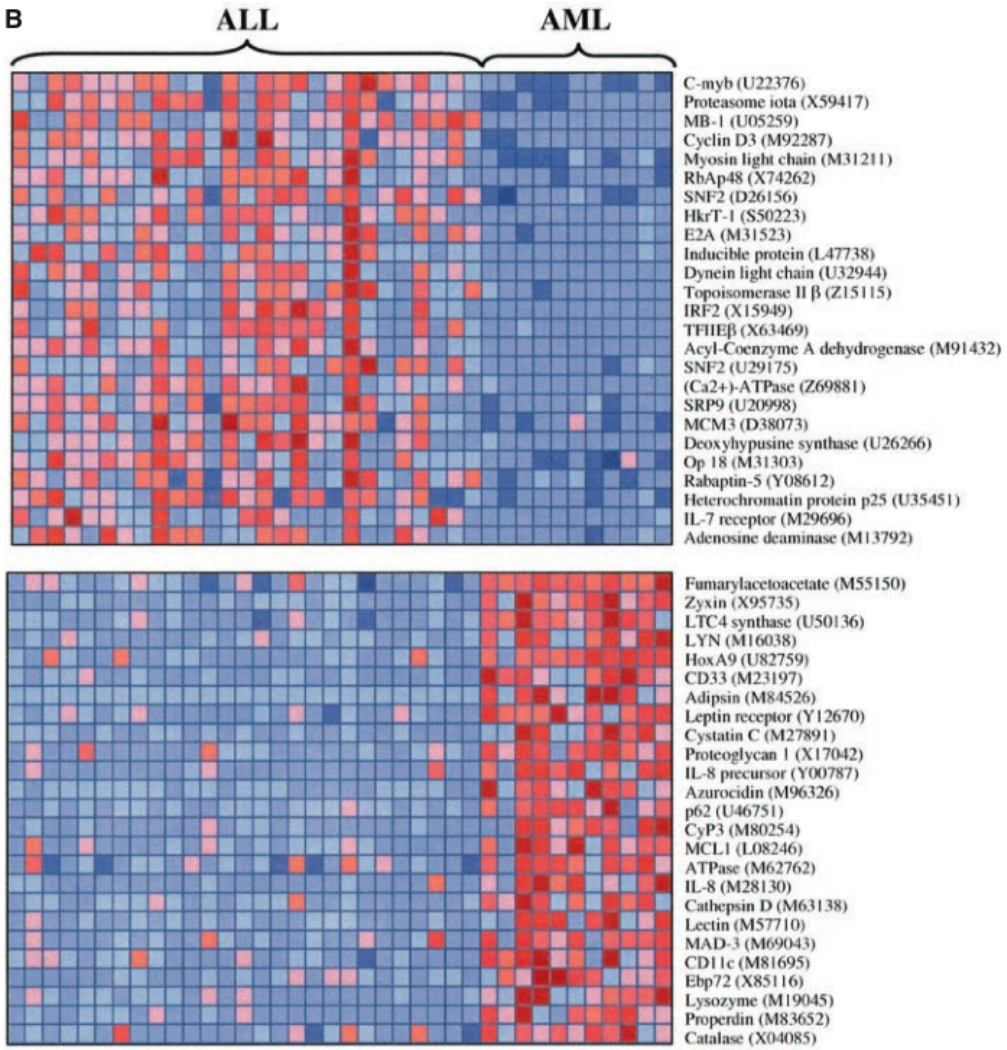
Outline

Outline

Outline

Outline

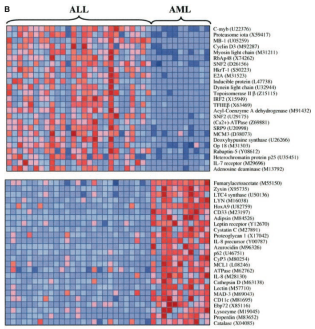
Motivation

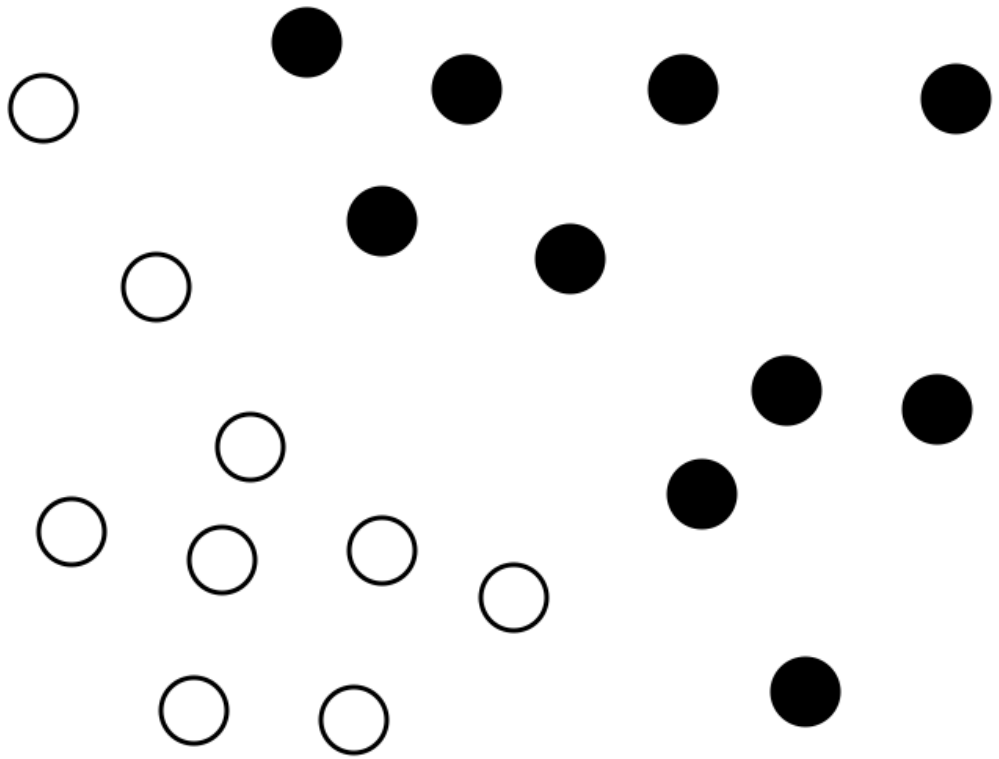


Goal

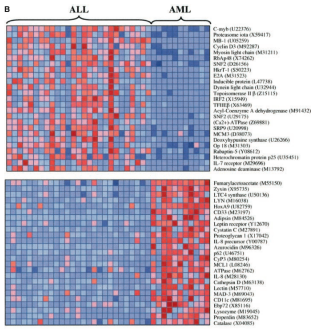
- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes

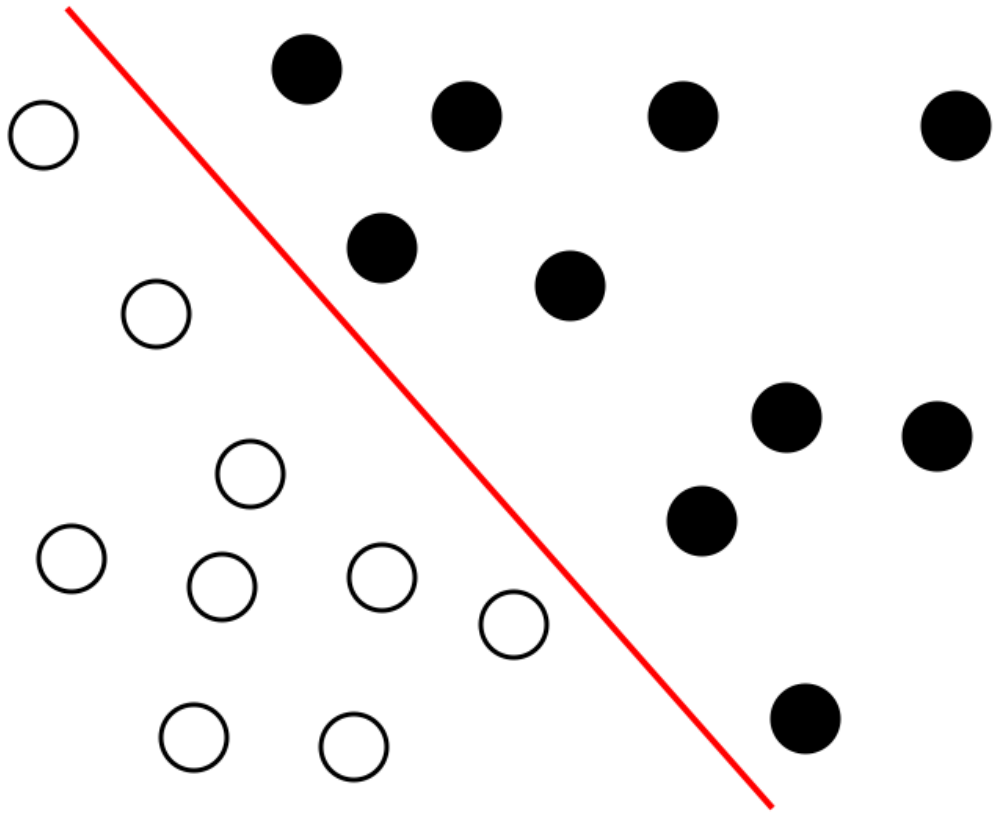
Pattern recognition, *aka* supervised classification



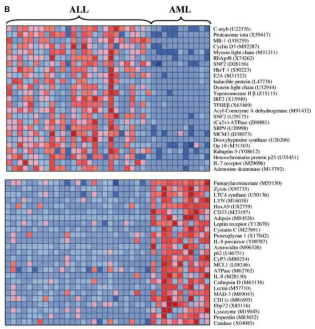


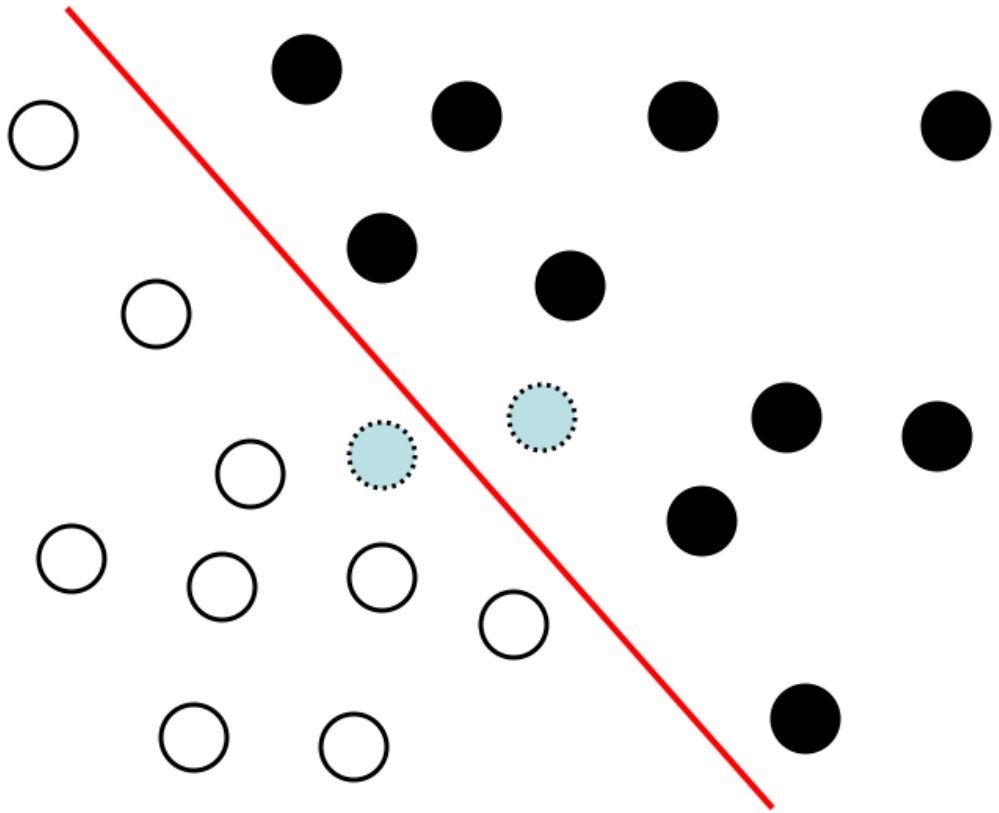
Pattern recognition, *aka* supervised classification



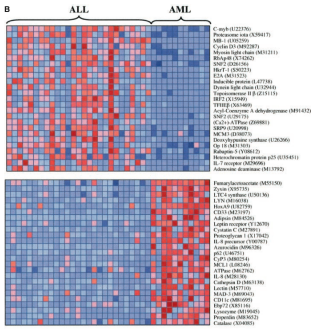


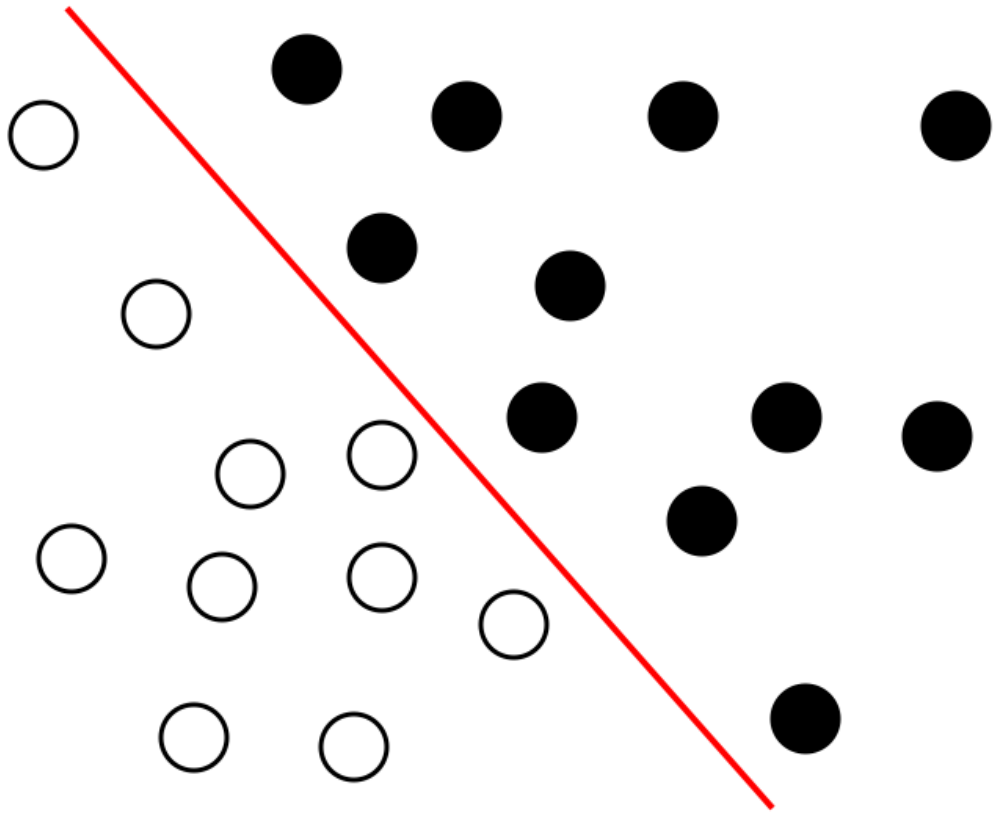
Pattern recognition, *aka* supervised classification



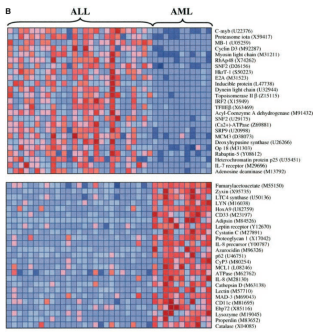


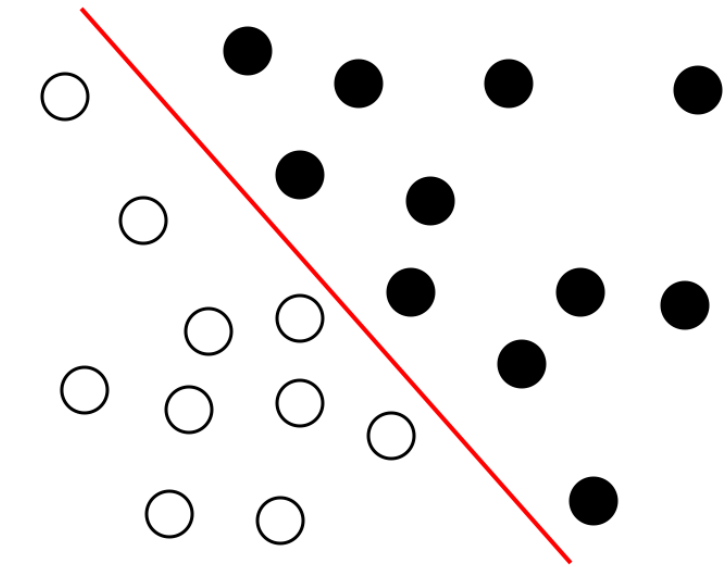
Pattern recognition, *aka* supervised classification





Pattern recognition, *aka* supervised classification





Challenges

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

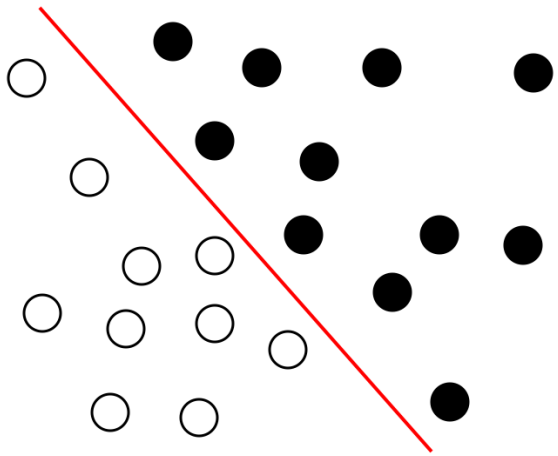
Linear classifiers

The model

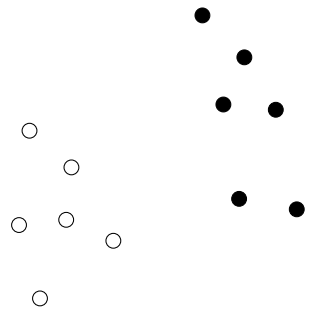
- Each sample is represented by a vector $x = (x_1, \ldots, x_p)$
- Goal: estimate a linear function:

$$f_\beta(x) = \sum_{i=1}^{p} \beta_i x_i + \beta_0 \ .$$

- Interpretability: the weight β_i quantifies the influence of feature i (but...)



Linear classifiers



Training the model

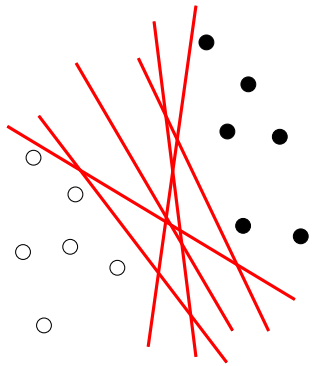
- Minimize an empirical risk on the training samples:

$$\min_{\beta \in \mathbb{R}^{p+1}} R_{emp}(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) \,,$$

- ... subject to some constraint on β, e.g.:

$$\Omega(\beta) \leq C \,.$$

Linear classifiers



Training the model

- Minimize an empirical risk on the training samples:

$$\min_{\beta \in \mathbb{R}^{p+1}} R_{emp}(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) \,,$$

- ... subject to some constraint on β, e.g.:

$$\Omega(\beta) \leq C \,.$$

Example : Norm Constraints

The approach

A common method in statistics to learn with few samples in high dimension is to constrain the Euclidean norm of β

$$\Omega_{ridge}(\beta) = \| \beta \|_2^2 = \sum_{i=1}^{p} \beta_i^2 \, ,$$

(ridge regression, support vector machines...)

Pros

- Good performance in classification

Cons

- Limited interpretation (small weights)
- No prior biological knowledge

Example : Feature Selection

The approach

Constrain most weights to be 0, i.e., select a few genes whose expression are sufficient for classification.

$$\Omega_{\text{Best subset selection}}(\beta) = \| \beta \|_0 = \sum_{i=1}^{p} 1(\beta_i > 0) \, .$$

This is usually a NP-hard problem, many greedy variants have been proposed (filter methods, wrapper methods)

Pros

- Good performance
- Biomarker selection
- Interpretability

Cons

- NP-hard
- Gene selection not robust
- No use of prior knowledge

Example : Sparsity inducing convex priors

The approach

Constrain most weights to be 0 through a convex non-differentiable penalty:

$$\Omega_{\text{LASSO}}(\beta) = \| \beta \|_1 = \sum_{i=1}^{p} | \beta_i | \ .$$

- Several variants exist, e.g., elastic net penalty ($\| \beta \|_1 + \| \beta \|_2$), ...)

Pros

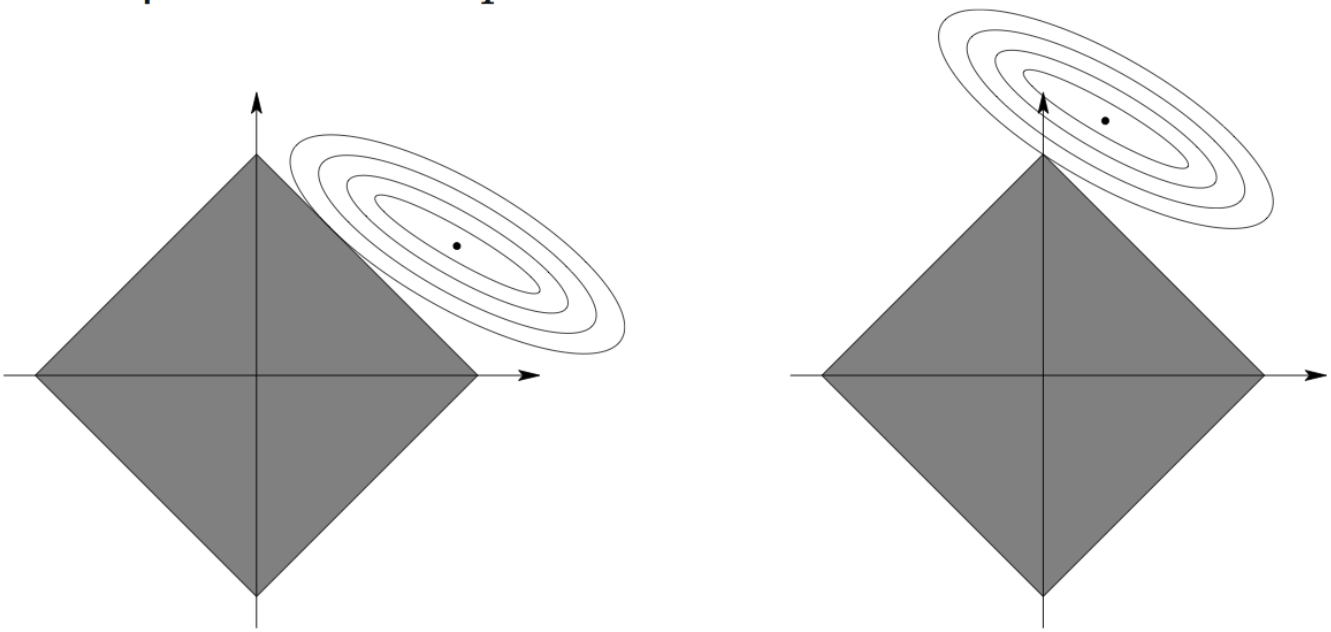
- Good performance
- Biomarker selection
- Interpretability

Cons

- Gene selection not robust
- No use of prior knowledge

Why LASSO leads to sparse solutions

Geometric interpretation with $p = 2$



Incorporating prior knowledge

The idea

- If we have a specific prior knowledge about the "correct" weights, it can be included in Ω in the contraint:

$$\text{Minimize } R_{emp}(\beta) \text{ subject to } \Omega(\beta) \leq C\,.$$

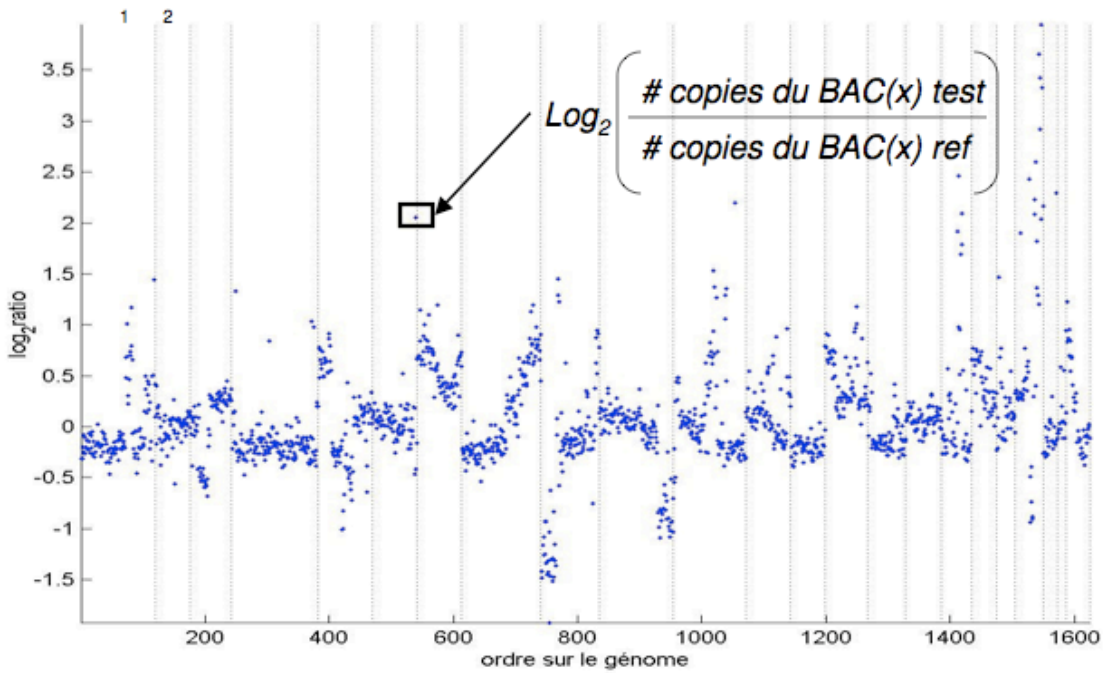
- If we design a convex function Ω, then the algorithm boils down to a convex optimization problem (usually easy to solve).
- Similar to priors in Bayesian statistics

Outline

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

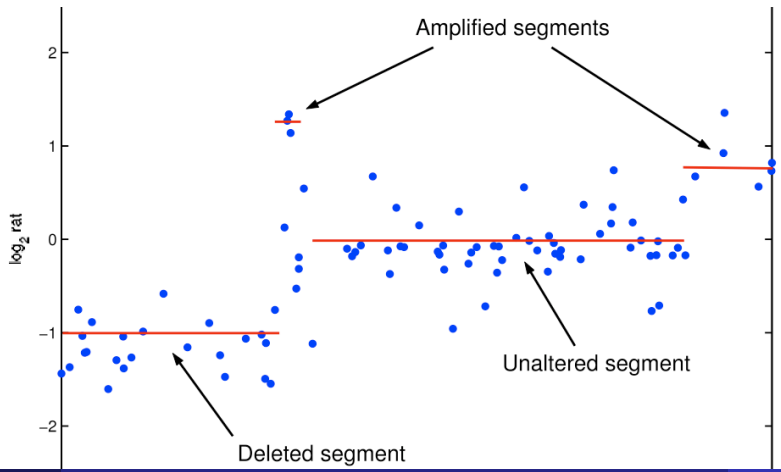
Classification of array CGH

Prior knowledge

- Let **x** be a CGH profile
- We focus on linear classifiers, i.e., the sign of :

$$f(\mathbf{x}) = \mathbf{x}^\top \beta \,.$$

- We expect β to be
  - sparse : only a few positions should be discriminative
  - piecewise constant : within a region, all probes should contribute equally



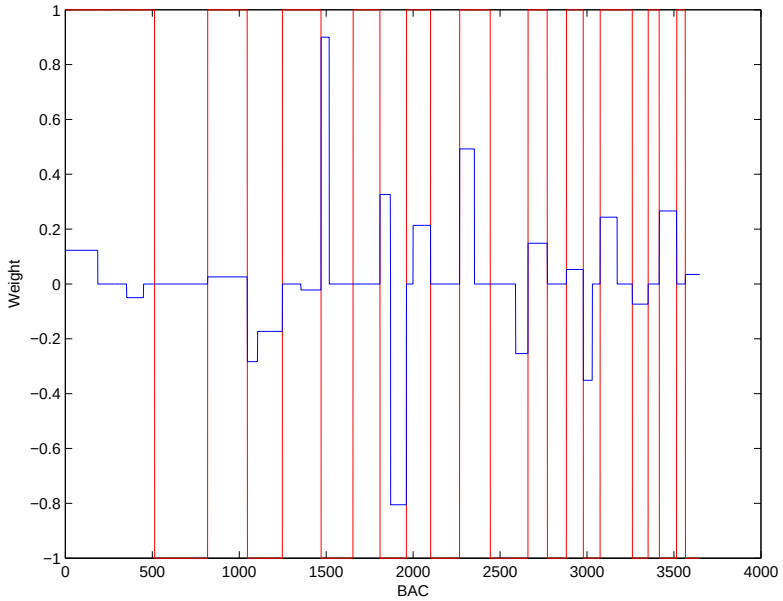
A penalty for CGH array classification

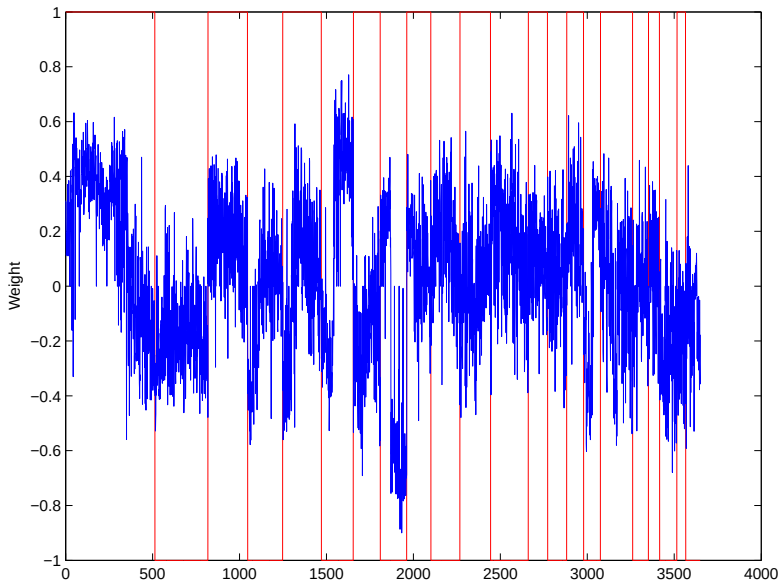
The fused LASSO penalty (Tibshirani et al., 2005)

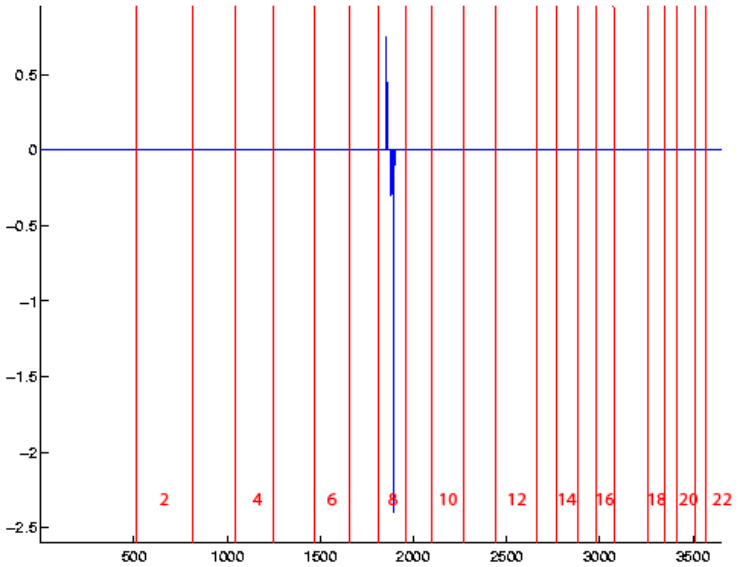
$$\Omega_{fusedlasso}(\beta) = \sum_i |\beta_i| + \sum_{i \sim j} |\beta_i - \beta_j| \,.$$

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions
- Combined with a hinge loss leads to a fused SVM (Rapaport et al., 2008);

Application: metastasis prognosis in melanoma

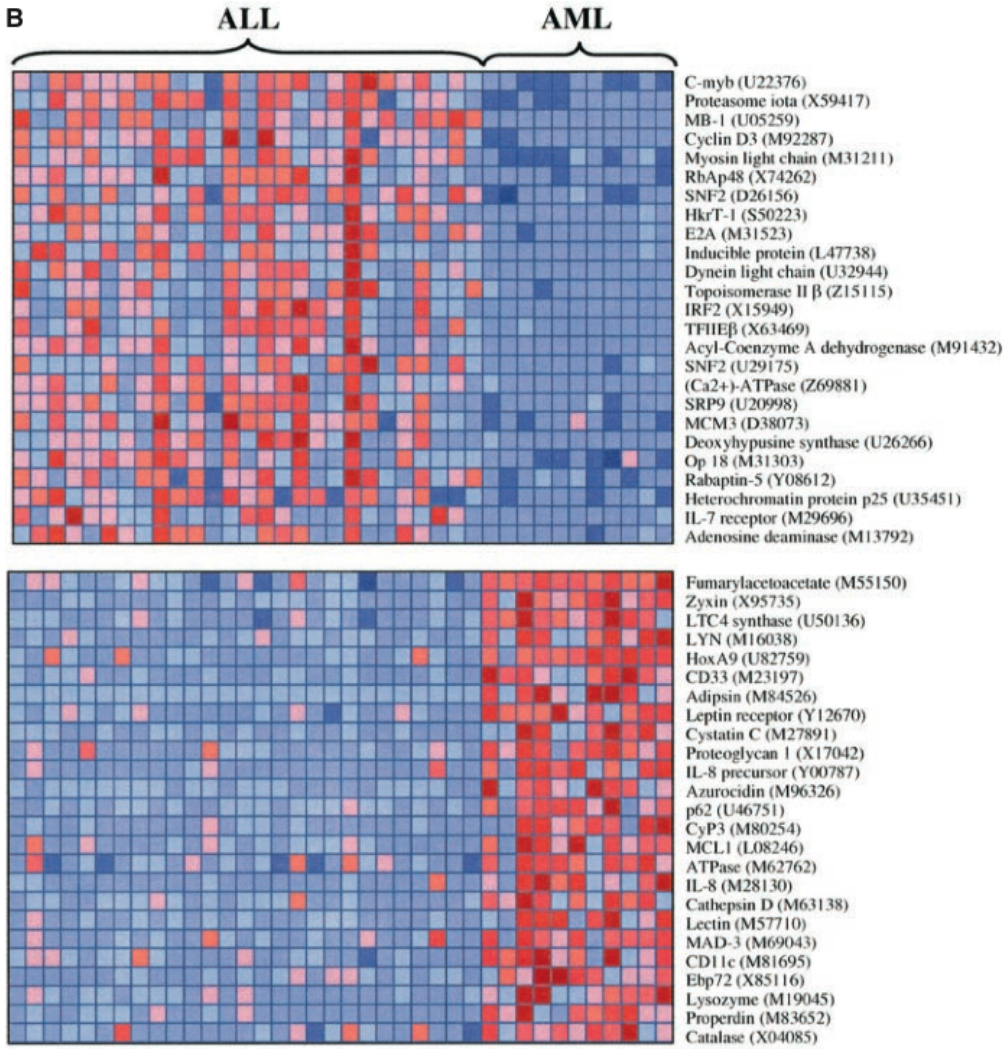




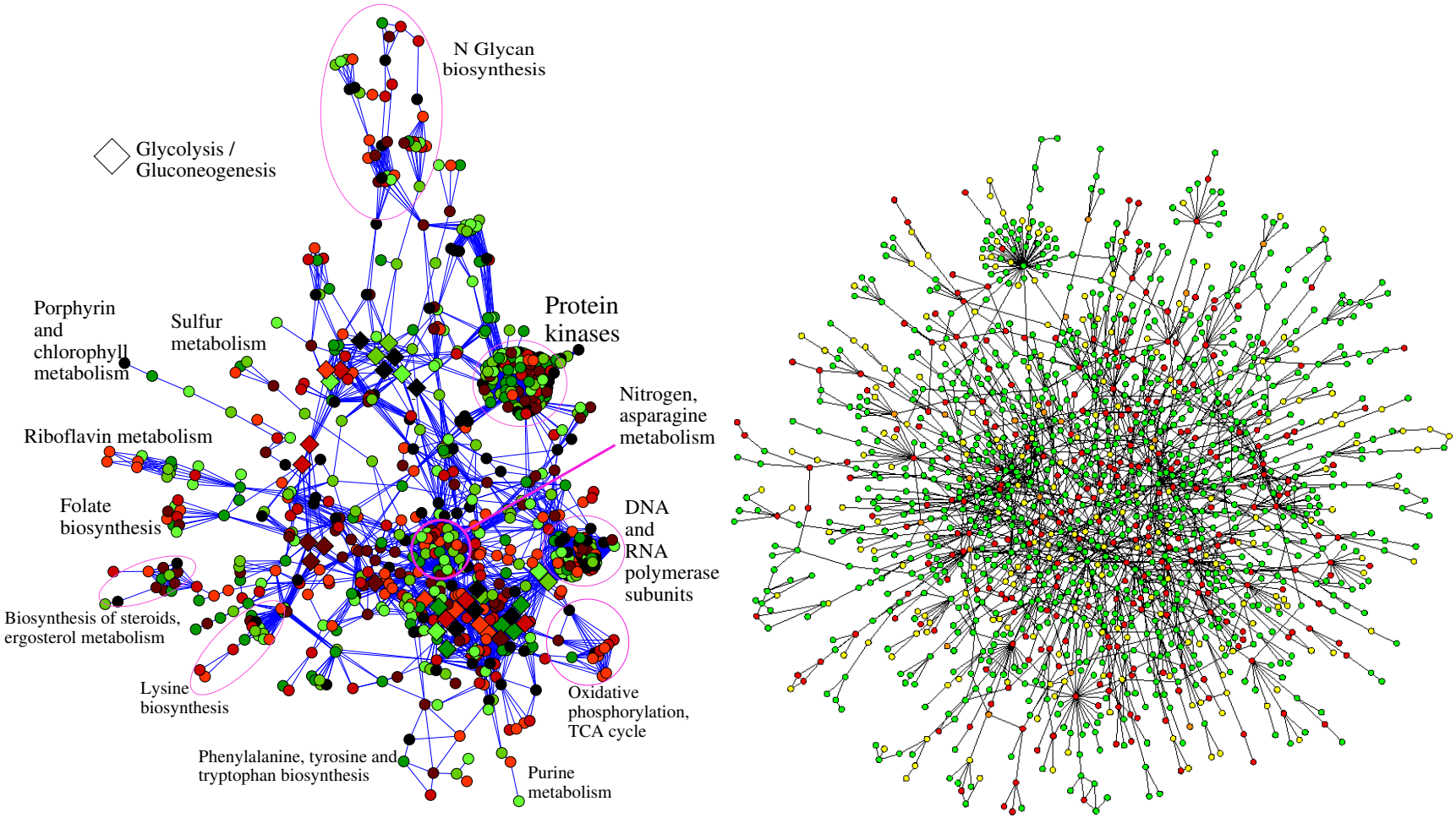


Outline

Tissue classification from microarray data



Gene networks

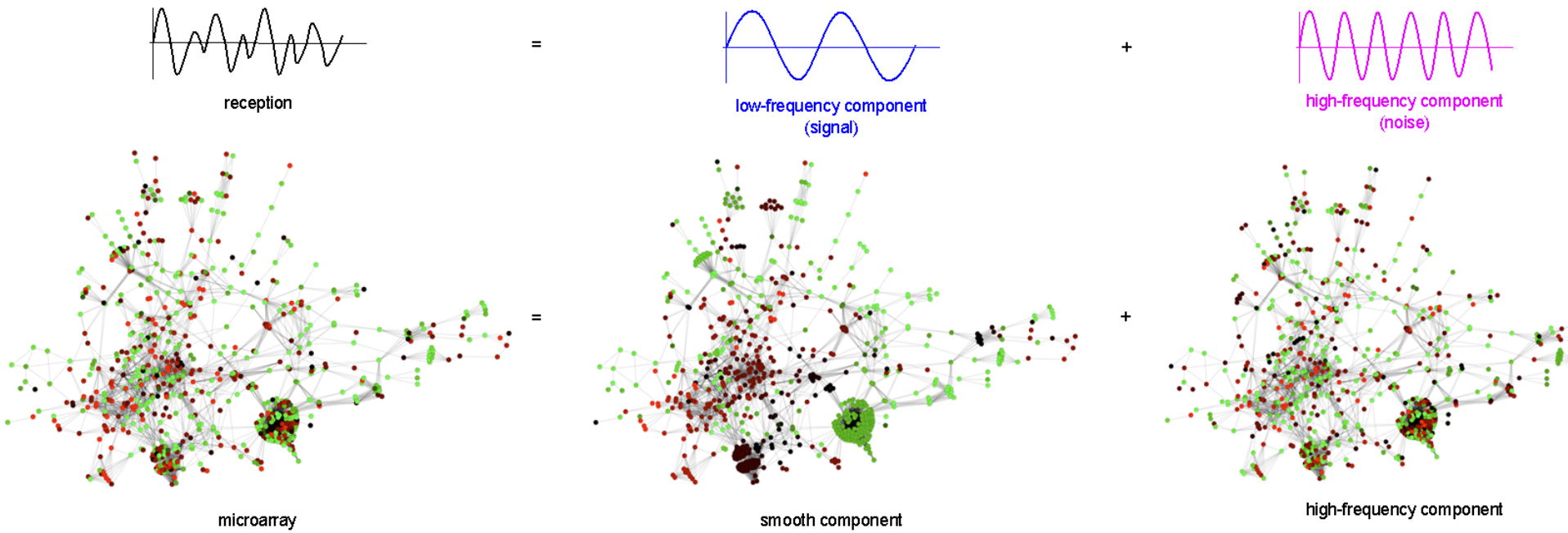


Gene networks and expression data

Motivation

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- Many pathways and protein-protein interactions are already known
- Hypothesis: the weights of the classifier should be "coherent" with respect to this prior knowledge

An idea

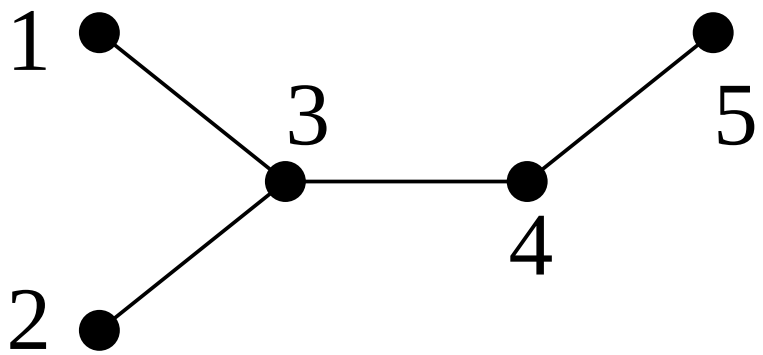


1. Use the gene network to extract the "important information" in gene expression profiles by Fourier analysis on the graph
2. Learn a linear classifier on the smooth components

Graph Laplacian

Definition

The Laplacian of the graph is the matrix $L = D - A$.



$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Fourier basis

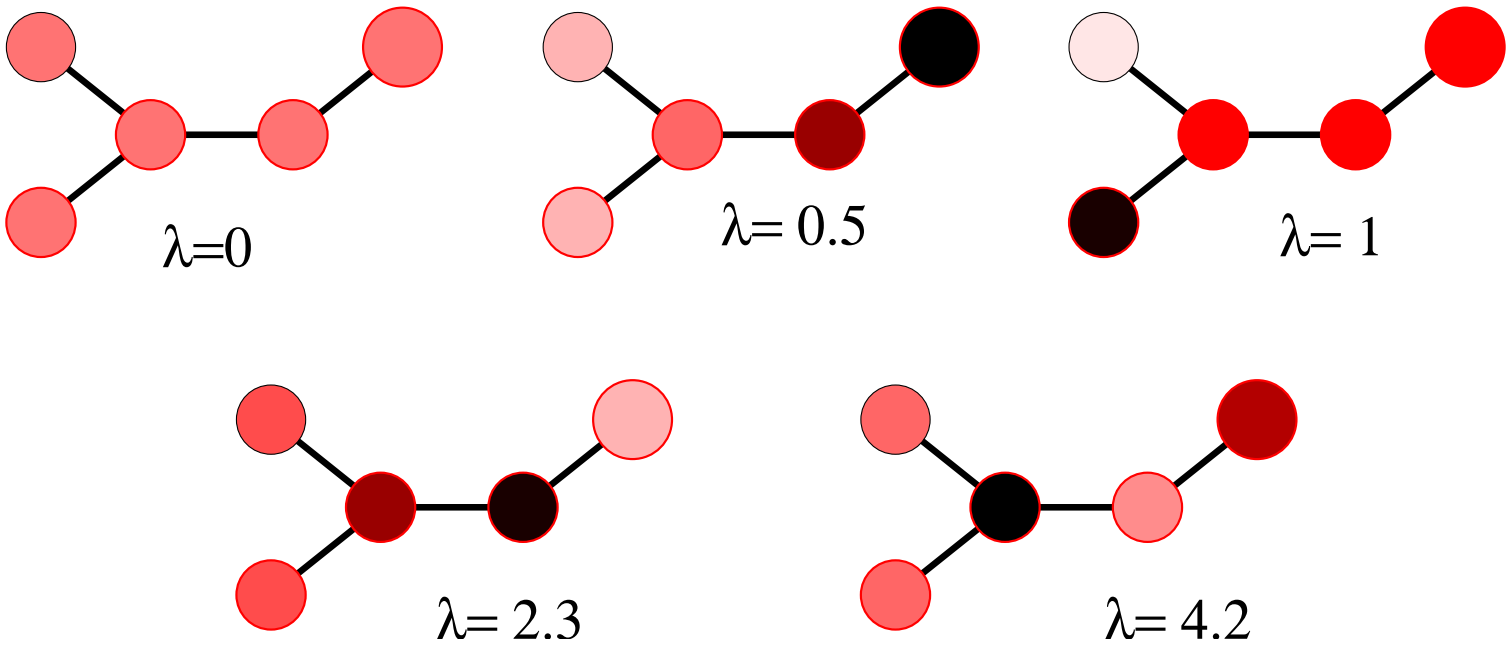
- L is positive semidefinite
- The eigenvectors $e_1, \ldots, e_n$ of L with eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$ form a basis called Fourier basis
- For any $f : V \to \mathbb{R}$, the Fourier transform of f is the vector $\hat{f} \in \mathbb{R}^n$ defined by:

$$\hat{f}_i = f^\top e_i \,, \quad i = 1, \ldots, n.$$

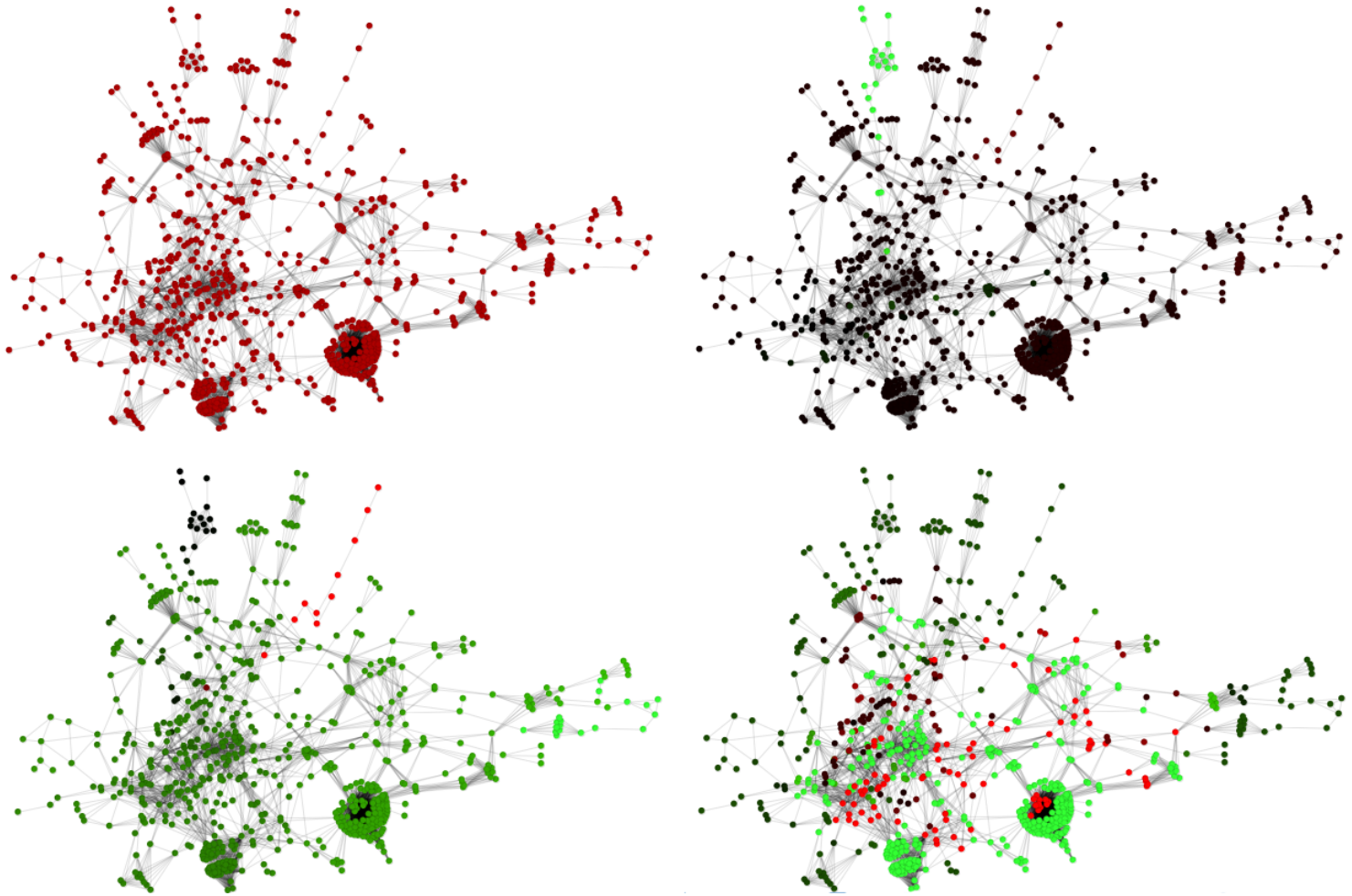
- The inverse Fourier formula holds:

$$f = \sum_{i=1}^{n} \hat{f}_i e_i \,.$$

Fourier basis



Fourier basis



Smoothing operator

Definition

- Let $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ be non-increasing.
- A smoothing operator S_ϕ transform a function $f : V \to \mathbb{R}$ into a smoothed version:

$$S_\phi(f) = \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i \,.$$

Smoothing operators

Examples

- Identity operator ($S_\phi(f) = f$):

$$\phi(\lambda) = 1\,, \quad \forall \lambda$$

- Low-pass filter:

$$\phi(\lambda) = \begin{cases} 1 & \text{if } \lambda \leq \lambda^*\,, \\ 0 & \text{otherwise.} \end{cases}$$

- Attenuation of high frequencies:

$$\phi(\lambda) = \exp(-\beta\lambda)\,.$$

Smoothing operators

Examples

- Identity operator ($S_\phi(f) = f$):

$$\phi(\lambda) = 1\,, \quad \forall \lambda$$

- Low-pass filter:

$$\phi(\lambda) = \begin{cases} 1 & \text{if } \lambda \leq \lambda^*\,, \\ 0 & \text{otherwise.} \end{cases}$$

- Attenuation of high frequencies:

$$\phi(\lambda) = \exp(-\beta\lambda)\,.$$

Smoothing operators

Examples

- Identity operator ($S_\phi(f) = f$):

$$\phi(\lambda) = 1\,, \quad \forall \lambda$$

- Low-pass filter:

$$\phi(\lambda) = \begin{cases} 1 & \text{if } \lambda \leq \lambda^*\,, \\ 0 & \text{otherwise.} \end{cases}$$

- Attenuation of high frequencies:

$$\phi(\lambda) = \exp(-\beta\lambda)\,.$$

Supervised classification and regression

Working with smoothed profiles

- Classical methods for linear classification and regression with a ridge penalty solve:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l\left(\beta^\top f_i, y_i\right) + \lambda \beta^\top \beta \,.$$

- Applying these algorithms on the smooth profiles means solving:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l\left(\beta^\top S_\phi(f_i), y_i\right) + \lambda \beta^\top \beta \,.$$

Link with shrinkage estimator

Lemma

This is equivalent to:

$$\min_{v \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l\left(v^\top f_i, y_i\right) + \lambda \sum_{i=1}^{p} \frac{\hat{v}_i^2}{\phi(\lambda_i)} \,,$$

hence the linear classifier v is smooth.

Proof

- Let $v = \sum_{i=1}^{n} \phi(\lambda_i) e_i e_i^\top \beta$, then

$$\beta^\top S_\phi(f_i) = \beta^\top \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i = f^\top v \,.$$

- Then $\hat{v}_i = \phi(\lambda_i)\hat{\beta}_i$ and $\beta^\top \beta = \sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)^2}$.

Link with shrinkage estimator

Lemma

This is equivalent to:

$$\min_{v \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l\left(v^\top f_i, y_i\right) + \lambda \sum_{i=1}^{p} \frac{\hat{v}_i^2}{\phi(\lambda_i)} \,,$$

hence the linear classifier v is smooth.

Proof

- Let $v = \sum_{i=1}^n \phi(\lambda_i) e_i e_i^\top \beta$, then

$$\beta^\top S_\phi(f_i) = \beta^\top \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i = f^\top v \,.$$

- Then $\hat{v}_i = \phi(\lambda_i)\hat{\beta}_i$ and $\beta^\top \beta = \sum_{i=1}^n \frac{\hat{v}_i^2}{\phi(\lambda_i)^2}$.

Smoothing kernel

Kernel methods (SVM, kernel ridge regression..) only need the inner product between smooth profiles:

$$\begin{aligned} K(f,g) &= S_\phi(f)^\top S_\phi(g) \\ &= \sum_{i=1}^{n} \hat{f}_i \hat{g}_i \phi(\lambda_i)^2 \\ &= f^\top \left(\sum_{i=1}^{n} \phi(\lambda_i)^2 e_i e_i^\top \right) g \\ &= f^\top K_\phi g \,, \end{aligned} \tag{1}$$

with

$$K_\phi = \sum_{i=1}^{n} \phi(\lambda_i)^2 e_i e_i^\top \,.$$

Examples

- For $\phi(\lambda) = \exp(-t\lambda)$, we recover the diffusion kernel:

$$K_\phi = \exp_M(-2tL)\,.$$

- For $\phi(\lambda) = 1/\sqrt{1+\lambda}$, we obtain

$$K_\phi = (L + I)^{-1}\,,$$

and the penalization is:

$$\sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)} = v^\top (L + I)\, v = \| v \|_2^2 + \sum_{i \sim j} (v_i - v_j)^2\,.$$

Examples

- For $\phi(\lambda) = \exp(-t\lambda)$, we recover the diffusion kernel:

$$K_\phi = \exp_M(-2tL)\,.$$

- For $\phi(\lambda) = 1/\sqrt{1+\lambda}$, we obtain

$$K_\phi = (L + I)^{-1}\ ,$$

and the penalization is:

$$\sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)} = v^\top (L + I)\, v = \| v \|_2^2 + \sum_{i \sim j} (v_i - v_j)^2\,.$$

Data

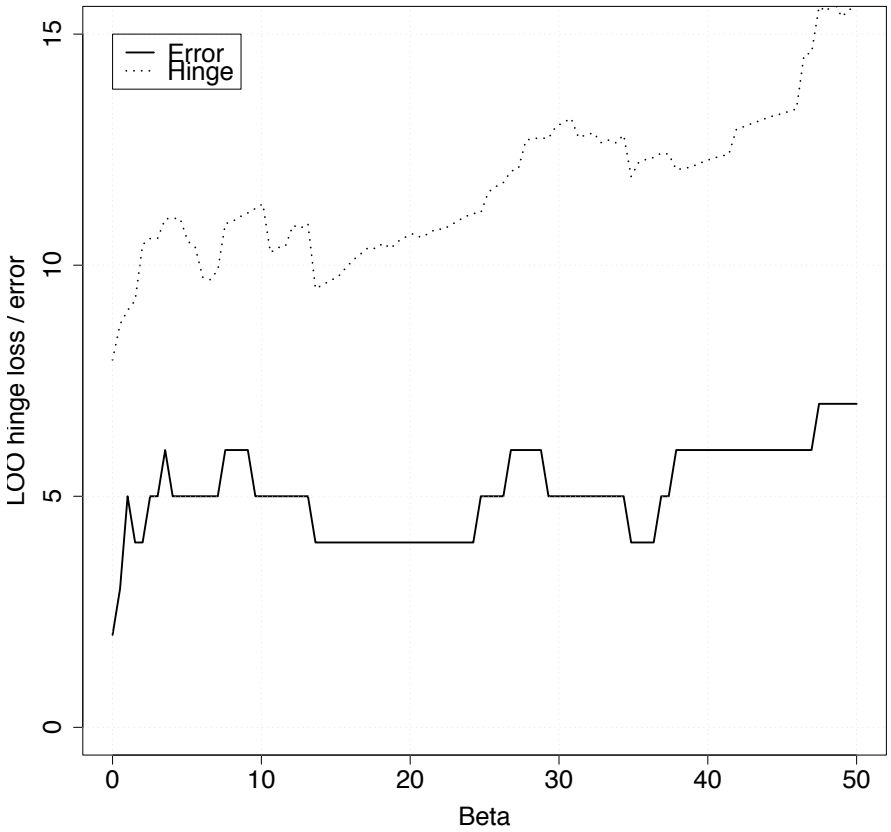
Expression

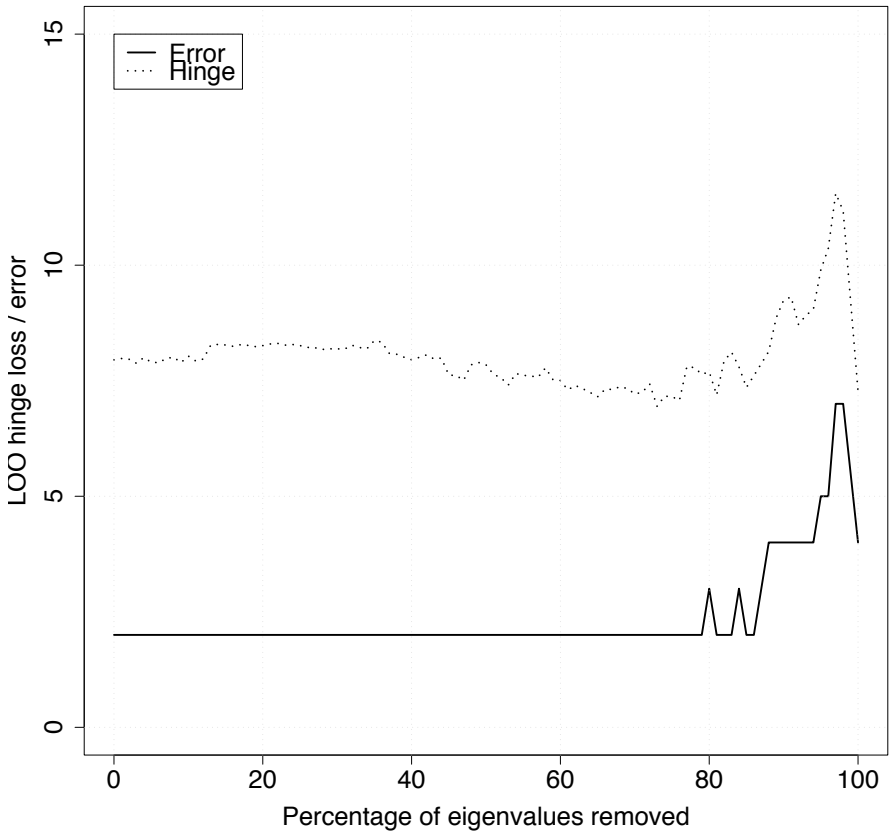
- Study the effect of low irradiation doses on the yeast
- 12 non irradiated vs 6 irradiated
- Which pathways are involved in the response at the transcriptomic level?

Graph

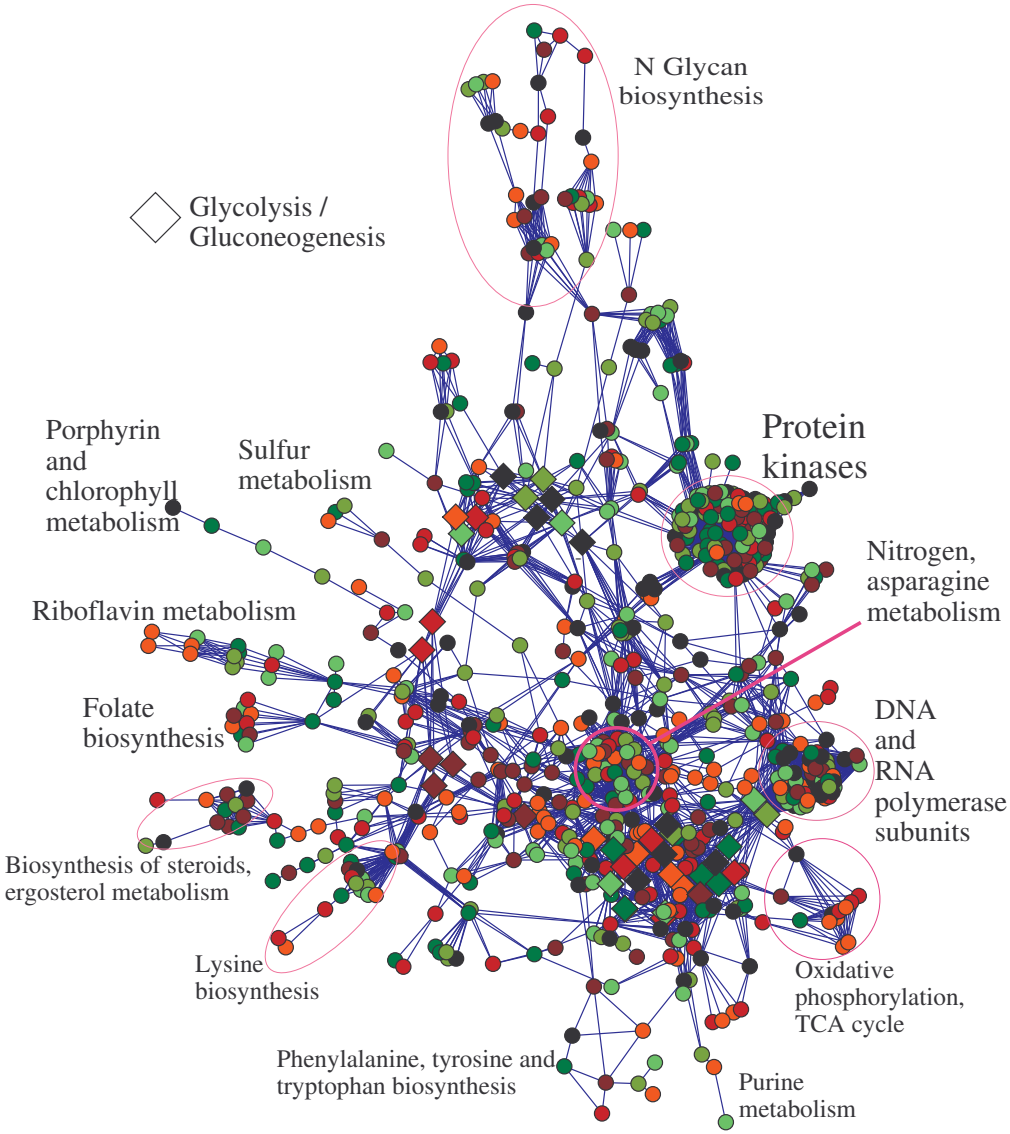
- KEGG database of metabolic pathways
- Two genes are connected is they code for enzymes that catalyze successive reactions in a pathway (metabolic gene network).
- 737 genes, 4694 vertices.

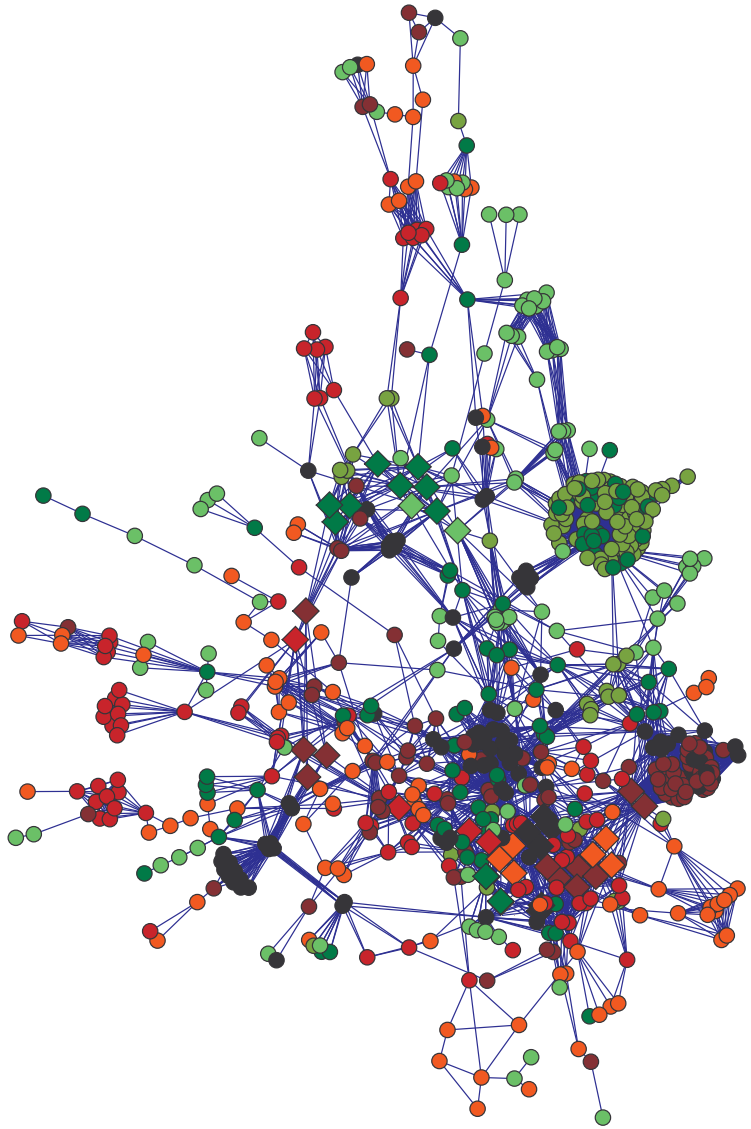
Classification performance



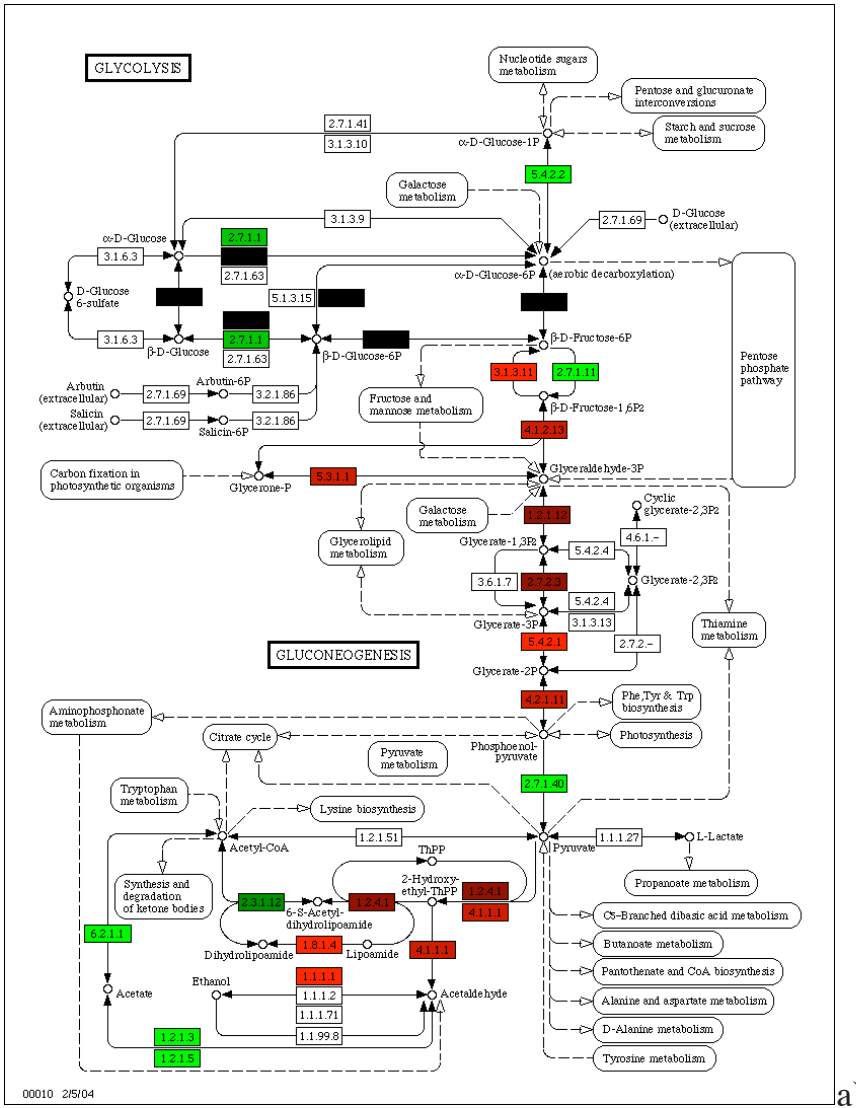


Classifier

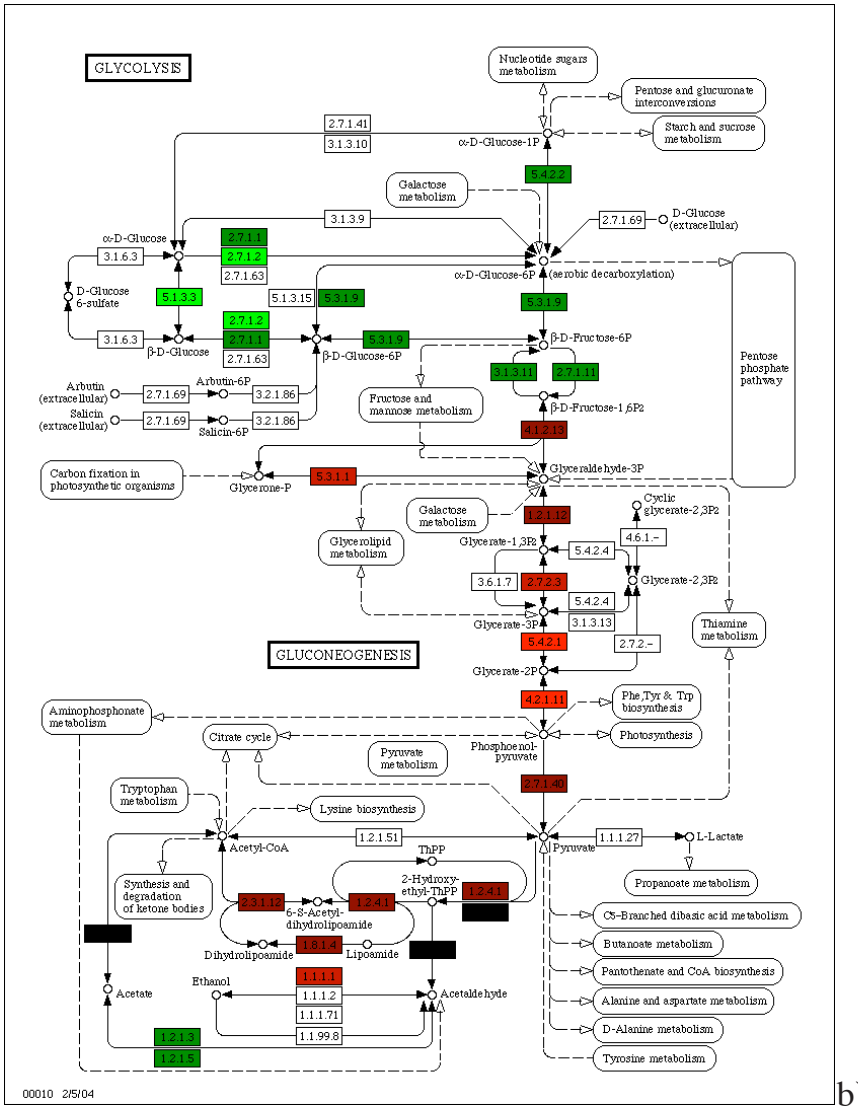

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism



Classifier



a)



b)

Other penalties

Prior hypothesis

Genes near each other on the graph should have similar weigths.

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

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

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

Other penalties

Prior hypothesis

Genes near each other on the graph should have similar weigths.

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

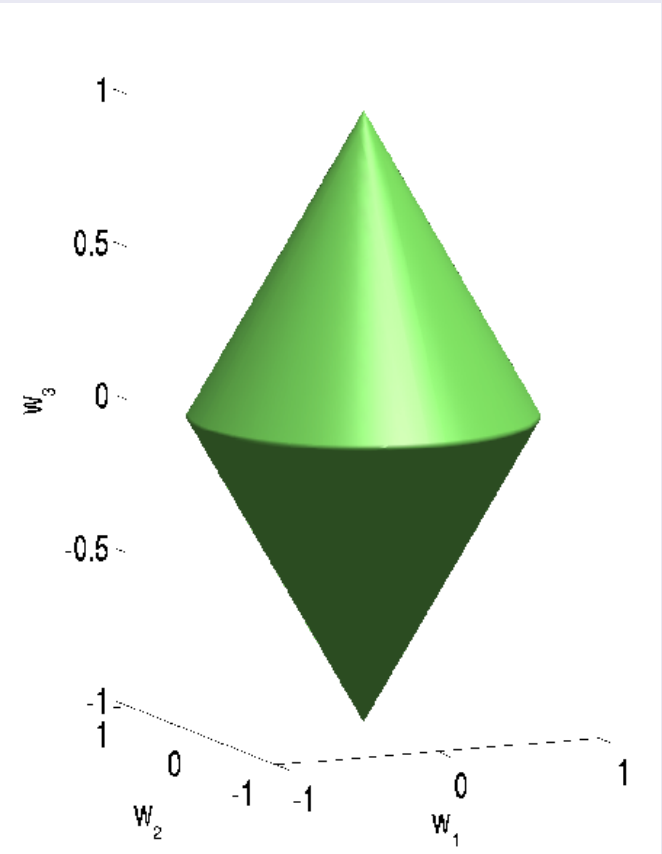
$$\Omega_{spectral}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \,,$$

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

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*.

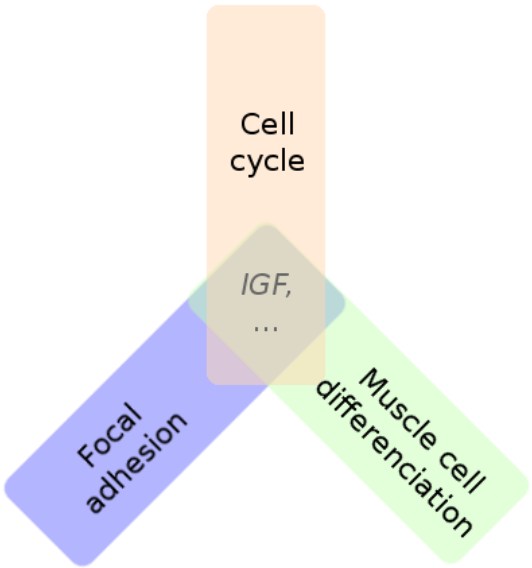


$$min_w L(w) + \lambda \left(\|(w_1, w_2)\|_2 + \|w_3\|_2 \right).$$

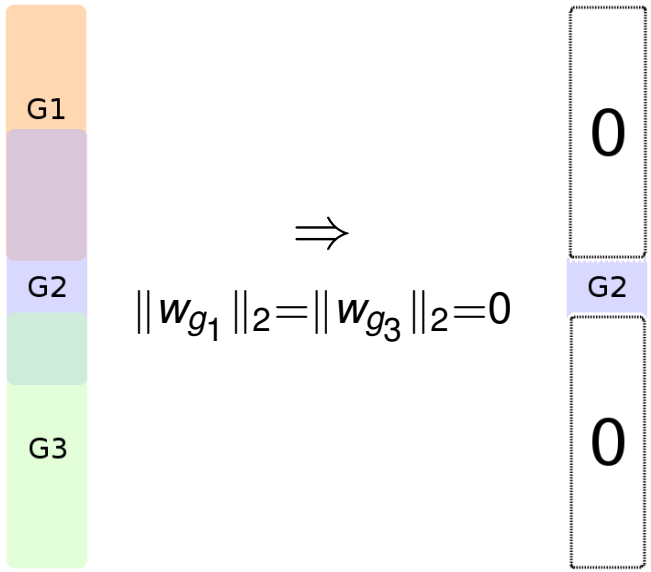
Biological markers for cancer

Issue of using the group-lasso

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



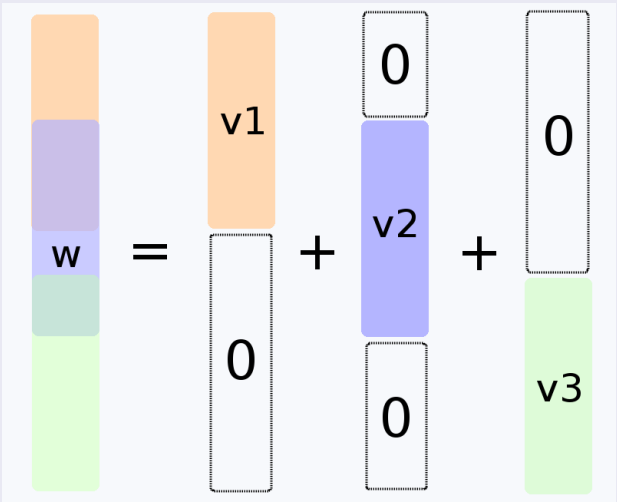
Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

Overlap norm

Overlap norm

Introduce latent variables v_g:

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Setting one v_g to 0 doesn't necessarily set to 0 all its variables in w.

Overlap norm

Overlap norm

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases} = \min_{w} L(w) + \lambda \Omega_{overlap}(w)$$

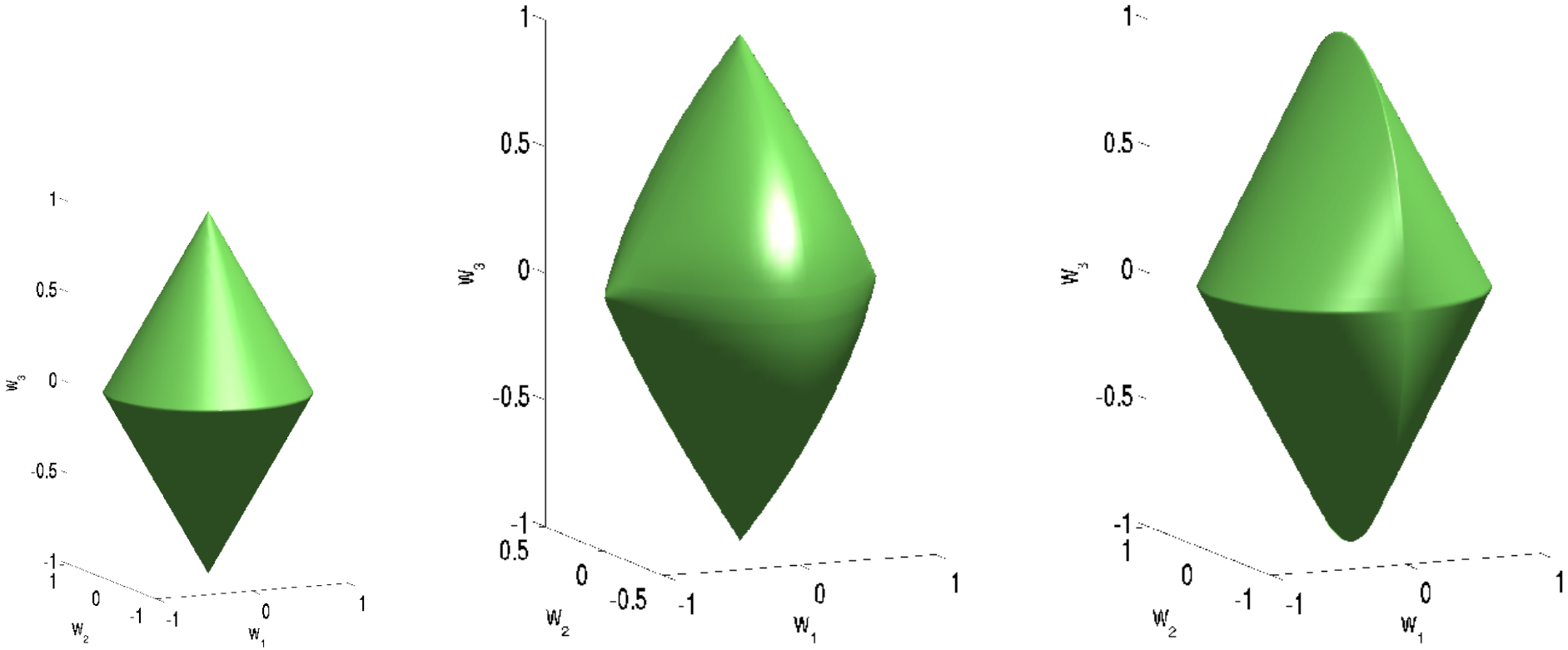
with

$$\Omega_{overlap}(w) \triangleq \begin{cases} \min_{v} \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases} \quad (*)$$

Property

- $\Omega_{overlap}(w)$ is a norm of w.
- $\Omega_{overlap}(.)$ associates to w a specific (not necessarily unique) decomposition $(v_g)_{g \in \mathcal{G}}$ which is the argmin of $(*)$.

Overlap and group unity balls



Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$ (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.

Overlap norm

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 $\Omega^{\mathcal{G}}_{\text{overlap}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{overlap}}(w)$.

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

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

.

Overlap norm

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 $\Omega^{\mathcal{G}}_{\text{overlap}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{overlap}}(w)$.

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

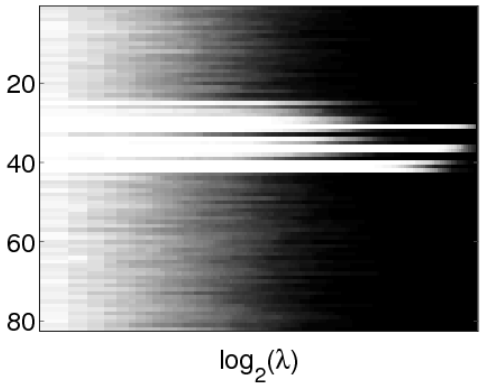
$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

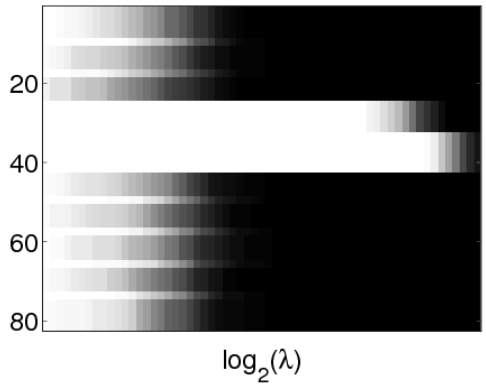
.

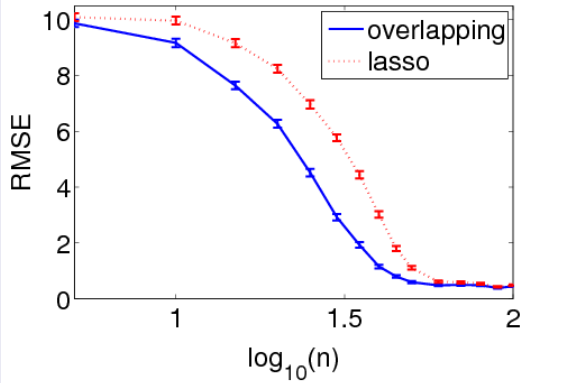
Results

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 4*th* and 5*th* groups.
- Learn from 100 training points.







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

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	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
♯ PATH.	$148, 58, 183$	$6, 5, 78$
PROP. PATH.	$0.32, 0.14, 0.41$	$0.01, 0.01, 0.17$

- Graph on the genes.

METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	$1.1, 1, 1.0$	$1.3, 1.4, 1.2$

Outline

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...)
- Natural extensions for data integration
- Extension to "structured statistical tests"?

People I need to thank

- Franck Rapaport, Emmanuel Barillot, Andrei Zynoviev, Laurent Jacob (Institut Curie / Mines ParisTech)
- Guillaume Obozinski (UC Berkeley / INRIA)

This presentation is supported by a JSPS Invitation Fellowship Program for Research in Japan, hosted by Tatsuya Akutsu (Kyoto University)