Including Prior Knowledge in Machine Learning for Genomic Data

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

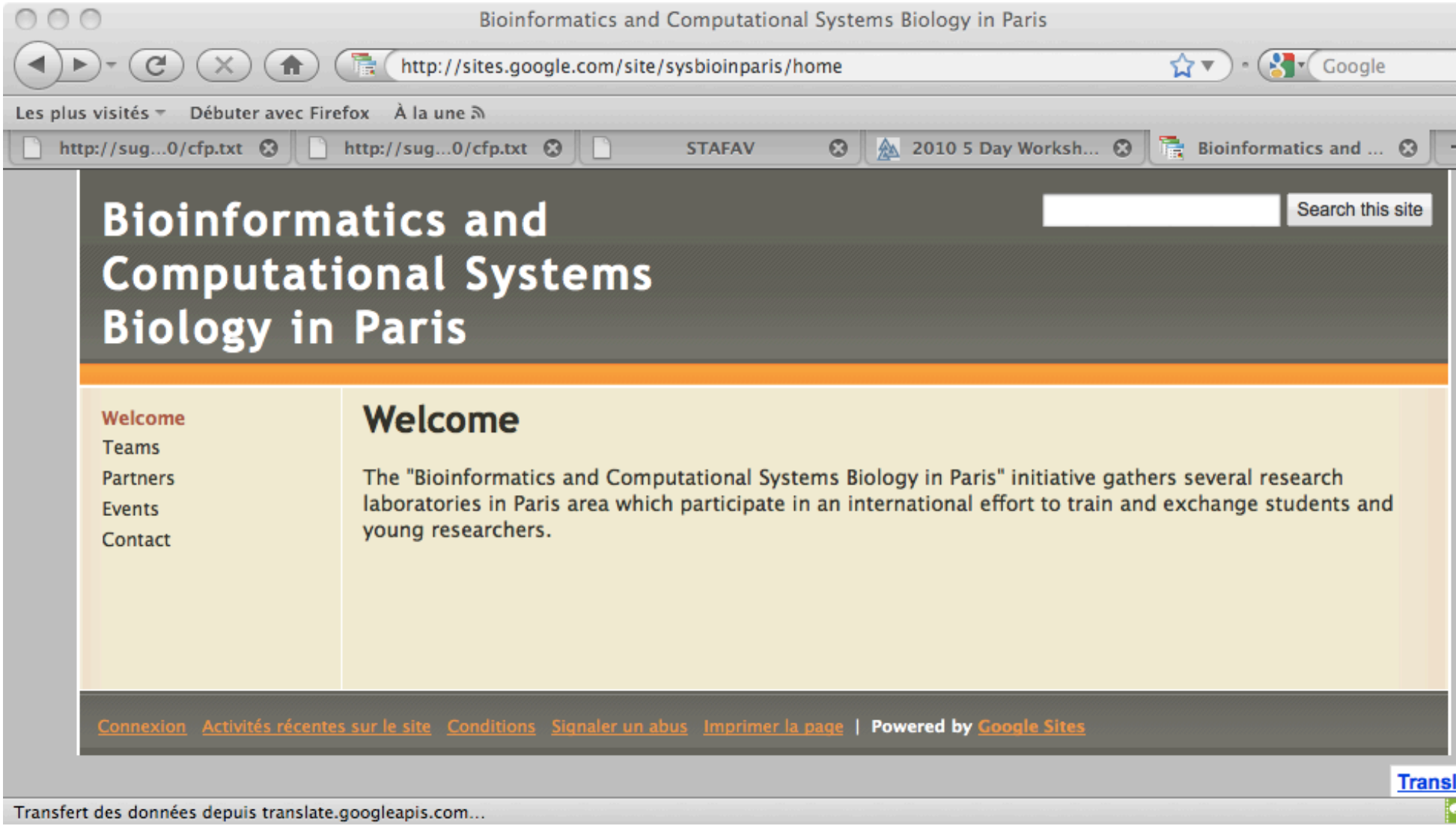
Mines ParisTech / Curie Institute / Inserm

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

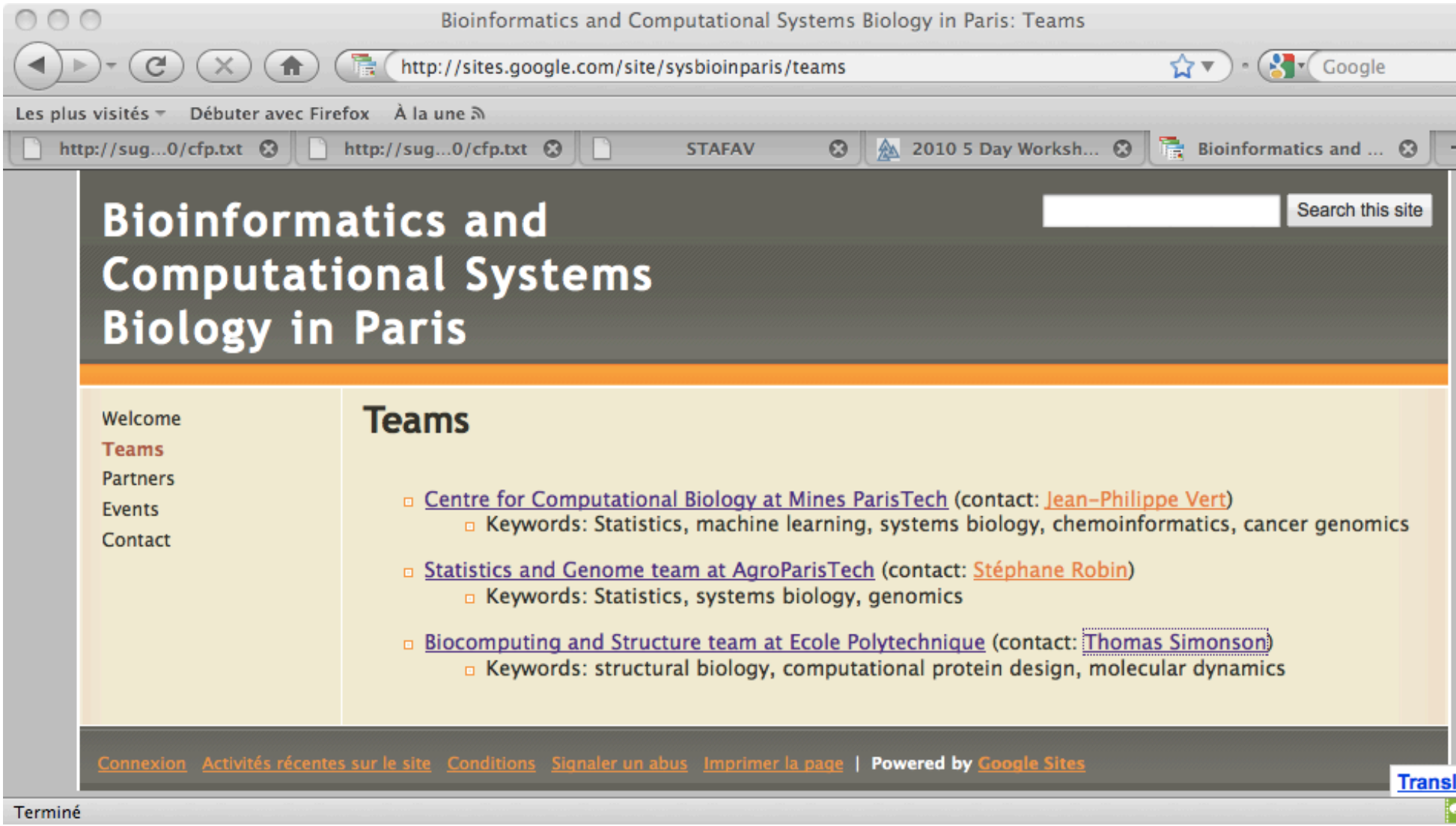
Outline

Outline

Bienvenue à Paris!

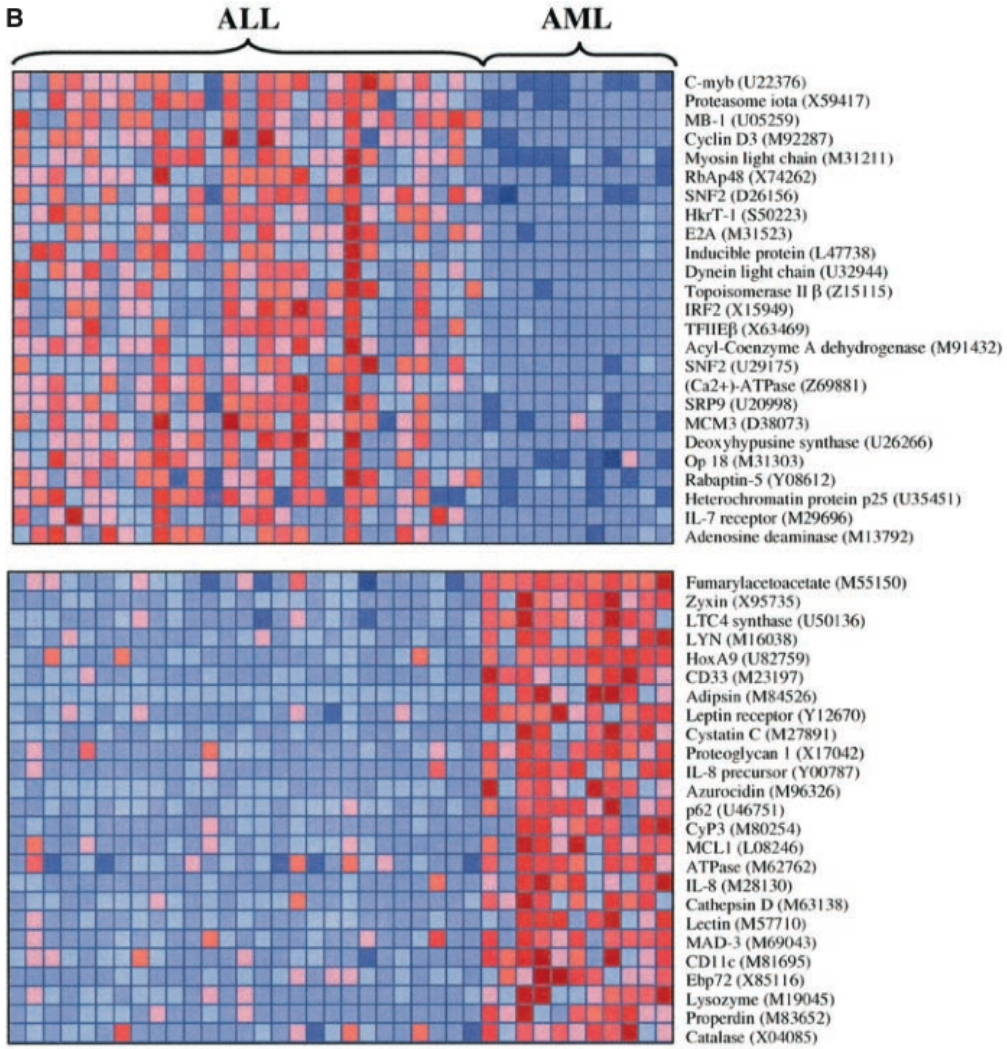


Current teams

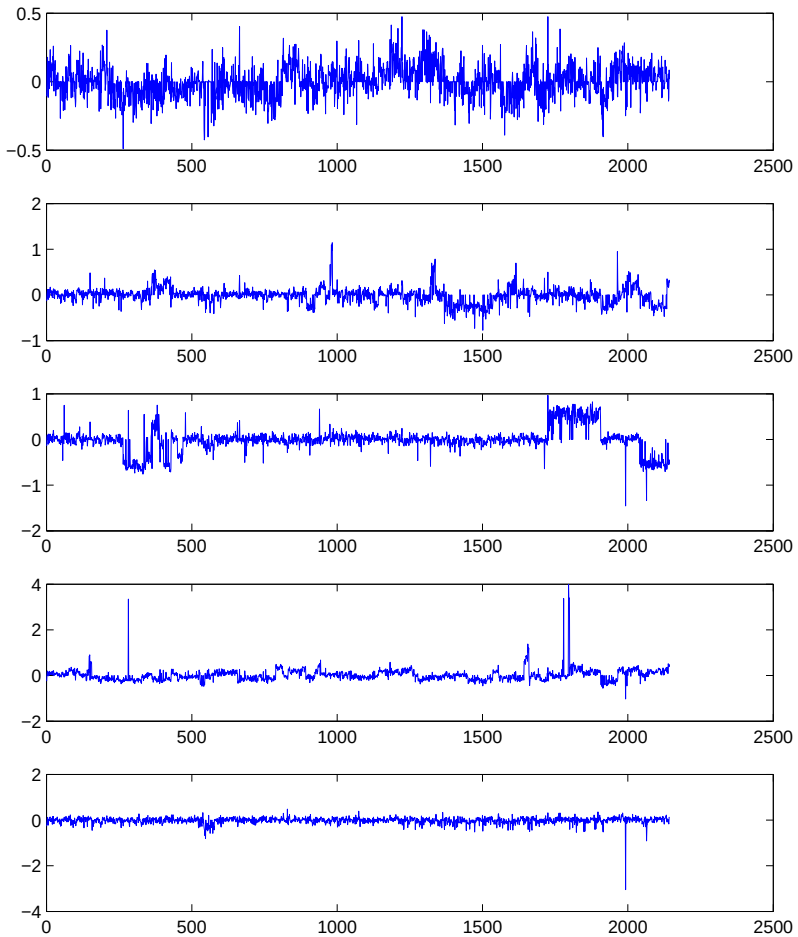


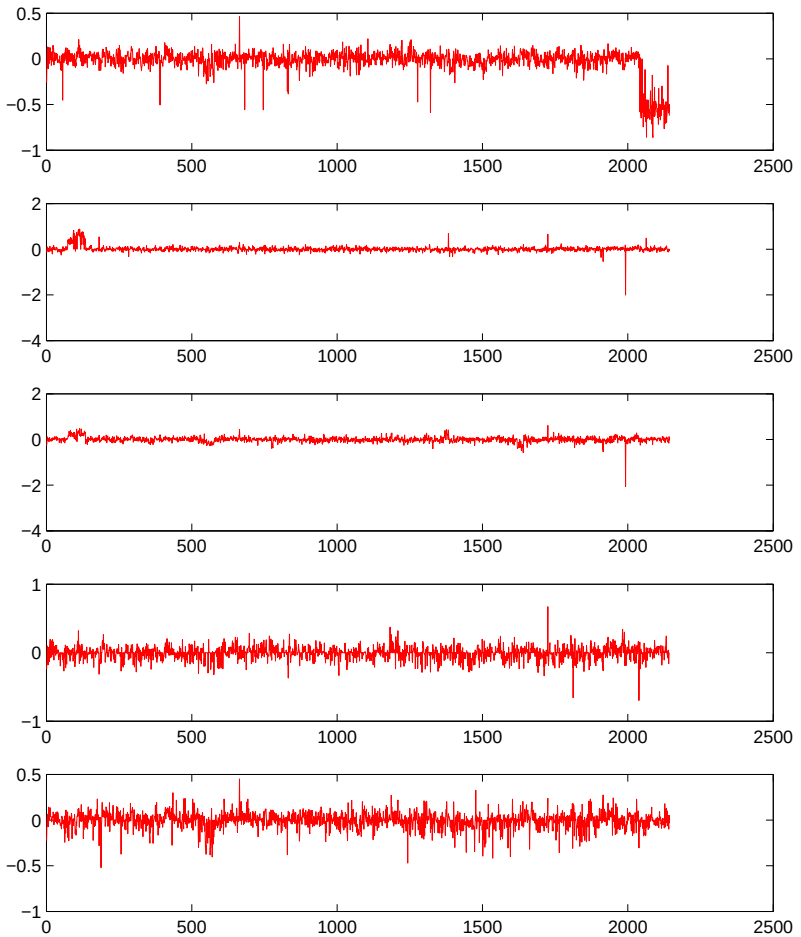
Outline

Cancer diagnosis

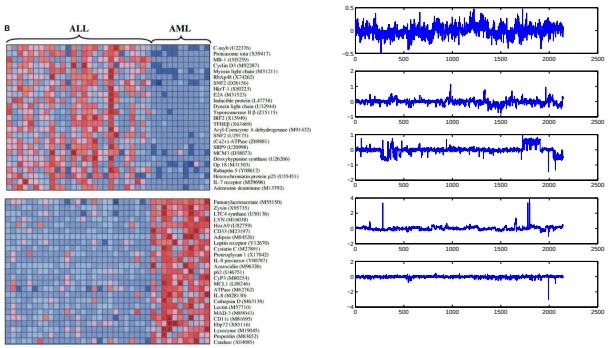


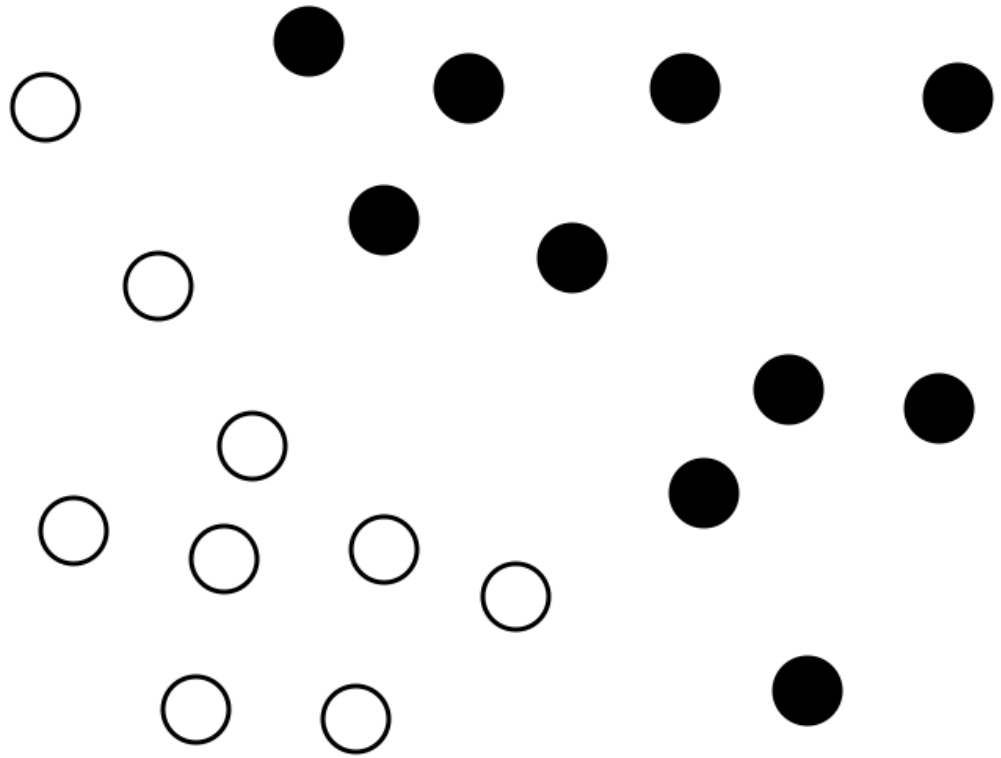
Cancer prognosis



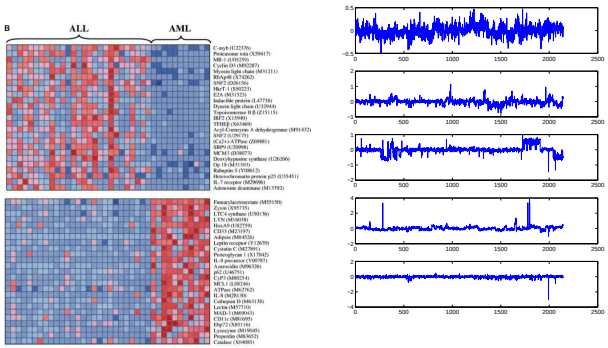


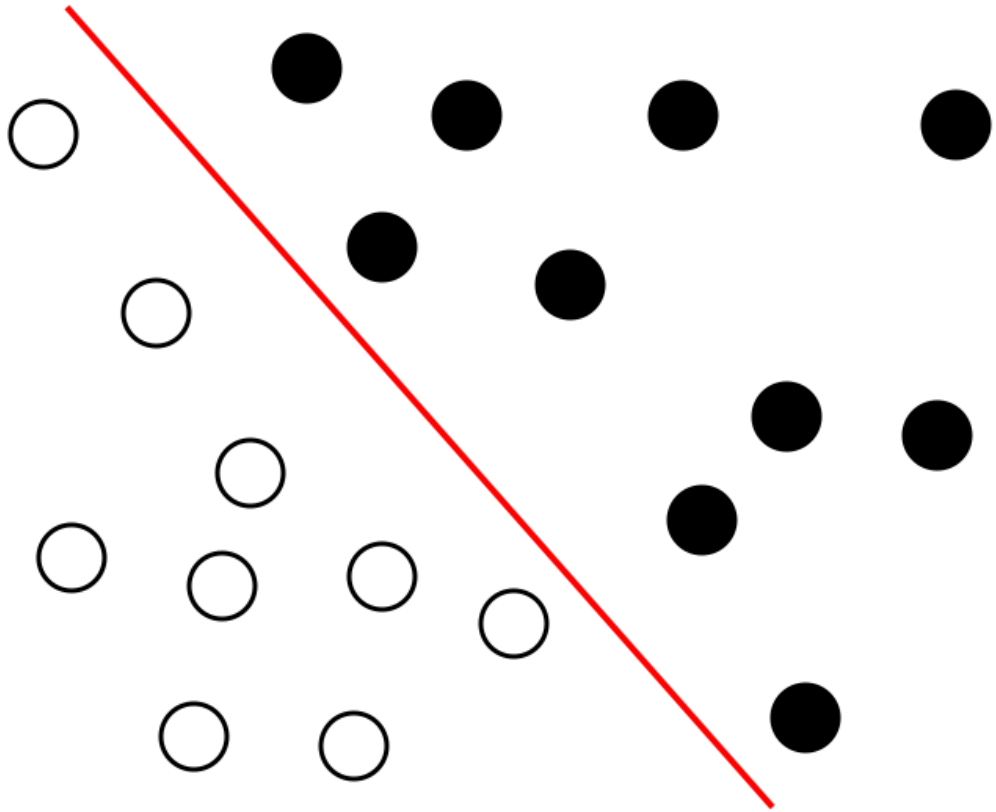
Pattern recognition, *aka* supervised classification



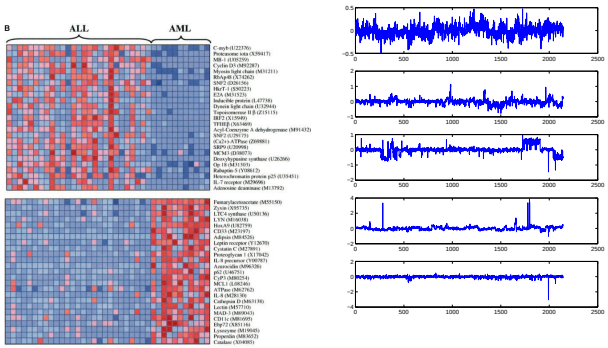


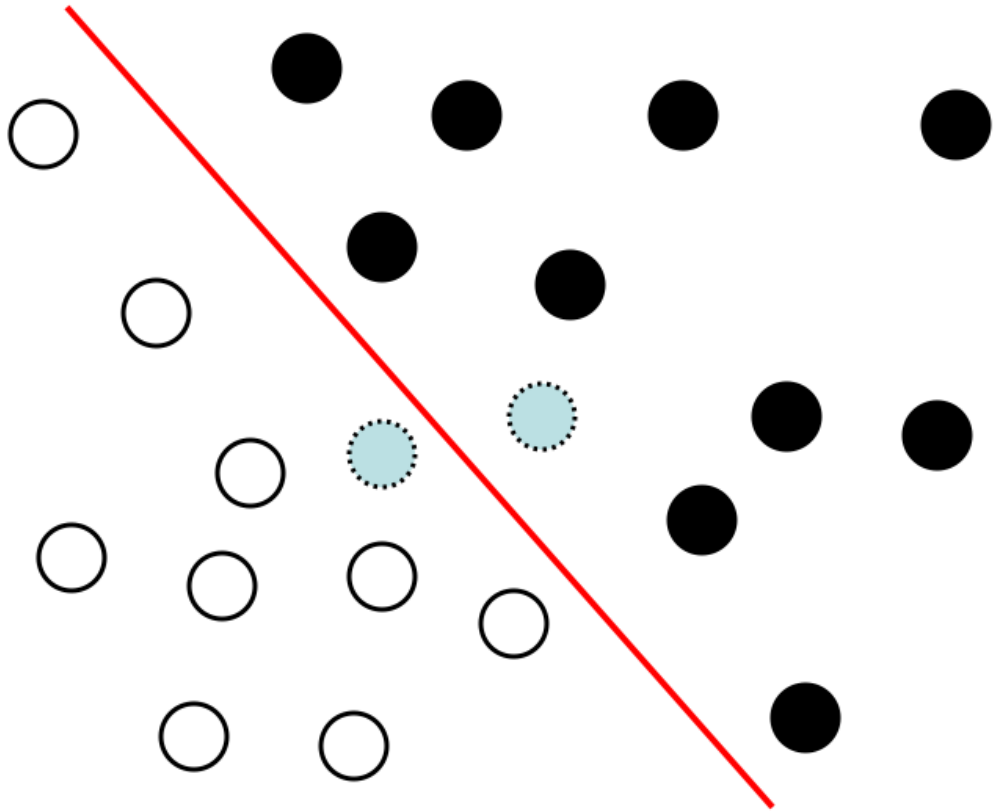
Pattern recognition, *aka* supervised classification



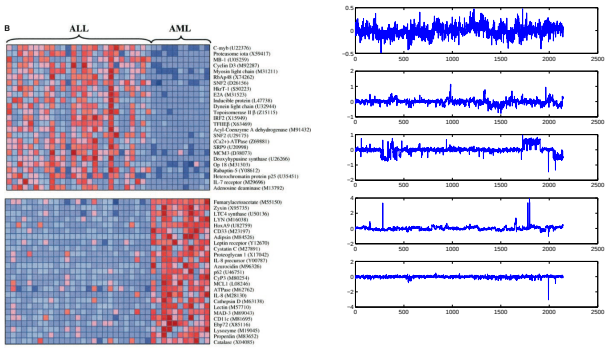


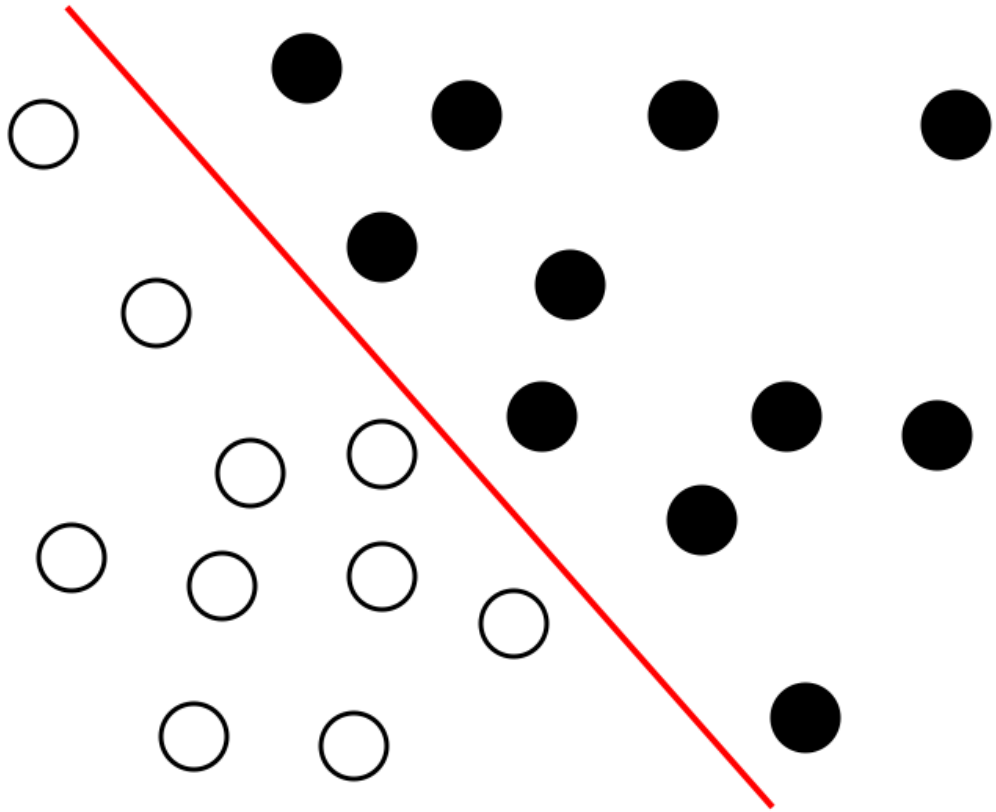
Pattern recognition, *aka* supervised classification



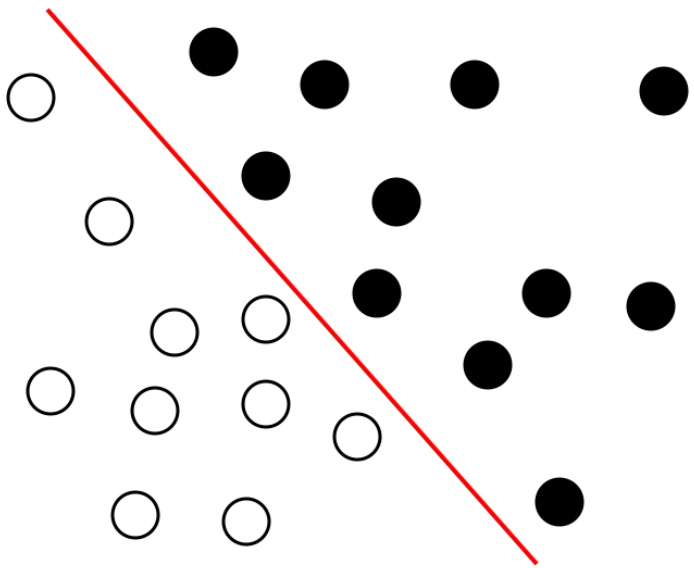


Pattern recognition, *aka* supervised classification





Pattern recognition, *aka* supervised classification



Challenges

- 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 minimium empirical risk, subject to some constraint:

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

Shrinkage estimators

- Define a large family of "candidate classifiers", e.g., linear predictors $f_\beta(x) = \beta^\top x$
- For 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 minimium empirical risk, subject to some constraint:

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

Shrinkage estimators

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

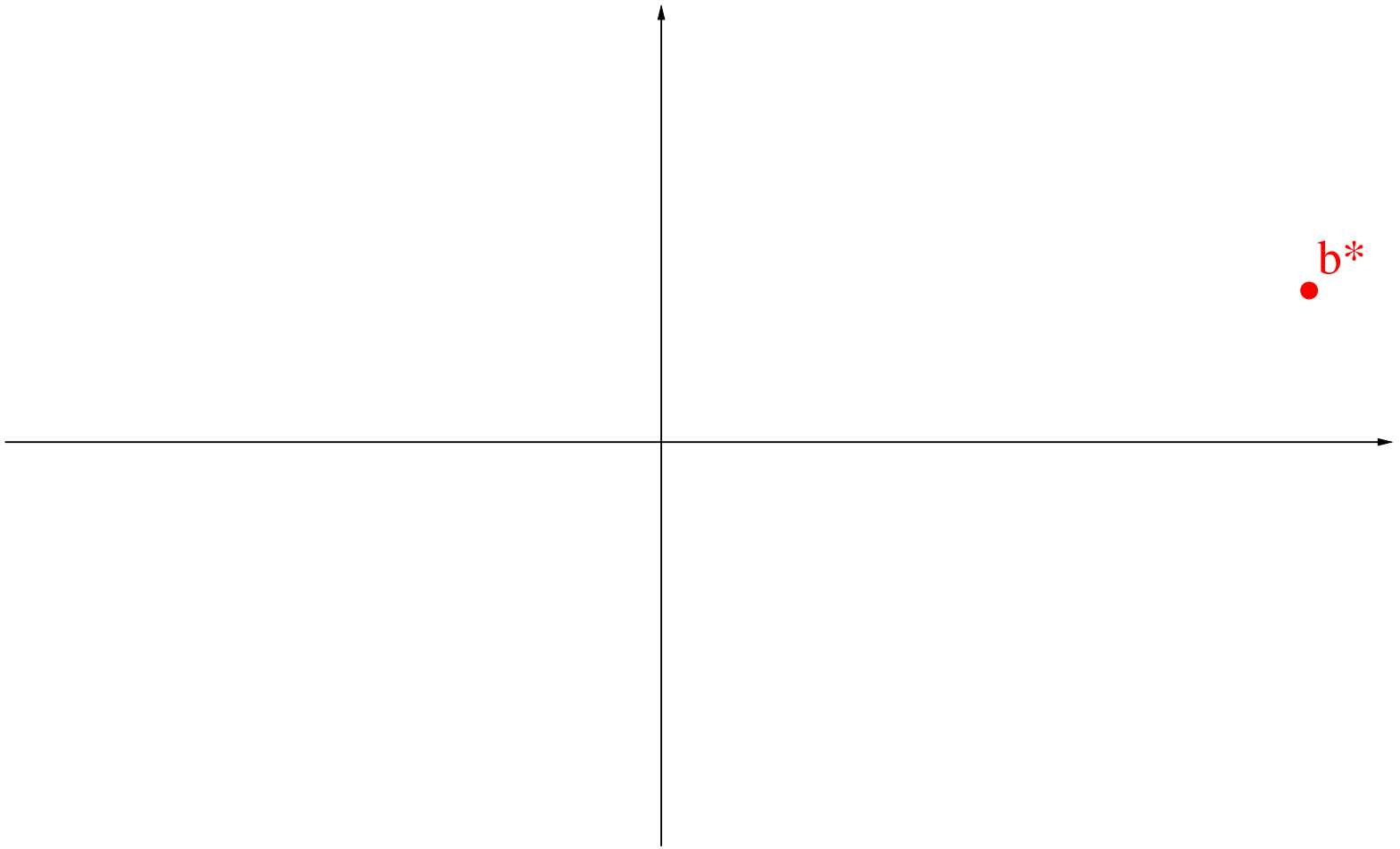
$$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) \,.$$

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

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

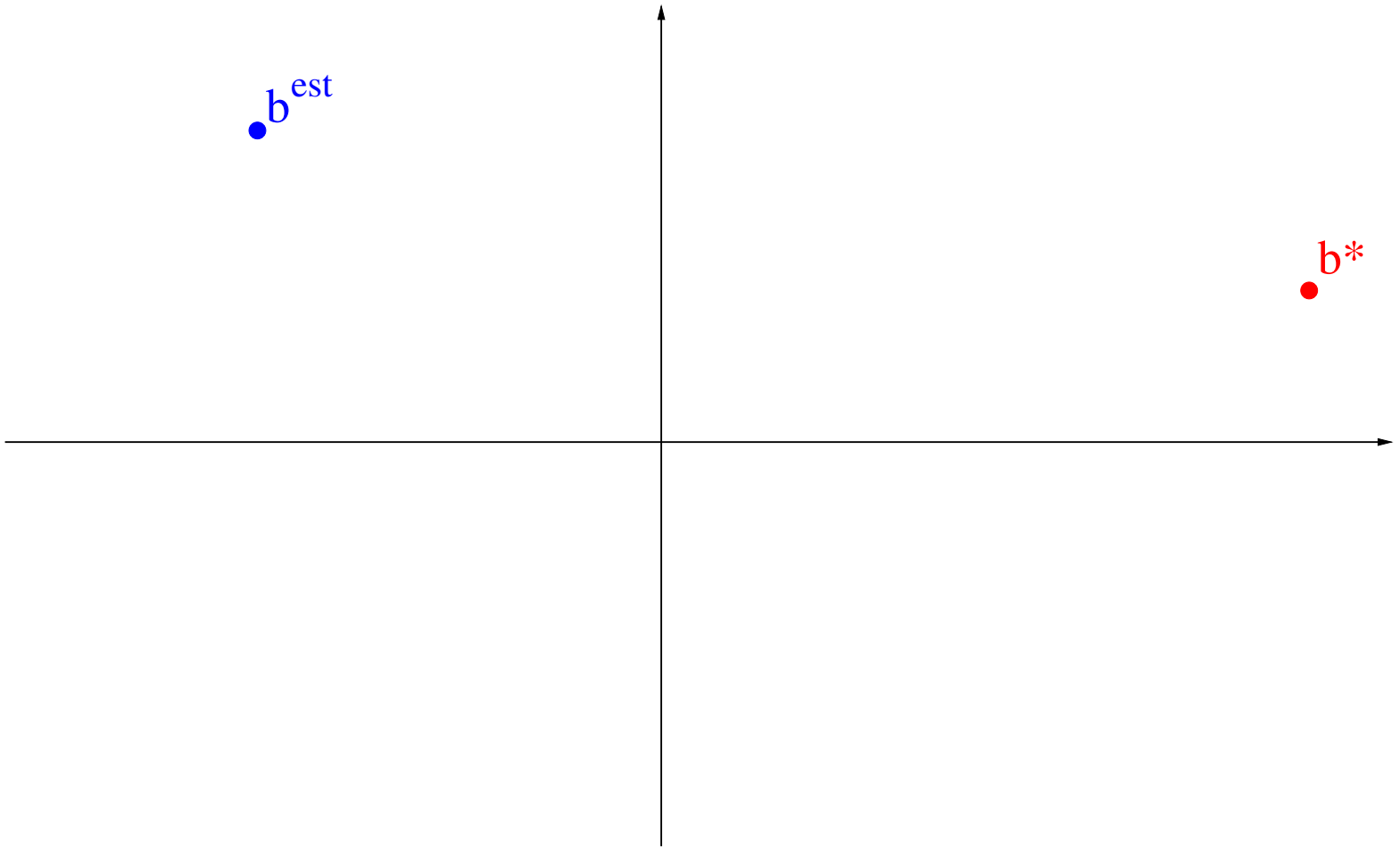
Why skrinkage classifiers?

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



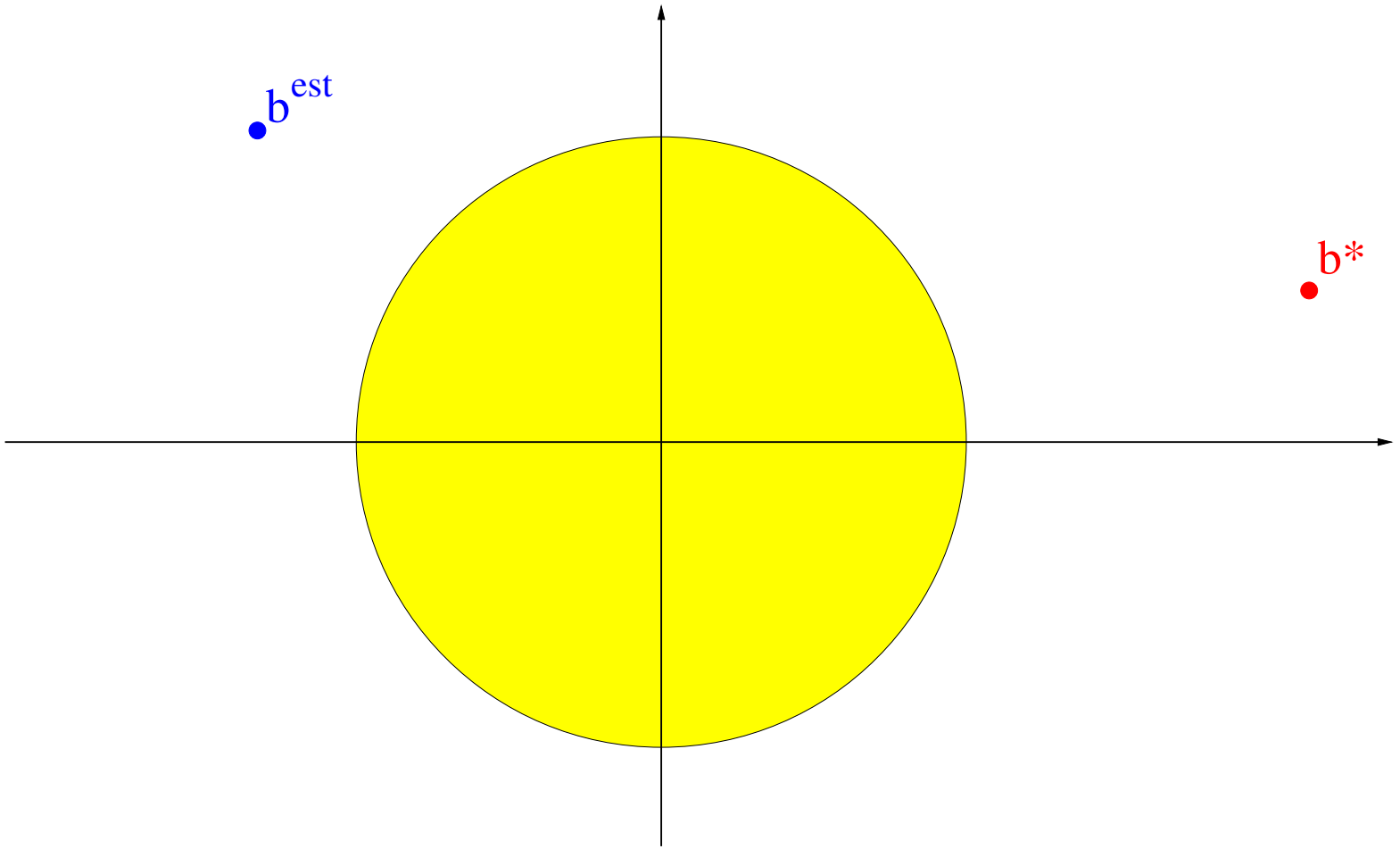
Why skrinkage classifiers?

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



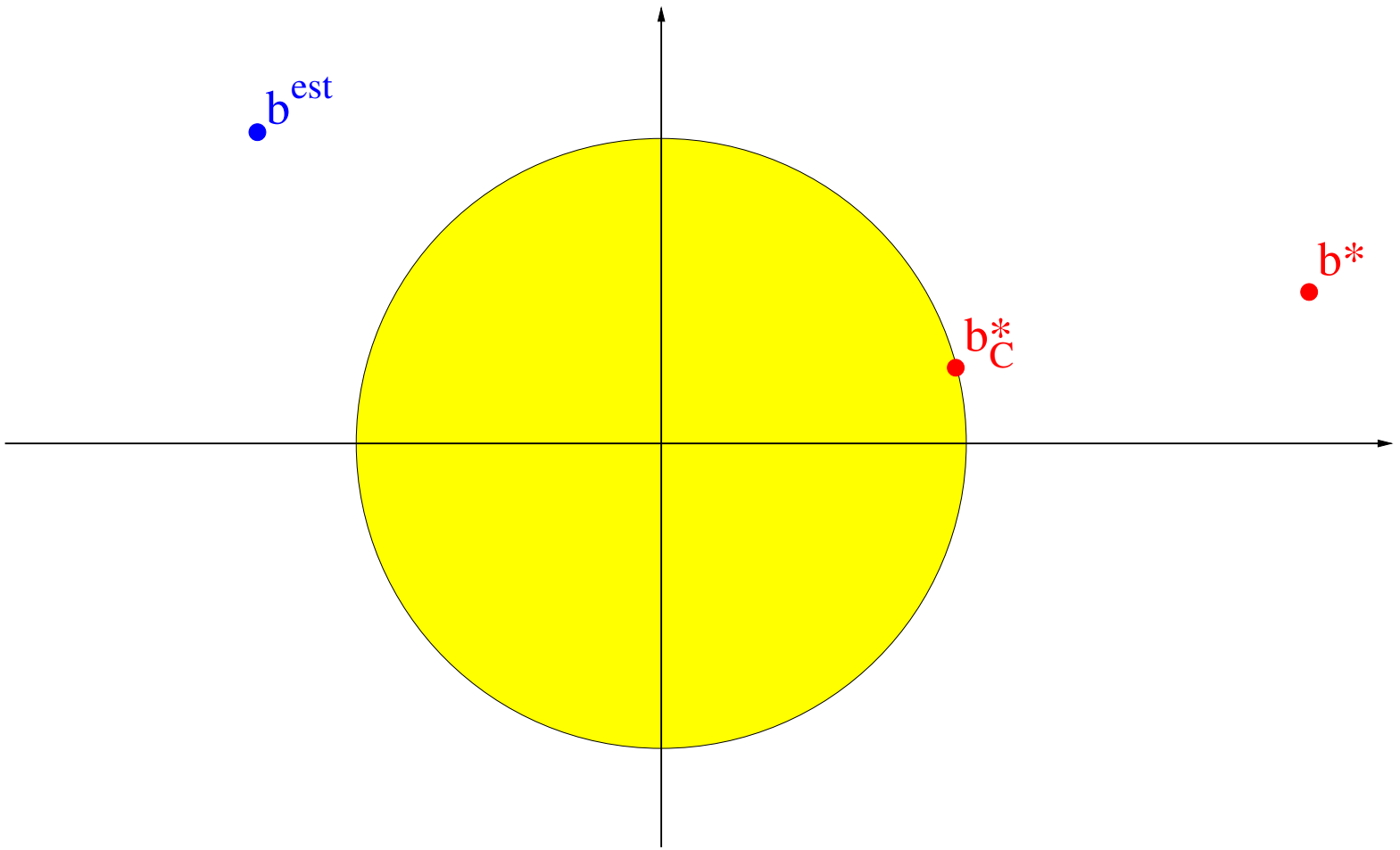
Why skrinkage classifiers?

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



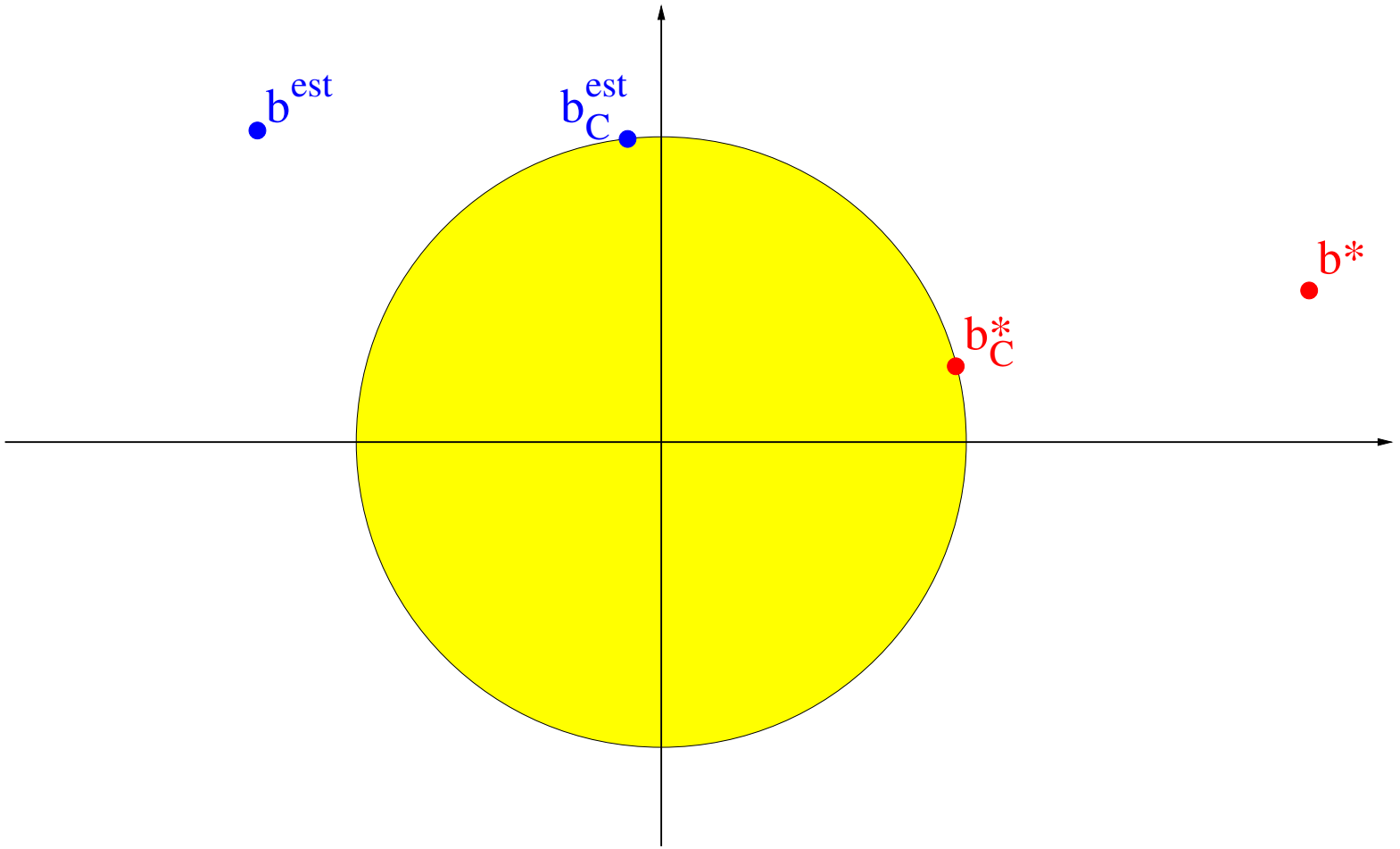
Why skrinkage classifiers?

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



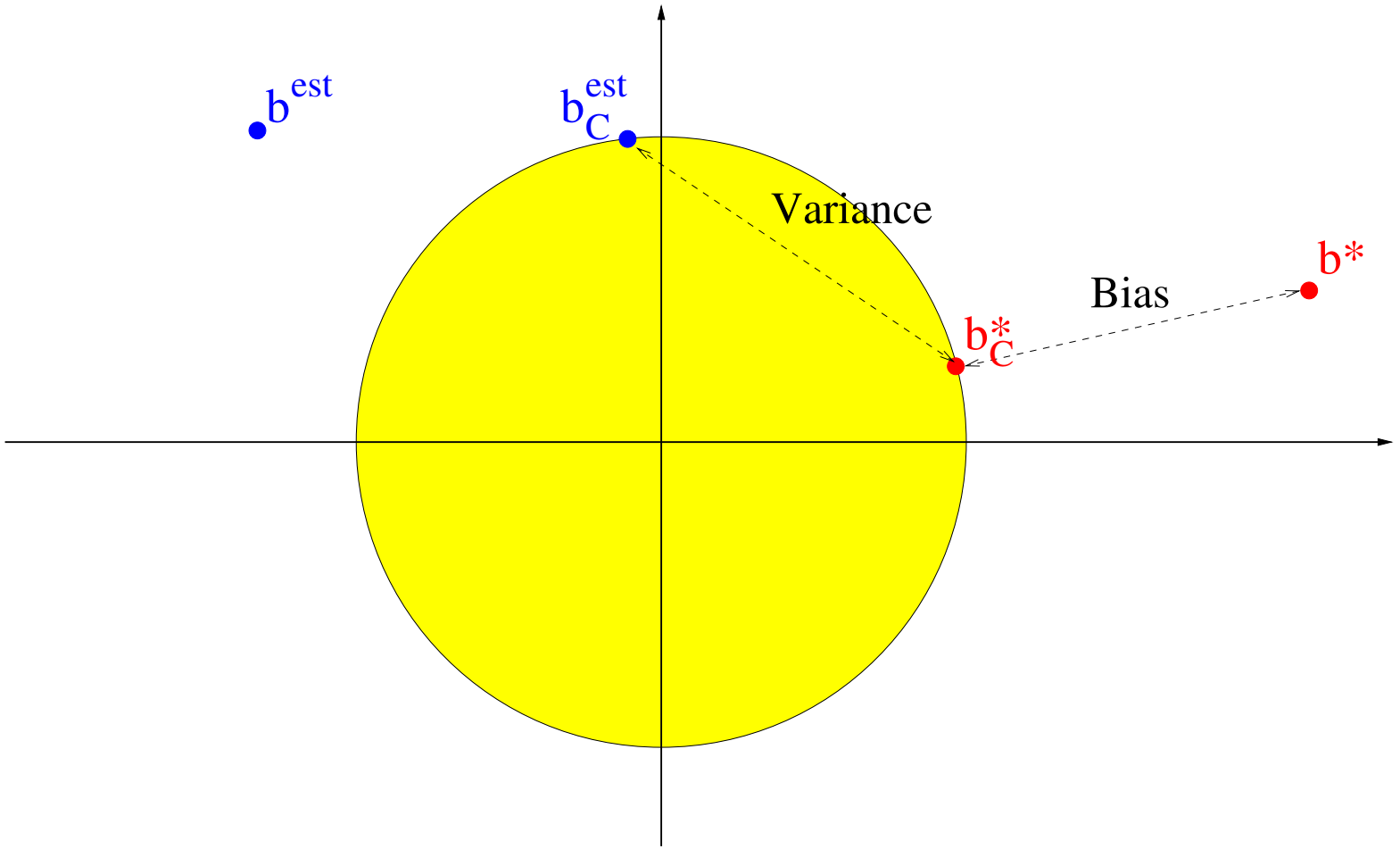
Why skrinkage classifiers?

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

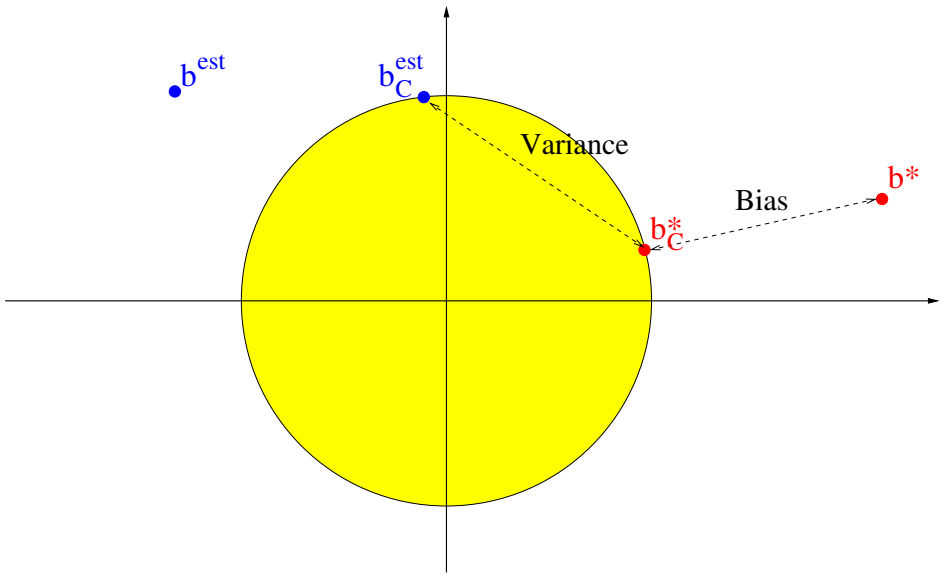


Why skrinkage classifiers?

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



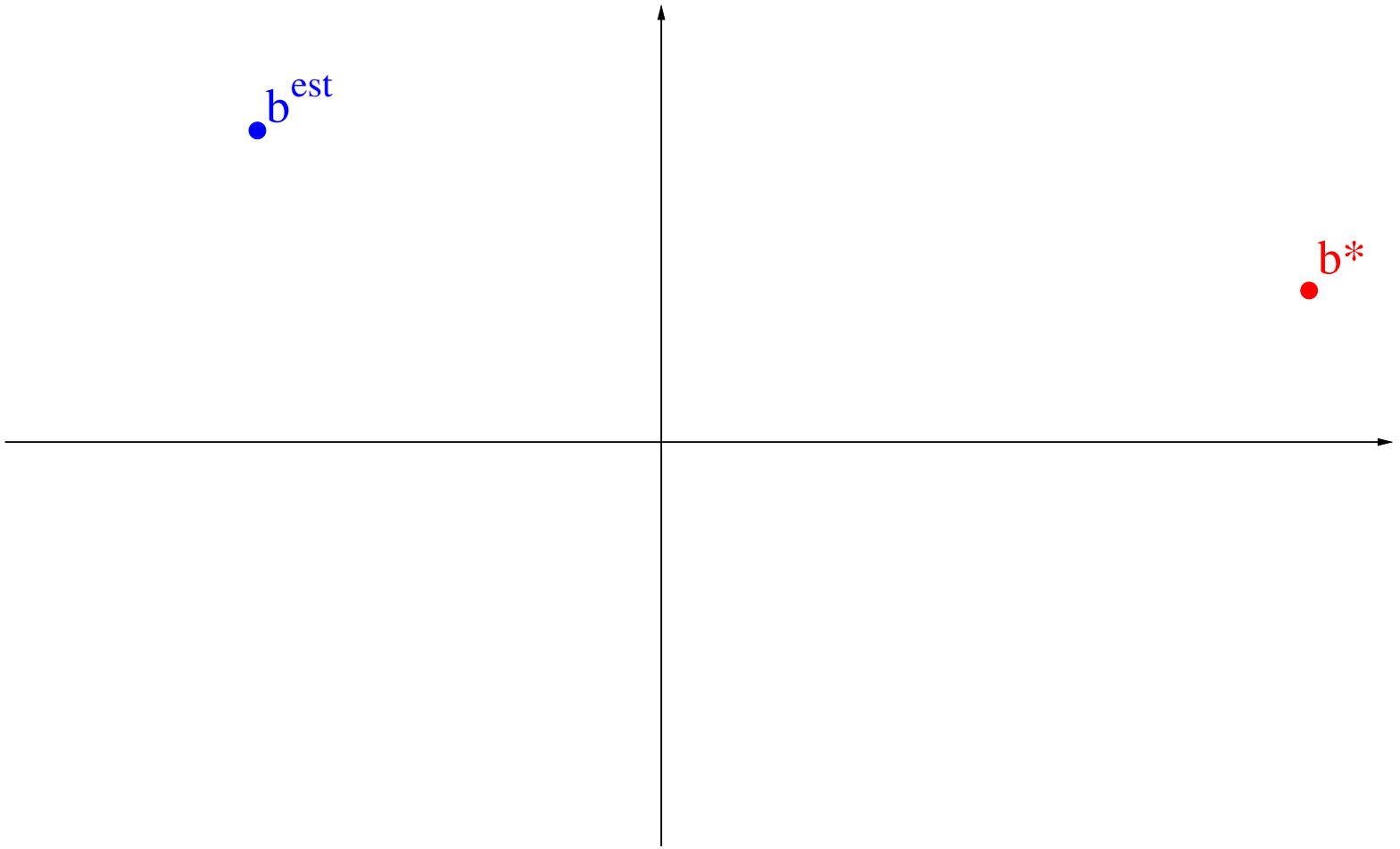
Why skrinkage classifiers?



- "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, ...)

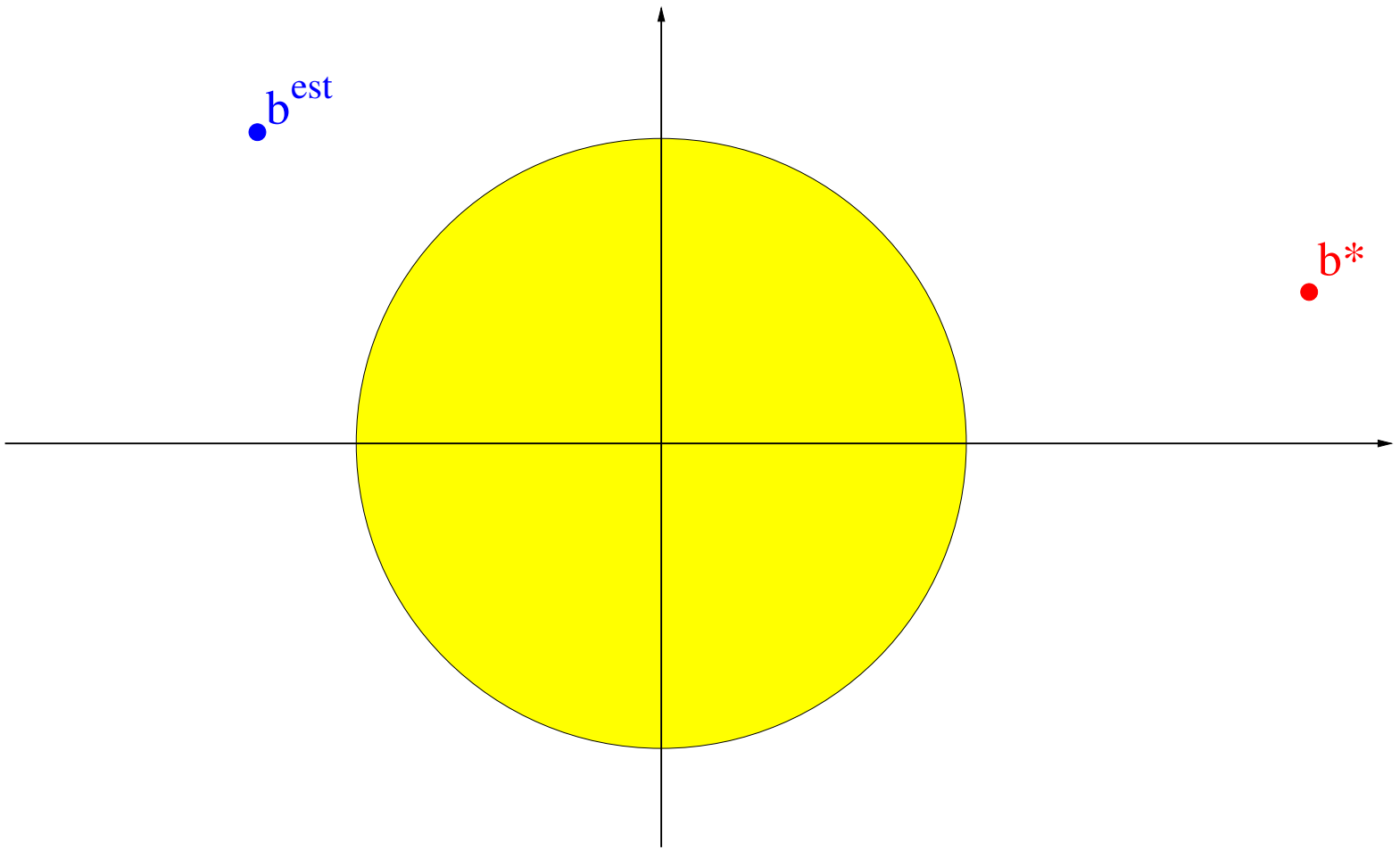
Including prior knowledge in the penalty?

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



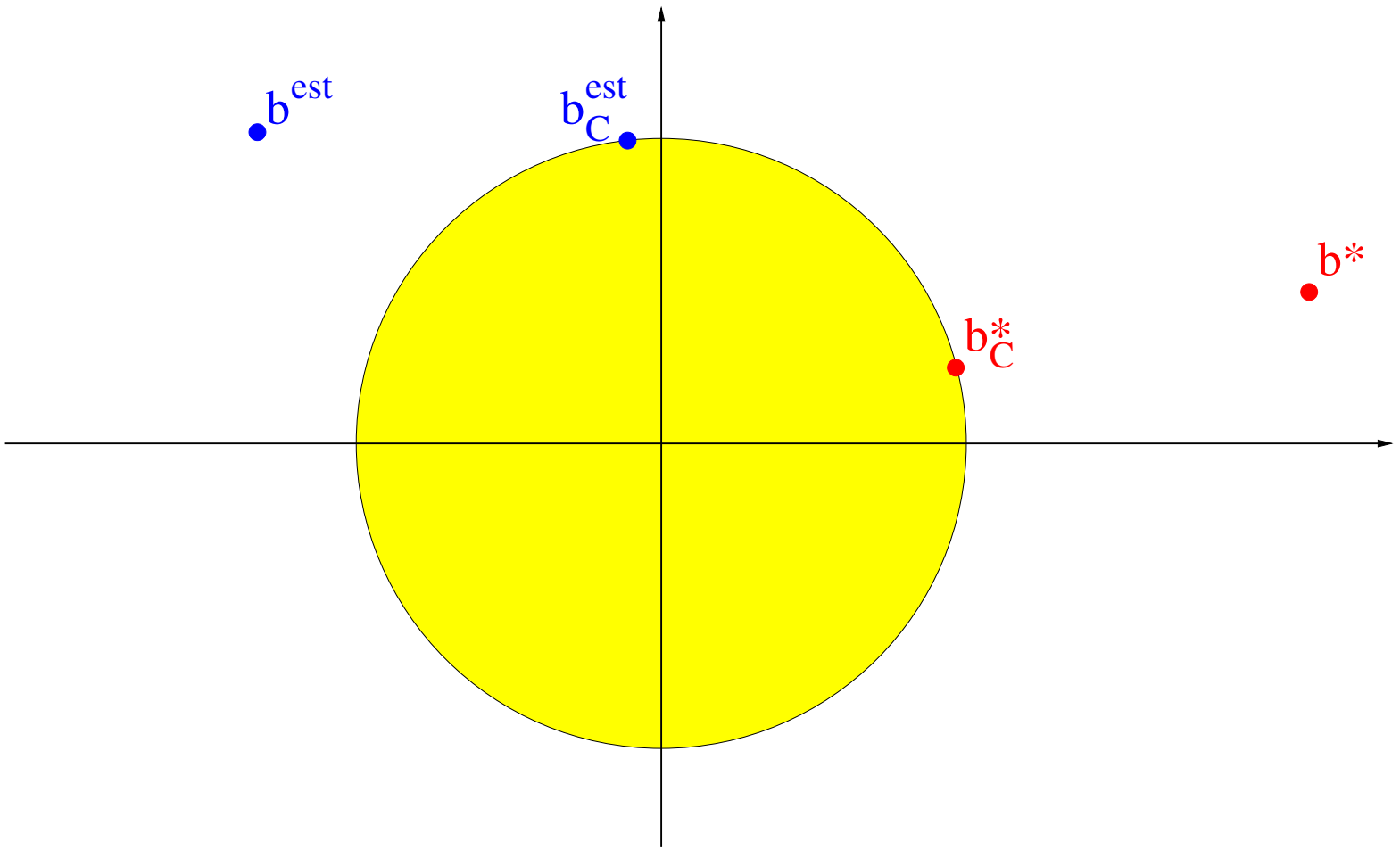
Including prior knowledge in the penalty?

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



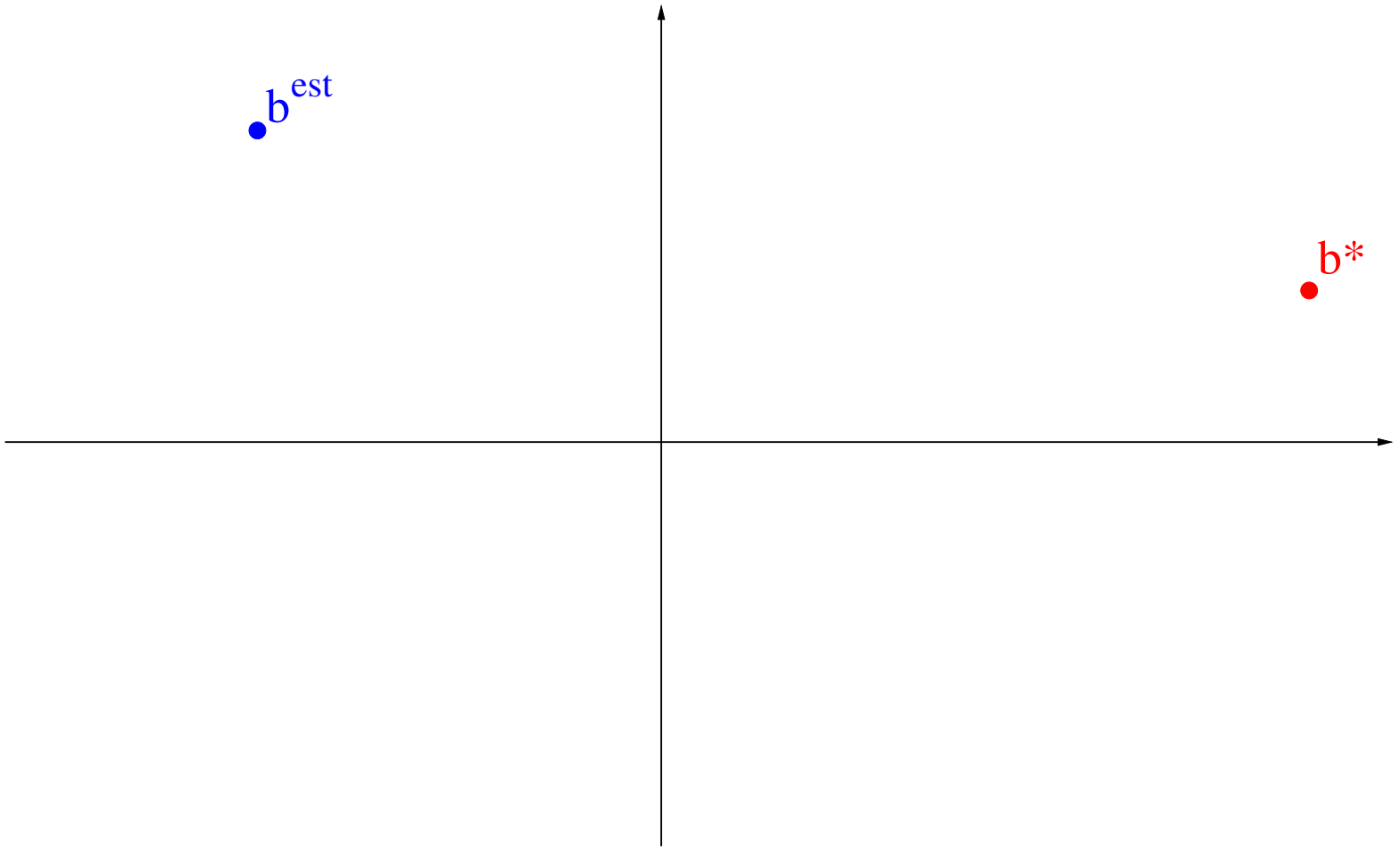
Including prior knowledge in the penalty?

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



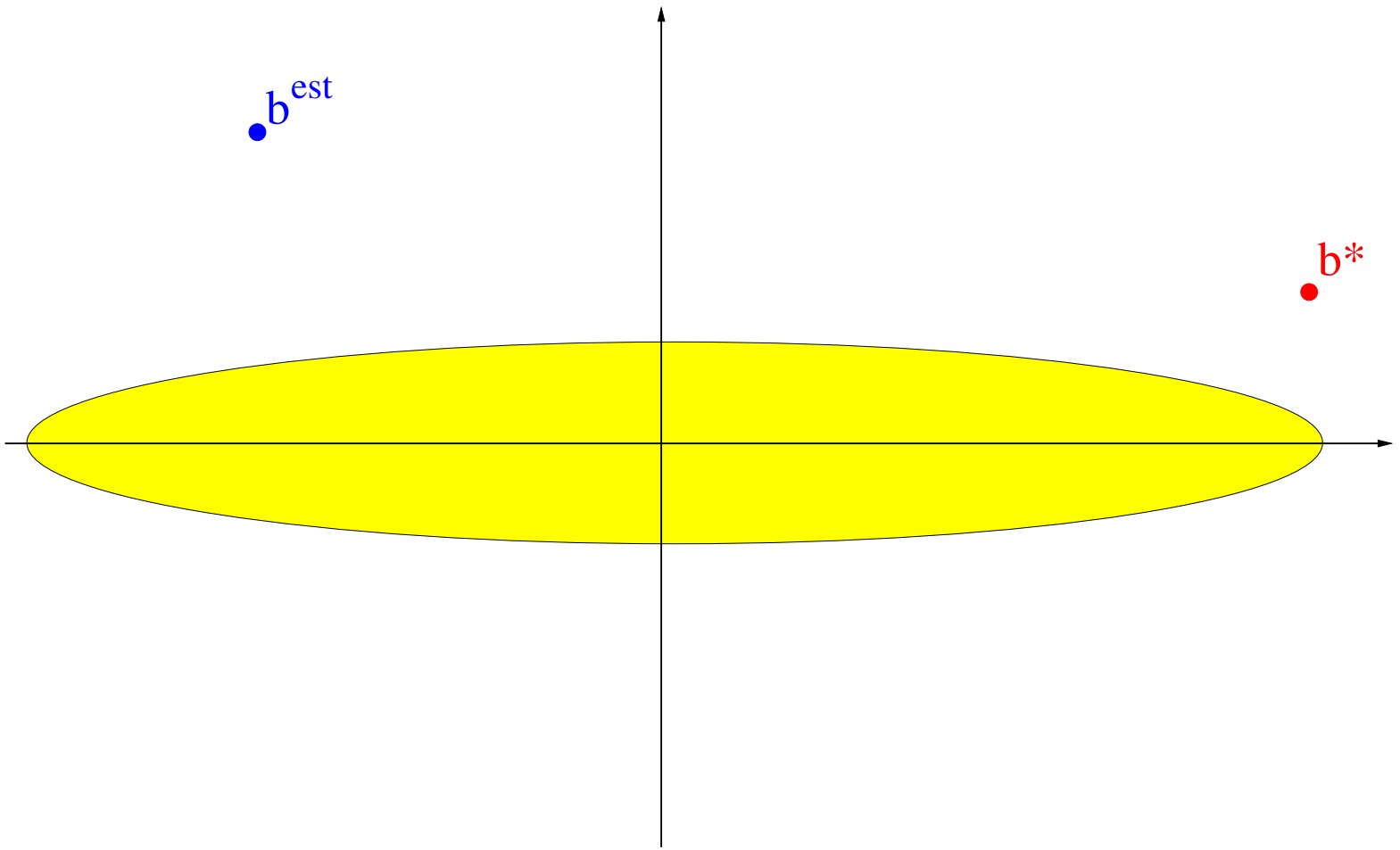
Including prior knowledge in the penalty?

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



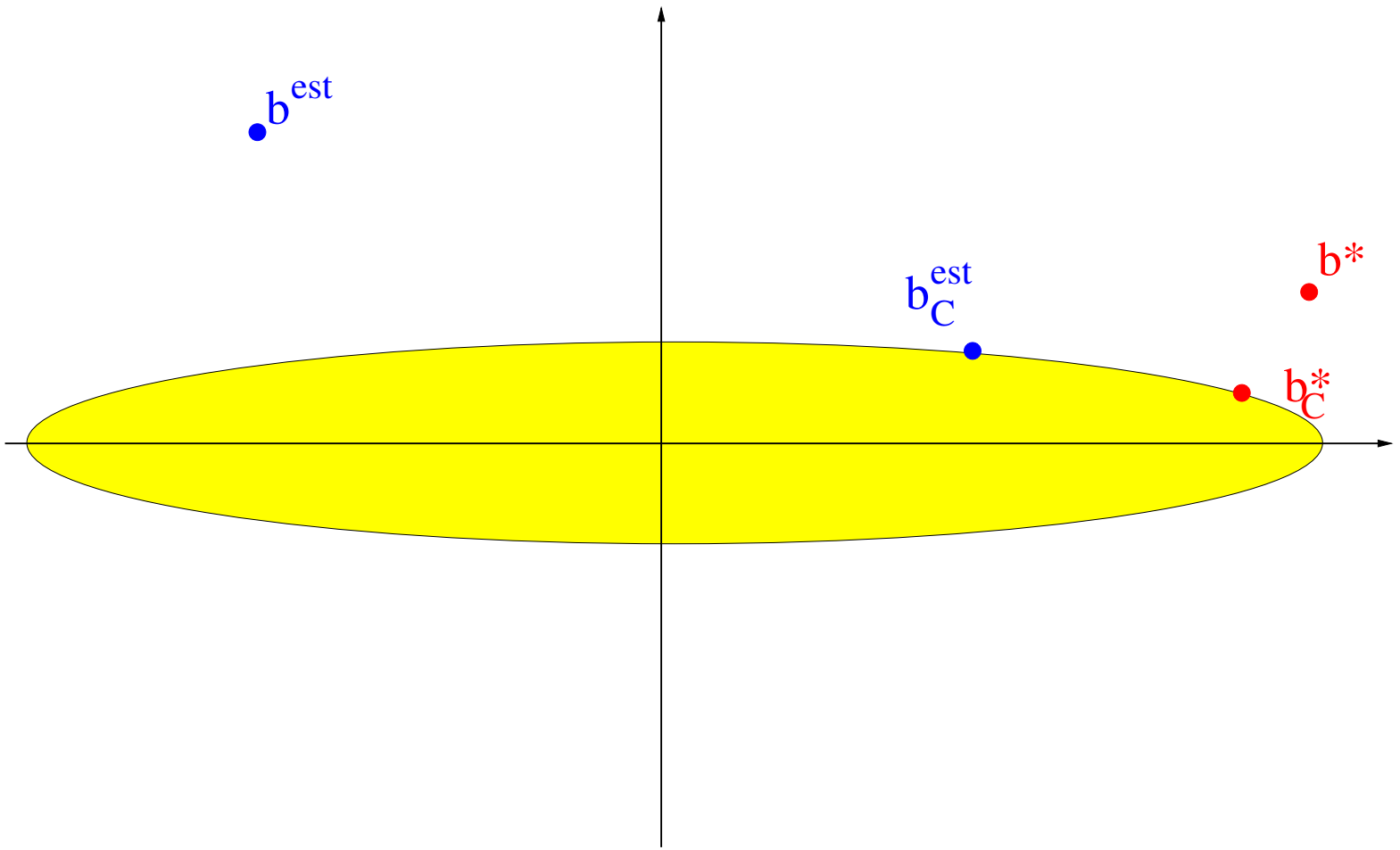
Including prior knowledge in the penalty?

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



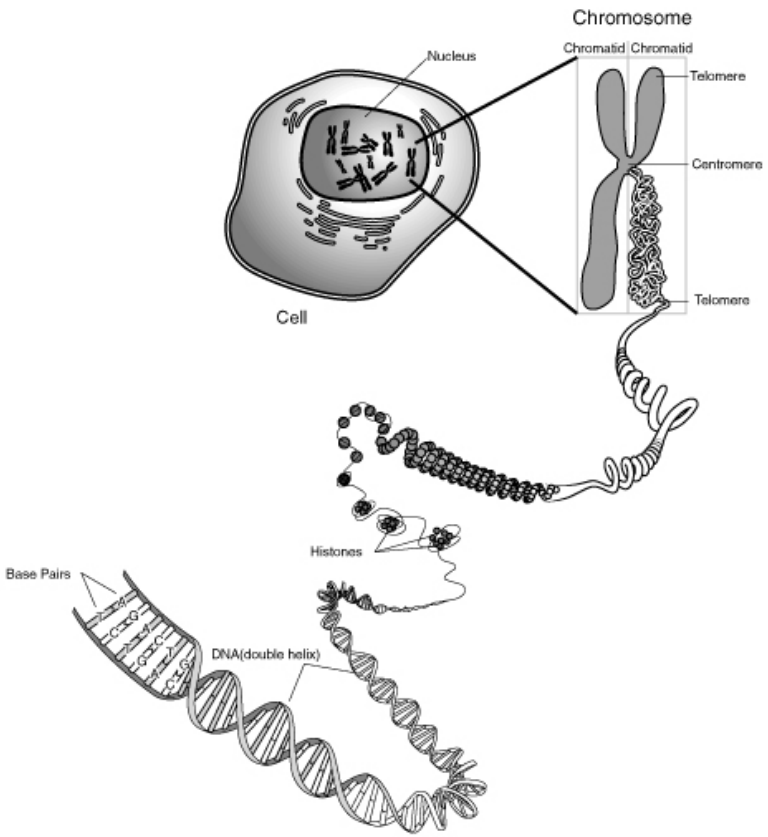
Including prior knowledge in the penalty?

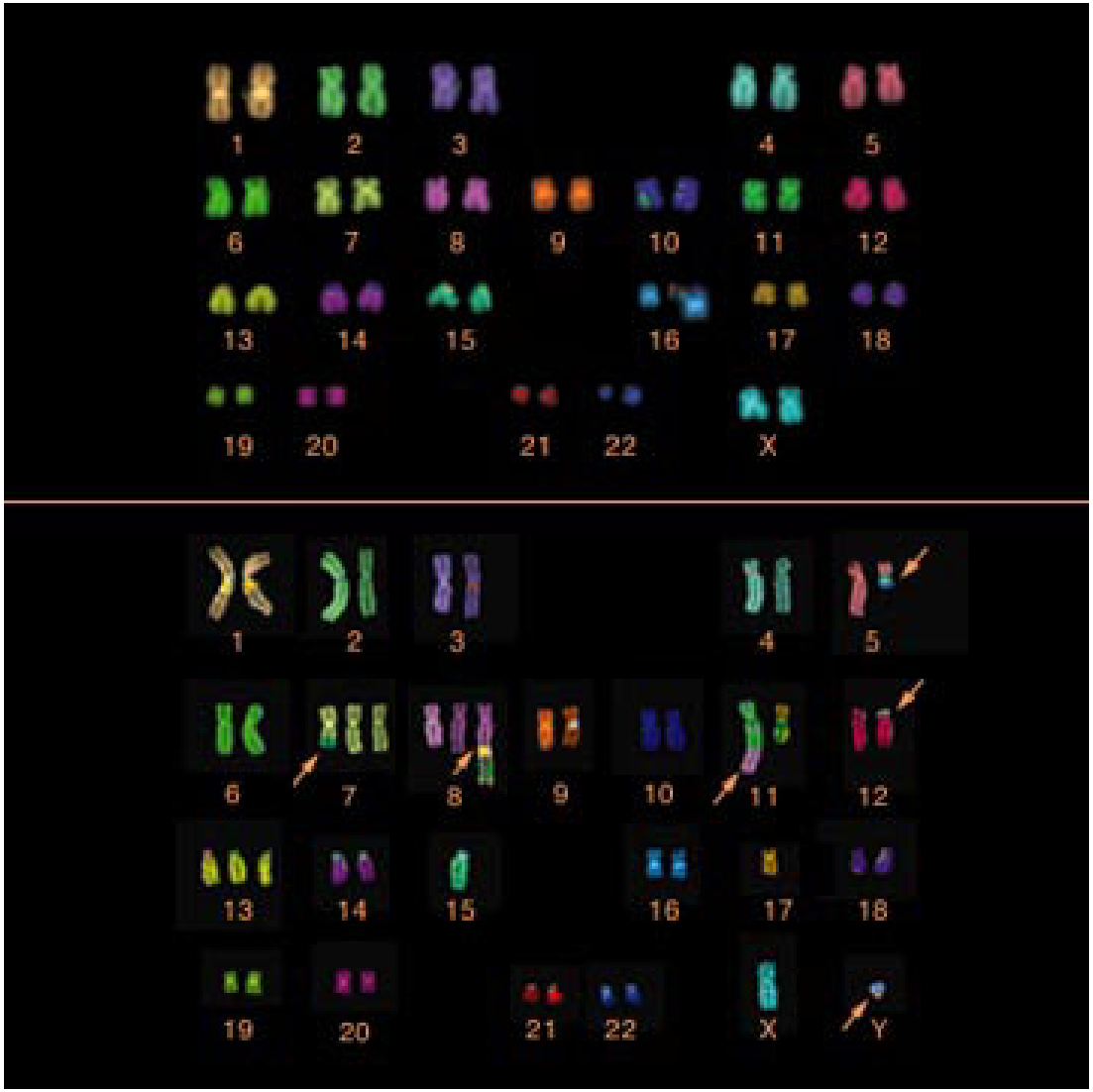
$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$



Outline

Chromosomic aberrations in cancer

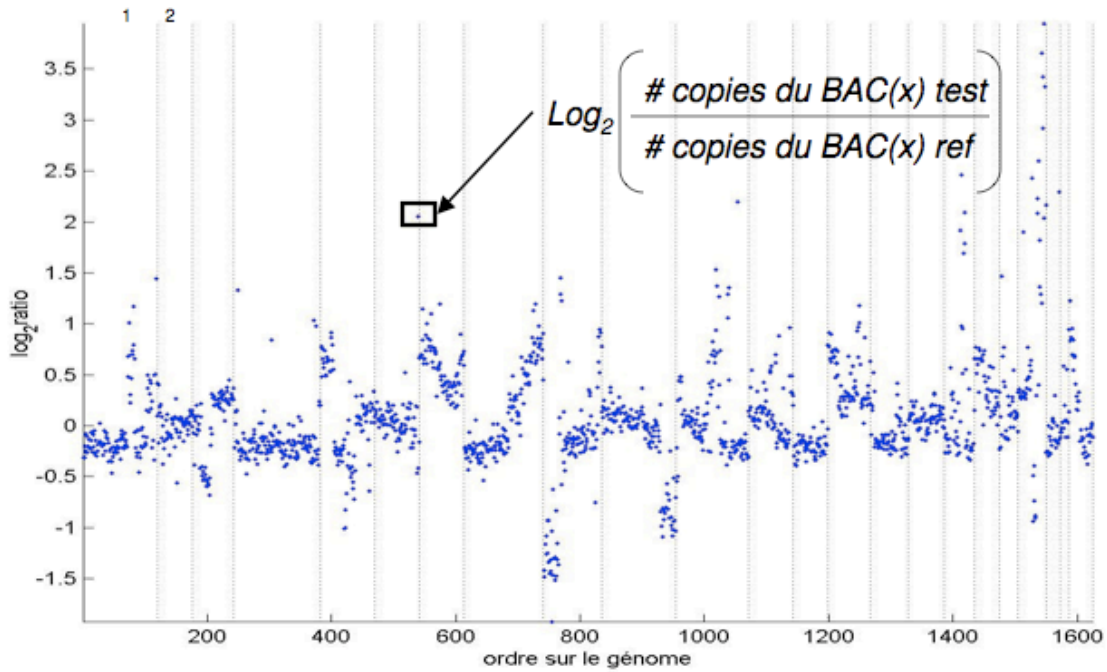




Comparative Genomic Hybridization (CGH)

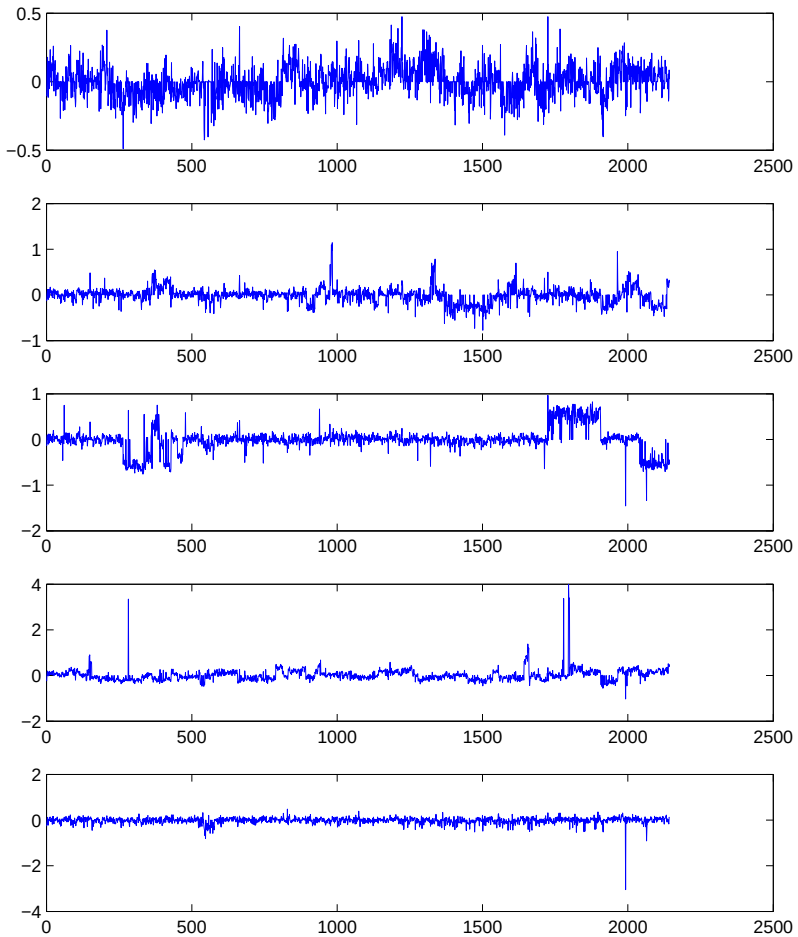
Motivation

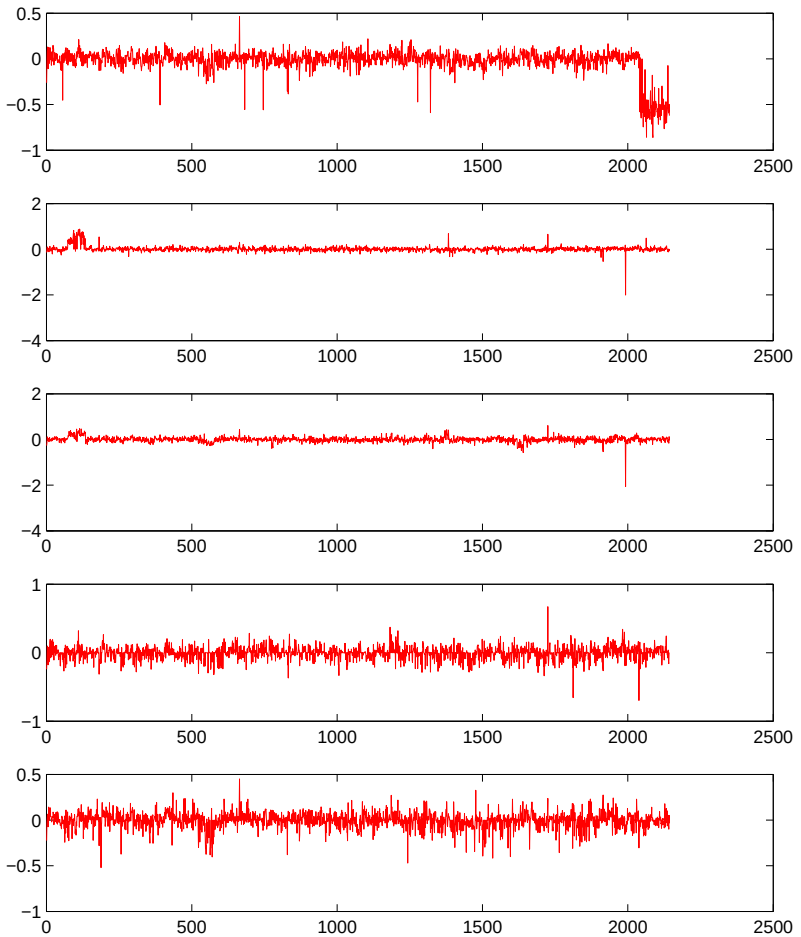
- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research
- Can we classify CGH arrays for diagnosis or prognosis purpose?



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

Aggressive vs non-aggressive melanoma





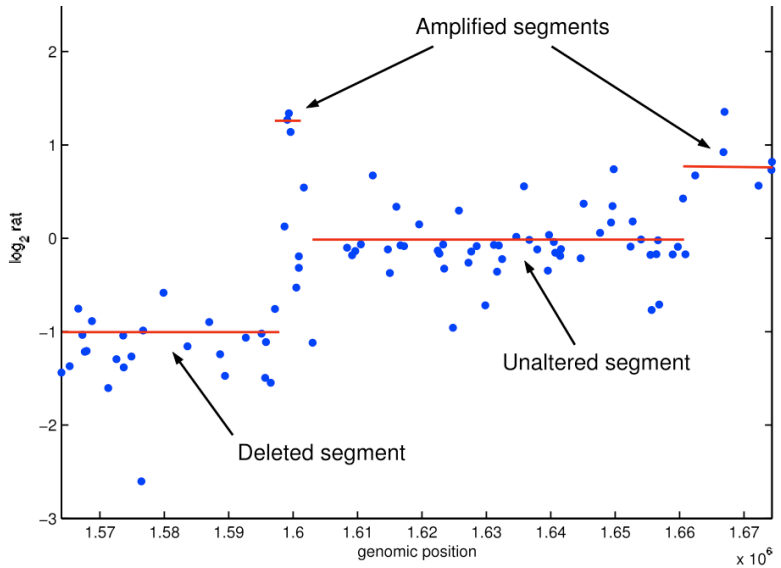
CGH array classification

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

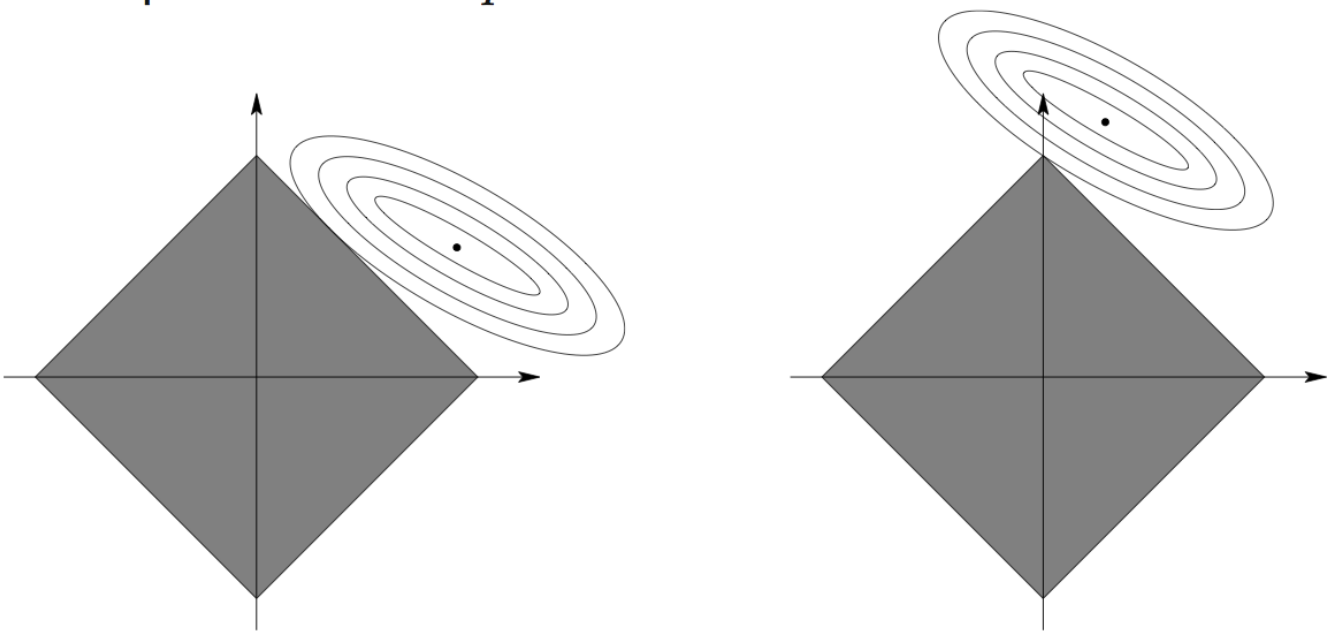
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

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

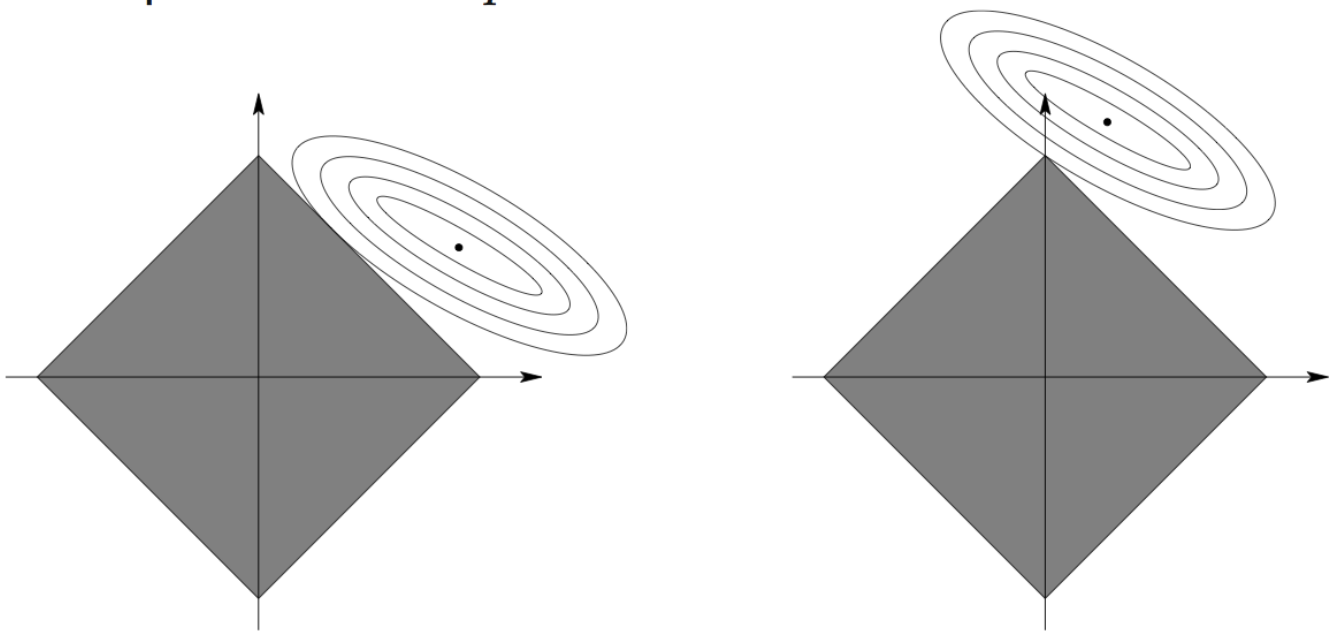
The variable fusion penalty (Land and Friedman, 1996)

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.

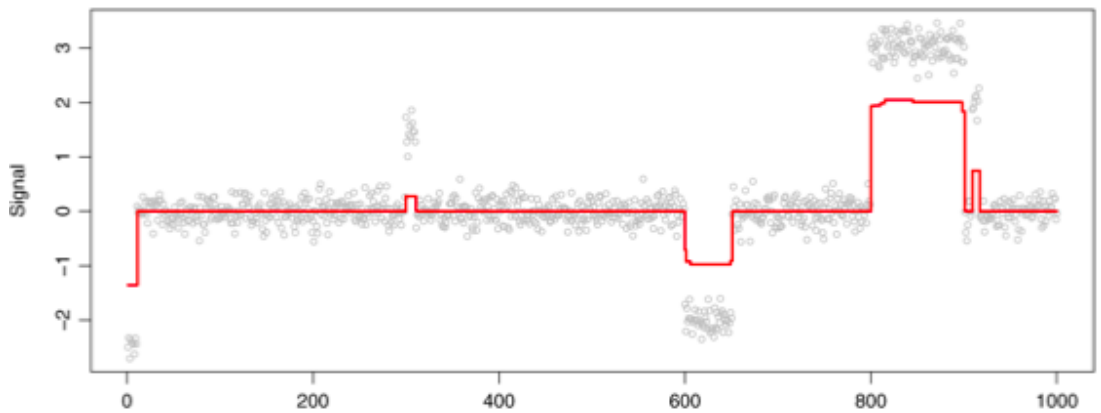
Geometric interpretation with $p = 2$



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)

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

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

Implementation

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

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

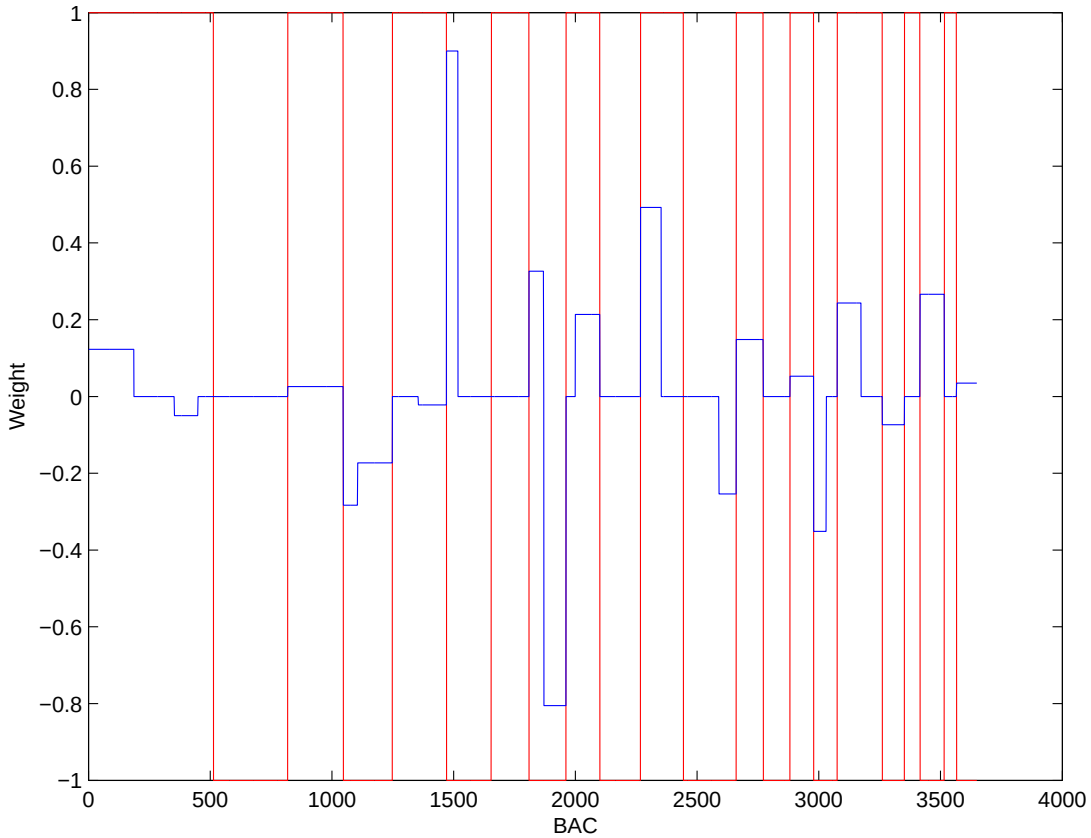
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

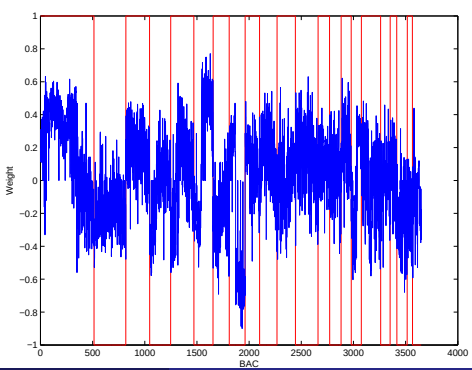
where ℓ is, e.g., the hinge loss $\ell(y, t) = max(1 - yt, 0)$.

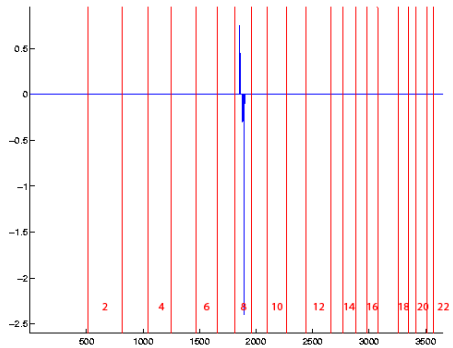
Implementation

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

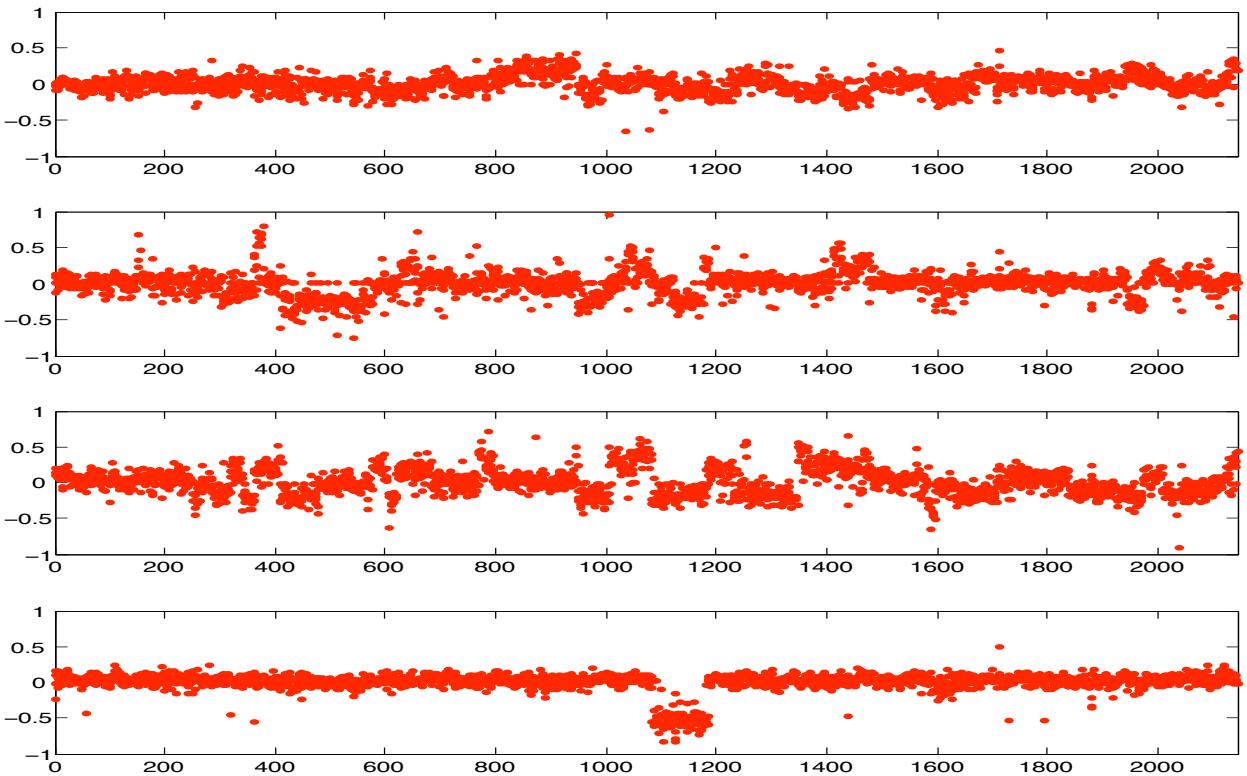
Example: predicting metastasis in melanoma





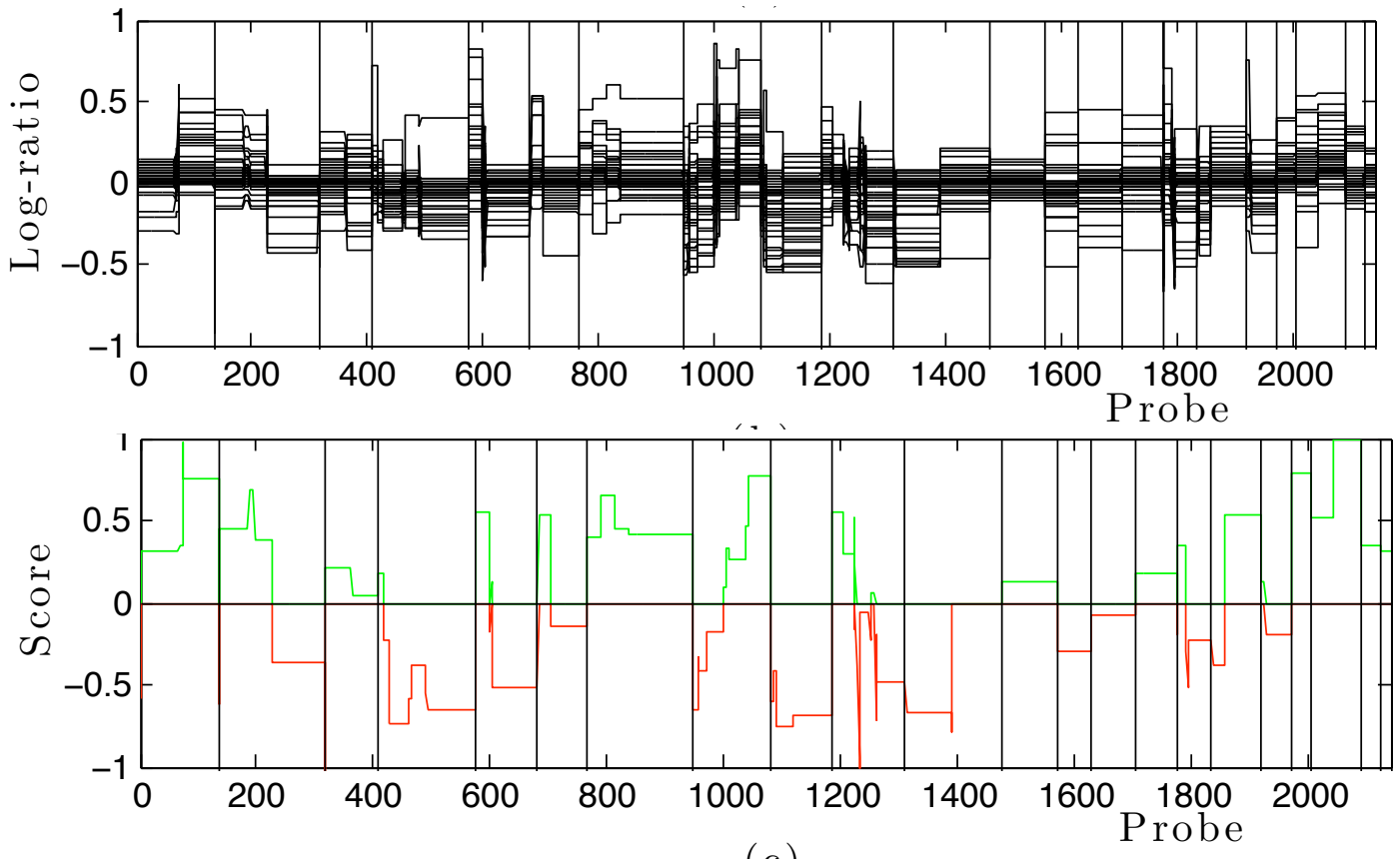


Extension: joint segmentation of many profiles



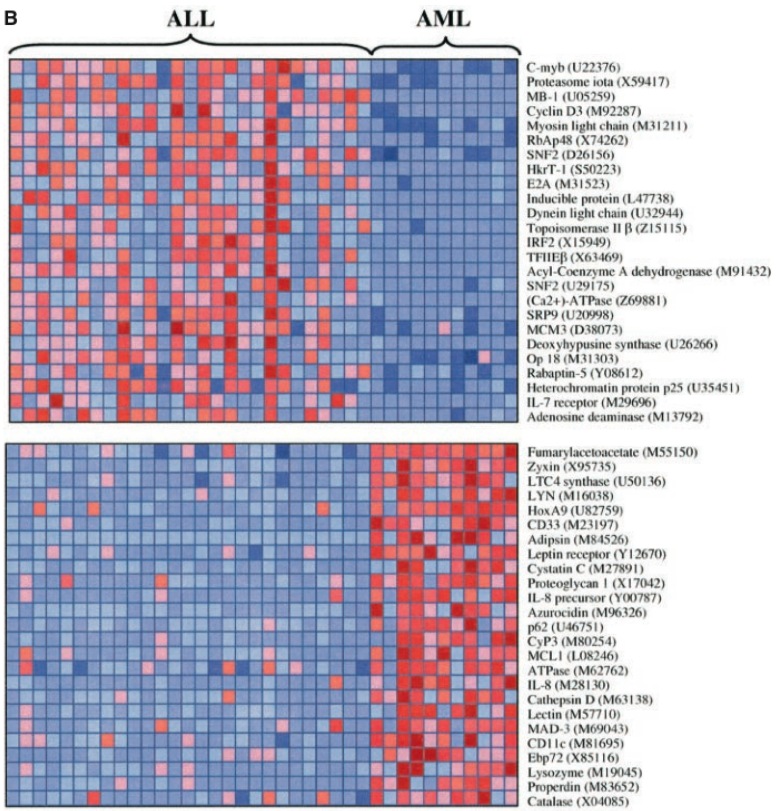
Fused group Lasso signal approximator

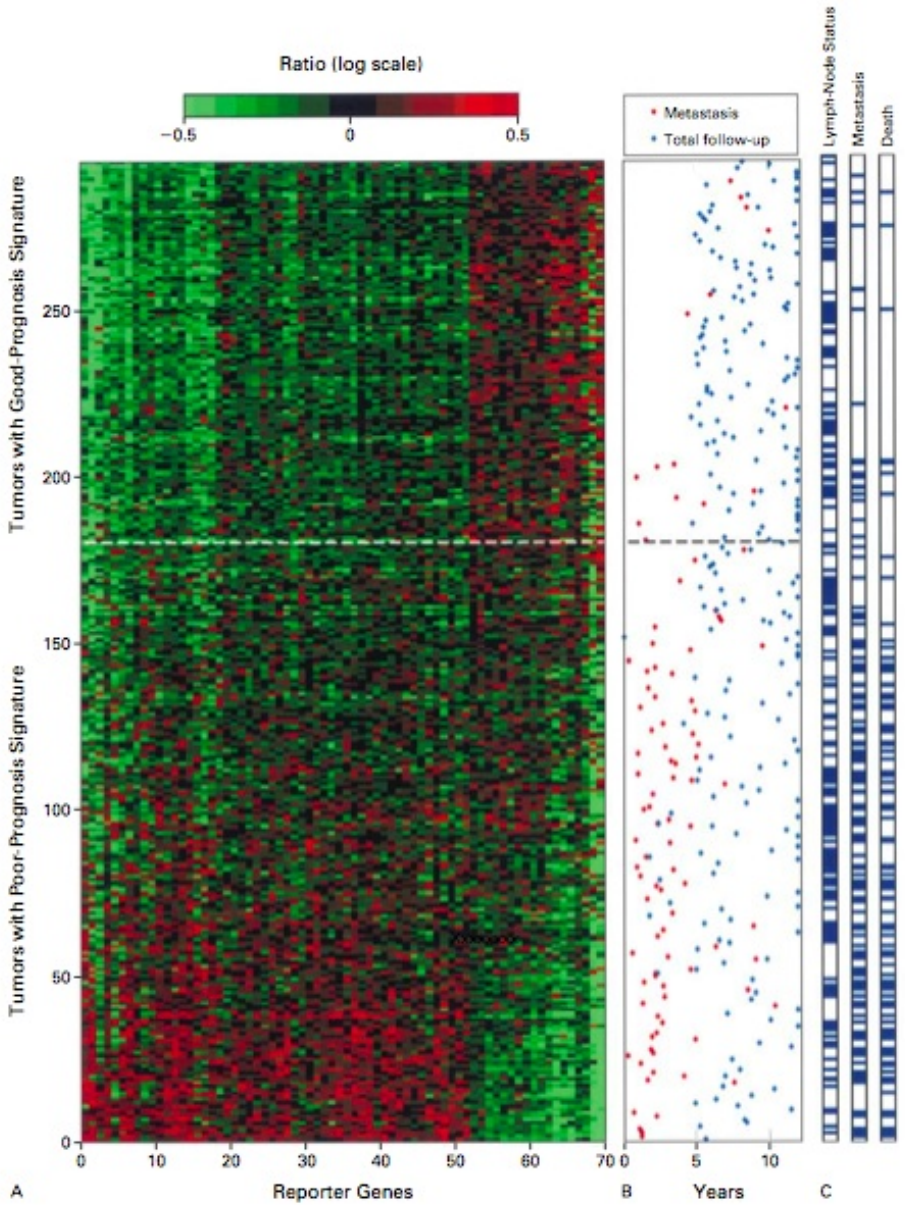
$$\min_{\beta \in \mathbb{R}^{n \times p}} \|Y - \beta\|^2 + \lambda \sum_{i=1}^{p-1} \|\beta_{i+1} - \beta_i\|$$



Outline

Molecular diagnosis / prognosis / theragnosis





Gene selection, signature

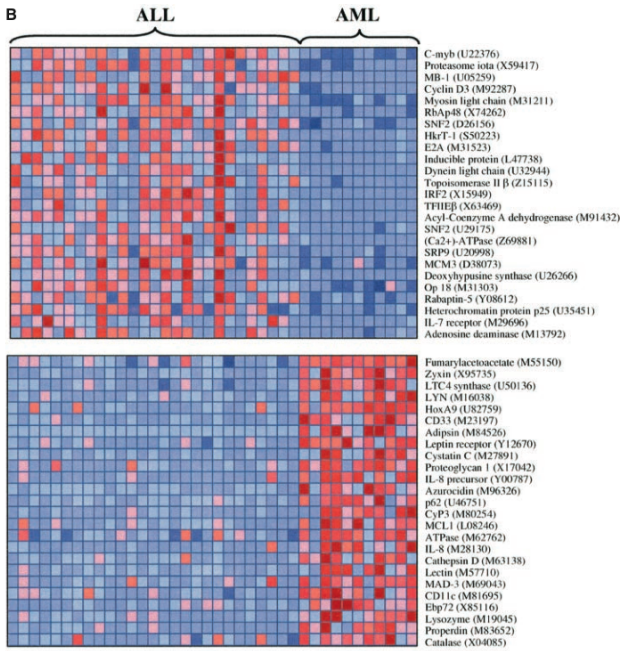
The idea

- 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.
- Satistics: this is one way to constrain the solution and reduce the complexity to allow learning.

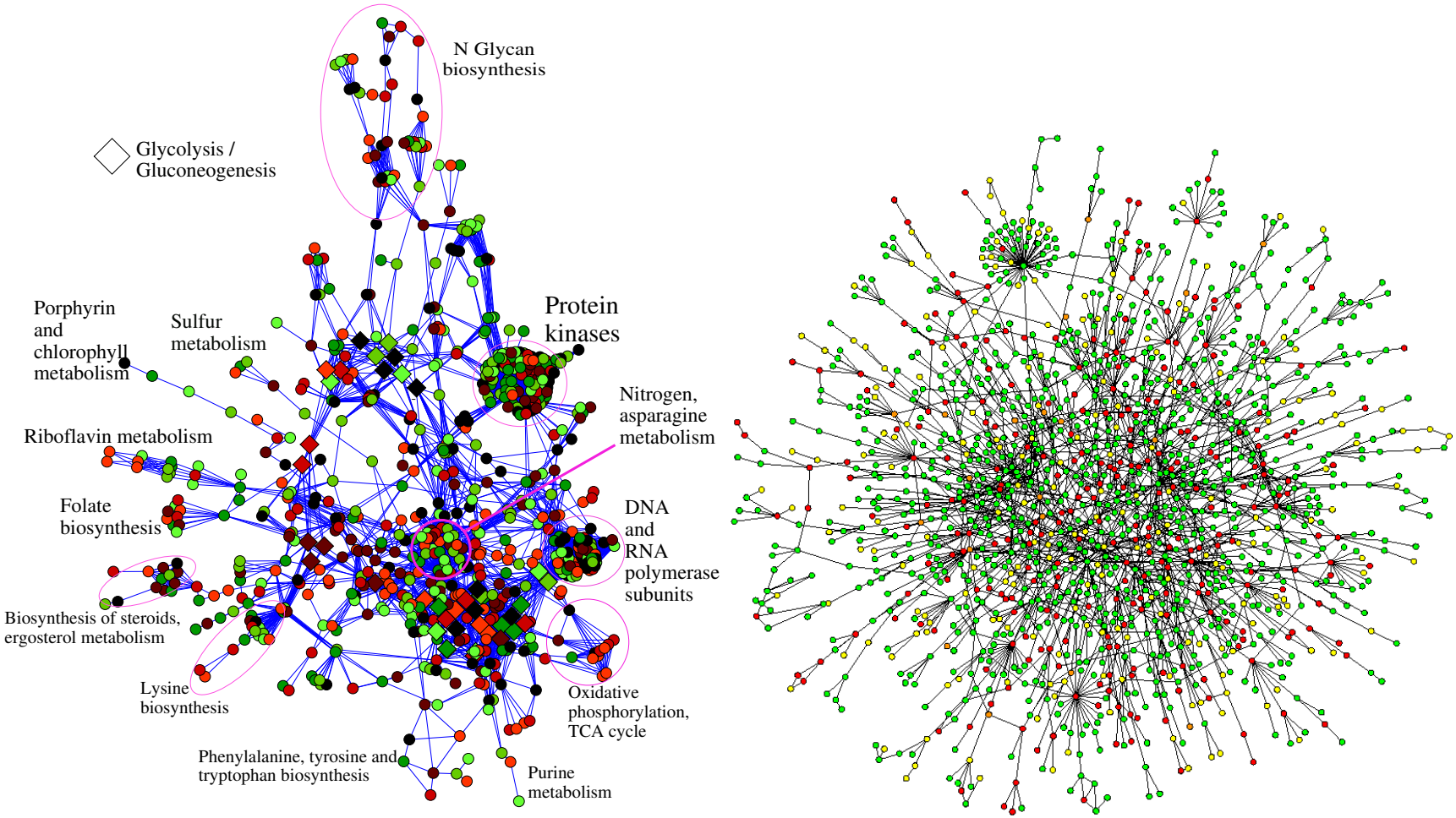
But...



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



Graph based penalty

Prior hypothesis

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

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

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

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

Graph based penalty

Prior hypothesis

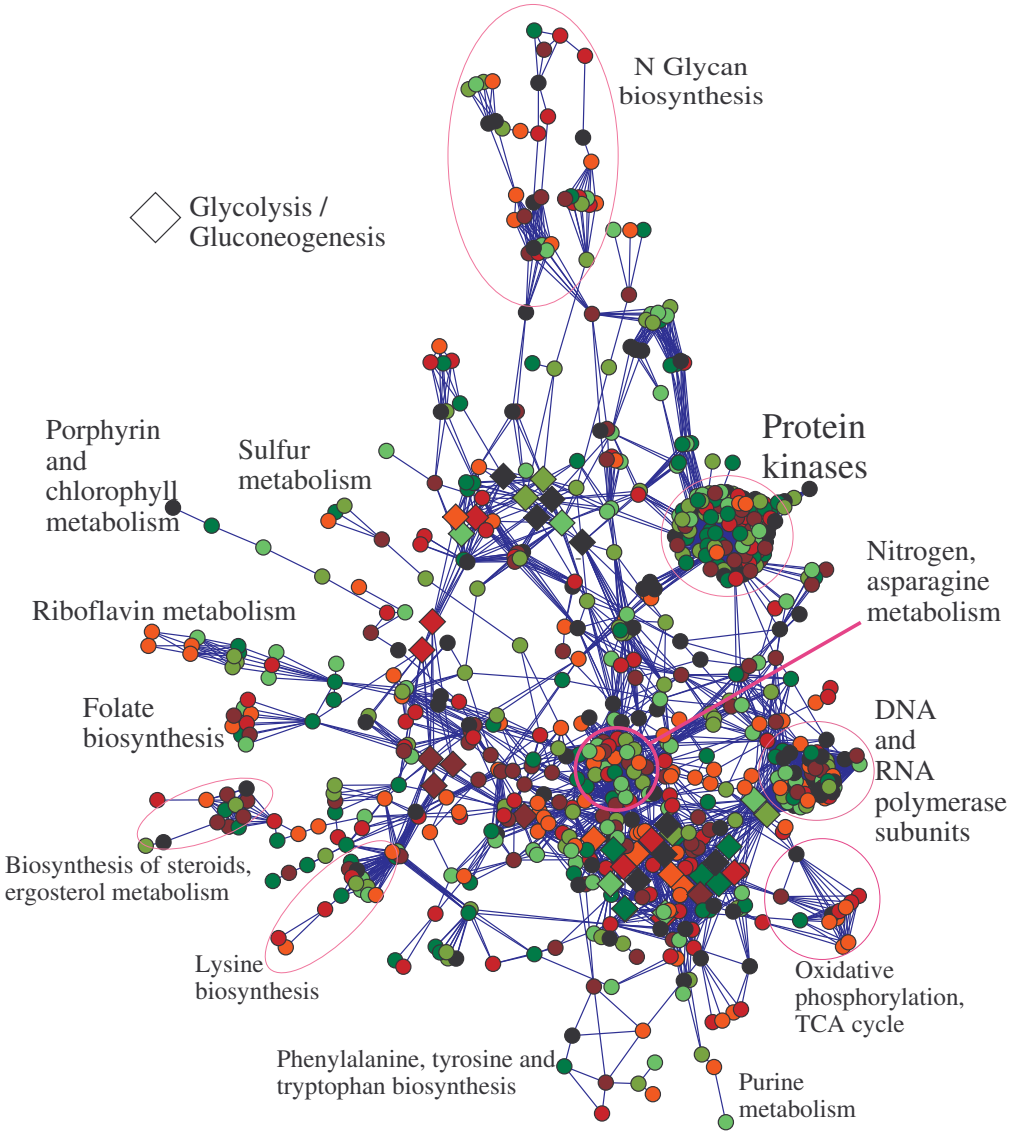
Genes near each other on the graph should have similar weigths.

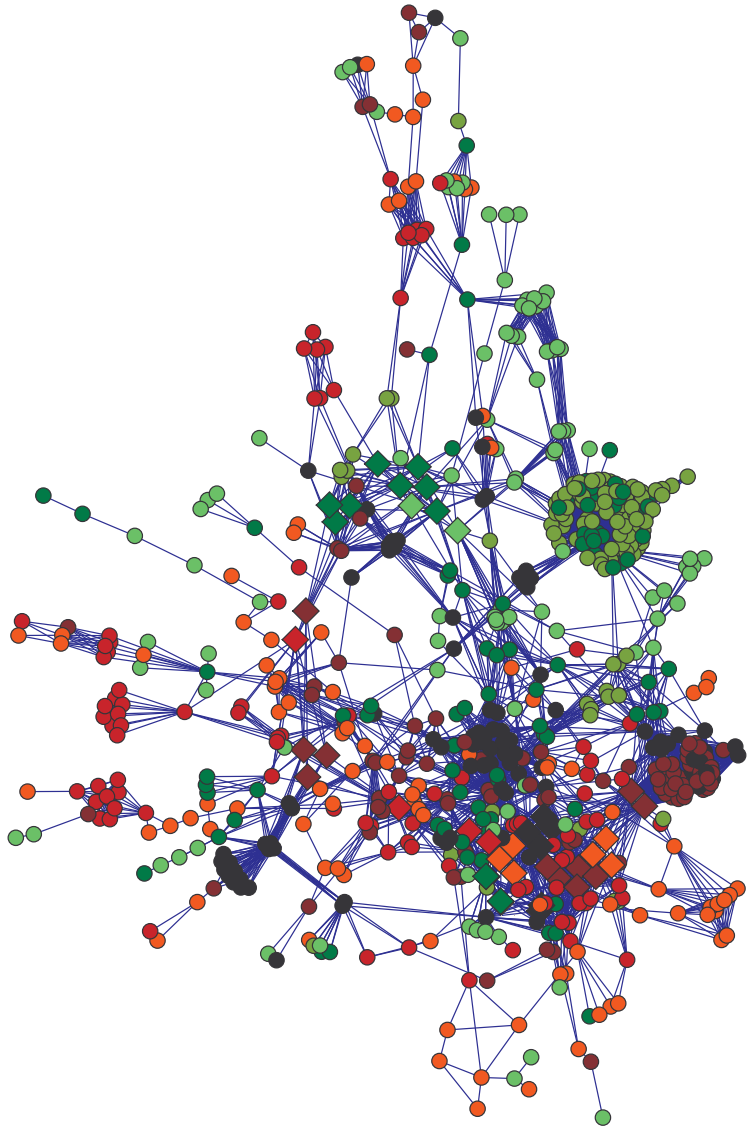
Two solutions (Rapaport et al., 2007, 2008)

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

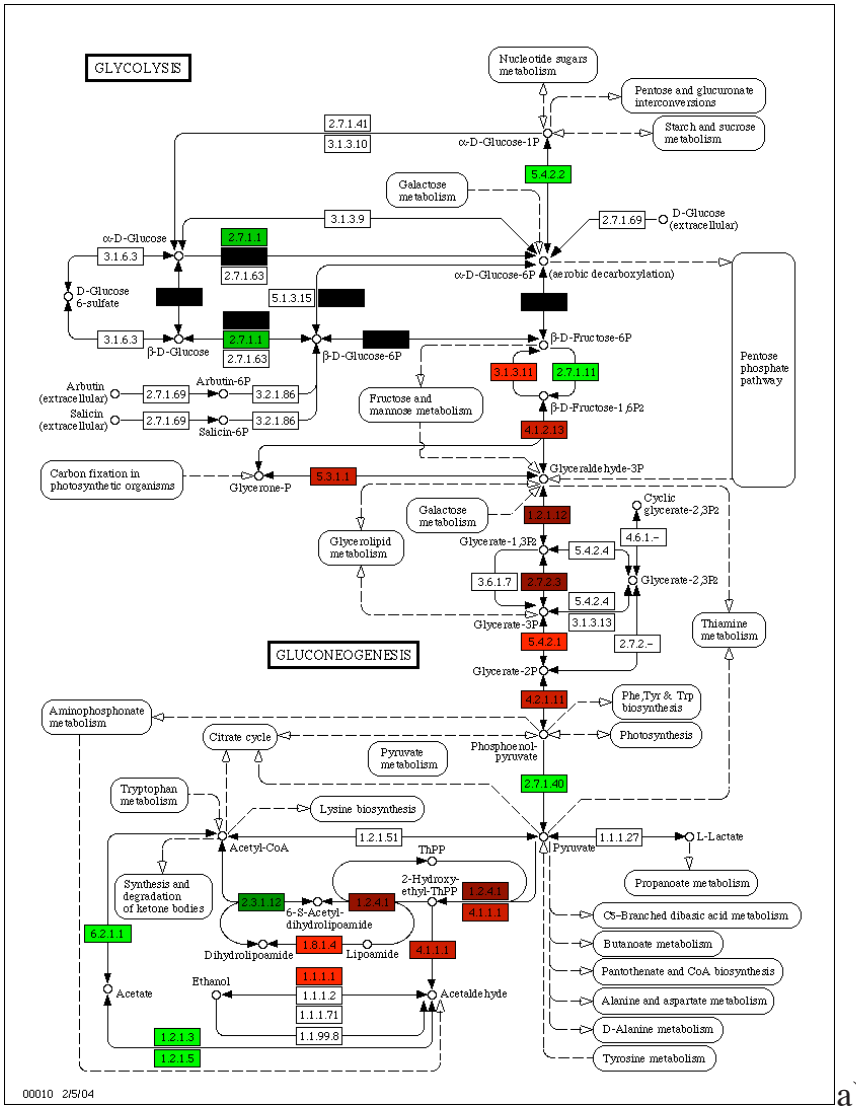
$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Classifiers

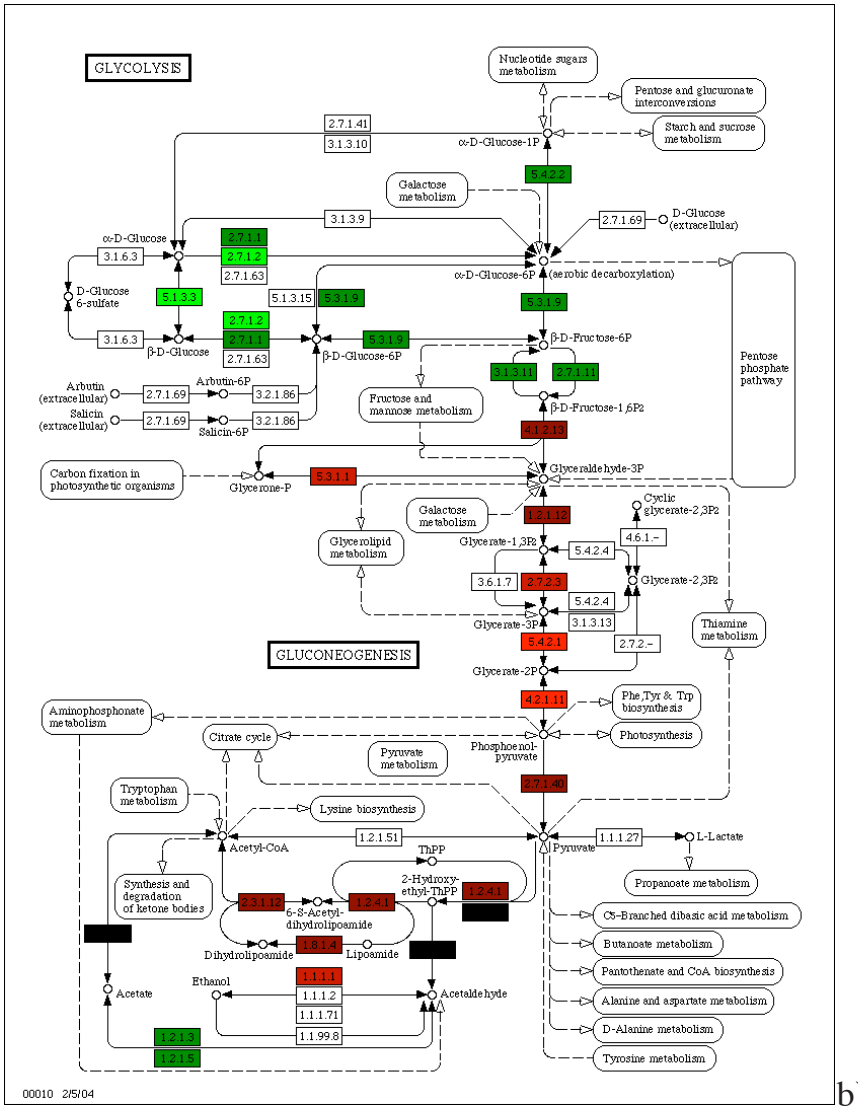




Classifiers

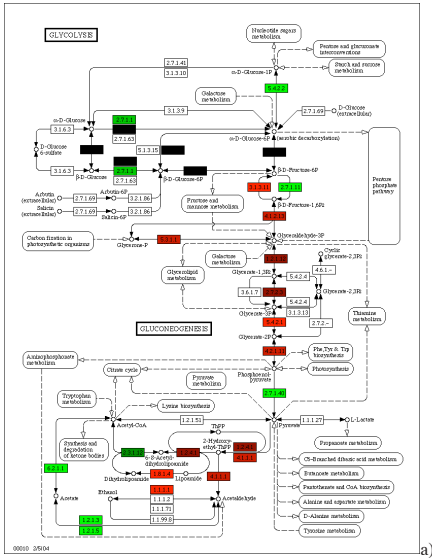


a)

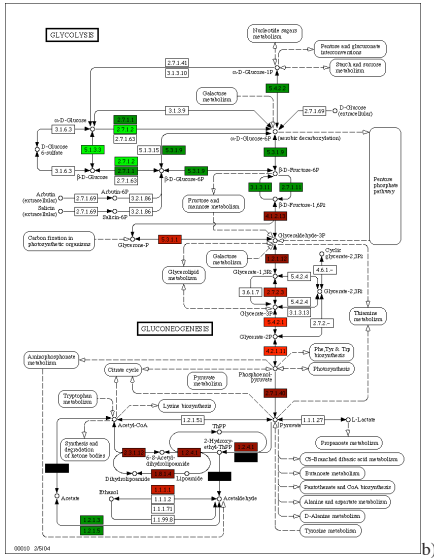


b)

Limits



a)



b)

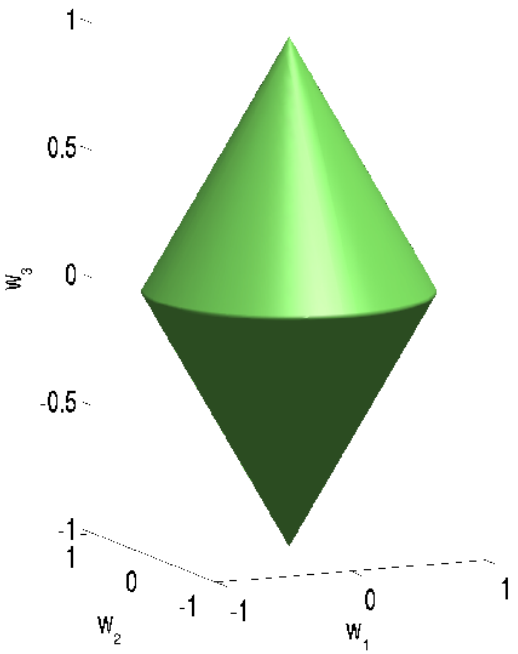
- We are happy to see pathways appear.
- However, in some cases, connected genes should have "opposite" weights (inhibition, pathway branching, etc...)
- How to capture pathways without constraints on the weight similarities?

Selecting pre-defined groups of variables

Group lasso (Yuan & Lin, 2006)

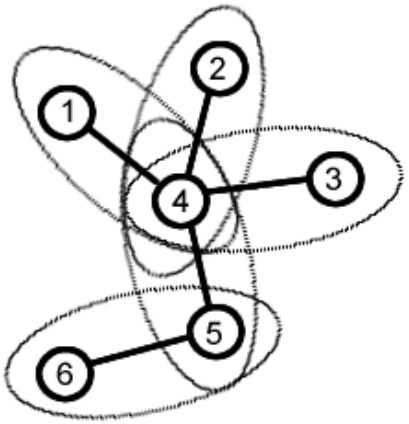
If groups of covariates are likely to be selected together, the ℓ_1/ℓ_2-norm induces sparse solutions *at the group level*:

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



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

Graph lasso

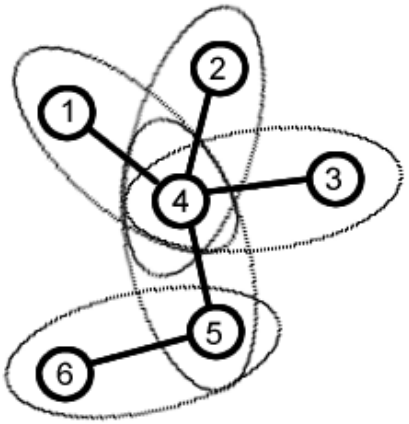


- Hypothesis: selected genes should form connected components on the graph
- Two solutions (Jacob et al., 2009):

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

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

Graph lasso

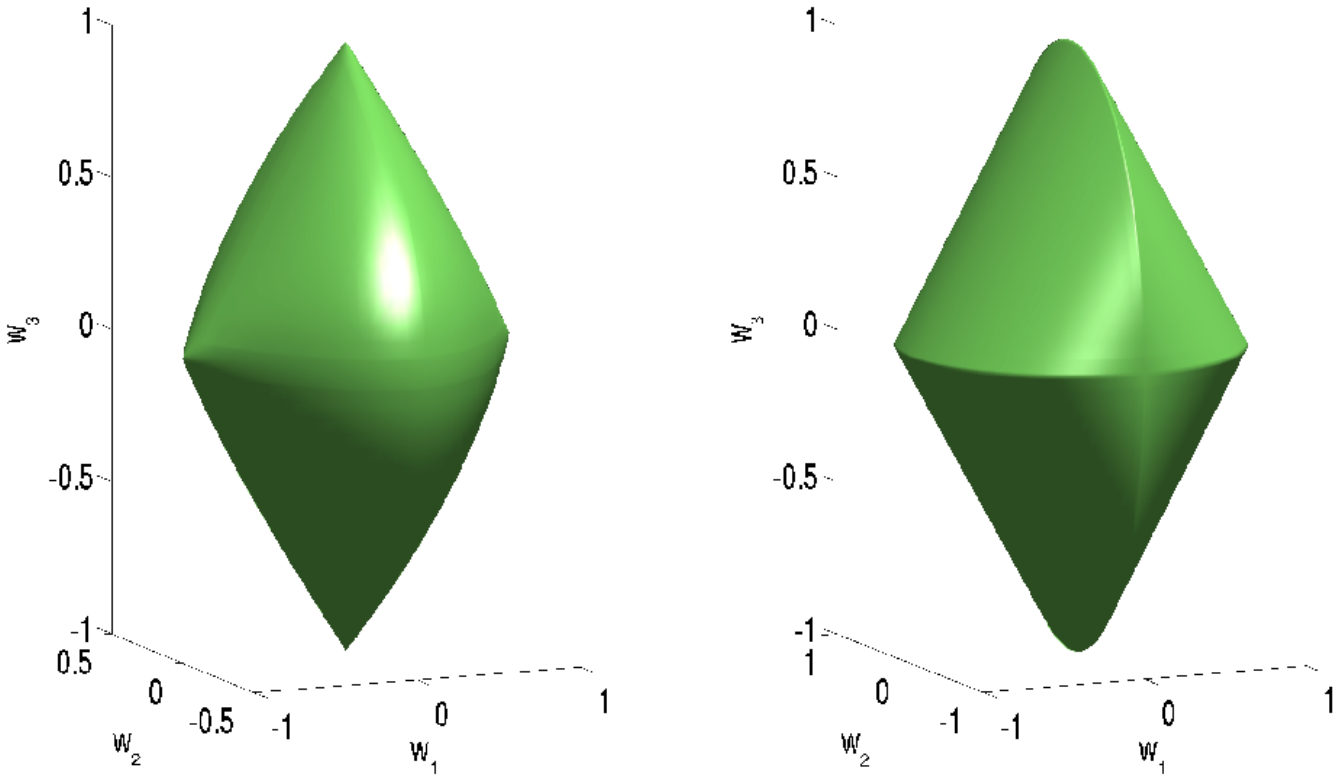


- Hypothesis: selected genes should form connected components on the graph
- Two solutions (Jacob et al., 2009):

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

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

Overlap and group unity balls



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

Summary: Graph lasso vs kernel

- Graph lasso:

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

  constrains the sparsity, not the values

- Graph kernel

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

  constrains the values (smoothness), not the sparsity

Preliminary results

Breast cancer data

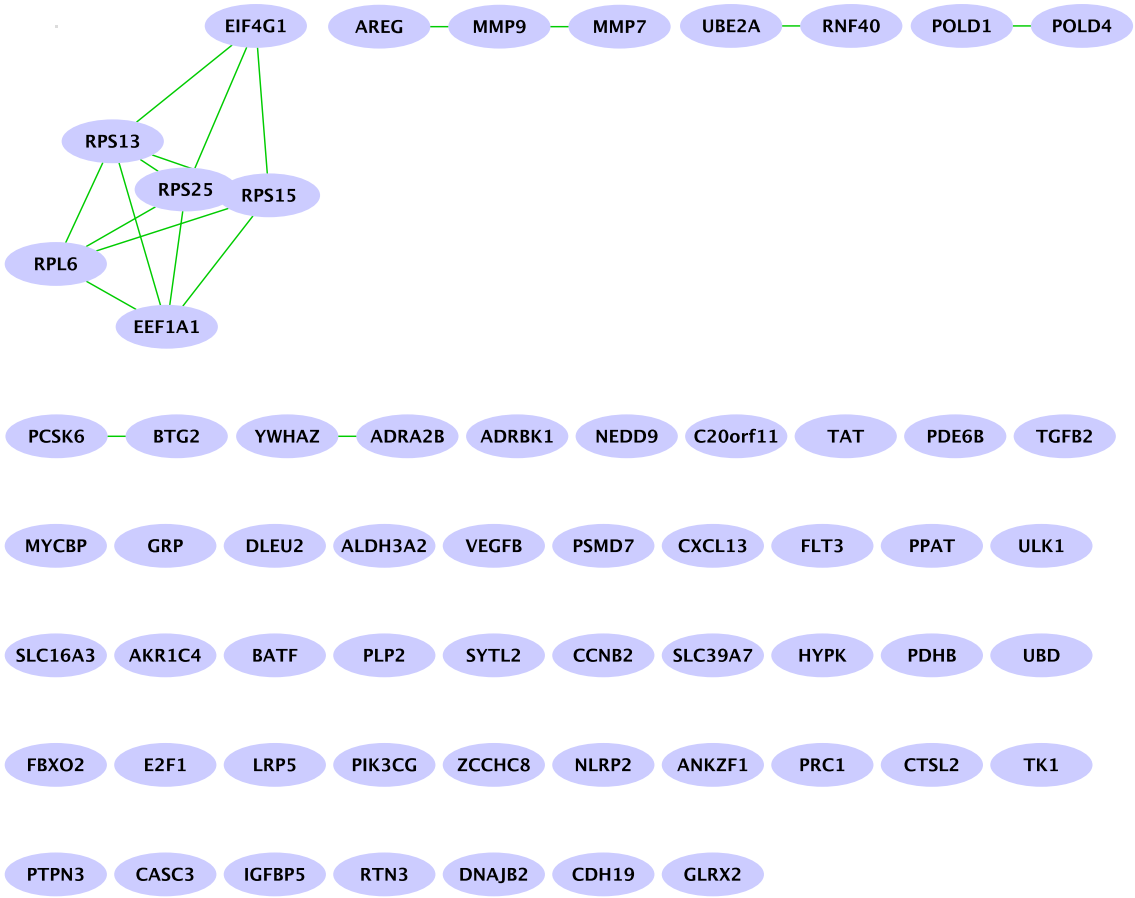
- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

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

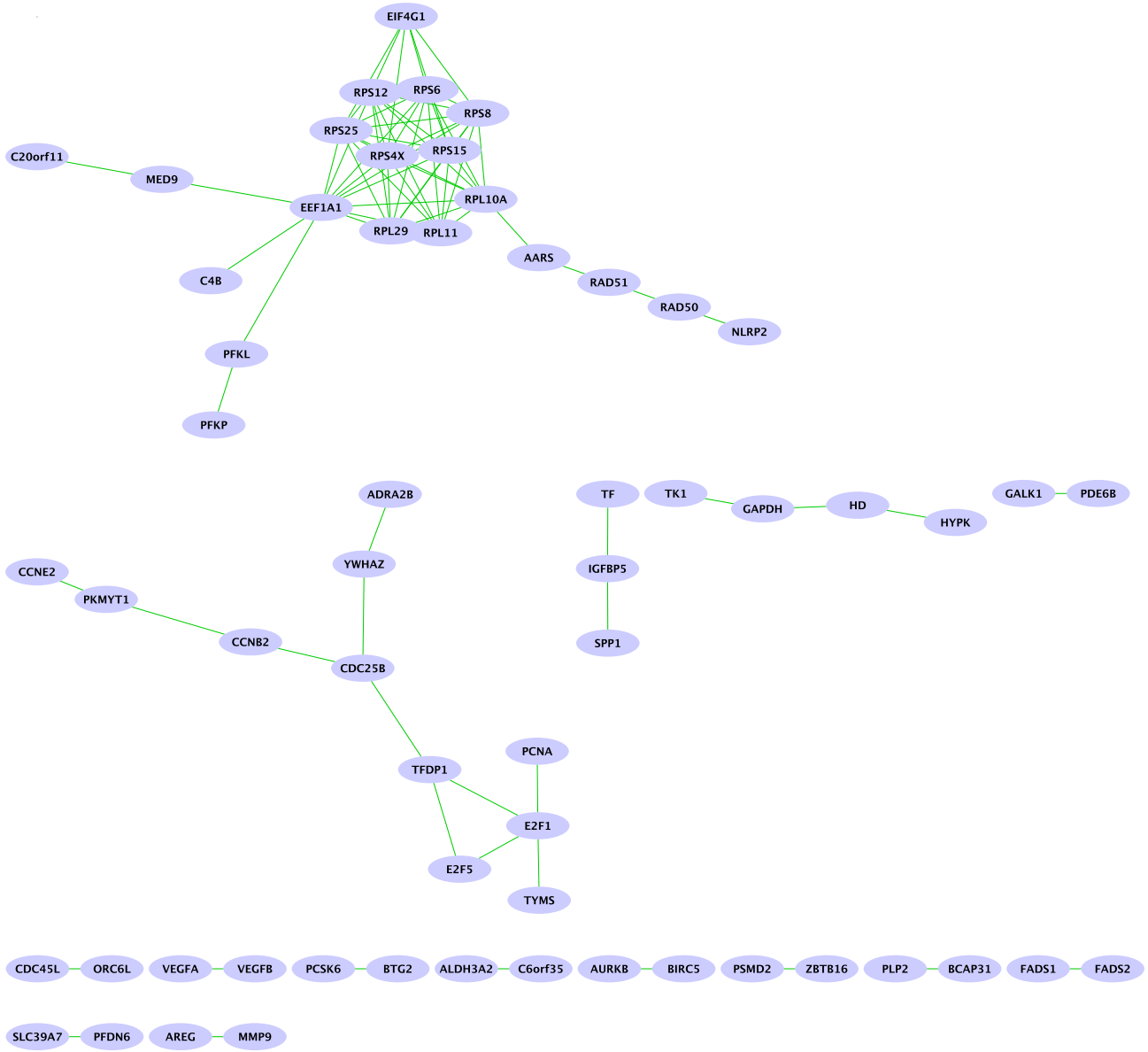
- Graph on the genes.

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

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)