Statistical inference for complex systems

Jean-Philippe Vert

Mines ParisTech / Curie Institute / Inserm
Paris, France

U900 lab meeting, Institut Curie, Sep 28, 2010.

Outline

The modeller vs statistician dilemma

Shrinkage classifiers

Examples

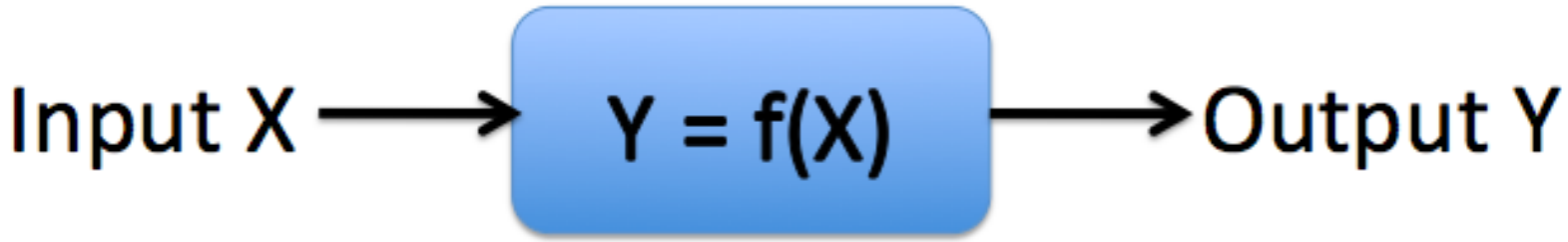
Conclusion

Outline

Some (interesting) complex systems

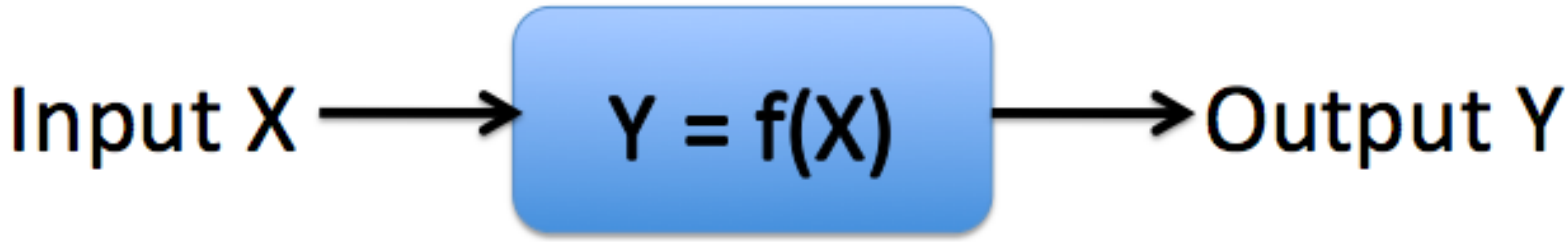
- Diagnosis/Prognosis/Theragnosis: $X =$ genome/transcriptome/... , $Y =$ tumor evolution / survival / response to therapy...
- Regulatory/signalling pathways: $X =$ perturbation (molecule, knock-out...), $Y =$ phenotype / expression level
- Genotype-phenotype relationship: $X =$ genome/mutations, $Y =$ a phenotype (disease, growth rate,..)
- QSAR/Virtual screening/chemogenomics: $X =$ molecule/perturbation, $Y =$ phenotypic cellular response

Modelling/inferring complex systems



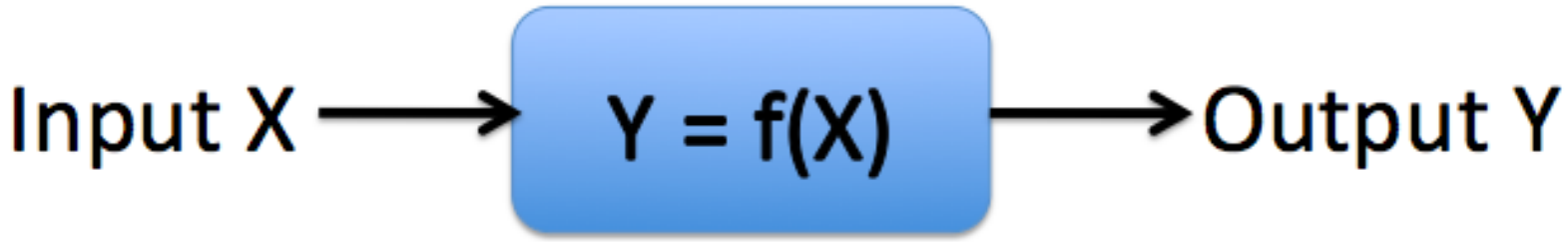
- A model is a human construct to help us better understand real world systems and make predictions
- Remember: all models are wrong, but some are useful (Box, 1987).
- How to make a model $f(x)$ from:
  - prior knowledge
  - observations $(X_i, Y_i)_{i=1,\ldots,n}$

General 2-step principle



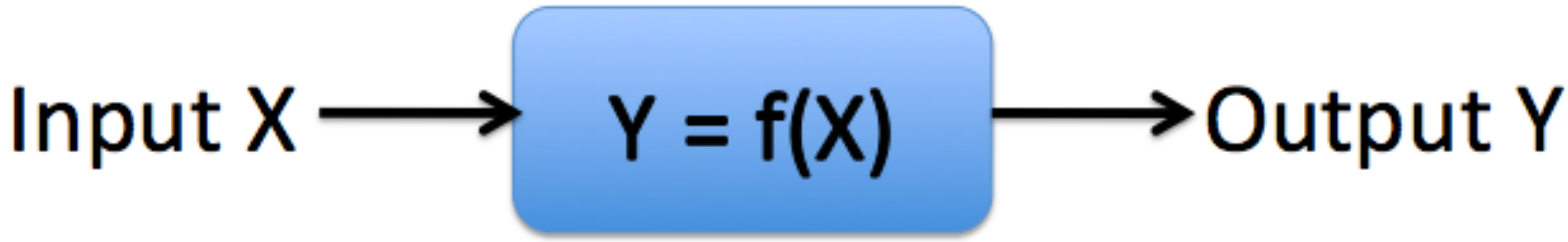
- Step 1 (modelling): define a family of candidate functions $\mathcal{F} = \{f : X \mapsto Y\}$
  - using prior knowledge
  - e.g., linear models, boolean networks, neural networks....
- Step 2 (inference): confront the model to the data to estimate one function $\hat{f} \in \mathcal{F}$
  - using statistical inference techniques, e.g., empirical risk miminization
  - using prior knwoledge/belief

General 2-step principle



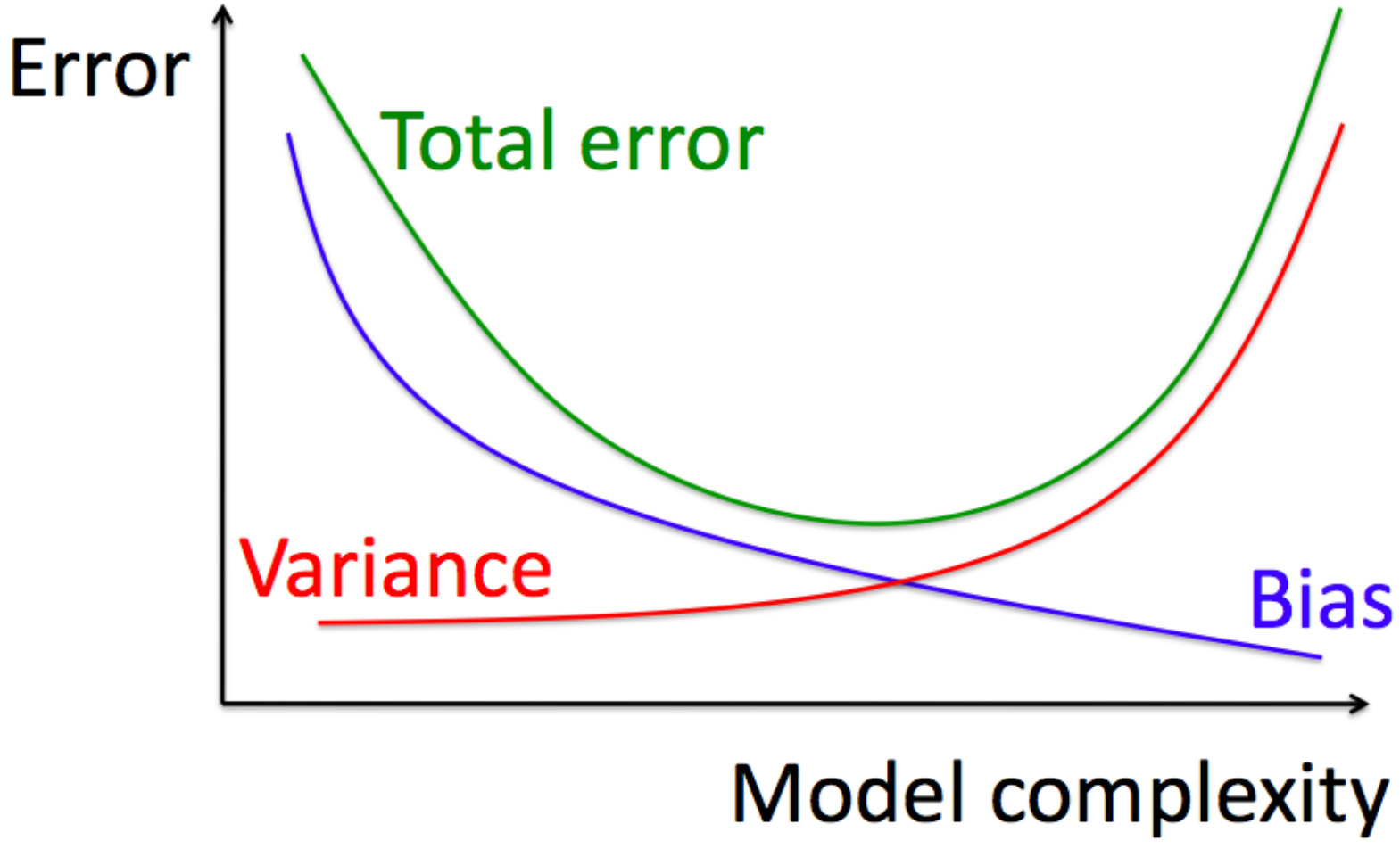
- Step 1 (modelling): define a family of candidate functions $\mathcal{F} = \{f : X \mapsto Y\}$
  - using prior knowledge
  - e.g., linear models, boolean networks, neural networks....
- Step 2 (inference): confront the model to the data to estimate one function $\hat{f} \in \mathcal{F}$
  - using statistical inference techniques, e.g., empirical risk miminization
  - using prior knwoledge/belief

General 2-step principle

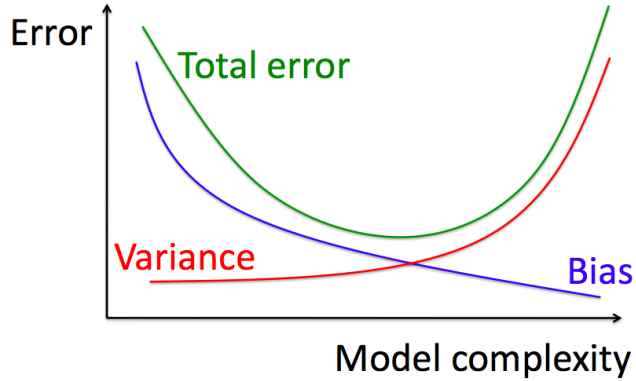


- Step 1 (modelling): define a family of candidate functions $\mathcal{F} = \{f : X \mapsto Y\}$
  - using prior knowledge
  - e.g., linear models, boolean networks, neural networks....
- Step 2 (inference): confront the model to the data to estimate one function $\hat{f} \in \mathcal{F}$
  - using statistical inference techniques, e.g., empirical risk miminization
  - using prior knwoledge/belief

The bias/variance trade-off



The modeller vs statistician dilemma



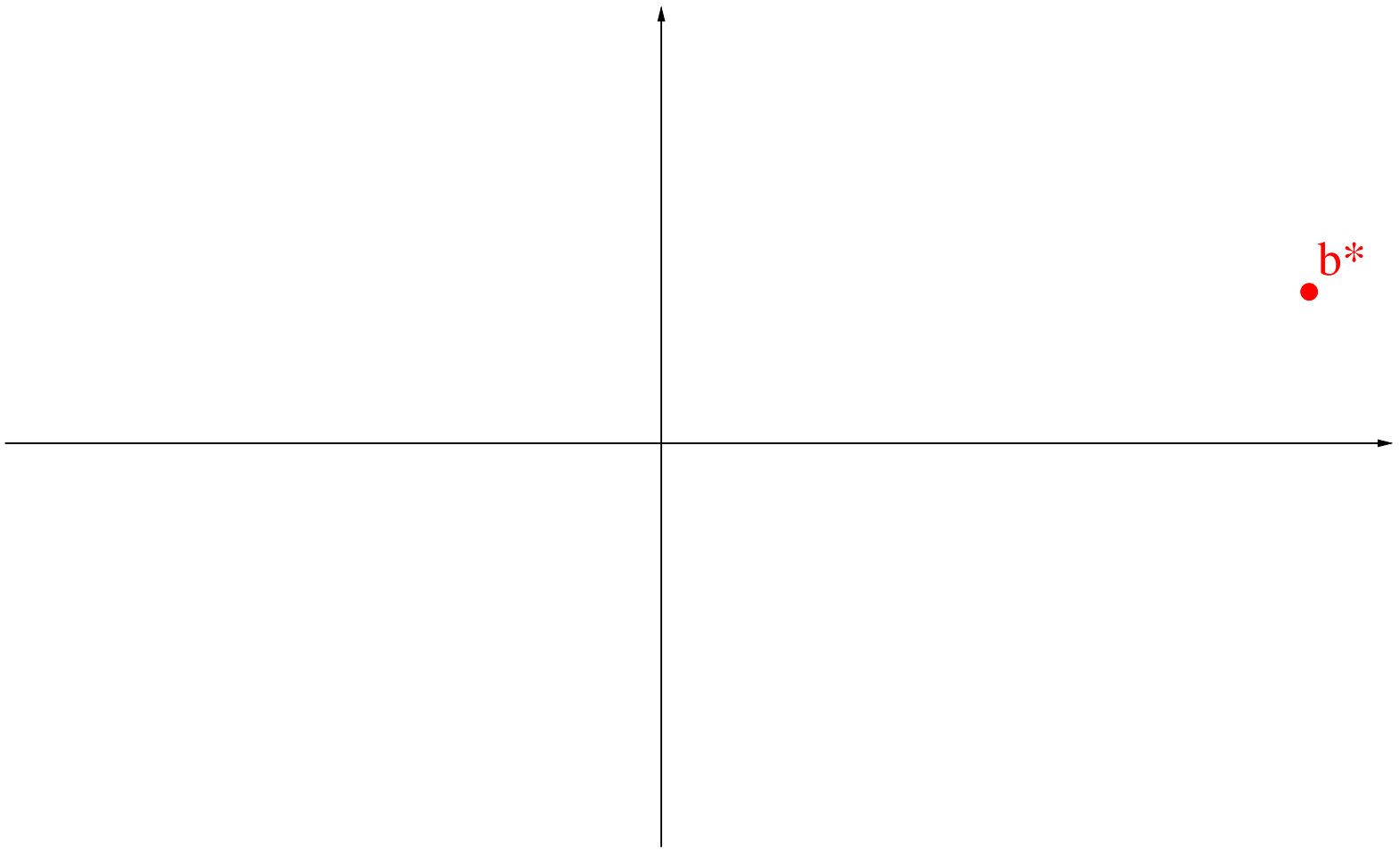
- Both steps must be taken into account to have a "good" model
- Modellers / experts usually focus more on making good models (deacreasing bias), and forget about estimation errors (variance)
- But we have often very few data compared to the complexity of realistic models $\Longrightarrow$ variance is likely to dominate!
- *Illustration: success of generic machine learning approaches (which intrinsically control the trade-off) vs knowledge-based models*
- Challenge: reconcile modellers and statisticians

Outline

Illustration

$$\min_f R(f)\,.$$

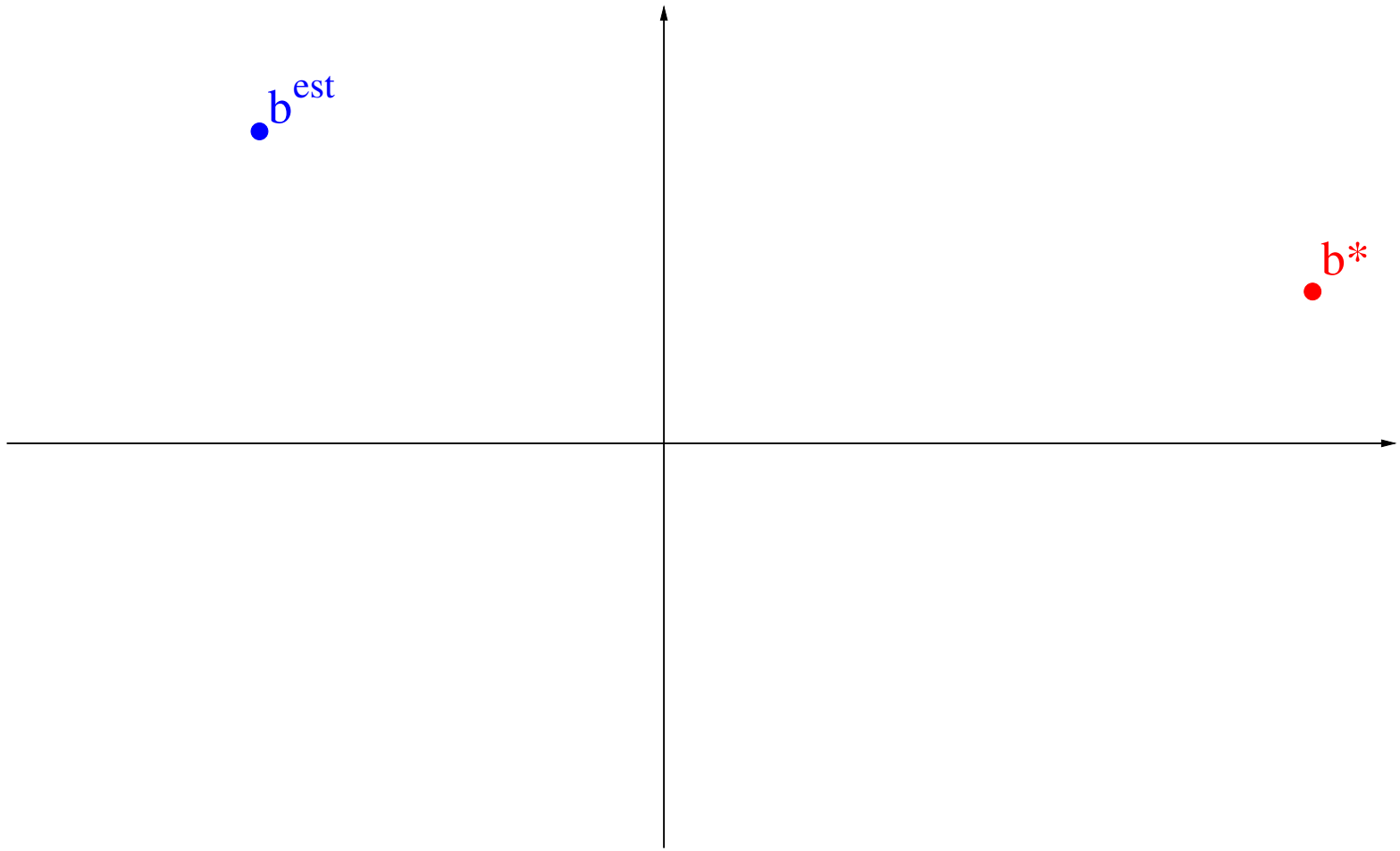
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Illustration

$$\min_f R(f)\,.$$

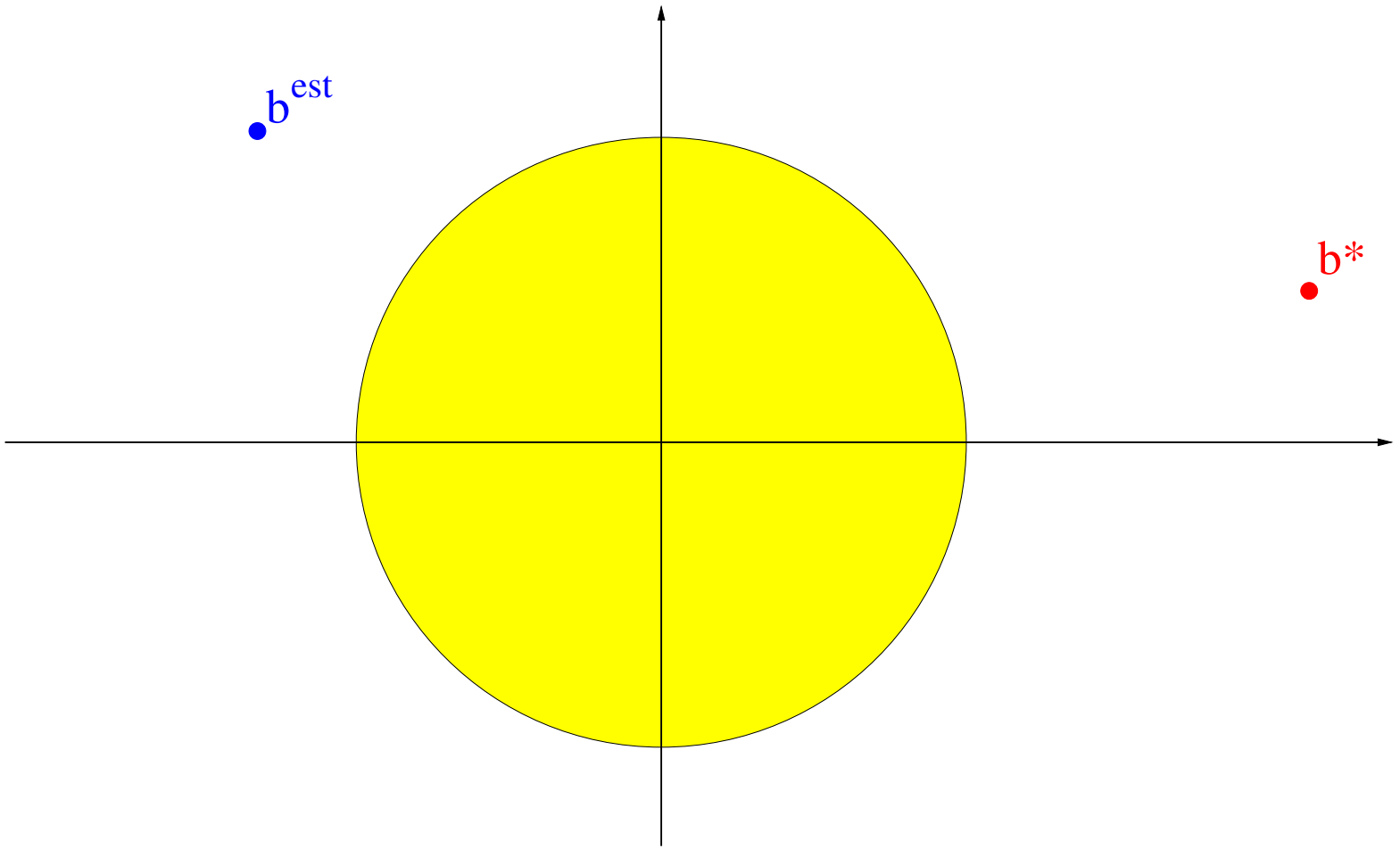
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Illustration

$$\min_f R(f)\,.$$

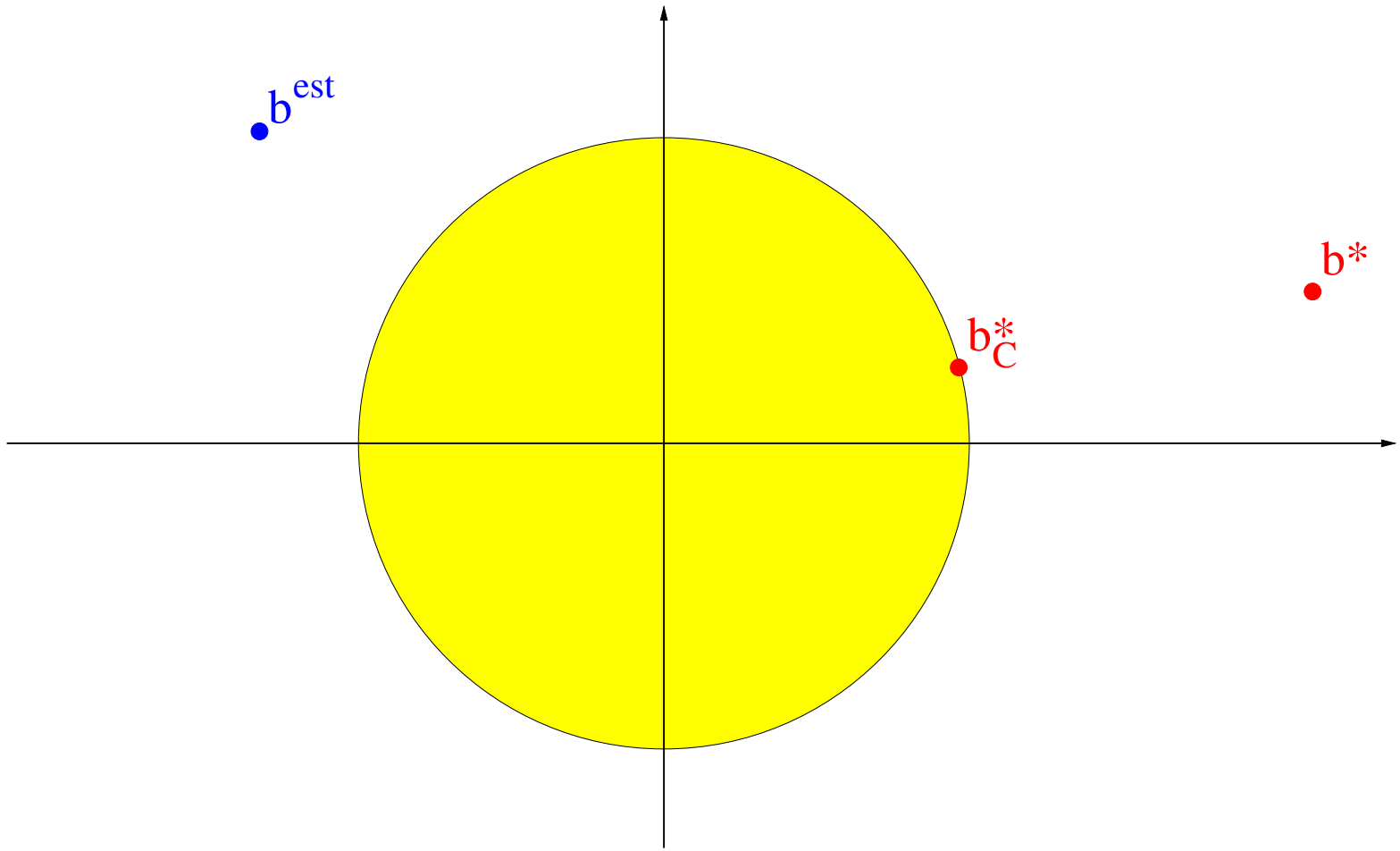
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Illustration

$$\min_f R(f)\,.$$

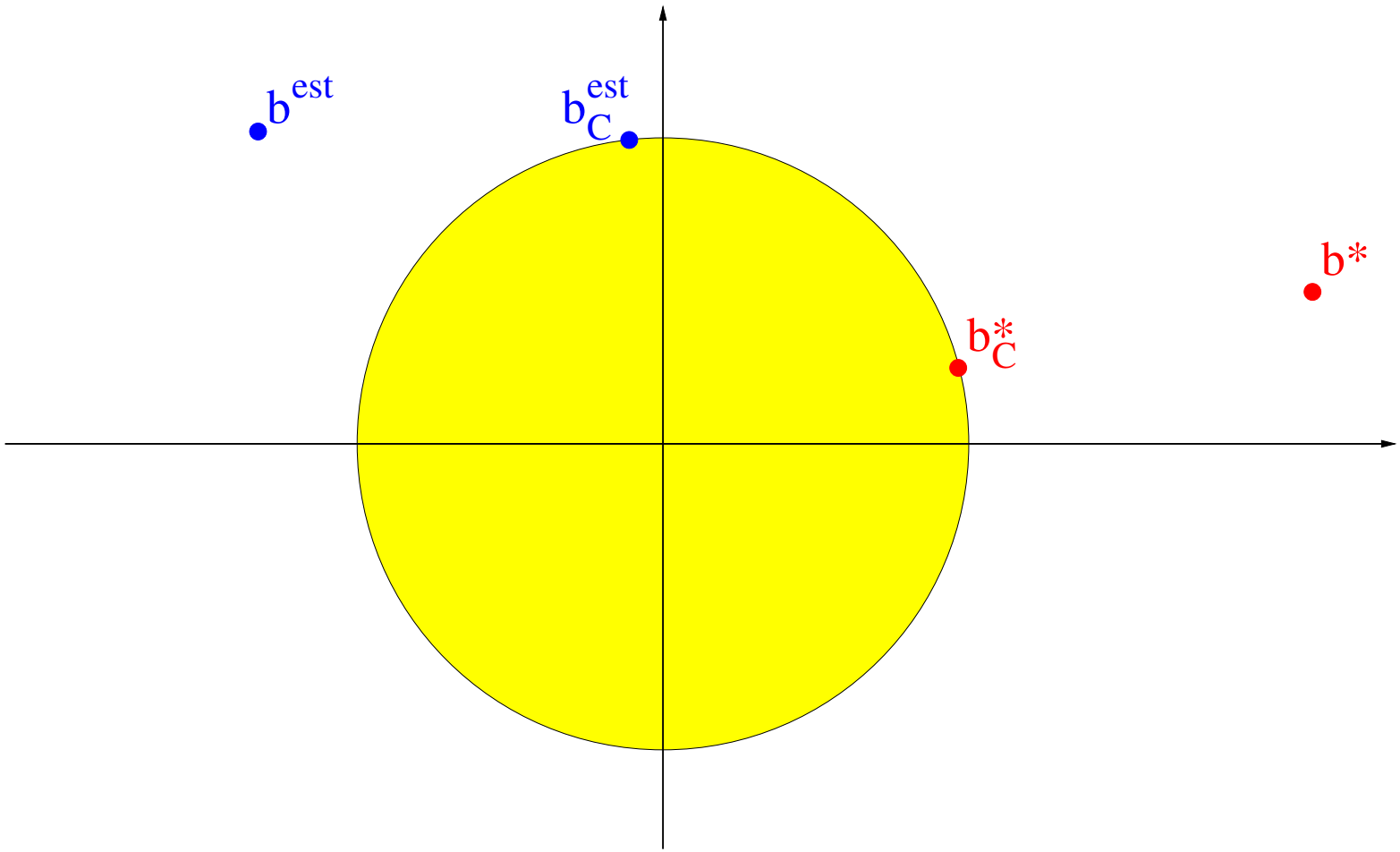
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Illustration

$$\min_f R(f)\,.$$

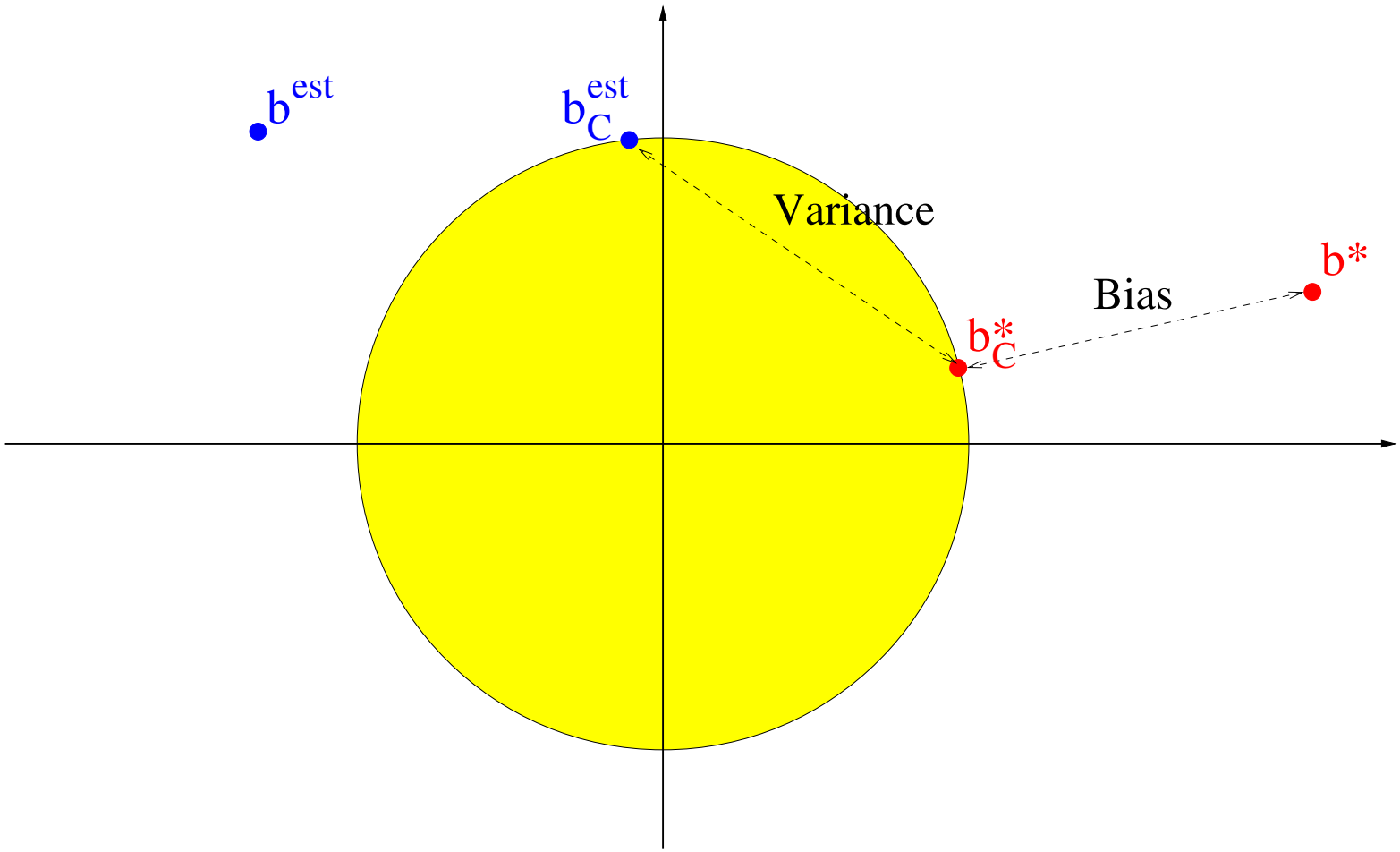
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Illustration

$$\min_f R(f)\,.$$

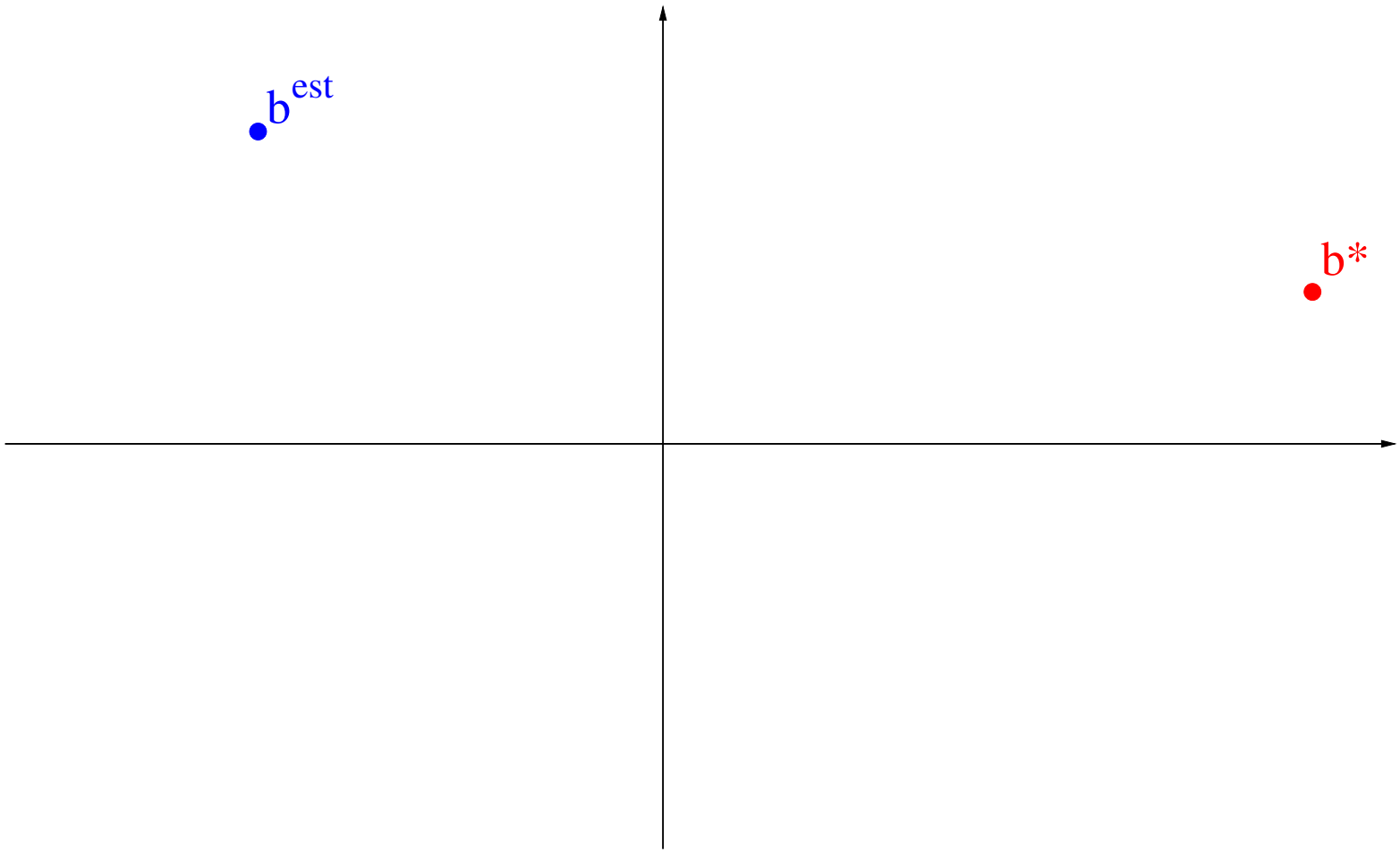
$$\text{subject to} \quad \Omega(f) \leq C\,. \tag{1}$$



Changing the penalty with prior knowledge

$$\min_f R(f)\,.$$

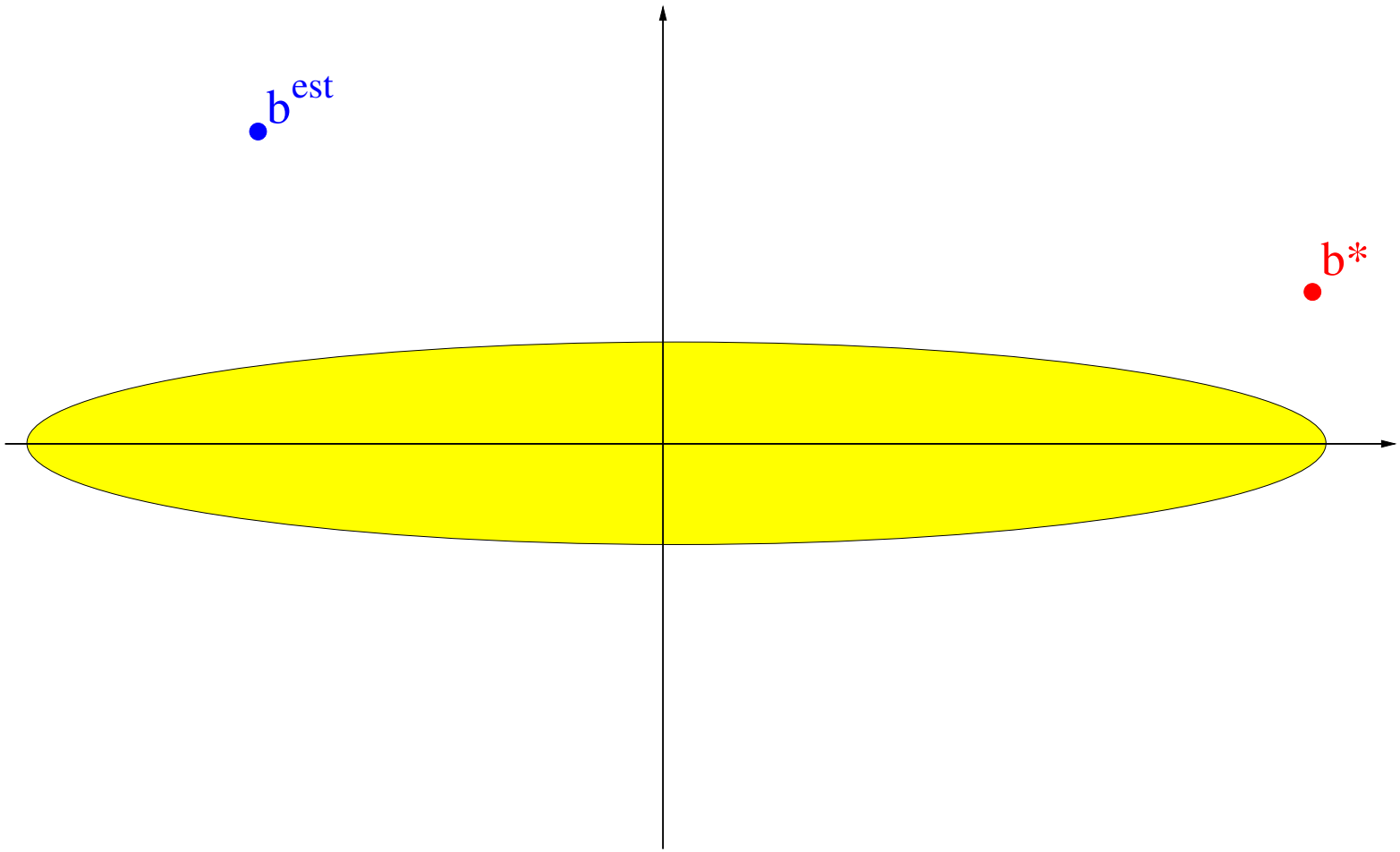
$$\text{subject to} \quad \Omega_{new}(f) \leq C\,. \tag{2}$$



Changing the penalty with prior knowledge

$$\min_f R(f)\,.$$

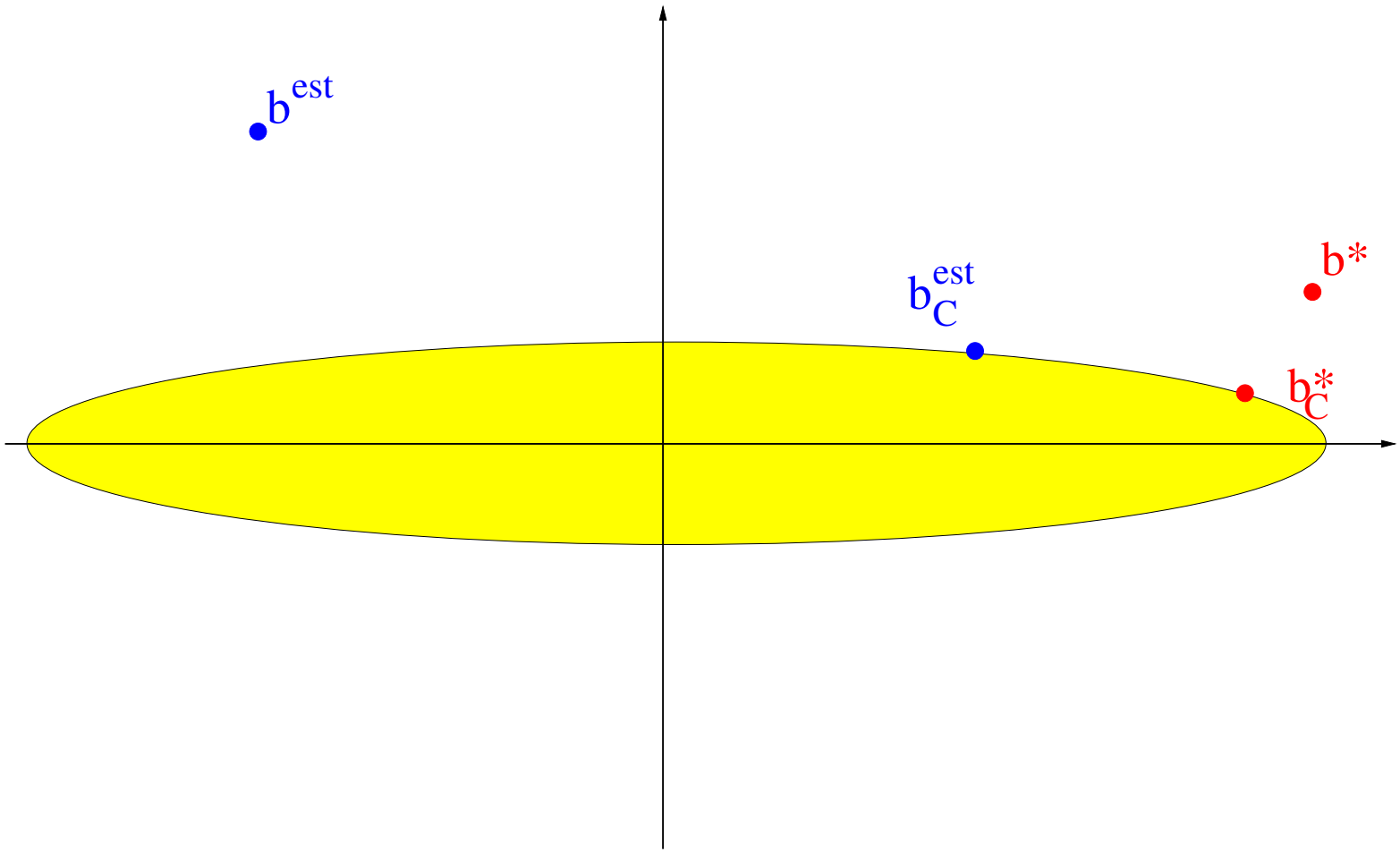
$$\text{subject to} \quad \Omega_{new}(f) \leq C\,. \tag{2}$$



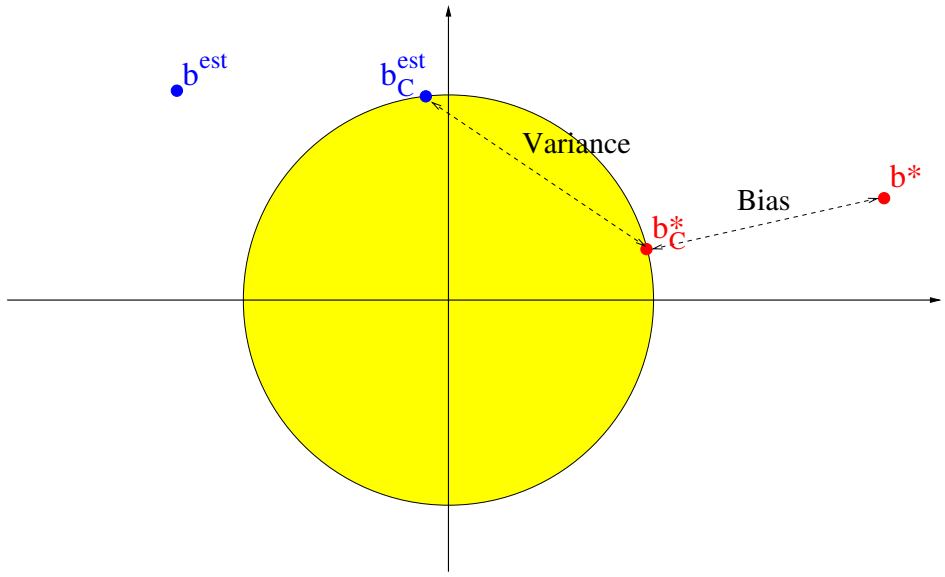
Changing the penalty with prior knowledge

$$\min_f R(f)\,.$$

$$\text{subject to} \quad \Omega_{new}(f) \leq C\,. \tag{2}$$



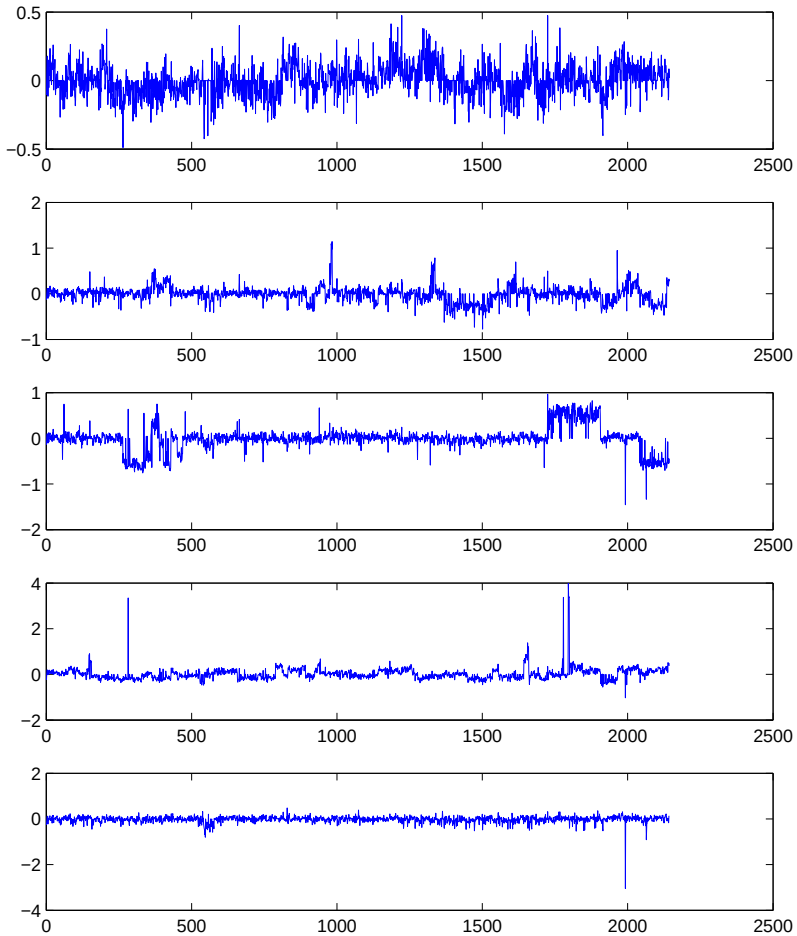
Summary

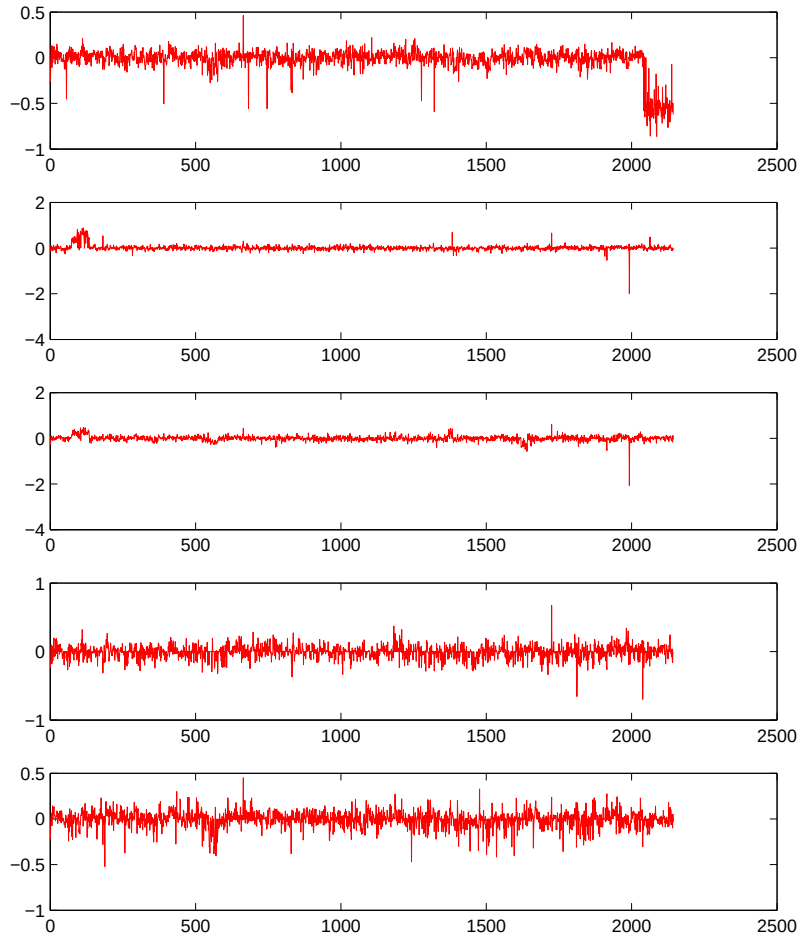


- Shrinkage method offer a principled approach to "Increases bias and decreases variance", and control the trade-off through C
- At the heart of many successful methods (SVM, Lasso, boosting)
- Changing $\Omega(f)$ may in addition decrease the bias without increasing the variance
- In practice: design (convex) penalties $\Omega(f)$ that encode prior knowledge

Outline

Classification of DNA copy number profiles





Aggressive vs non-aggressive melanoma

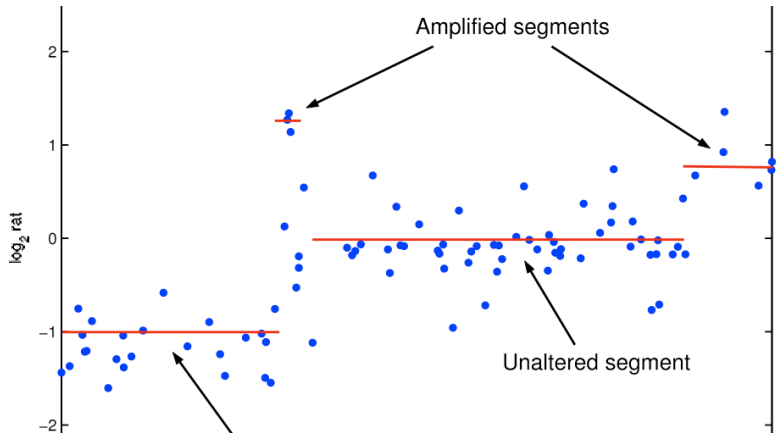
CGH array classification

Prior knowledge

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

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

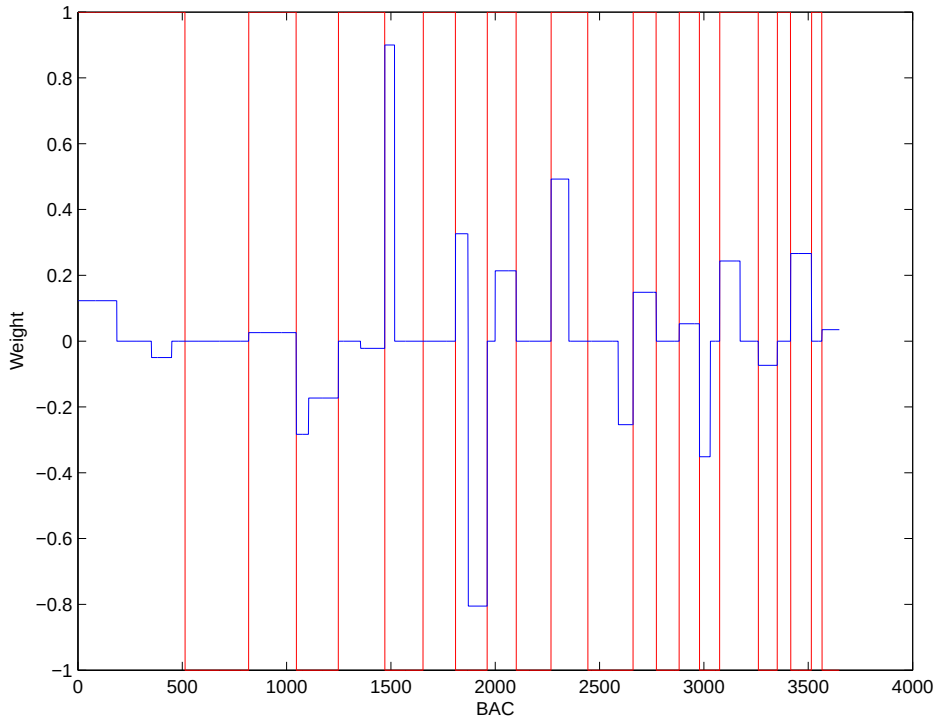
- We expect β to be
  - sparse : not all positions should be discriminative
  - piecewise constant : within a selected region, all probes should contribute equally



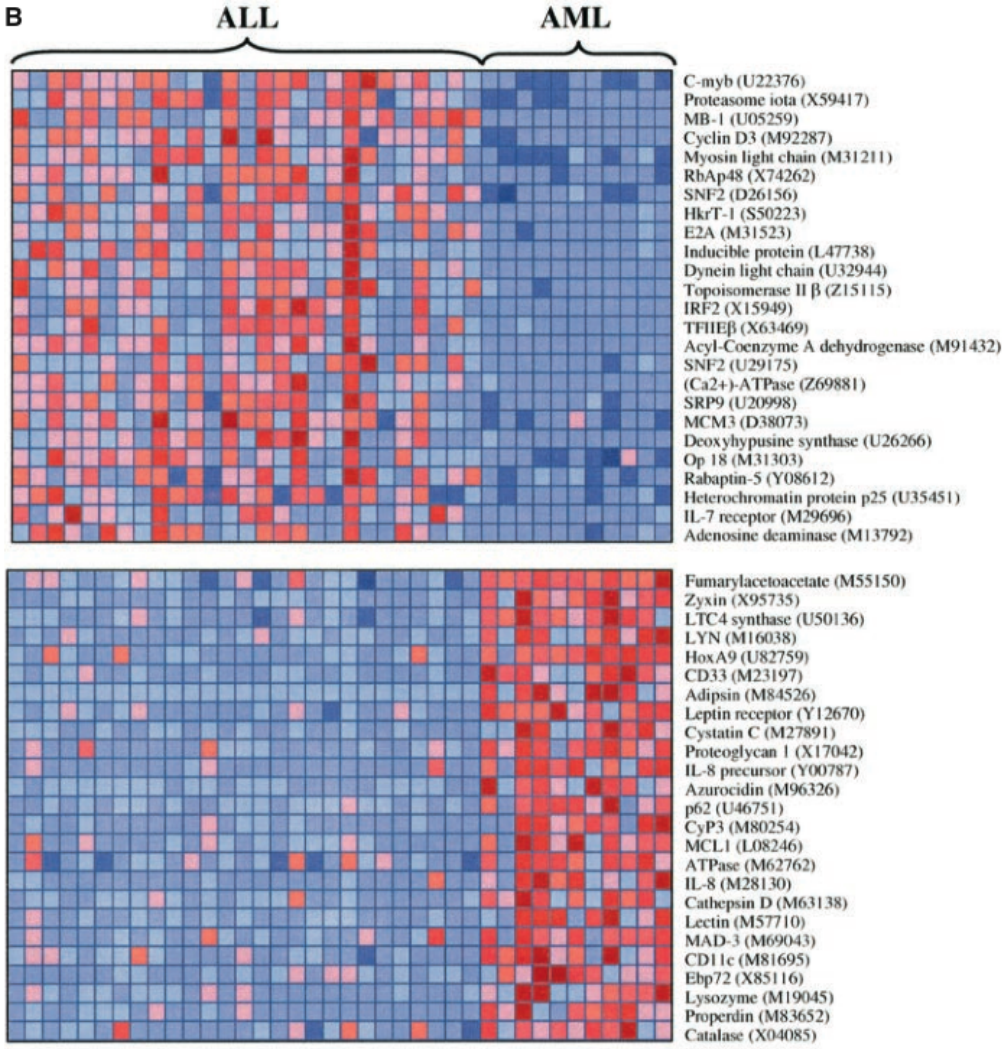
A solution (Rapaport et al., 2008)

$$\Omega_{fusedlasso}(\beta) = \sum_{i=1}^{p} |\beta_i| + \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

- First term promotes sparse solution (Lasso penalty)
- Second term promotes piecewise constant solutions



Tissue classification from microarray data (diagnosis/prognosis...)



Goal

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

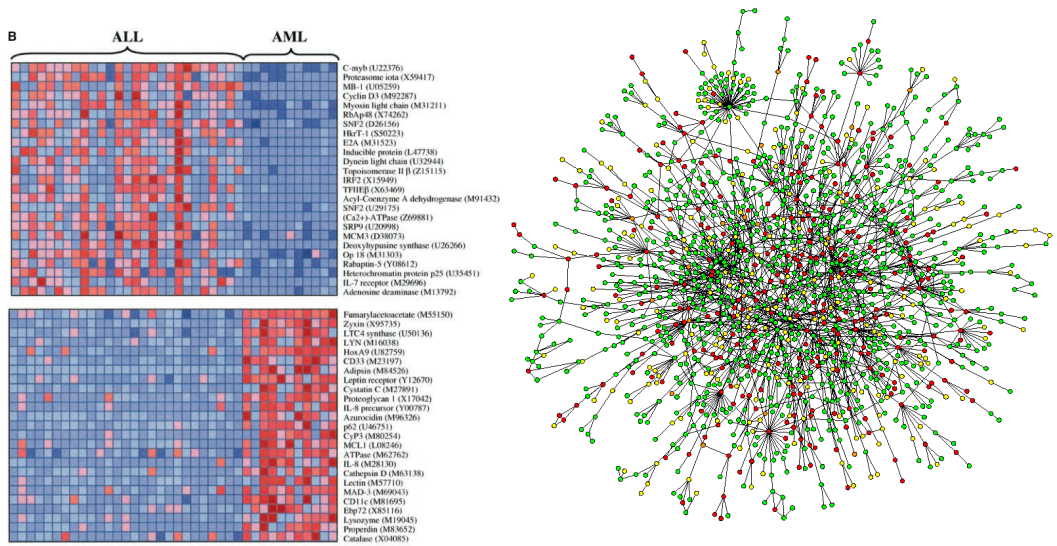
Difficulty

- Large dimension
- Few samples

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



Graph-based penalties: smooth classifiers

Prior hypothesis

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

- Smooth weights on the graph (Rapaport et al., 2007)

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

- Gene selection + Piecewise constant on the graph

$$\Omega_{fused}(\beta) = \sum_{i \sim j} | \beta_i - \beta_j | + \sum_{i=1}^{p} | \beta_i |$$

- Gene selection + smooth on the graph

$$\Omega_{mix}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} | \beta_i |$$

Graph-based penalties: smooth classifiers

Prior hypothesis

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

- Smooth weights on the graph (Rapaport et al., 2007)

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

- Gene selection + Piecewise constant on the graph

$$\Omega_{fused}(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$

- Gene selection + smooth on the graph

$$\Omega_{mix}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\, \beta_i \,|$$

Graph-based penalties: smooth classifiers

Prior hypothesis

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

- Smooth weights on the graph (Rapaport et al., 2007)

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

- Gene selection + Piecewise constant on the graph

$$\Omega_{fused}(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$

- Gene selection + smooth on the graph

$$\Omega_{mix}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\, \beta_i \,|$$

Graph-based penalties: smooth classifiers

Prior hypothesis

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

- Smooth weights on the graph (Rapaport et al., 2007)

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

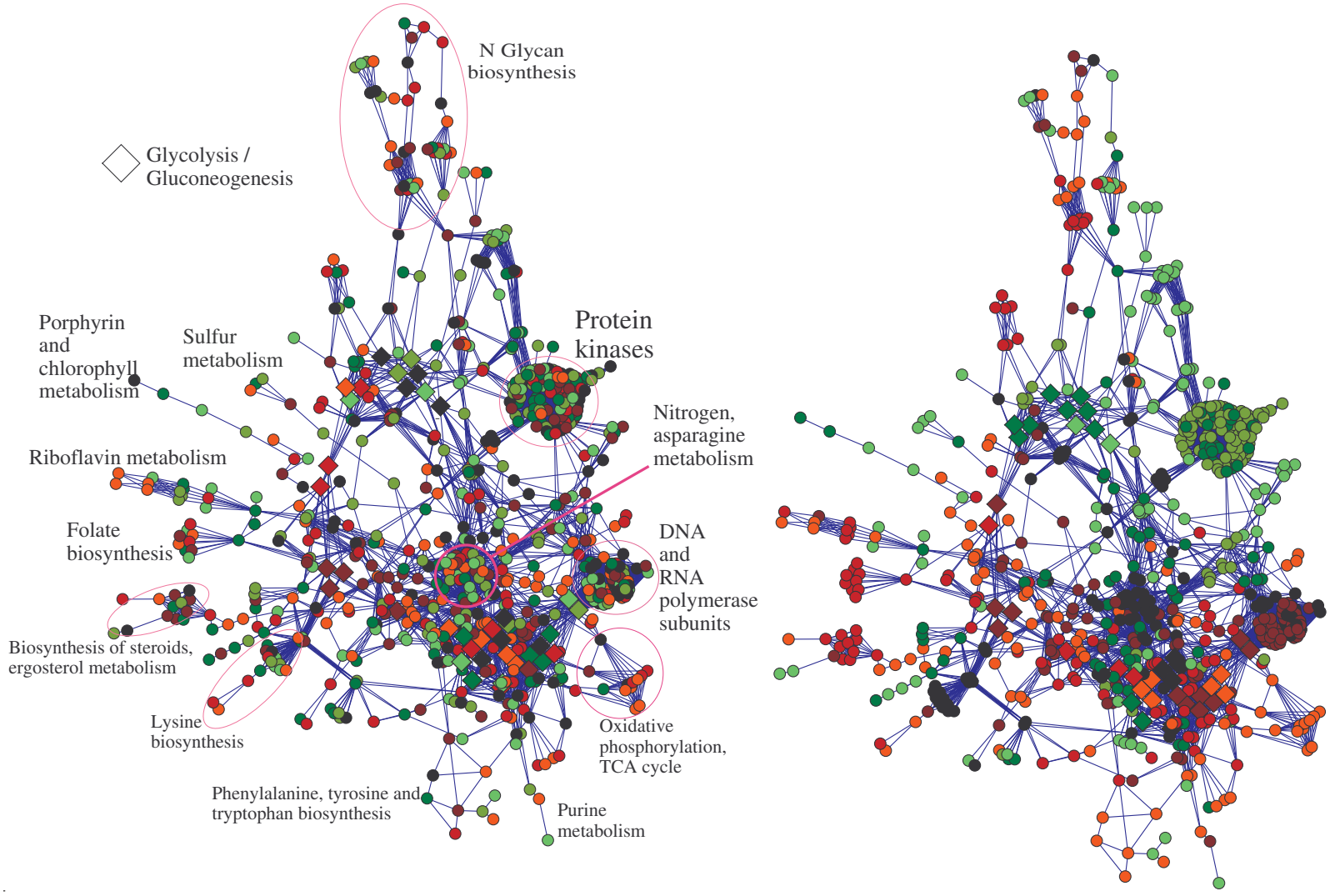
- Gene selection + Piecewise constant on the graph

$$\Omega_{fused}(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$

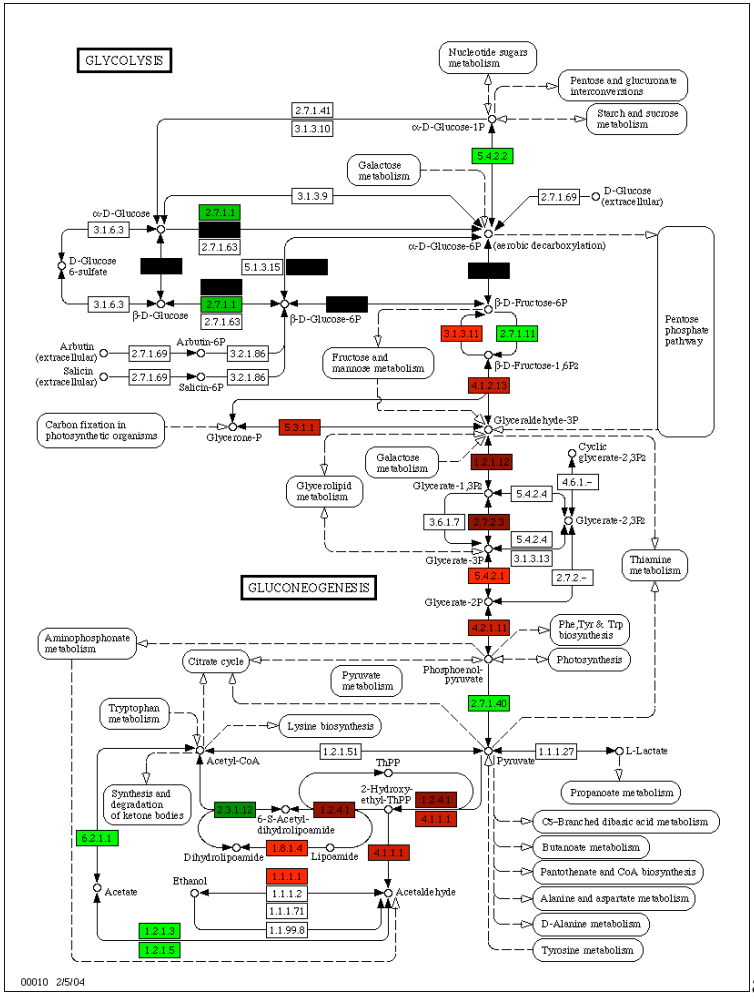
- Gene selection + smooth on the graph

$$\Omega_{mix}(\beta) = \sum_{i \sim j} \left(\beta_i - \beta_j\right)^2 + \sum_{i=1}^{p} |\, \beta_i \,|$$

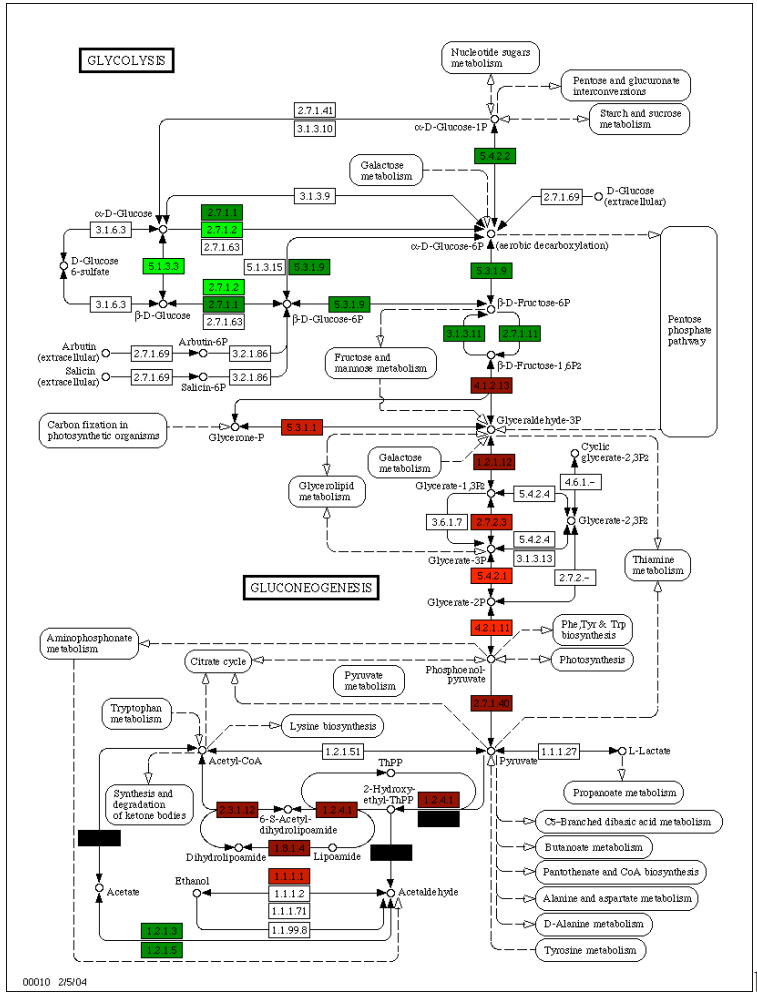
Illustration



Illustration



a)

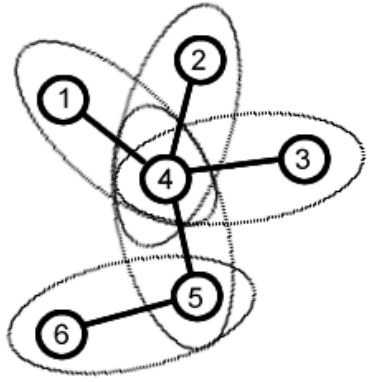


b)

Graph-based penalty: structured feature selection

Prior hypothesis

Selected genes should form connected components on the graph



Two solutions (Jacob et al., 2009):

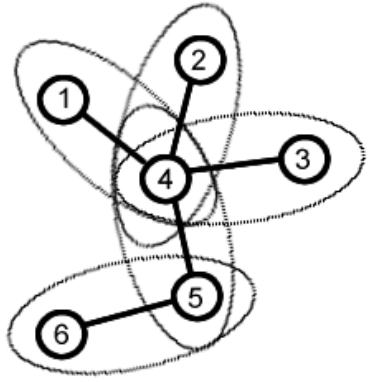
$$\Omega_{group}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

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

Graph-based penalty: structured feature selection

Prior hypothesis

Selected genes should form connected components on the graph

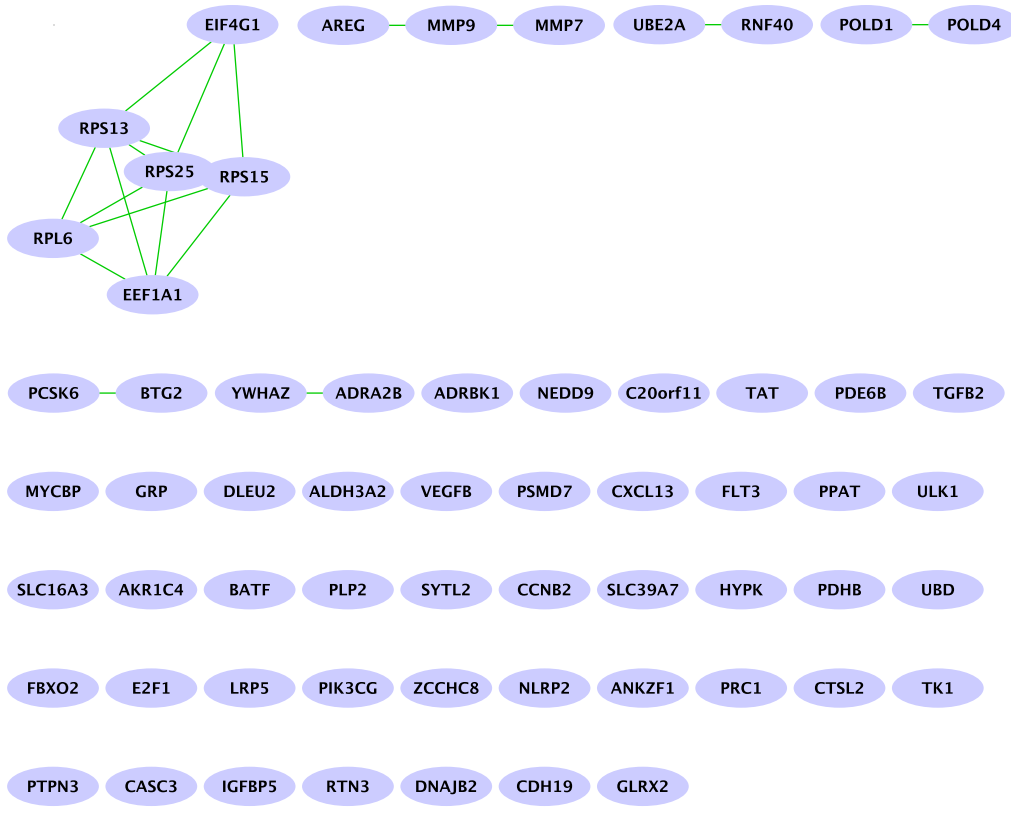


Two solutions (Jacob et al., 2009):

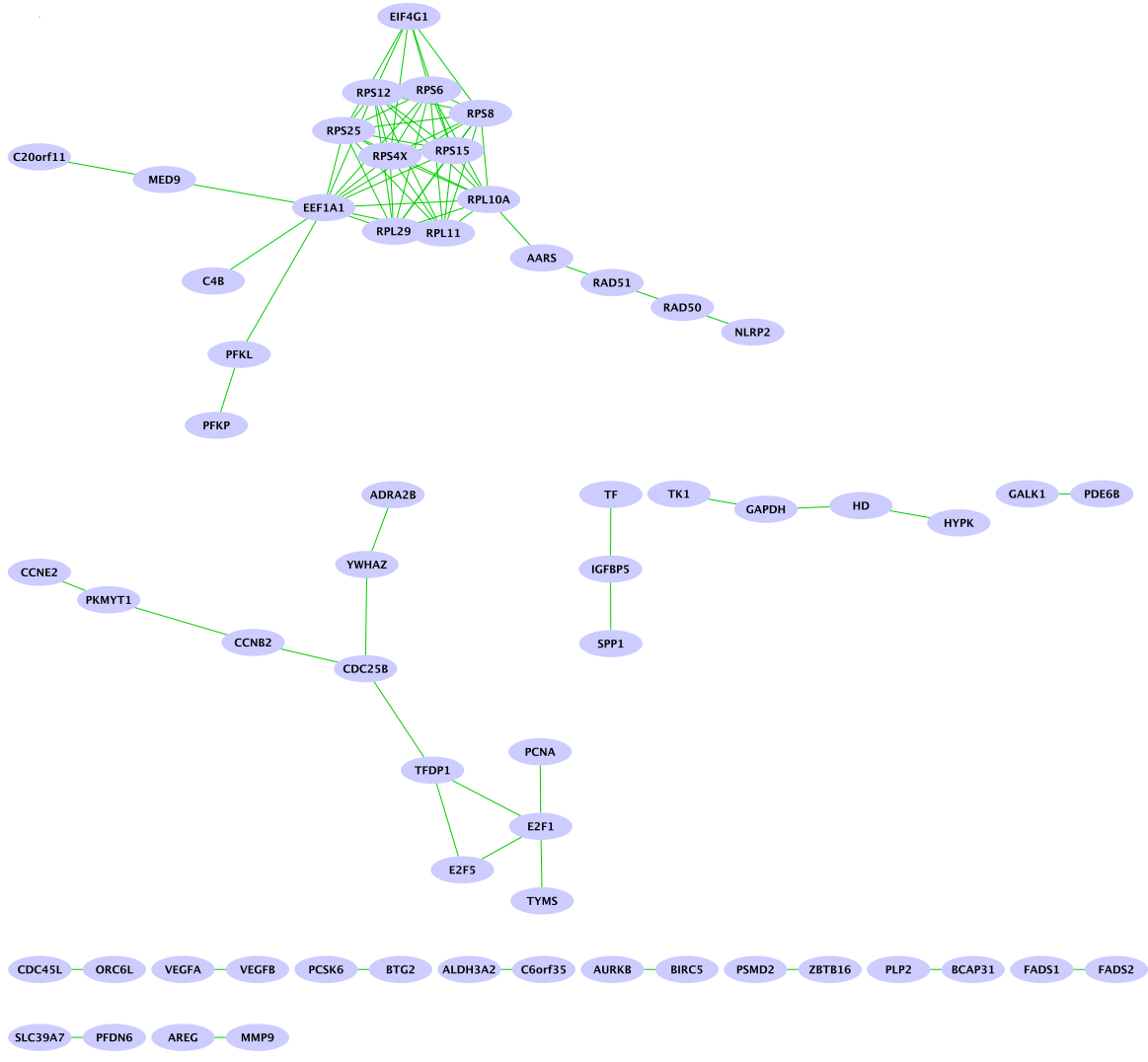
$$\Omega_{group}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

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

Classical gene selection (Lasso)



Graph-based gene selection



Outline

Conclusion

- Controlling the bias/variance trade-off is key! Better to work with wrong but simple models if variance dominates...
- Shrinkage methods provide a convenient strategy to control this trade-off and include prior knowledge
- Important challenges:
  - Enforcing bias/variance control with complex models (eg, dynamic equations in systems biology)?
  - To what extent can we extract knowledge from the estimated model?

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)