On stability and interpretability of gene expression signatures

Prediction of breast cancer outcome

Anne-Claire Haury and Jean-Philippe Vert

Mines ParisTech, INSERM U900, Institut Curie

MLSB, October 15, 2010

Outline

Motivation

Prediction of breast cancer outcome

- Develop a tool to assist breast cancer prognosis
- Avoid preventive chemotherapy when not needed

Gene expression signature

- Data: primary site tumor expression arrays
- Among the genome, find the few (50-100) genes responsible for metastasis
- Challenge: high-dimensional data, typically tens/hundreds of examples and thousands of features

Background

- Little overlap between published signatures from different groups.
- However: biological interpretation of the signature is more robust than the genes themselves.
- Good predictive accuracy but not transferable across datasets.
- Meta-studies show that, for a given method, signatures lack robustness to perturbations of the training set within a given dataset.
- Ensemble Methods (resampling, bootstrap) improve stability and performance in some cases.

Background

- Little overlap between published signatures from different groups.
- However: biological interpretation of the signature is more robust than the genes themselves.
- Good predictive accuracy but not transferable across datasets.
- Meta-studies show that, for a given method, signatures lack robustness to perturbations of the training set within a given dataset.
- Ensemble Methods (resampling, bootstrap) improve stability and performance in some cases.

Background

- Little overlap between published signatures from different groups.
- However: biological interpretation of the signature is more robust than the genes themselves.
- Good predictive accuracy but not transferable across datasets.
- Meta-studies show that, for a given method, signatures lack robustness to perturbations of the training set within a given dataset.
- Ensemble Methods (resampling, bootstrap) improve stability and performance in some cases.

Background

- Little overlap between published signatures from different groups.
- However: biological interpretation of the signature is more robust than the genes themselves.
- Good predictive accuracy but not transferable across datasets.
- Meta-studies show that, for a given method, signatures lack robustness to perturbations of the training set within a given dataset.
- Ensemble Methods (resampling, bootstrap) improve stability and performance in some cases.

Background

- Little overlap between published signatures from different groups.
- However: biological interpretation of the signature is more robust than the genes themselves.
- Good predictive accuracy but not transferable across datasets.
- Meta-studies show that, for a given method, signatures lack robustness to perturbations of the training set within a given dataset.
- Ensemble Methods (resampling, bootstrap) improve stability and performance in some cases.

Aim of this work

Compare algorithms...

- Feature selection algorithms : Filters, Wrappers, Embedded Methods
- Assess the gain from Ensemble methods

... in light of relevant criteria

- Predictive performance: will the signature give accurate predictions?
- Stability: can we trust the list of genes?
- Interpretability: can we trust the biological interpretation of the signature?

Outline

Data

Four public breast cancer datasets from the same technology (Affymetrix U133A):

GEO Reference	♯ genes	♯ samples	♯ positives
GSE1456	12,065	159	40
GSE2034	12,065	286	107
GSE2990	12,065	125	49
GSE4922	12,065	249	89

Feature selection methods

Type	Examples	Characteristics
Filters	T-test Wilcoxon rank-sum test Entropy maximization	Univariate, fast Only depend on the data Do not use loss function
Wrappers	SVM RFE Greedy FS	Search for the best subset Computationaly expensive
Embedded	Lasso Elastic Net	Similar to wrappers Search guided by constraint

Each of these algorithms returns a ranked list of genes to be thresholded.

Ensemble Methods

- Run each algorithm B times (e.g. B=50)
- Get B ranked lists of genes $(r^b)_{b=1\ldots B}$ where r_g^b is the rank of gene g after run b
- Aggregate these lists and get a score for each gene:

$$S(g) = \frac{1}{B}\sum_{b=1}^{B} exp\{-\alpha r_g^b\}$$

  where α controls the weight given to the top-list genes.
- Sort by decreasing order and threshold to get final signature

Evaluation

Predictive performance

- Run FS algorithm, either single run or ensemble, get a signature.
- Restrict data to the signature.
- Classify (SVM, KNN, Nearest Neighbors, Naive Bayes...)

Stability

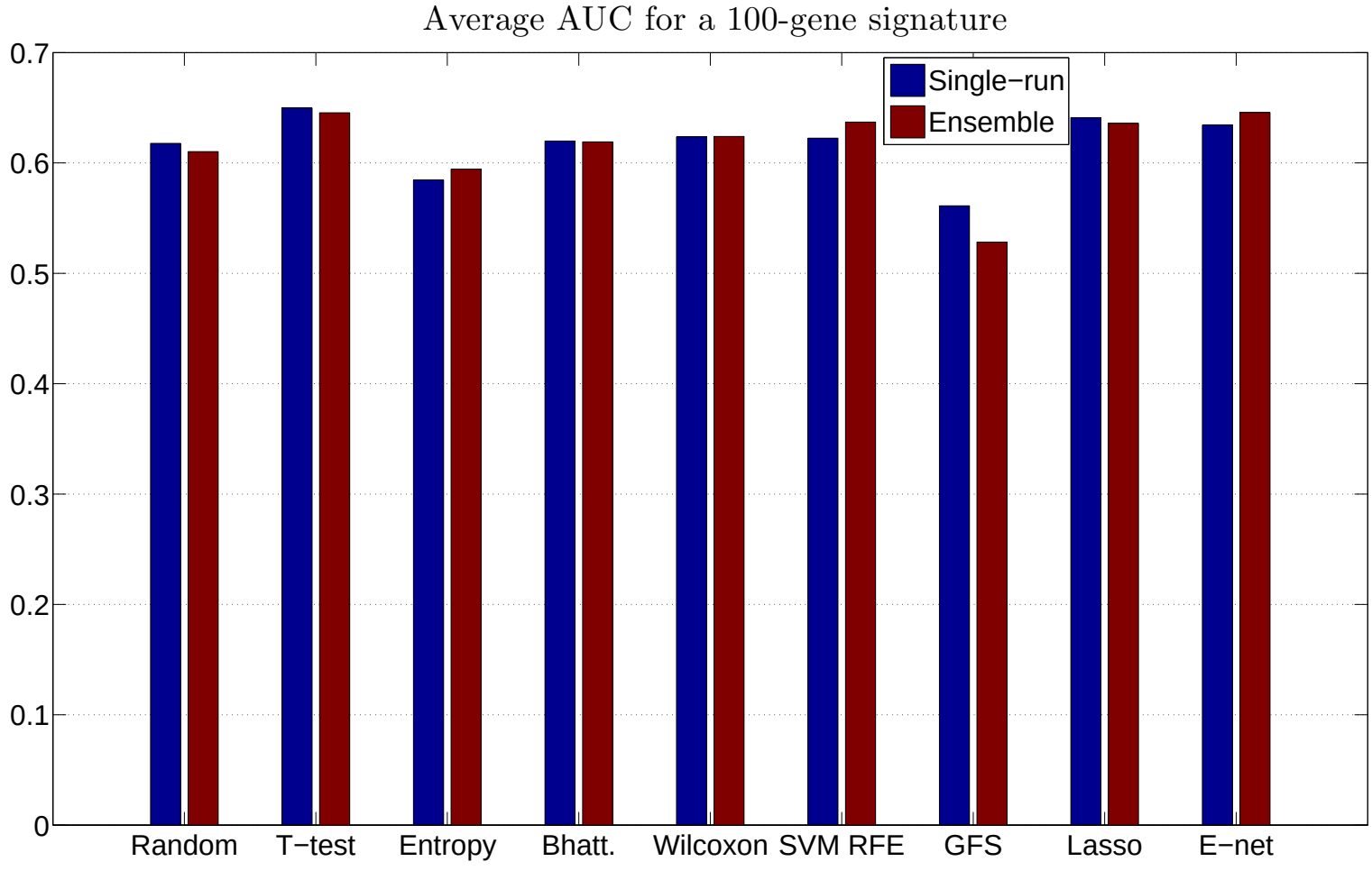
- Select two training sets (with or without examples in common)
- Learn two signatures
- Evaluate overlap

Interpretability

- Select two training sets (with or without examples in common)
- Learn two signatures
- Evaluate overlap in terms of Gene Ontology Biological Processes

Outline

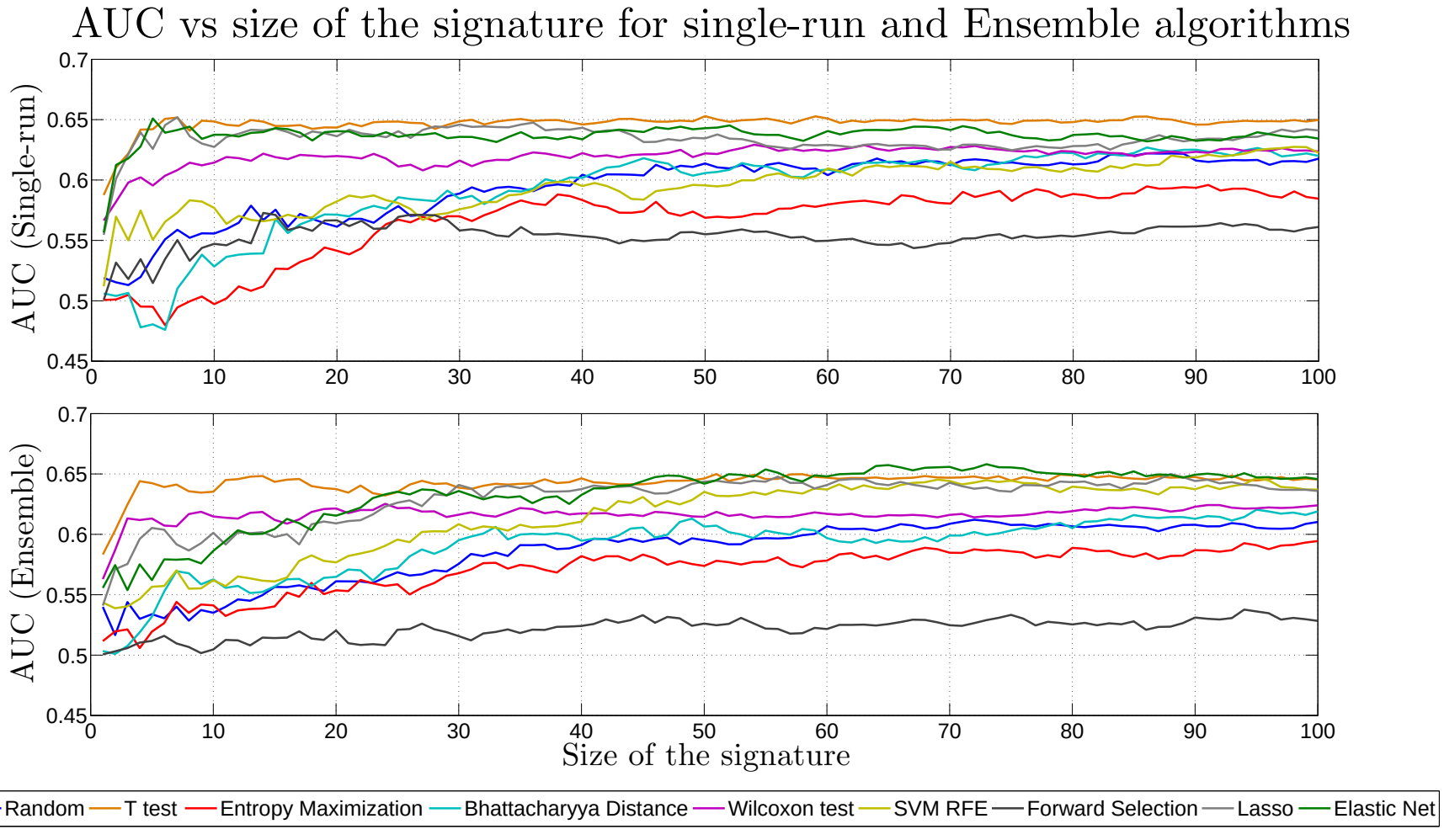
Predictive performance



Classifier used: Nearest centroids (best overall classifier among those tested).

The best AUC is obtained by Lasso, Elastic Net and T-test.

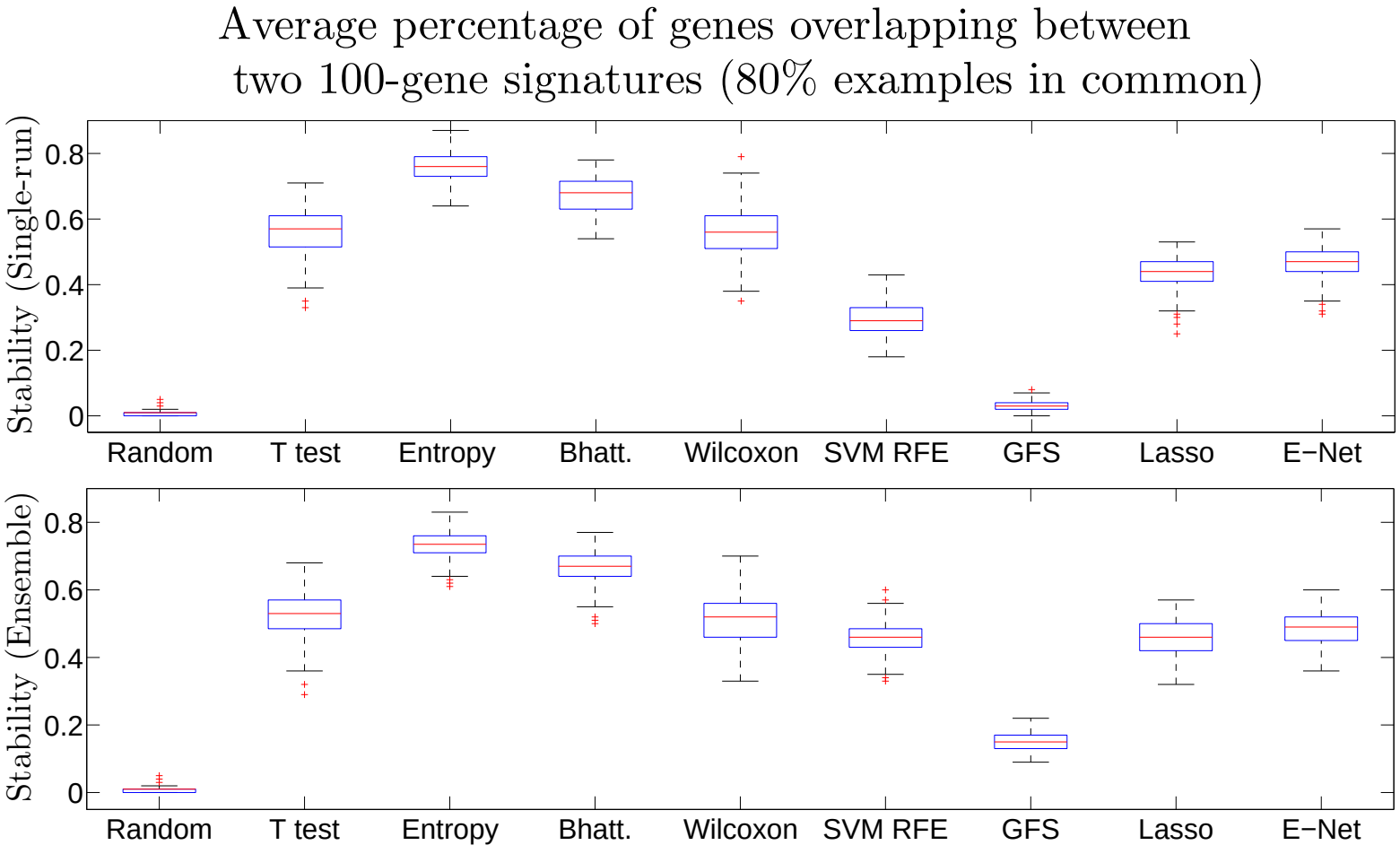
Predictive performance



Classifier used: Nearest centroids (best overall classifier among those tested).

The best AUC is obtained by Lasso, Elastic Net and T-test.

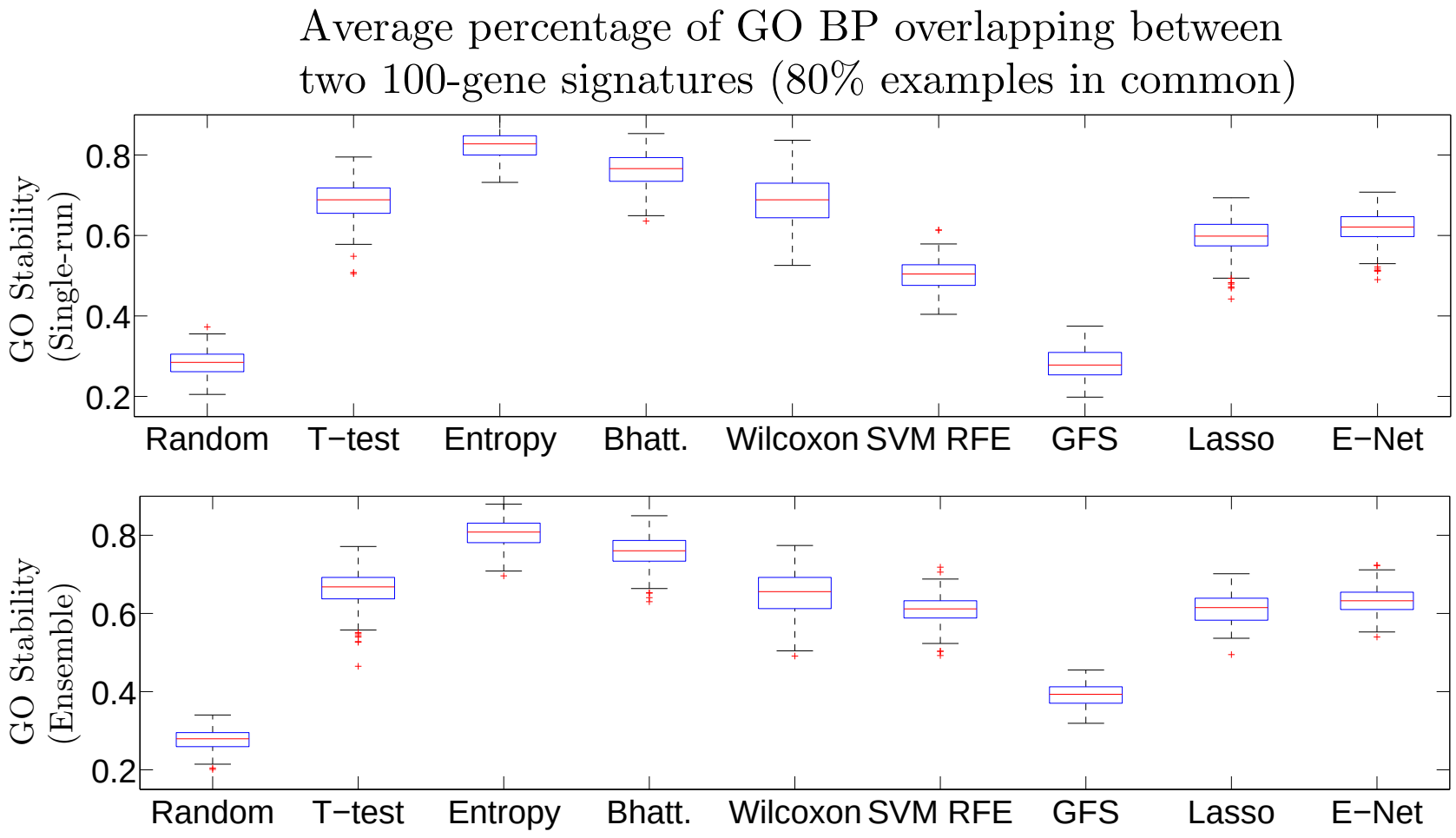
Stability



Filters output more stable signatures.

Ensemble methods improve stability for wrappers/embedded methods.

Interpretability



Filter methods are more stable in terms of GO interpretability.

Outline

Conclusions

Simple works best

- The most stable/interpretable signatures were output by filters.
- They can achieve similar predictive accuracy than more complex procedures.

Ensemble methods are not helpful

- A small gain in stability is observed for wrappers/embedded methods.
- But still not as good as the best single-run method.
- No significant gain in predictive accuracy.

Comments on stability

- Even with small perturbations of the data, stability remains low.
- **Stability and GO stability are necessary but not sufficient to trust a signature.**

Discussion

Why do simple methods work best?

- They select correlated genes, whereas wrappers/embedded methods tend to pick very different genes.
- They assume no particular hypotheses: Occam's razor principle?
- The model we choose is simplistic: high bias, low variance?

Is is a statistically impossible problem?

- $n << p$ problem: well-known to be tricky
- But some methods are designed to overcome the problem (e.g. Lasso and other L_1-penalized classifiers)
- Remaining issue: theoretical conditions do not hold (e.g. low correlation between signal and noise)
- We are only in a sparse setting if we assume uniqueness of the signature.

Acknowledgments

Jean-Philippe Vert

Fabien Reyal

Laurent Jacob