Statistically interpretable importance indices for Random Forests

Jérôme Paul Pierre Dupont

Machine Learning Group
Université catholique de Louvain

June 6, 2014

Feature selection to understand complicated problems

Typical example : personalized medicine

- genomic data $\rightarrow$ thousands variables
- clinical data $\rightarrow$ a few tens variables

Physicians want to **understand**

- good predictive model : not enough
- provide a reduced set of **relevant** variables

Feature selection ?

$$X^{n \times p} = \begin{pmatrix} & & \\ & \cdots & \\ & & \end{pmatrix} \rightarrow Y^{n \times 1}$$

n data points
p variables
Y labels vector

Feature selection ?

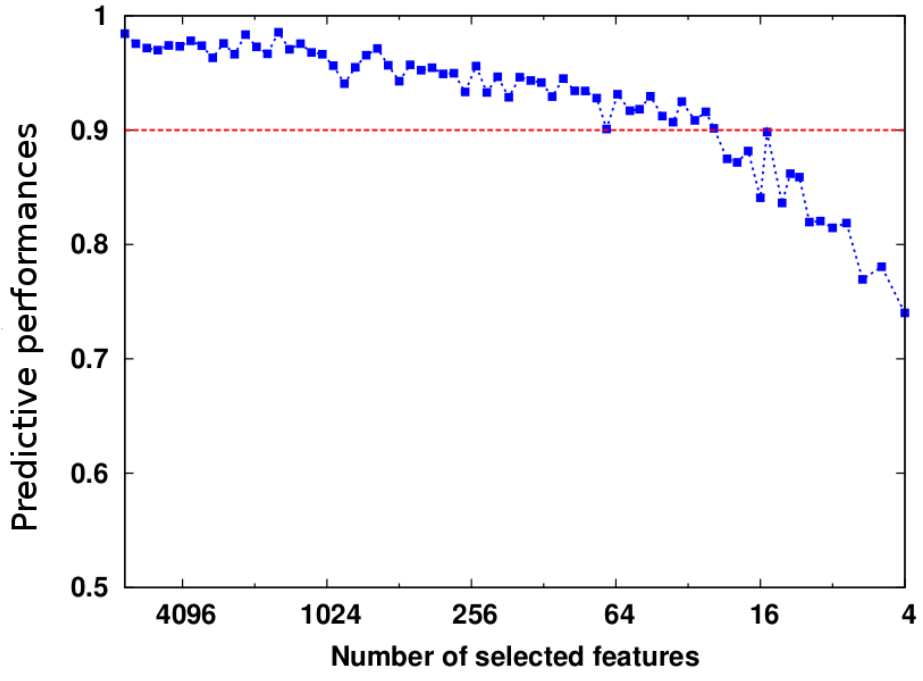
$$X' = \begin{pmatrix} \\ \cdots \\ \\ \end{pmatrix} \rightarrow Y^{n \times 1}$$

n data points
small subset of variables
Y labels vector

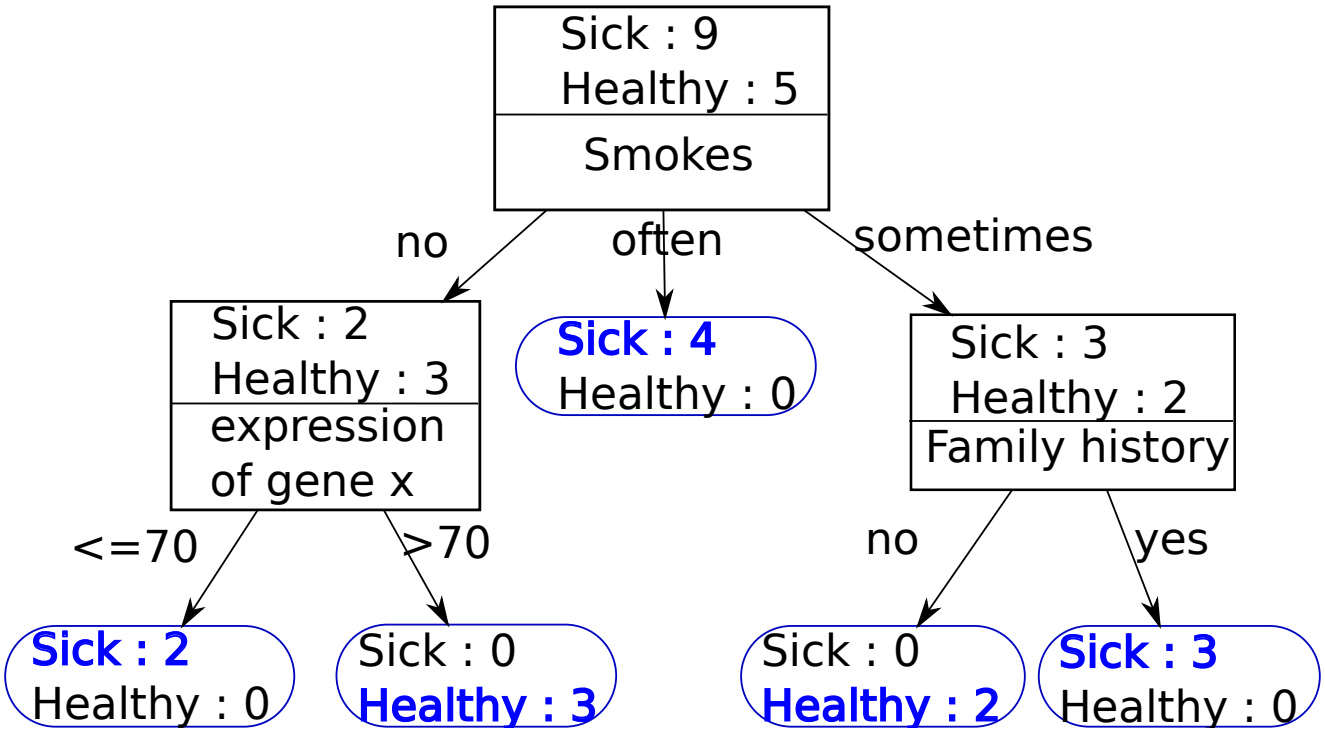
Feature selection ?

$$X' = \begin{pmatrix} & \\ \cdots & \\ & \end{pmatrix} \rightarrow Y^{n \times 1}$$

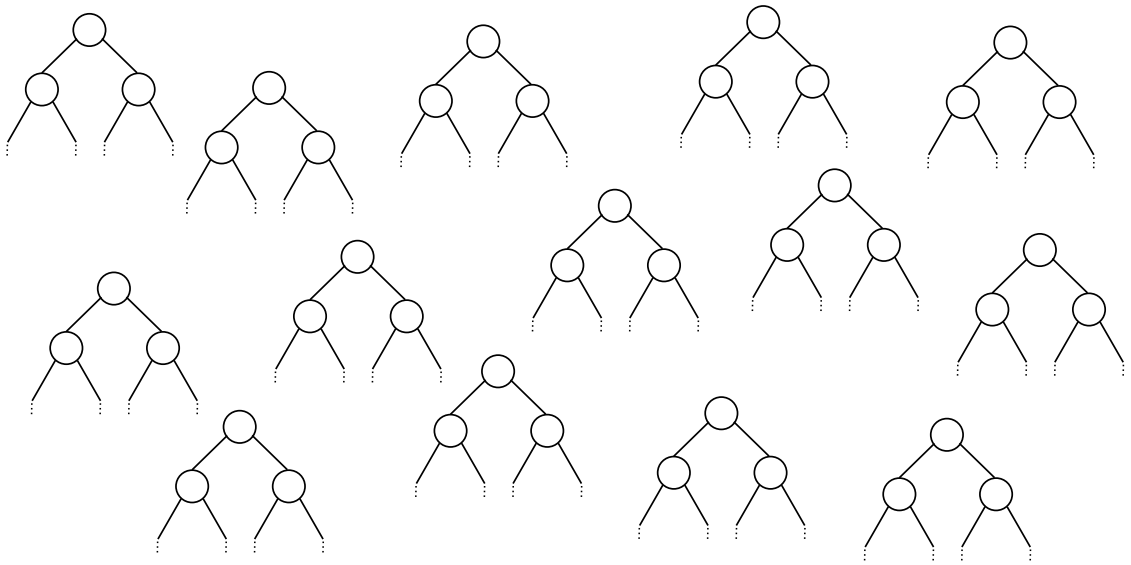
n data points
small subset of variables
Y labels vector



Random Forest : a decision tree based approach



Random Forest : a decision tree based approach



Random Forest : T trees

- each tree k
  - grown from a bootstrap sample B_k of the n data points
  - corresponding out-of-bag $\overline{B_k}$
- final prediction by a majority vote

Outline

Feature selection from RF

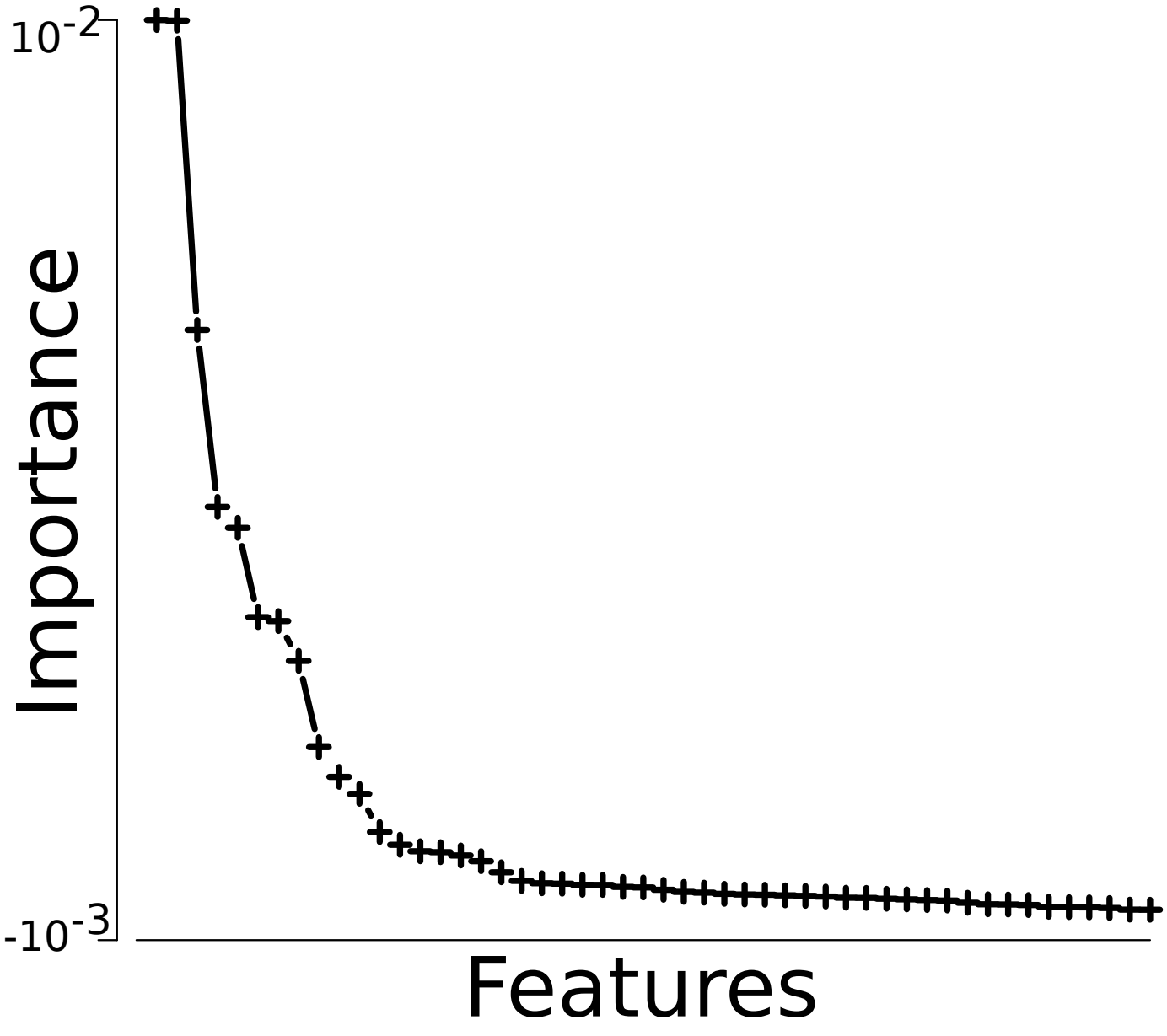
Importance of x_j = average decrease in accuracy over all trees when randomly permuting its values $\rightarrow \tilde{x}_j$

$$J_a(x_j) = \frac{1}{T}\sum_{k=1}^{T}\left(ACC_k(\overline{B}_k) - ACC_k(\overline{B}_k^{\tilde{x}_j})\right)$$

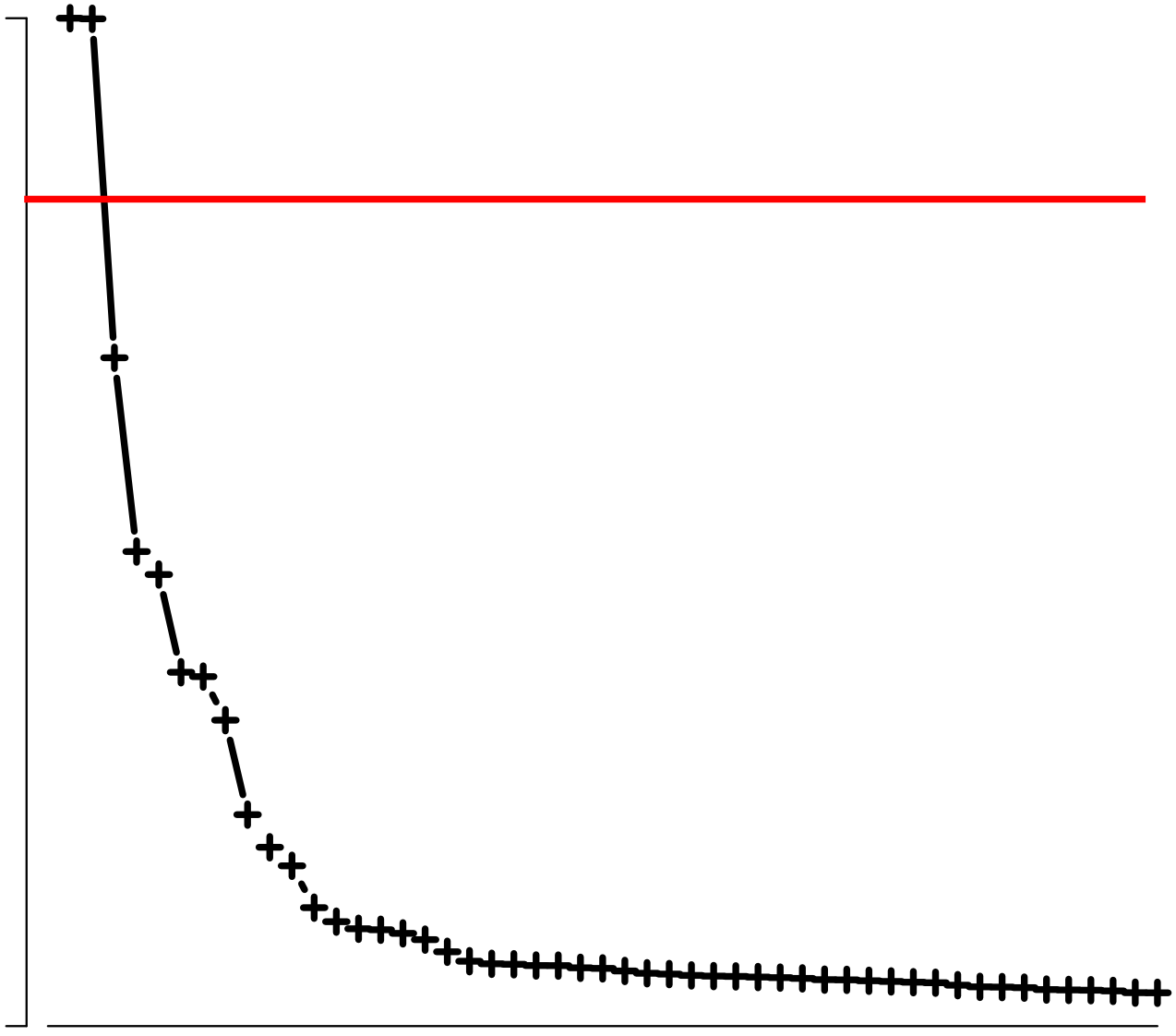
where

- $\overline{B}_k$ = out-of-bag of k-th tree
- $\overline{B}_k^{\tilde{x}_j}$ = out-of-bag of k-th tree with x_j permuted
- $ACC_k(\cdot)$ = accuracy of k-th tree's predictions

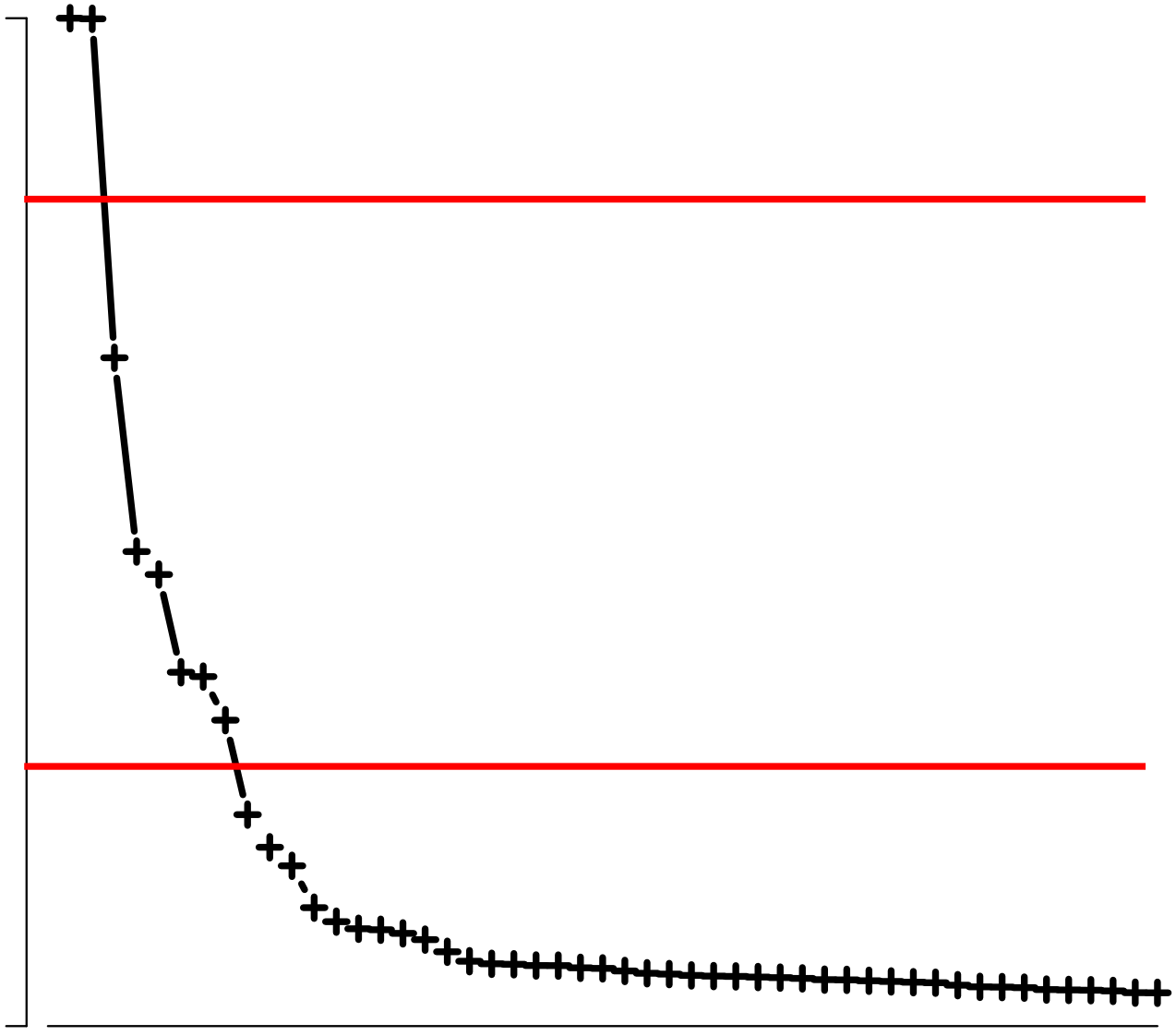
Highlighting significant features is complex



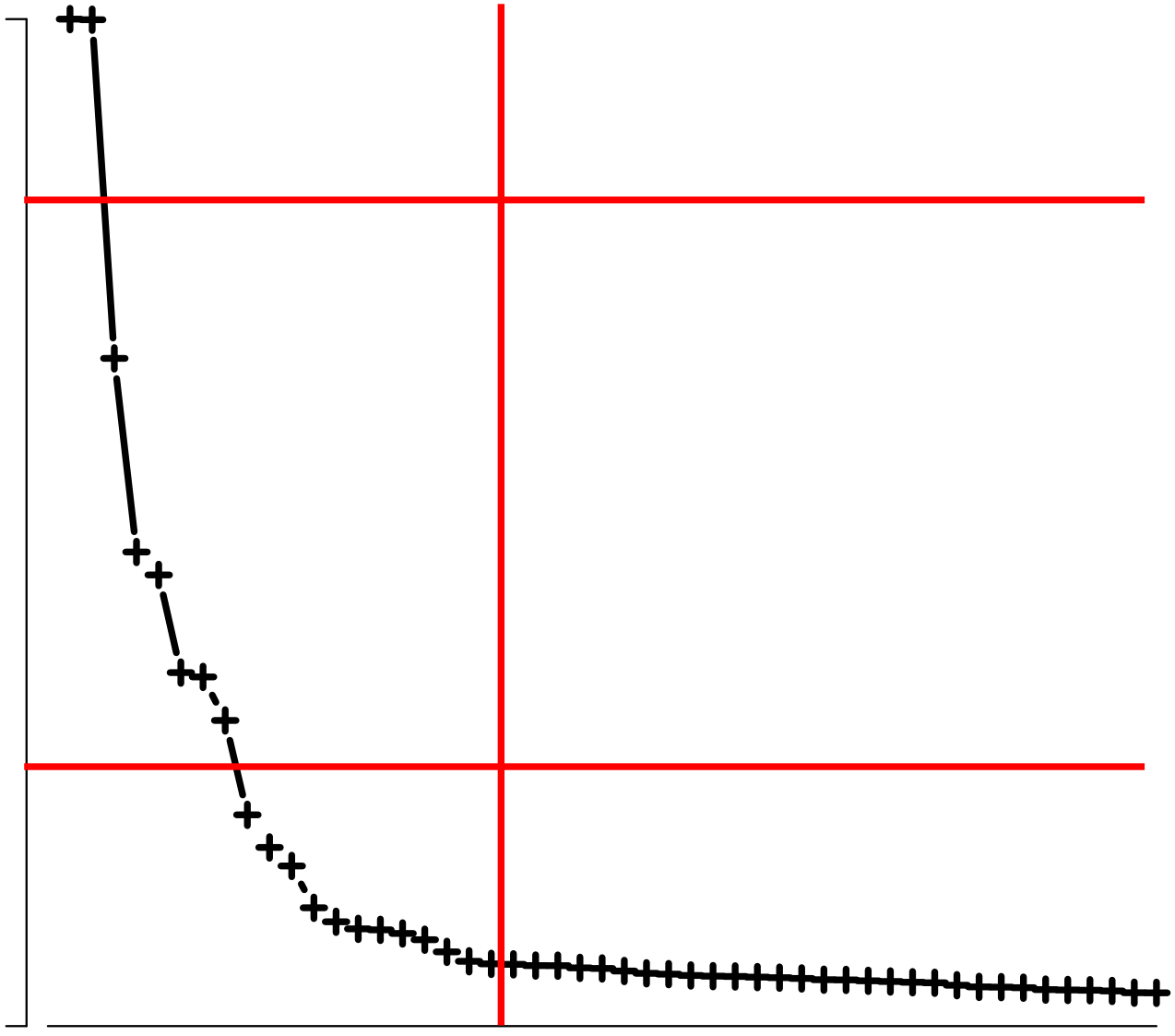
Highlighting significant features is complex



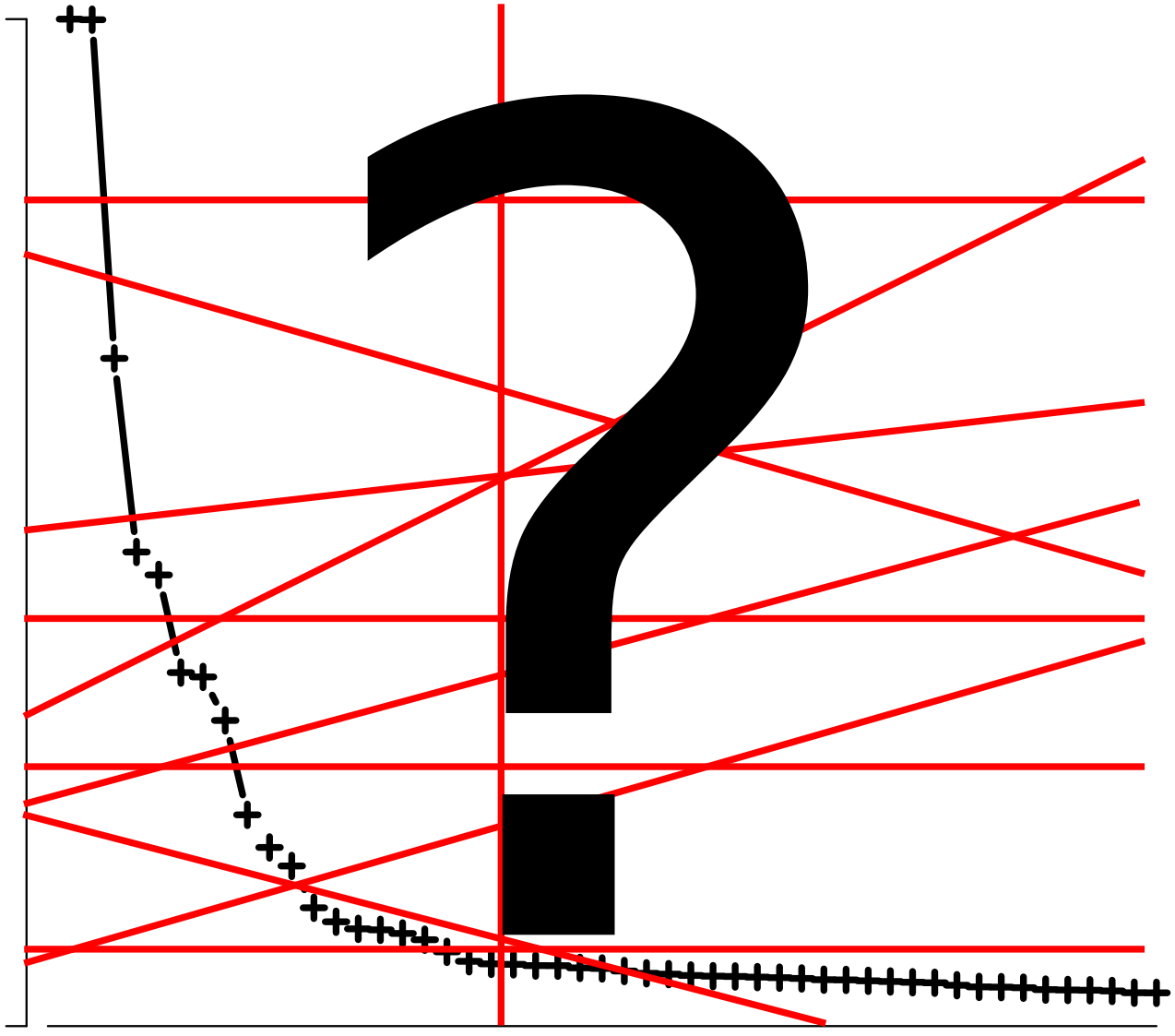
Highlighting significant features is complex



Highlighting significant features is complex



Highlighting significant features is complex



Outline

Assess change in class vote distribution

Permuting a variable may change prediction $\rightarrow$ class vote distribution

Analysis of changes in class vote distribution

- estimated on out-of-bag samples
- assessed by

  Pearson's χ^2 test
Kolmogorov-Smirnov test $\Big\} \rightarrow$ p-value $\rightarrow$ interpretability

J_{χ^2}: Confusion matrices instead of accuracies

For k-th tree, classification of

- out-of-bag $\overline{B}_k$

true \ pred	0	1
0	a_k	b_k
1	c_k	d_k

J_{χ^2}: Confusion matrices instead of accuracies

For k-th tree, classification of

- out-of-bag $\overline{B}_k$

true \ pred	0	1
0	a_k	b_k
1	c_k	d_k

- $\overline{B}_k^{\tilde{x}_j}$ with x_j permuted

true \ pred	0	1
0	$a_k^{\tilde{x}_j}$	$b_k^{\tilde{x}_j}$
1	$c_k^{\tilde{x}_j}$	$d_k^{\tilde{x}_j}$

J_{χ^2}: Confusion matrices instead of accuracies

For k-th tree, classification of

- out-of-bag $\overline{B}_k$

true \ pred	0	1
0	a_k	b_k
1	c_k	d_k

- $\overline{B}_k^{\tilde{x}_j}$ with x_j permuted

true \ pred	0	1
0	$a_k^{\tilde{x}_j}$	$b_k^{\tilde{x}_j}$
1	$c_k^{\tilde{x}_j}$	$d_k^{\tilde{x}_j}$

↓ ↙

class vote	is x_j permuted: No	Yes
	a_k	$a_k^{\tilde{x}_j}$
	b_k	$b_k^{\tilde{x}_j}$
	c_k	$c_k^{\tilde{x}_j}$
	d_k	$d_k^{\tilde{x}_j}$

defined for

- one tree
- one variable

J_{χ^2}: χ^2 test to assess differences in class vote distributions

Global class vote distribution for variable x_j:

class vote	is x_j permuted: No	Yes
	a	$a^{\tilde{x}_j}$
	b	$b^{\tilde{x}_j}$
	c	$c^{\tilde{x}_j}$
	d	$d^{\tilde{x}_j}$

$$= \sum_{k=1}^{T} \left(\begin{array}{c|cc} \text{class vote} & \text{is } x_j \text{ permuted: No} & \text{Yes} \\ \hline & a_k & a_k^{\tilde{x}_j} \\ & b_k & b_k^{\tilde{x}_j} \\ & c_k & c_k^{\tilde{x}_j} \\ & d_k & d_k^{\tilde{x}_j} \end{array} \right)$$

$J_{\chi^2}(x_j) =$ p-val of Pearson's χ^2 test
Low p-value $\rightarrow x_j$ is important
. . . corrected for multiple testing

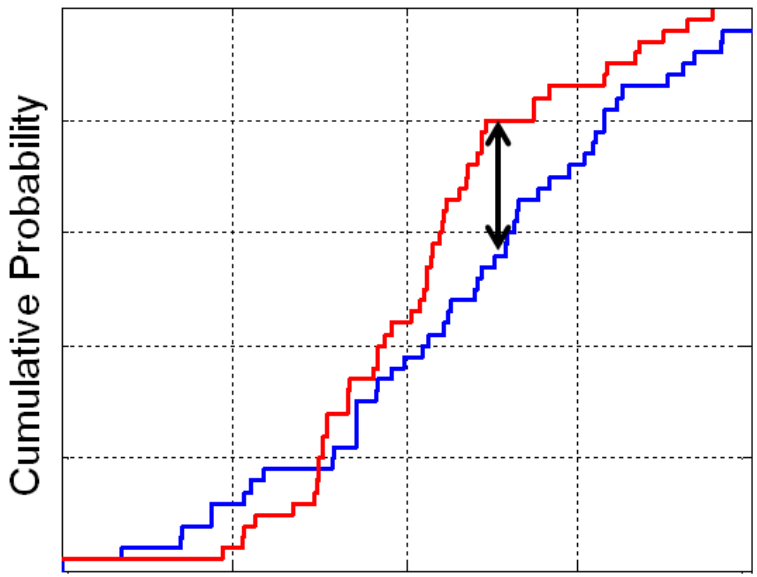
J_{ks}: KS-test to assess differences in accuracy distributions

For k-th tree, classification accuracy of

- $ACC_k(\overline{B}_k)$: out-of-bag
- $ACC_k(\overline{B}_k^{\tilde{x}_j})$: out-of-bag with x_j permuted

Non-parametric Kolmogorov-Smirnov test with T observations for each distribution

$J_{ks}(x_j) =$ p-val of KS-test
Low p-value $\rightarrow x_j$ is important
. . . corrected for multiple testing



Outline

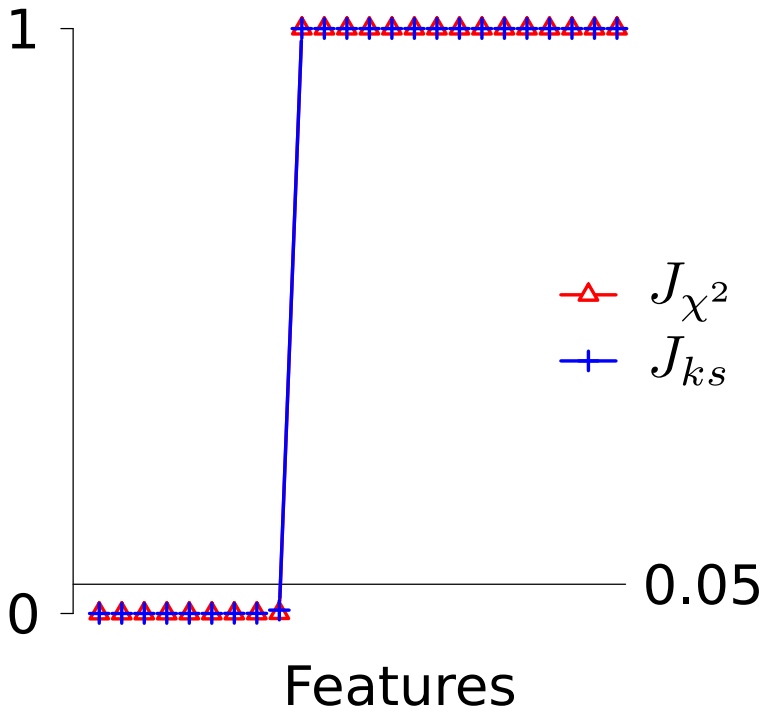
Are J_{χ^2} and J_{ks} able to identify important variables ?

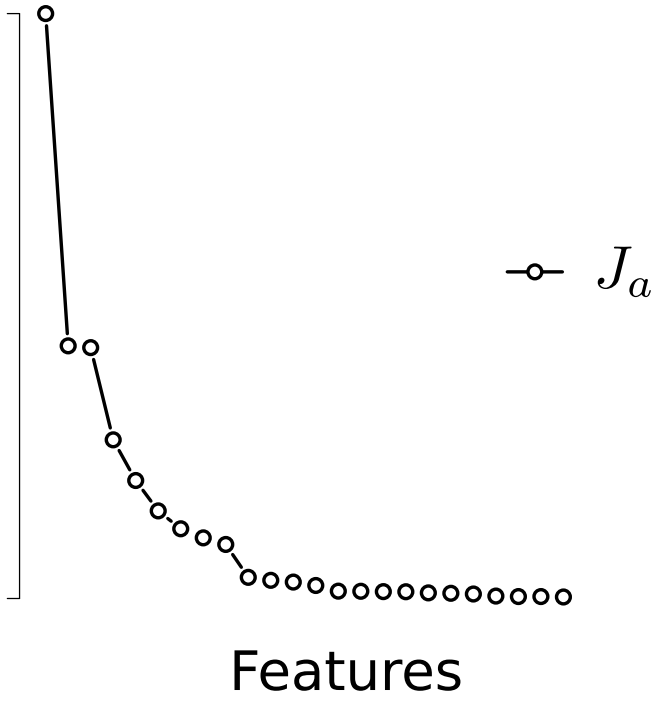
Artificial dataset

- $X \in \mathbb{R}^{500 \times 110} \sim \mathcal{N}(0, 1)$
- $w \in \mathbb{R}^{110}$
  - $w[1\!:\!10] \sim \mathcal{U}(0.5, 1) \rightarrow$ relevant features
  - $w[11\!:\!110] = 0$
- $y = sign(Xw)$

Threshold = 0.05 after correction for multiple testing

J_{χ^2} and J_{ks} are able to highlight relevant variables





Require a large amount of trees (here 10,000)
to gain confidence that features are relevant

Comparison to a recent alternative : 1probe

Build N rankings with J_a
on a dataset with an additional random feature $x_{rand} \sim \mathcal{N}(0, 1)$

p-values that x_j is important

$$= \frac{\#\text{ times } x_{rand} \text{ is better ranked than } x_j}{N}$$

Huynh-Thu, V. A. A., Saeys, Y., Wehenkel, L., & Geurts, P. (2012).
Statistical interpretation of machine learning-based feature importance scores for biomarker discovery.
Bioinformatics (Oxford, England), 28, 1766–1774.

Predictive performances with only significant variables

$200\times$ resamplings

- train $=$ 90% data
  - rank variables with a forest of T trees
  - $s =$ set of significant variables : p-val < 0.05
  - train a final model using only features in s
- predict on remaining 10% data

$$BCR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

Name	Class priors	p
Arrhythmia	245/185	262
Musk1	269/207	166
Golub	25/47	7129
Lymphoma	22/23	4026
Prostate	52/50	6033

Comparison for same number of trees

		T or $N \times T$	s	BCR
arrhythmia	J_{χ^2}	1000	6.17	0.78
	J_{ks}	1000	2.62	0.70
	1Probe	100x10	0.26	0.52
	J_{χ^2}	5000	24.19	0.84
	J_{ks}	5000	22.33	0.84
	1Probe	100x50	4.92	0.74
musk1	J_{χ^2}	1000	12.15	0.81
	J_{ks}	1000	5.06	0.73
	1Probe	100x10	0.01	0.50
	J_{χ^2}	5000	63.23	0.90
	J_{ks}	5000	60.52	0.89
	1Probe	100x50	6.51	0.76

Comparison for same number of trees : genomic datasets

		T or $N \times T$	s	BCR
golub	J_{χ^2}	10,000	10.80	0.96
	1Probe	100x100	0.42	0.64
	J_{χ^2}	100,000	40.83	0.97
	1Probe	100x1,000	67.00	0.97
lymphoma	J_{χ^2}	10,000	4.83	0.93
	1Probe	100x100	0.18	0.54
	J_{χ^2}	100,000	27.72	0.94
	1Probe	100x1,000	36.45	0.94
prostate	J_{χ^2}	10,000	7.89	0.94
	1Probe	100x100	2.98	0.92
	J_{χ^2}	100,000	41.52	0.94
	1Probe	100x1,000	50.20	0.94

Conclusion

Statistically interpretable indices that measure class vote distribution shifts

- p-value $\rightarrow$ natural threshold to highlight important variables
- multivariate
- require an order of magnitude less trees than recent alternatives

Future work

- other ensemble methods
- study size effect

http://www.info.ucl.ac.be/~jpaul/