Feature Selection

B. Ghattas & G. Oppenheim

Université d'Aix-Marseille - Université Paris-Sud

badihghattas@gmail.com, georges.oppenheim@gmail.com

Outline

Plan

Motivation

- High dimension, dimension reduction,
  - Reducing computation time and memory
  - Reducing the ration p/n necessary for some techniques...
  - Reducing expenses...
  - Increasing readability and/or Interpretability...
- Reducing Noise
- Increasing accuracy ?

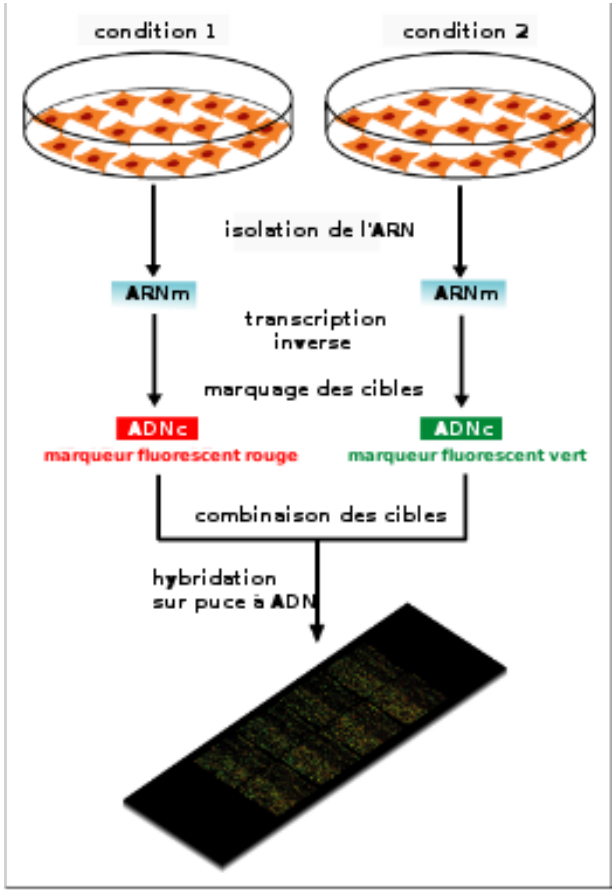
Motivation

- High dimension, dimension reduction,
  - Reducing computation time and memory
  - Reducing the ration p/n necessary for some techniques...
  - Reducing expenses...
  - Increasing readability and/or Interpretability...
- Reducing Noise
- Increasing accuracy ?

Motivation

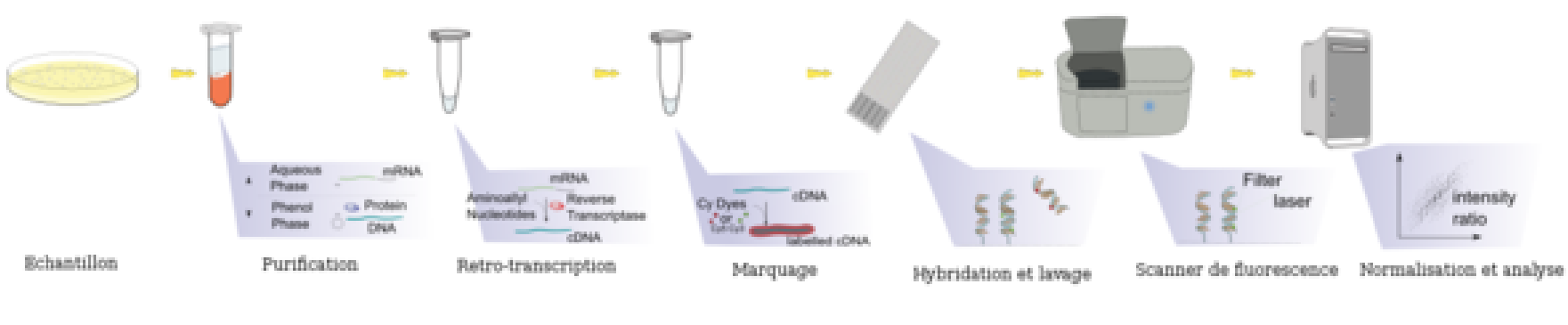
- High dimension, dimension reduction,
  - Reducing computation time and memory
  - Reducing the ration p/n necessary for some techniques...
  - Reducing expenses...
  - Increasing readability and/or Interpretability...
- Reducing Noise
- Increasing accuracy ?

Application: Microarray data



Steps for acquiring microarray data

Different steps are needed befor getting the dataset. They may depend on the type of microarray used...



Application: Microarray data

A particular attention for situations where $n << p$ (Sparse data), In Microarray data, typically $n \sim 100, p \sim 10^5$

Y
+1
...
-1
+1

	g_1	...	...	g_p
C_1				
...				
...				
C_n				

Which are the genes that give the best discrimination between the presence and absence of a cancer ?

What is feature selection ?

It may be seen as a combinatorial problem....

Suppose we have a set $S = \{X_1, ..., X_p\}$ variables and a score $J(S)$ coomputed from observations of these variables....

We wish to find a subset $S_{p'} \in S$ of p' variables, $p' << p$ selected among those of S, such that $J(S_{p'}) \sim J(S)$.

The number p' may be fixed or not ...

J may be ralated only to S, or to any supervised learning question with respect to a variable Y.

Different approaches...

Selecting a subset of variables is a NP hard problem, even when its cardinal is fixed in advance.
Some approaches do not need a specific statistical learning model...Others, are based on a specific regression or classification model...like SVM, CART, Random Forests

- Filters
- Wrappers
- Embedded

Different approaches...

Selecting a subset of variables is a NP hard problem, even when its cardinal is fixed in advance.
Some approaches do not need a specific statistical learning model...Others, are based on a specific regression or classification model...like SVM, CART, Random Forests

- Filters
- Wrappers
- Embedded

Different approaches...

Selecting a subset of variables is a NP hard problem, even when its cardinal is fixed in advance.
Some approaches do not need a specific statistical learning model...Others, are based on a specific regression or classification model...like SVM, CART, Random Forests

- Filters
- Wrappers
- Embedded

Filters

Filter type methods

- select variables regardless of the model.
- They are based only on general features like the correlation with the variable to predict.
- They suppress the least interesting variables.

- These methods are particularly effective in computation time and robust to overfitting.
- However, filter methods tend to select redundant variables because they do not consider the relationships between variables.

Therefore, they are mainly used as a pre-process method.

Wrappers

Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. The two main disadvantages of these methods are :

- The increasing overfitting risk when the number of observations is small.
- The significant computation time when the number of variables is large.

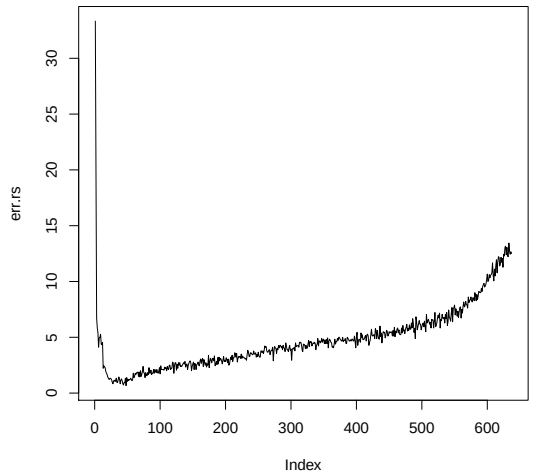
Embedded

Embedded methods combine the advantages of both previous methods. The learning algorithm takes advantage of its own variable selection algorithm. So, it needs to know preliminary what a good selection is, which limits their exploitation.
Examples of theses approaches are LASSO, L1-SVM, ...

- Less computationally expensive
- Less prone to overfitting

A three steps approach

- First order the variables.
- Next introduce them sequentiallay within the model monitoring its performance evolution.
- Localize the optimal number of variables to keep in the model.



Performance Evaluation

Given a FS approach.. How can we evaluate its performence ? We need for that a criterion...

- Natural evaluation when we are in a supervised learning framework....
- If not... The choice of the criterion may be not evident ....

Plan

Some univariate approaches

In a supervised context, the main idea is to measure the intensity of the *link* between each input variable X_j and the target variable Y. Such measures depend on the nature of the specific target variable, (descrete or continuous) and the input variables.

- Both continuous: Correlation, sometimes Mutual Information.
- Both discrete: χ^2, Mutual Information.
- Y discrete, X continuous: Fisher Discriminative Score, T-tests, multiple testing with corrections....

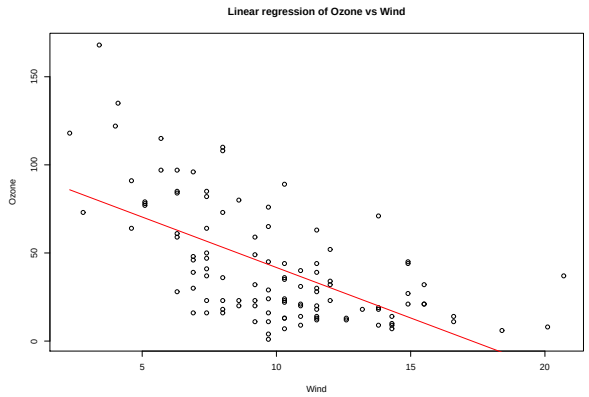
In an Unsupervised context,very few approaches exist:
Results...

- Are (relatively) robust against overfitting
- May fail to select the most "useful" features

Pearson Linear Correlation

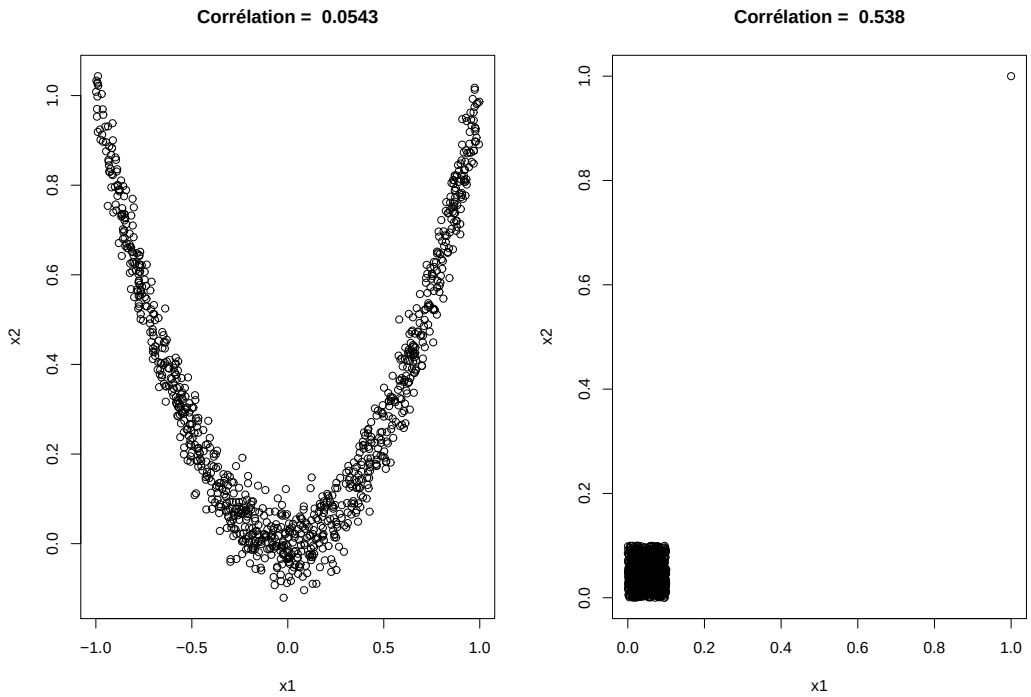
$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{\frac{1}{n}Cov(X,Y)}{s_x s_y}$$

The covariance may be also written: $\frac{1}{n}\sum x_i y_i - \overline{xy}$.



Defaults of the linear correlation

Non Correlation does no imply Independence
Correlation is very sensitive to extremes and outliers



Two qualitative variables, Example

$n = 120$ observations of variables
Sex (2 labels, "H" and "F")
Eyes' colors (3 labels, "B", "M" and "V").

	B	M	V	
F	20	15	19	54
H	21	16	29	66
	41	31	48	120

Dividing by n:

	B	M	V
F	0.167	0.125	0.158
H	0.175	0.133	0.242

Two qualitative variables, Cross tables

$$X = Sex \in \{a_1 = "H", a_2 = "F"\}$$
$$Y = Eyes' \ colors, \in \{b_1 = "'B"', b_2 = "'M"', b_3 = "'V"'\}$$

	b_1	b_2	...	b_j	...	b_c	Marge colonne
a_1	n_{11}	n_{12}				n_{1c}	n_{1+}
a_2	n_{21}					n_{2c}	n_{2+}
.							
a_i				n_{ij}			n_{i+}
.							
a_l	n_{l1}					n_{lc}	n_{l+}
Marge ligne	n_{+1}	n_{+2}		n_{+j}		n_{+c}	n

n_{ij} number of times where $X = a_i$ and $Y = b_j$.

Conditional distributions

Divide each line by its sum:

	B	M	V
F	0.370	0.278	0.352
H	0.318	0.242	0.439

Divide each column by its sum:

	B	M	V
F	0.488	0.484	0.396
H	0.512	0.516	0.604

Joint and conditional distributions

Denote the frequency table : $\frac{n_{ij}}{n}$.
Each line divided by its sum gives the lines profiles : $\{\frac{n_{ij}}{n_{i+}}\}_{j=1..c}$.
Each column divided by its total gives columns profiles: $\{\frac{n_{ij}}{n_{+j}}\}_{i=1..l}$

Remarks

Line profiles are estimation of the distribution of Y knowing a fixed value of X, $P[Y|X = a_i]$.
Column profiles are estimation of the distribution of X knowing a fixed value of Y, $P[X|Y = b_j]$

Mutual Information

For two discrete variables X and Y taking each a finite number of values in the sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, the mutual information of X and Y is defined by:

$$I_{(X,Y)} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{xy} log \left(\frac{p_{xy}}{p_x p_y} \right)$$

where $p_x = P[X = x]$, $p_y = P[Y = y]$, and $p_{xy} = P[X = x, Y = y]$

- $I(X, Y) \geq 0$ for each X and Y.
- $I(X, Y) = 0$ if and only if the random variables X and Y are independent.
- $I(X, Y) = I(Y, X)$

MI, Example

	B	M	V	
F	20	15	19	54
H	21	16	29	66
	41	31	48	120

P_{xy}

	B	M	V
F	0.167	0.125	0.158
H	0.175	0.133	0.242

$$P_x = (\tfrac{54}{120}, \tfrac{66}{120})$$

$$P_y = (\tfrac{41}{120}, \tfrac{31}{120}, \tfrac{48}{120})$$

$$MI(X,Y) = 0.167 log\left(\frac{0.167}{\frac{54}{120} \times \frac{41}{120}}\right) + 0.125 log(\left(\frac{0.125}{\frac{54}{120} \times \frac{31}{120}}\right) + \ldots$$

Expected frequencies

Those are the expected frequencies in case of independence of the two crossed criteria

	B	M	V	Total
F	$\frac{41 \times 54}{120}$			54
H				66
Total	41	31	48	120

	B	M	V
F	18.4	14	21.6
H	22.6	17	26.4

Deviation from independence

recall: Two events A et B are independent If $P(A \text{ et } B) = P(A) * P(B)$.

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = \frac{n_{+j} n_{i+}}{n}$ are the expected frequencies.
This index is positive, closer it is to zero more the variables may be suspected to be independent.
When both variables are binary we may compute also

- True and False positive and negative rates
- Sensitivity and specificity...

Example....

Expected frequencies E_{ij}:

	B	M	V
F	18.4	14	21.6
H	22.6	17	26.4

Deviation $n_{ij} - E_{ij}$:

	B	M	V
F	1.55	1.05	−2.6
H	−1.55	−1.05	2.6

Quadratic differences $(n_{ij} - E_{ij})^2$:

	B	M	V
F	2.4	1.1	6.76
H	2.4	1.1	6.76

Normed quadratic deviation $(n_{ij} - E_{ij})^2 / E_{ij}$:

Table: Khi2 = 0.949

	B	M	V
F	0.130	0.0790	0.313
H	0.107	0.0647	0.256

X continuous, Y binary: Fisher Discriminative Score

Y is discrete, but X is continuous.. *This index is appropriate only when the target variable is binary.*

$$FDS(k) = \left| \frac{\mu_k^+ - \mu_k^-}{\sigma_k^+ + \sigma_k^-} \right| \quad ; \quad k = 1, 2, \ldots p,$$

where $\mu_k^{\pm}$ is the mean of the k^{th} variable for positive and negative groups, and $\sigma k^{\pm}$ is the *sd*.
The variable maximizing this score may be considered as the most important.

Using T-tests, or ANOVA

Test if the averages of two subgroups are equal. Choose one decision among the two hypothesis:

$$\begin{cases} H_0 : & m_x = m_y \\ \text{vs} & \\ H_1 : & m_x \neq m_y, \end{cases}$$

To do that fix a risk $\alpha = 5\%$ and use the statistic which gives a good information about the deviation between the means and whose distribution is known under H_0:

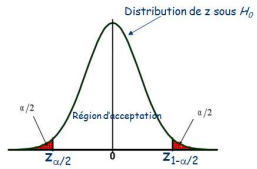
$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

Compute this statistic.

T test...

X	0.77	0.33	-2.00	0.79	-1.50	-1.60	-0.73	-0.14	0.20	1.60	-2.40	2.00	0.49	-0.83	-1.40	-0.11	0.17	0.05
Y	H	H	F	H	F	F	H	H	H	H	F	F	F	H	H	H	H	H

signif	F	H
Means	−0.67	0.042
SD	1.60	0.770



Place the empirical value and its symmetric on the curve. The $p - value$ is the area outside the interval...

Ordering the variables

A lower p-value indicates more confidence in the rejection of the hypothesis of equal means.

Multivariate methods...Filters vs wrappers

The main goal is to **rank subsets of useful features**.

- Filters select a subset once...
- Wrappers Select a subset, estimate a model performance, and loops over both steps.

The main danger is over-fitting with intensive search.

Sequential search

- Sequential Forward Selection (**SFS**) and Backward (**SBS**), Inserting sequentially till a stopping rule is satisfied.
- **GSFS(g)**: generalized sequential forward selection Ű try at each step to include a subset of g features among $(p-k)$ remaining. More trainings at each step $(\binom{g}{p-k})$, but fewer steps.
- **PTA(l,r)**: plus l , take away r at each step, run SFS l times then SBS r times.
- Floating search (**SFFS** and **SBFS**): One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far. Any time, if a better subset of the same size was already found, switch abruptly.

- Advantage: Do not need a specific model, but a monotonic criterion over a set of variables.
- Drawbacks: computational complexity, depend on the order of variables in the data.

Embedded methods

Make use of a statistical learning algorithm (SVM, CART, RF, ...).
Data is generally split in three parts:

- Learning sample, to learn the model
- Validation sample, to validate the choice of the features at each step
- Testing sample, to estimate the performance of the model.

Plan

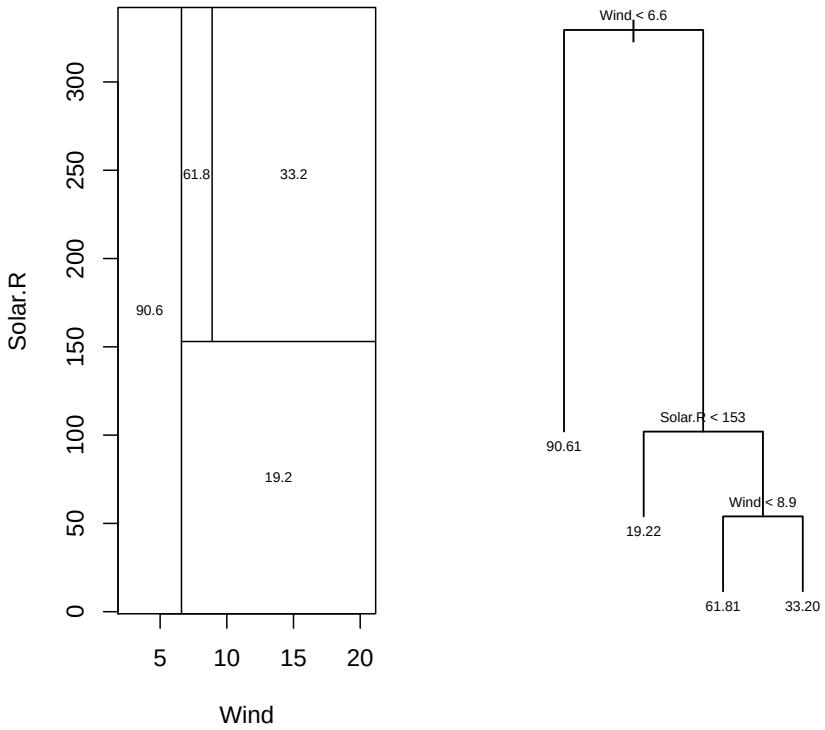
Ranking variables within the model learning process

Variables ranking may be done through the model estimation...
Some supervised models suggest a measure of variables importance related to the model learning...

- CART
- Random Forests
- Linear models

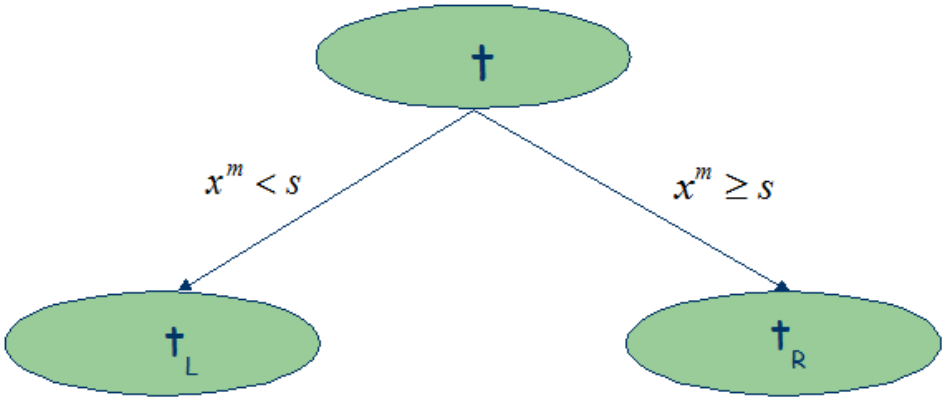
Once variables are ranked a selection process may be used using this ranking ...
Some approaches combine both steps, ranking+selection, simultaneously:
That's typically based on penalization: LASSO, LARS, L1 SVM, ...
Similar ideas exist also in an unsupervised context: L1-kmeans, .

Example


Solar.R
Wind
0
50
100
150
200
250
300
5
10
15
20
61.8
33.2
90.6
19.2
Wind < 6.6
Solar.R < 153
Wind < 8.9
90.61
19.22
61.81
33.20

2 stages: Maximal Tree and Pruning

All the observations are in the root node.



Splitting rule: one variable and a threshold. How to do ?
Use the deviance to measure the heterogeneity of a node:

$$R(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2$$

Optimal Splits: minimize the children's deviance

Minimize total new nodes Heterogeneity. Let s be a split of the form: $x^m < a$,

$$\Delta R(s,t) = R(t) - (R(t_L) + R(t_R)) \geq 0$$

$$\Delta R(s,t) = max_{s \in \Sigma} \Delta R(s,t)$$

In classification,

$$R(t) = -\sum_{j \in J} p_j(t) log(p_j(t))$$

where $p_j(t)$ prior probability for each class j in t.

Substitution splits

Let s be any apparent split of a node t of the actual tree, splitting t into t_L and t_R. Let s_j a split over the j^{th} variable chosen from the set S_j of all the possible splits (for continuous variables the splits having the form $x_j < a$ or $x_j > a$). s_j gives rise to two sub nodes t'_L and t'_R. The probability for an observation to be at the left for both splits is:

$$p(t_L \cap t'_L) = \frac{\#\{t_L \cap t'_L\}}{n_t}$$

The probability that both splits send an observation to the left is:

$$p_{LL}(s, s_j) = \frac{p(t_L \cap t'_L)}{p(t)}$$

p_{RR} maybe defined equivalently.

Substitution splits and Variable Importance

The probability that s_j predicts well s is

$$p(s, s_j) = p_{LL} + p_{RR}$$

$\tilde{s}_j$ is a substitution split for s if

$$p(s, \tilde{s}_j) = \max_{S_j} p(s, s_j)$$

The importance of variable j is given by:

$$I(X_{.j}) = \sum_t \Delta(R(\tilde{s}_j, t))$$

which the total of the deviance reduction induced if each split in the tree was replaced by the substitution split over $X_{.j}$.

Handling Missing values

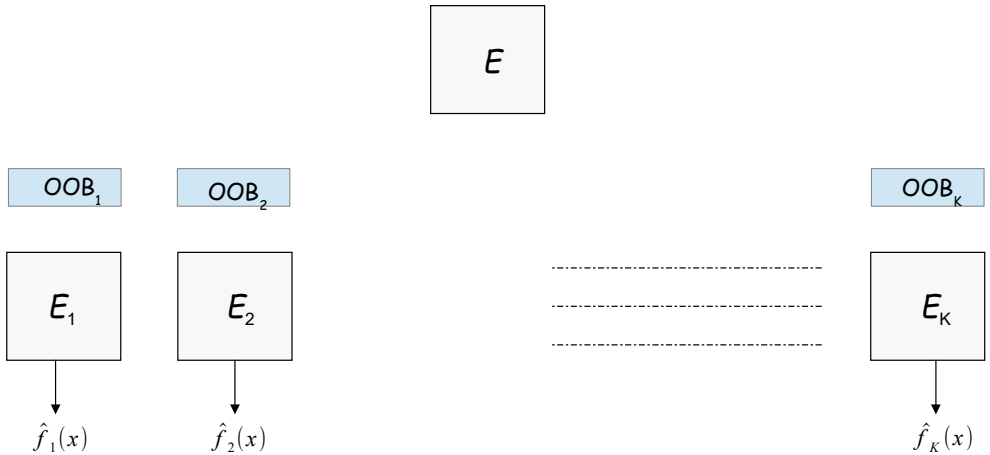
For the prediction case, one may use substitution splits to follow a path into the tree when the observation of the apparent variable is missing.
Another possibility is to use the empirical proportion of observations at each side of a node in order to choose the direction to follow.

Random Forests, (L. Breiman, 2001)

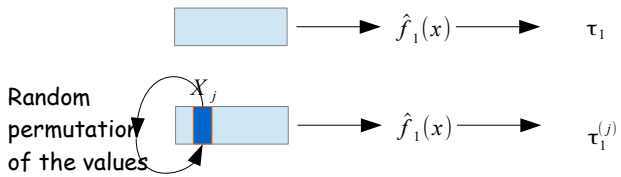
- K bootstrap Samples, keeping the out of bag samples.
- Construct a Maximum tree over each one, using best split over very few variables randomly selected.
- Don't prune.
- Aggregate trees using mean (regression) or majority vote (classification).

"*Random Input*" uses one variable at each split.
"*Random Features*" uses a linear combination of variables with randomly selected coefficients.
Weak trees + weak correlation between trees (between their predictions) $\rightarrow$ Powerful learner.

Variables importance



For each variable j, for each OOB :



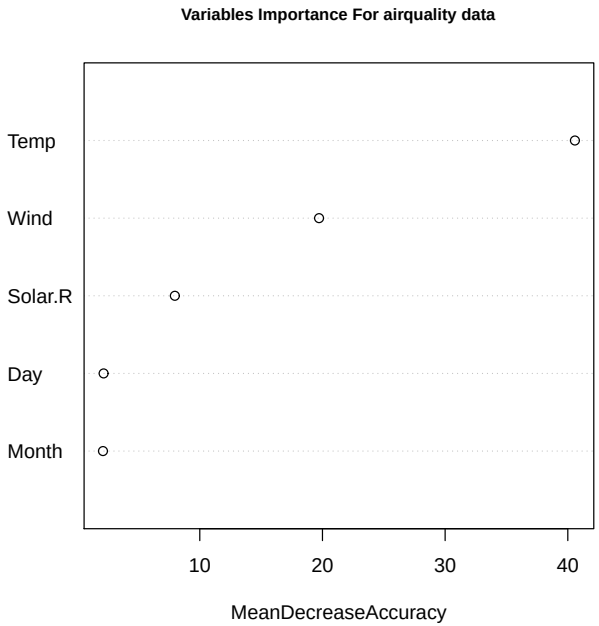
Compute the relative differences : $\frac{\tau_k^{(j)} - \tau_k^{(j)}}{\tau_k}$ Then Importance of variable j : $\frac{1}{K}\sum_{k=1}^{K} \frac{\tau_k^{(j)} - \tau_k^{(j)}}{\tau_k}$

Variables importance

Based on OOB samples, and difference in the performance of a tree when the values of one variable are randomly permuted.

- Consider the prediction error τ_k of the k^{th} tree of the forest over the OOB Sample.
- Permute randomly the values of X_j in the OOB sample and use the modified sample for prediction.
- Measure the prediction error for the modified sample $\tau_k^{'}(j)$
- The Importance measure for variable j is :
  $I(j) = \frac{1}{K} \sum_{k=1}^{K} \frac{\tau_k - \tau_k^{'}(j)}{\tau_k}$

Variables importance - Example



Variables importance- Comments

- Insensitive to the nature of the resampling used (bootstrap samples with or without replacement).
- Stable in presence of correlations between variables.
- Invariant to normalization (using standard deviation of $Z_i(j)$
- Stable w.r.t. data perturbations. Bootstrapping VI is unnecessary.

A linear model

$$y = f(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Parameters $\beta = (\beta_0, ..., \beta_p)$ are estimated using Least Squares, that is their optimal values minimizes the MSE:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})]^2$$

LASSO, Penalization

The MSE criterion is Penalized: A constraint is added over the coefficients values, type L_1 constraint:

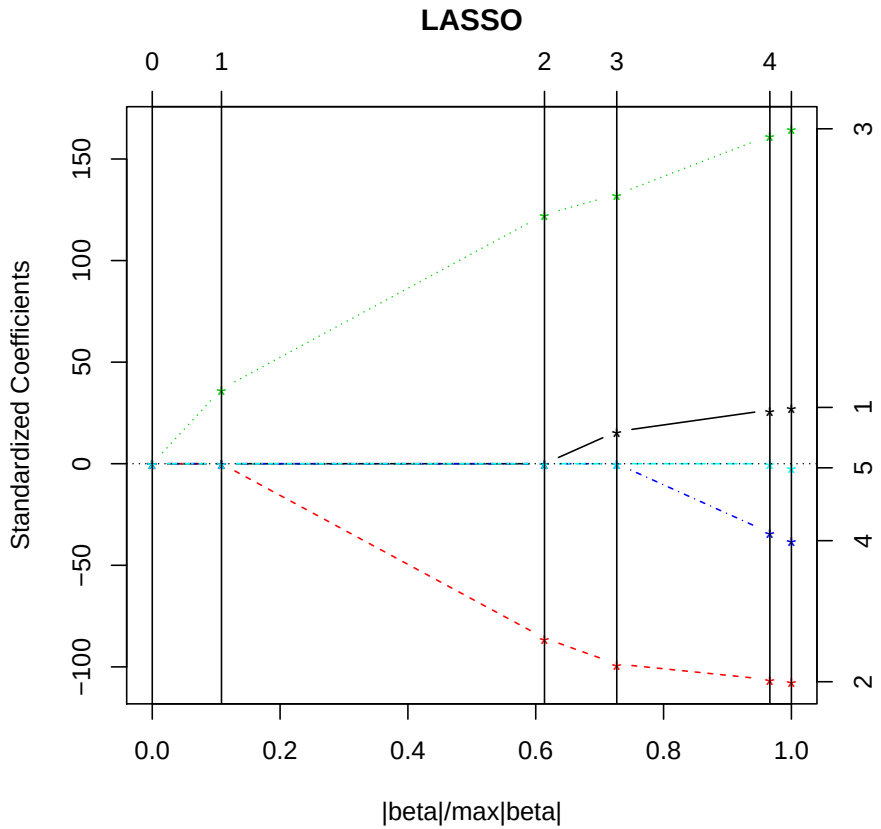
$$\widehat{\beta}(\lambda) = \mathit{argmin}_{\beta} \left\{ \mathit{MSE} + \lambda \left\| \beta \right\|_1 \right\}$$

where

$$\left\| \beta \right\|_1 = \sum_{j=1}^{p} |\beta_j|$$

where $\lambda > 0$ is a regularization parameter.
The sequence $\widehat{\beta}(\lambda), 0 < \lambda < \infty$ is called the *path*.
For $\lambda = \infty$ all the coefficients are equal to zero.
Increasing λ sets more coefficients to zero.

Variables importance - Example



Variables importance

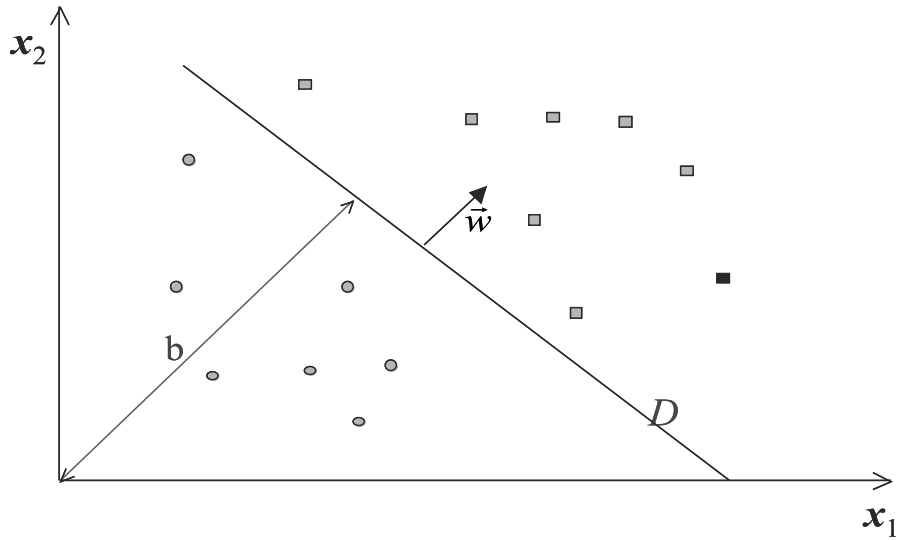
- Use B=500 bootstrap samples.
- Compute the optimal GLM-penalized model, and keep it's coefficients.
- The importance of variable j is the absolute value of it's coefficient's bootstrap mean $\hat{\beta}_j^B$.
- Variables whose coefficient bootstrap mean is zero won't be used for comparisons.

Linear Separation, binary case

$\mathcal{S}$= *n i.i.d.* sample of $(\mathcal{X}, \mathcal{Y}) \subseteq (\mathbb{R}^p, \{-1, +1\})$

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq (\mathcal{X} \times \mathcal{Y})^n .$$

We look for a function: $f(x) = sign(\langle w.x \rangle + b)$



Risk Bounds

- **Radius-margin bound:** For the LOO error estimation (Vapnik [10])

$$\mathcal{L} \leqslant \frac{R^2}{\gamma^2} = R^2 \left\| w^* \right\|^2, \tag{1}$$

$\mathcal{L}$ is the number of misclassified observations by LOO, γ the margin, R radius of the smallest ball covering $\mathcal{S}$.

- **Span bound:** Vapnik and Chapelle [11].

$$\mathcal{L} \leqslant \sum_{i \in sv} \alpha_i^* S_i^2, \tag{2}$$

where the *span* S_i is the distance between the support vectors x_i and a set of constrained linear combination of the other SV.

Scores

Three scores are commonly used :

The weight vector score: $W = \|w^*\|^2$

The Radius score: $RW = R^2 \|w^*\|^2$

The Span Score: $Spb = \sum_{i=1}^{n} \alpha_i^* S_i^2$

Each score may be computed at different orders :

- *"zero-order"* : The value of the score computed omitting that variable.
- *"difference-order"* difference betwwen the score using that variable and its value without it.
- *"first-order"* is the derivative of the score w.r.t. to artificial weights.

We use Bootstrap mean estimates for each score.

Plan

SVM-RFE (Guyon *et al.* [5],Rakotomamonjy [9])

- While there are still variables
  - Learn an SVM and sort variables using the score $||w||^2$ by differences.
  - Estimate its misclassification error.
  - Eliminate half of the variables, the least important if there are more then 100 kept.
- For the last 100 variables, eliminate them recursively one by one.

Ben Ishak et al procedure

$D =$ Learning sample. $B = 200$ Number of bootstrap samples.
Compute the score(D, B) to get a hierarchy $X^{(1)}, \ldots, X^{(p)}$.
For $k = 1, \ldots, p$
    For $l = 1, \ldots, 50$
        Randomly split with stratification $D = A_l \cup T_l$
        A_l is the learning sample T_l the test sample.
        $M_l^k = f\left(X^{(1)}, \ldots, X^{(k)}, A_l\right)$
        $Er_l^k = Test\left(M_l^k, T_l\right)$
    $Er^k = \frac{1}{50} \sum_{l=1}^{50} Er_l^k$
$kopt = Arg \min_k \left\{Er^k\right\}$.

Plan

Toys

$Y \in \{-1, 1\}$ following a uniform distribution.
With probability 0.7,

$$x_i \sim yN(i, 1), i = 1, 2, 3$$

$$x_i \sim yN(0, 1), i = 4, 5, 6$$

else

$$x_i \sim yN(0, 1), i = 1, 2, 3$$

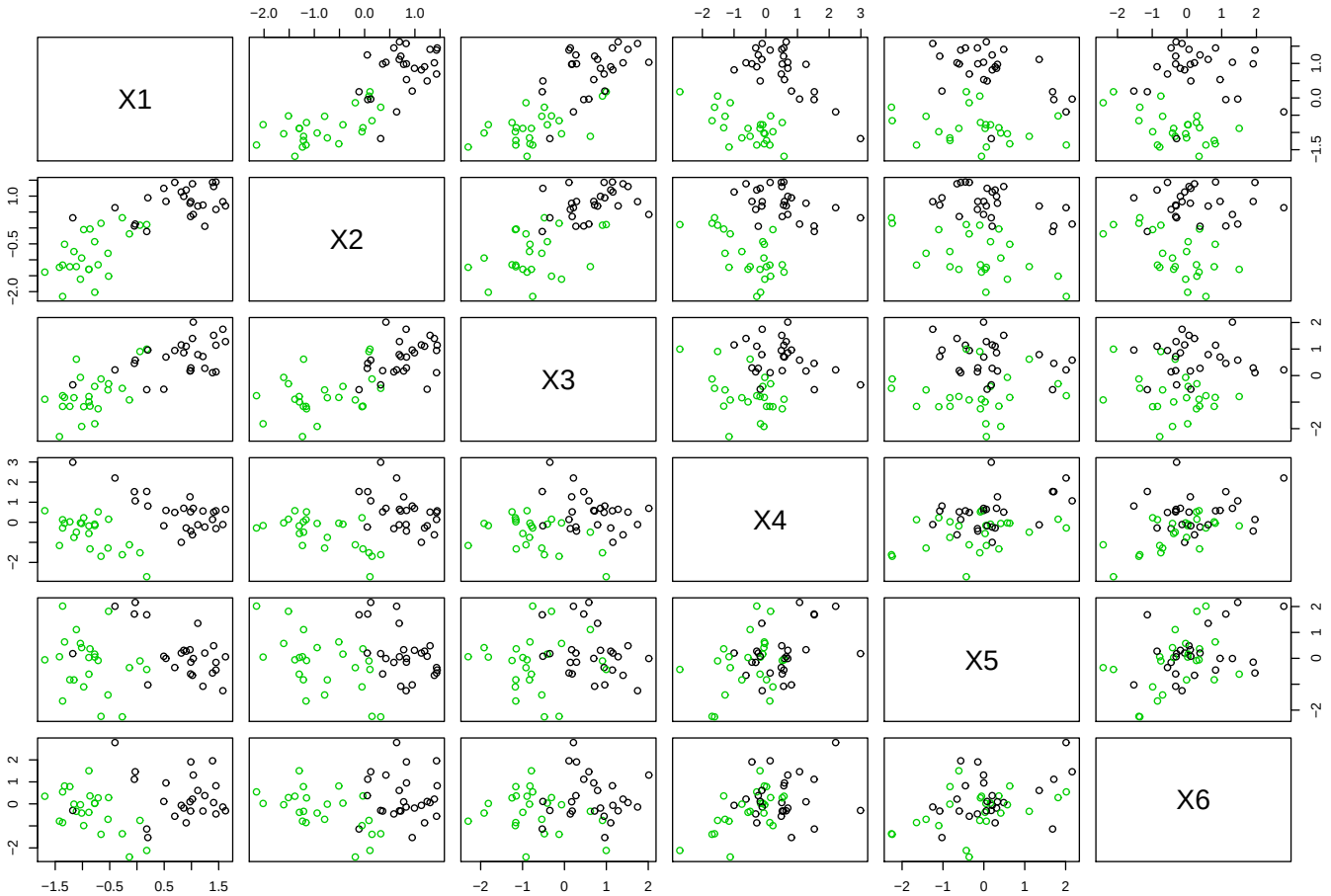
$$x_i \sim yN(i - 3, 1), i = 4, 5, 6$$

For the other variables:

$$x_i \sim N(0, 20), i = 7 \ldots, p$$

These data points are linearly separable with high probability, decreasing with the sample size.

Toys data [12]



Hierarchy, varying n

Rank where 4, 5 and 6 important variables appeared in the hierarchy. We have used $p = 200, B = 200,\ n = 50, 100, 200$.

n/Score	FDS	∂W	∂RW	∂Spb	RF	$GLMpath$
50	4	4	4	4	4	4
	6	5	5	5	6	5
	13	17	16	12	12	8
100	4	4	4	4	4	4
	5	5	5	5	5	5
	6	7	6	6	6	6
200	4	4	4	4	4	4
	5	5	5	5	5	5
	6	6	6	6	9	6

Hierarchy, varying p

$n = 50, B = 200, p = 500, 1000.$

p/Score	FDS	∂W	∂RW	∂Spb	RF	$GLMpath$
500	4	4	4	4	5	4
	5	7	7	5	12	5
	18	13	12	11	42	6
1000	4	4	4	4	4	4
	34	33	32	31	205	35
	173	194	202	224	206	38

Rank Correlations

200 observations, 200 variables.

	∂W	∂RW	∂Spb	RF	$GLMpath$
FDS	0.467	0.390	-0.216	0.180	0.542
∂W	1	0.685	-0.410	0.132	0.944
∂RW		1	-0.267	0.205	0.682
∂Spb			1	0.056	-0.484
RF				1	0.161

50 observations, 1000 variables.

	∂W	∂RW	∂Spb	RF	$GLMpath$
FDS	0.918	0.873	0.604	0.093	0.705
∂W	1	0.925	0.664	0.074	0.725
∂RW		1	0.622	0.073	0.702
∂Spb			1	0.083	0.567
RF				1	0.086

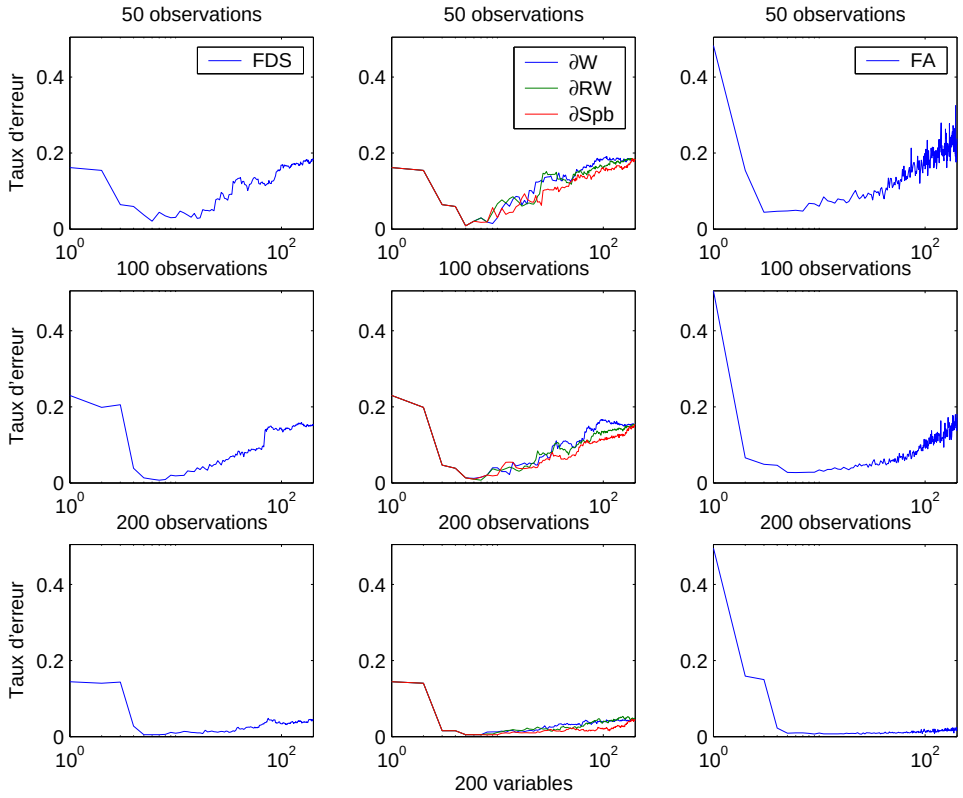
Performances

$Score/(n,p)$	(50,200)	(100,200)	(200,200)	(50,500)	(50,1000)
FDS	0.0208(6)	0.0072(7)	0.0048(7)	0.0044(5)	0.0084(5)
∂W	0.0084(5)	0.012(6)	0.0048(7)	0.008(7)	0.0084(5)
∂RW	0.0084(5)	0.0072(7)	0.0048(7)	0.008(7)	0.0076(6)
∂Spb	0.0084(5)	0.0096(6)	0.0044(8)	0.0044(5)	0.0084(5)
$SVM-RFE$	0.0476(8)	0.016(8)	0.006(4)	0.0132(8)	0.0104(4)
$GLMpath$	0.0188(1)	0.0252(3)	0.0074(4)	0.008(4)	0.0192(2)
RF	0.044(3)	0.0272(6)	0.0064(25)	0.0252(12)	0.0656(4)

Table: 50 stratified test, or CV (glmpath).

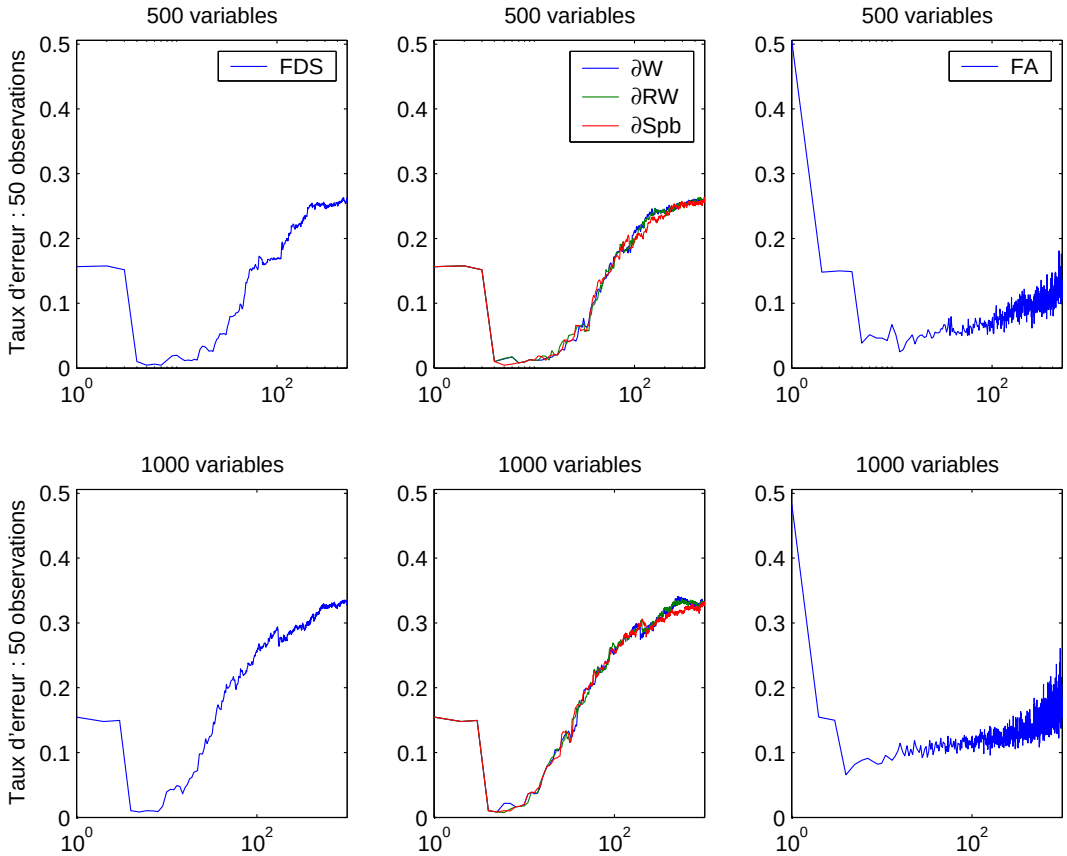
Toys data [12]

sample size effects, 50 stratified test samples, $p = 200$.



Toys data [12]

Number of variables effect. $p = 500, 1000$, $n = 50$.

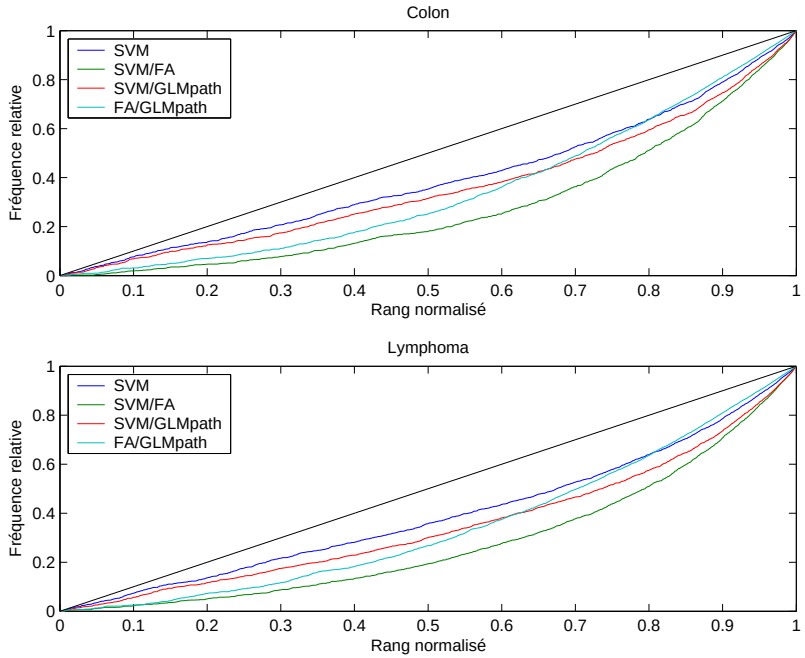


Data sets

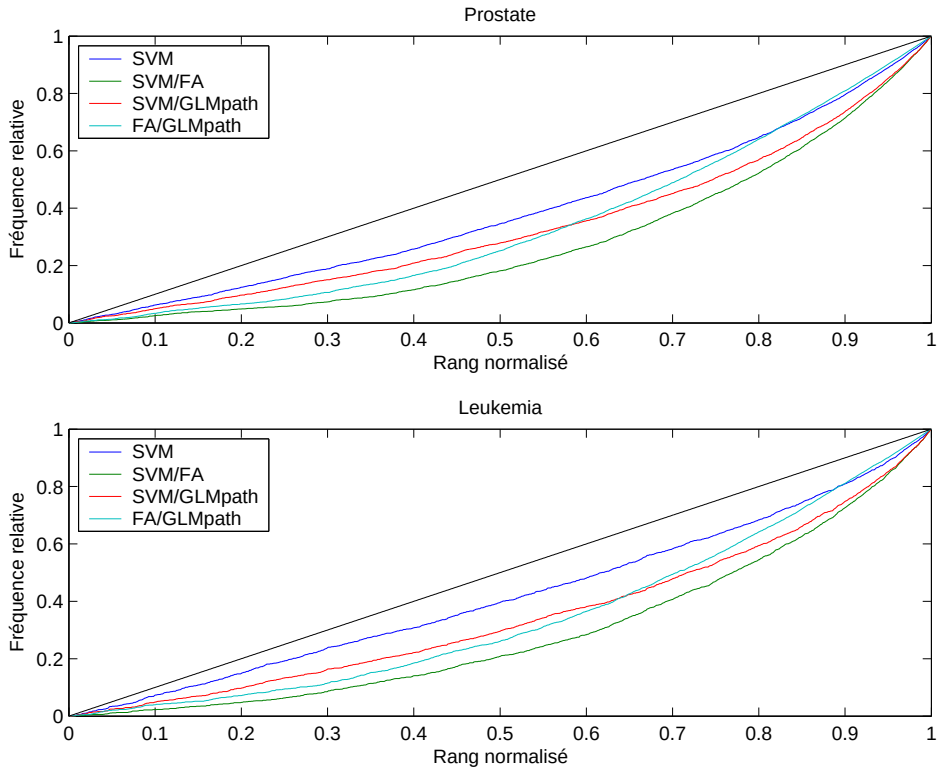
Data	*p*	*learning*	*test*	n *+1/-1*
Colon	2000	62	–	22/40
Lymphoma	4026	96	–	62/34
Prostate	12600	102	–	52/50
Leukemia	7129	38	34	27/11 - 20/14

Hierarchies comparison

0-coefficients: Colon-999, Lymphoma-1376, Leukemia-1190, Prostate-2234.
x-axis: Normalized rank.
y-axis: Proportion of common variables for the compared methods.



Hierarchies comparison 2



Common Variables

Comparison / data	Colon	Lymphoma	Prostate	Leukemia
SVM	37	37	32	30
SVM/GLMpath	33	26	24	21
SVM/RF	4	9	12	9
RF/GLMpath	10	12	16	21

Table: Number of common variables within the top 50

Results, real data sets

Score/Data	Colon	Lympoma	Prostate	Leukemia
FDS	0.1219(3)	0.0436(200)	0.0371(315)	0.0882(7)
∂W	0.0009(31)	0(186)	0.0269(83)	0.1176(2)
∂RW	0.0029(33)	0(60)	0.0269(902)	0.0882(22)
∂Spb	0.0029(34)	0.0006(118)	0.0109(45)	0.1176(11)
$SVM - RFE$	0.0057(32)	0(64)	0(64)	0.0882(1)
GLMpath	0.064(2)	0(3)	0(3)	0(1)
RF	0.0962(55)	0.0588(73)	0.0554(7)	0.0588(103)

Colon: 0.17, Lymphoma: 0.06, Prostate: 0.075, Leukemia: 0.20588.

Bias Selection

D data set, B Number of bootstrap samples

Partition D with stratification, $D_1, ..., D_{10}$.

Set $D_{-j} = D - D_j$.

For $j = 1, \ldots, 10$

Score(D_{-j},B) and use the hierarchy $X^{(1)}, ..., X^{(p)}$

For $k = 1, \ldots, p$

$M^k = f\left(X^{(1)}, ..., X^{(k)}\right)$

$Er^k = Test_{RS}(M^k, D_{-j})$

$kopt_j = Argmin_k\{Er^k\}$

$er_j =$ Mean error of M^{kopt_j} over D_j.

Compute $\bar{er} = \frac{1}{10}\sum_{j=1}^{10} er_j$.

Results

Data	Colon	Lymphoma	Prostate
FDS	0.1595(15.1)	0.1233(83.7)	0.0882(126.4)
∂W	0.233 (35.1)	0.051 (86.5)	0.054 (756.6)
∂RW	0.214 (43.3)	0.042 (71)	0.053 (573.3)
∂Spb	0.197 (31.8)	0.073 (70.5)	0.052 (95.5)
$SVM - RFE$	0.1452(26.4)	0.0878(16.8)	0.0582(43.2)
GLMpath	0.1809 (1.3)	0.0522 (2.8)	0.05909 (1.6)
RF	0.106 (49.8)	0.052 (65.9)	0.059 (81)

Bibliography

- L. Breiman. Random forests. *Machine Learning Journal*, 45:5-32, 2001.
- O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3) : 131-159.
- *Feature Extraction, Foundations and Applications*. I. Guyon et al, Eds. Springer, 2006.
- I. Guyon and A. Elisseff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 : 1157-1182, 2003.
- I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3) : 389-422, 2002.
- P. Langley. Selection of relevant features in machine learning. *In AAAI Fall Symposium on Relevance*, pages 140-144, New Orleans, 1994.
- P. McCullagh and J. Nelder. *Generalized Linear Models*. CHAPMAN & HALL/CRC, Boca Raton, 1989.
- M. Y. Park and T. Hastie. L_1 Regularization Path Algorithm for Generalized Linear Models. *Technical report*, Stanford University, February 2006.
- A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3 : 1357-1370, 2003.
- V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12 : 9, 2000.
- J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3 : 1439-1461, 2003.