Random Forests: some methodological insights

Robin Genuer[1] **Jean-Michel Poggi**[1,2] Christine Tuleau[3]

[1] Université Paris-Sud, Mathématique, Bât. 425, 91405 Orsay, France

[2] Université Paris Descartes, France

[3] Université Nice Sophia-Antipolis, France

1. Introduction
   - CART
   - Bagging
   - Random Forests
   - Data Sets
2. Method parameters
   - Regression
   - Classification
3. Variable Importance
   - Definition
   - Behavior
4. Variable Selection
   - Two objectives
   - Procedure
   - Examples

Random Forests

- introduced by L. Breiman in 2001
- ensemble methods, Dietterich (1999) and (2000)
- popular and very efficient algorithm of statistical learning, based on model aggregation ideas, for both classification and regression problems.

We consider a learning set $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ made of n i.i.d. observations of a random vector (X, Y).

Vector $X = (X^1, ..., X^p)$ contains explanatory variables, say $X \in \mathbb{R}^p$, and $Y \in \mathcal{Y}$ where $\mathcal{Y}$ is either a class label or a numerical response.

For classification problems, a classifier t is a mapping $t : \mathbb{R}^p \to \mathcal{Y}$ while for regression problems, we suppose that $Y = s(X) + \varepsilon$ and s is the so-called regression function.

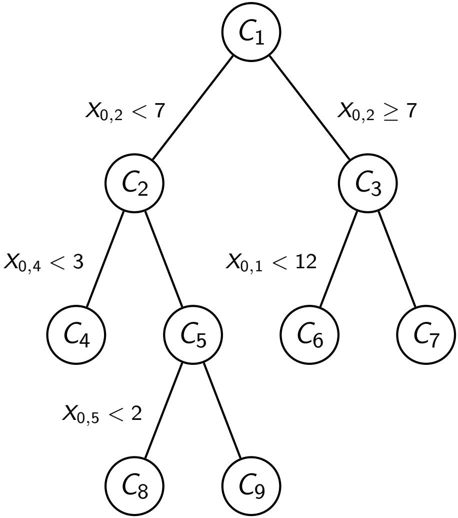
CART

CART (Classification And Regression Trees, (Breiman 1984)) can be viewed as the base rule of a random forest.
Recall that CART design has two main stages:

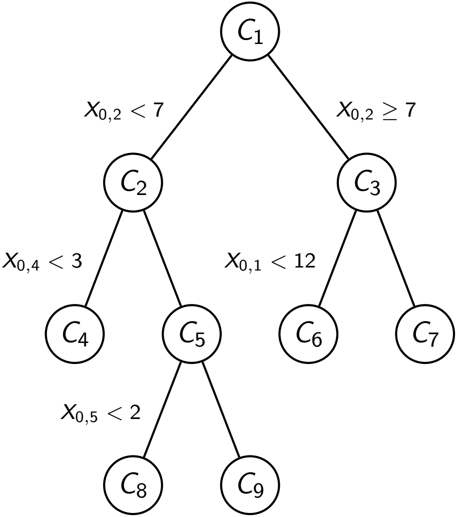
- maximal tree construction to build the family of models
- pruning for model selection

With CART, we get a classifier or an estimate of the regression function, which is a piecewise constant function obtained by partitioning the predictor's space

CART



CART



$$
\begin{aligned}
h(x) \quad = \quad & C_4 \mathbb{1}_{x_2<7,x_4<3} \\
& + C_8 \mathbb{1}_{x_2<7,x_4\geq 3,x_5<2} \\
& + C_9 \mathbb{1}_{x_2<7,x_4\geq 3,x_5\geq 2} \\
& + C_6 \mathbb{1}_{x_2\geq 7,x_1<12} \\
& + C_7 \mathbb{1}_{x_2\geq 7,x_1\geq 12}
\end{aligned}
$$

Growing step, stopping rule:

- do not split a pure node
- do not split a node containing less than `nodesize` data

Pruning step:

- the maximal tree overfits the data
- an optimal tree is pruned subtree of the maximal tree which realizes a good trade-off between the variance and the bias of the associated model
  Penalized criterion:

$$crit_{\alpha}(T) = R_n(f, \hat{f}_{|T}, \mathcal{L}_n) + \alpha \frac{|\tilde{T}|}{n}$$

where $R_n(f, \hat{f}_{|T}, \mathcal{L}_n) = \frac{1}{n} \sum_{(X_i, Y_i) \in \mathcal{L}_n} (Y_i - \hat{f}_{|T}(X_i))^2$ and where

$|\tilde{T}|$ is the number of leaves of the tree T

Denote by

- μ the marginal distribution of X
- $\|.\|$ the $\mathbb{L}^2(\mathbb{R}^p, \mu)$-norm
- $\tilde{f}$ is the final estimator given by CART

A typical result (Gey, Nedelec (2005))

There exist C_1, C_2, C_3 nonnegative constants such that:

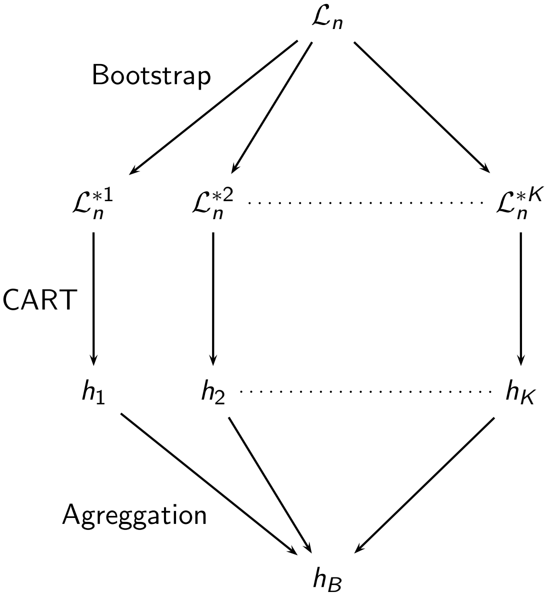
$$\mathbb{E}\left[\|\tilde{f} - f\| |\mathcal{L}_1\right] \leq C_1 \inf_{T \preceq T_{max}} \left[\inf_{u \in S_T} \|u - f\| + \sigma^2 \frac{|\tilde{T}|}{n_1}\right] + \frac{C_2}{n_1} + C_3 \frac{\ln n_1}{n_2}$$

where S_T is the set of piecewise constant functions defined on the partition $\tilde{T}$

Bagging (Breiman 1996)

Bootstrap **agg**regat**ing**

$$h_B(x) = \frac{1}{K}\sum_{k=1}^{K} h_k(x)$$



$\mathcal{L}_n$
Bootstrap
$\mathcal{L}_n^{*1}$
$\mathcal{L}_n^{*2}$
$\mathcal{L}_n^{*K}$
CART
h_1
h_2
h_K
Agreggation
h_B

CART-RF

We define CART-RF as the variant of CART consisting to select at random, at each node, `mtry` variables, and split using only the selected variables. The maximal tree obtained is not pruned.

`mtry` is the same for all nodes of all trees in the forest.

Random forest (Breiman 2001)

To obtain a random forest we proceed as in bagging. The difference is that we now use the CART-RF procedure on each bootstrap sample.

OOB = Out Of Bag.

OOB error

Consider a forest. For one data (X_i, Y_i), we only keep the classifiers h_k built on a bootstrap sample which does not contain (X_i, Y_i), and we aggregate these classifiers. We then compare the predicted label we get to the real one Y_i.

After doing that for each data (X_i, Y_i) of the learning set, the OOB error is the proportion of misclassified data .

To avoid unsignificant sampling effects, each OOB error is actually the mean of OOB errors over 10 runs.

R package:

- seminal contribution of Breiman and Cutler (early update in 2005)
- described in Liaw, Wiener (2002)

Focus on the `randomForest` procedure whose main parameters are:

- `ntree`, the number of trees in the forest;
- `mtry`, the number of variables randomly selected at each node.

Data Sets

Name	Observations	Variables	Classes
Ionosphere	351	34	2
Diabetes	768	8	2
Sonar	208	60	2
Votes	435	16	2
Ringnorm	200	20	2
Threenorm	200	20	2
Twonorm	200	20	2
Glass	214	9	6
Letters	20000	16	26
Sat-images	6435	36	6
Vehicle	846	18	4
Vowel	990	10	11
Waveform	200	21	3

Table: Standard ($n \gg p$) classification problems - data sets

Name	Observations	Variables	Classes
BostonHousing	506	13	
Ozone	366	12	
Servo	167	4	
Friedman1	300	10	
Friedman2	300	4	
Friedman3	300	4	

Table: Standard ($n \gg p$) regression problems - data sets

Name	Observations	Variables	Classes
Adenocarcinoma	76	9868	2
Colon	62	2000	2
Leukemia	38	3051	2
Prostate	102	6033	2
Brain	42	5597	5
Breast	96	4869	3
Lymphoma	62	4026	3
Nci	61	6033	8
Srbct	63	2308	4
toys data	100	100 to 1000	2
PAC	209	467	
Friedman1	100	100 to 1000	
Friedman2	100	100 to 1000	
Friedman3	100	100 to 1000	

Table: High dimensional ($n \ll p$) problems - data sets for classification at the top, and for regression at the bottom

Regression

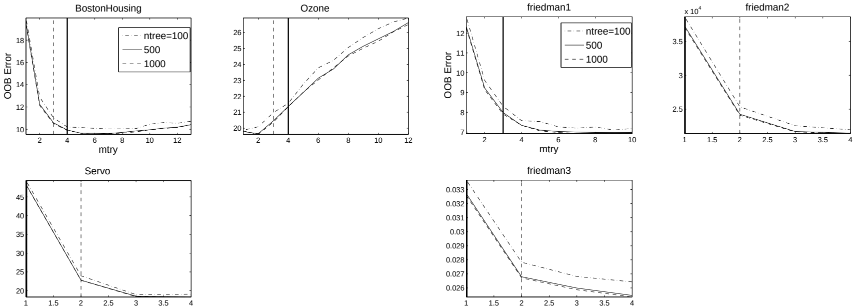


Figure: Standard regression real (left) and simulated (right) data sets

Vertical solid line $mtry = p/3$ (default value), dashed line $mtry = \sqrt{p}$

- the OOB error maximal for $mtry = 1$ then decreases quickly (except for the ozone data) then as soon as $mtry > \sqrt{p}$, the error is stable

- the choice $mtry = \sqrt{p}$ gives often lower OOB error than $mtry = p/3$, and the gain can be important. So the default value seems to be often suboptimal, especially when $\lfloor p/3 \rfloor = 1$

- $ntree = 500$ is convenient, but $ntree = 100$ leads to comparable results

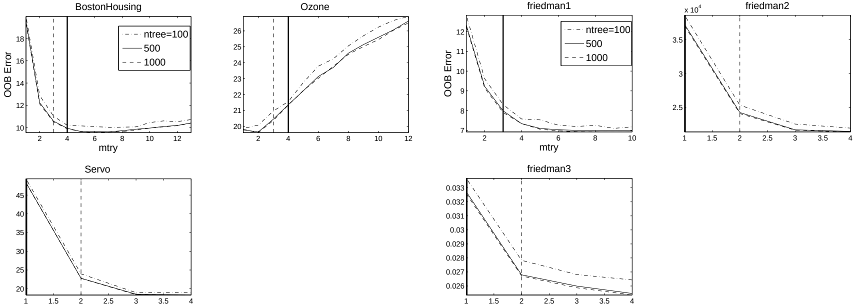


Figure: Standard regression real (left) and simulated (right) data sets

So, for standard ($n >> p$) regression problems, it seems that there is no improvement by using random forests with respect to unpruned bagging (obtained for $mtry = p$)

A high dimensional regression simulated data set

- Example built by adding extra noisy variables to the Friedman1 model defined by:

  $$Y = 10\sin(\pi X^1 X^2) + 20(X^3 - 0.5)^2 + 10X^4 + 5X^5 + \epsilon$$

  where $X^1, \dots, X^5$ are independent and uniformly distributed on $[0, 1]$ and $\epsilon \sim \mathcal{N}(0, 1)$
- So we have 5 variables related to the response Y, the others being noise (independent and uniformly distributed on $[0, 1]$)
- We set $n = 100$ and let p vary

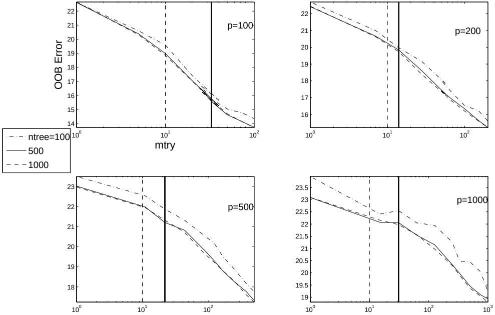


Figure: High dimensional regression simulated data set Friedman1. x-axis is in log scale, vertical solid line $mtry = p/3$, dashed line $mtry = \sqrt{p}$

- OOB error decreases while *mtry* increases
- while p increases, both OOB errors of unpruned bagging ($mtry = p$) and random forests with default value of *mtry* increase
- unpruned bagging performs better than RF (gain $\simeq 25\%$)
- $mtry = \sqrt{p}$ gives worse results than $mtry = p/3$
- $ntree = 500$ is convenient, but $ntree = 100$ is sufficient

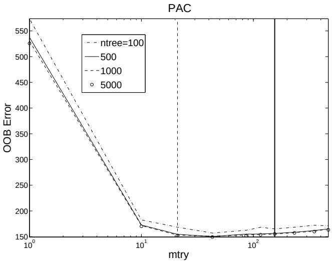


Figure: High dimensional regression PAC data. The x-axis is in log scale
- vertical solid line $mtry = p/3$, vertical dashed line $mtry = \sqrt{p}$
- General behavior is quite similar except for the shape: as soon as $mtry > \sqrt{p}$, the error remains the same instead of still decreasing
- *In considered simulated datasets*: $\widetilde{p}$ the number of true variables is very small compared to p. Often *in real datasets*, the proportion $\dfrac{\widetilde{p}}{p}$ of true variables is larger
- For high dimensional ($n << p$) regression problems, unpruned bagging seems to perform better than random forests

Classification

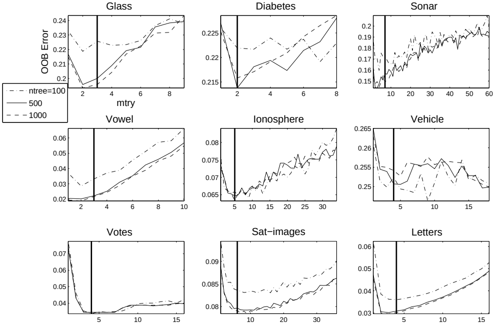


Figure: Standard classification real data sets - vertical solid line $mtry = \sqrt{p}$

- the default value $mtry = \sqrt{p}$ is convenient

- the default value $ntree = 500$ is sufficient while a much smaller one $ntree = 100$ is not

- the errors for $mtry = 1$ and for $mtry = p$ (corresponding to the unpruned bagging) are of the same "large" order of magnitude

- the minimum is reached for the value $\sqrt{p}$. The gain $\simeq$ 30 or 50%

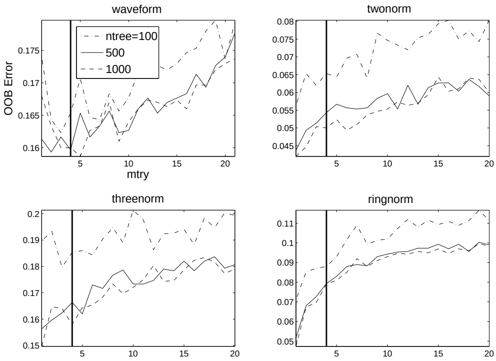


Figure: Standard classification: 4 simulated data sets - vertical solid line $mtry = \sqrt{p}$

- $ntree = 500$ is sufficient and, except for the ringnorm already pointed out as a somewhat special dataset (see Cutler, Zhao (2001)) the value $mtry = \sqrt{p}$ is a good choice

- the general shape of the error curve is quite different compared to real datasets: the error increases with *mtry*. So for these four examples, *the smaller mtry, the better*

Classification

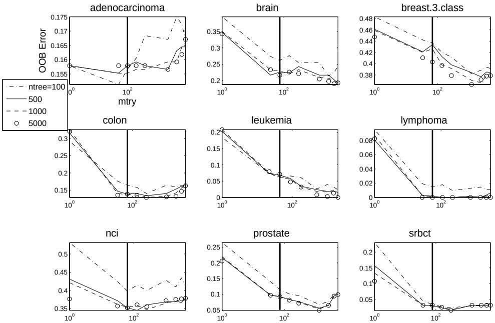


Figure: High dimensional classification: 9 real data sets.

- the default value $ntree = 500$ is sufficient

- general shape: it decreases in general and the minimum value is obtained or is close to the one reached using $mtry = p$ (unpruned bagging). The difference with standard problems is notable, why? *When p is large, $mtry$ must be sufficiently large* to preserve a high probability to capture important variables (highly related to the response) for defining the splits of the RF

Classification

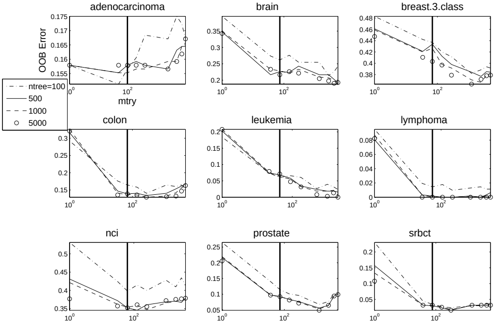


Figure: High dimensional classification: 9 real data sets.

The default value $mtry = \sqrt{p}$ is still reasonable from the OOB error viewpoint but of course, since $\sqrt{p}$ is small with respect to p, it is a very attractive value from a computational perspective (notice that the trees are not too deep since n is not too large)

"Toys data", Weston *et al.* (2003)

an interesting equiprobable two-class problem, $Y \in \{-1, 1\}$, with 6 true variables, the others being noise:

- two near independent groups of 3 significant variables (highly, moderately and weakly correlated with response Y)
- an additional group of noise variables, uncorrelated with Y

Model defined through the conditional distributions of the X^i for $Y = y$:

- for 70% of data, $X^i \sim y\mathcal{N}(i, 1)$ for $i = 1, 2, 3$ and $X^i \sim y\mathcal{N}(0, 1)$ for $i = 4, 5, 6$
- for the 30% left, $X^i \sim y\mathcal{N}(0, 1)$ for $i = 1, 2, 3$ and $X^i \sim y\mathcal{N}(i - 3, 1)$ for $i = 4, 5, 6$
- the other variables are noise, $X^i \sim \mathcal{N}(0, 1)$ for $i = 7, \dots, p$

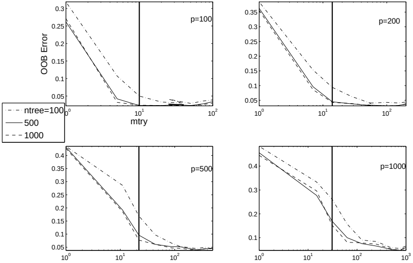


Figure: High dimensional classification simulated data set: toys data for 4 values of p. The x-axis is in log scale - vertical solid line $mtry = \sqrt{p}$
- for $p = 100$ and $p = 200$, the error decreases hugely until $mtry$ reaches $\sqrt{p}$ and then remains constant, so the default values work well and perform as well as unpruned bagging (even if the true dimension $\tilde{p} = 6 << p$)
- for larger values of p ($p \geq 500$), the shape of the curve is close to the one for high dimensional real data sets
- finally, for high dimensional classification problems, our conclusion is that it may be worthwhile to choose $mtry$ larger than $\sqrt{p}$

1 Introduction
- CART
- Bagging
- Random Forests
- Data Sets

2 Method parameters
- Regression
- Classification

3 Variable Importance
- Definition
- Behavior

4 Variable Selection
- Two objectives
- Procedure
- Examples

- The quantification of the variable importance (VI) is an important issue in many applied problems complementing variable selection by interpretation issues.
  For linear regression case, see various variance decomposition based indicators in Grömping (2006) and (2007)
- In the RF framework, permutation importance indices are preferred to total decrease of node impurity measures already introduced in Breiman *et al.* (1984)
- Little investigation is available about RF variable importance. Some interesting remarks in Strobl *et al.* (2007) and (2008), Ishwaran (2007), Archer *et al.* (2008) but they do not answer crucial questions like:
  - the importance of a group of variables
  - the behavior of VI in presence of highly correlated variables

Variable importance

Let $j \in \{1, \ldots, p\}$. For each classifier h_k we consider the corresponding OOB sample and permute at random the j-th variable values of these data. Then we compute the OOB error of h_k with these modified OOB data.
The variable importance of the j-th variable is defined as the increase of OOB error after permutation.

The more the increase of OOB error is, the more important is the variable

Subsequent VI boxplots are based on 50 runs.

"Toys data", Weston *et al.* (2003)

Interesting equiprobable two-class problem, $Y \in \{-1, 1\}$, with 6 true variables, the others being noise:

- two near independent groups of 3 significant variables (highly, moderately and weakly correlated with response Y)
- forward reference to the VI plots (left side of next figure) allows to note that the importance of the variables 1 to 3 is much higher than the one of variables 4 to 6
- an additional group of noise variables, uncorrelated with Y

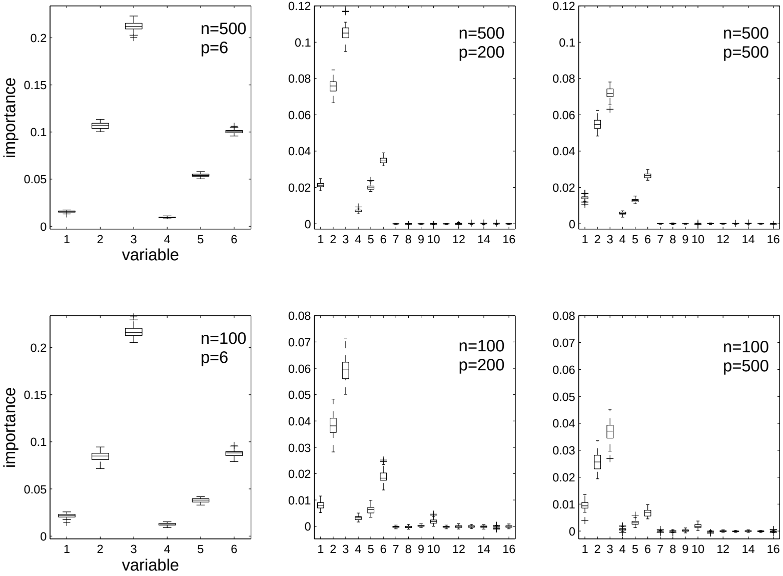


Figure: Variable importance sensitivity to *n* and *p*

Behavior

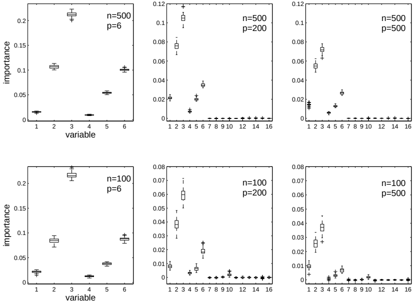


Figure: Variable importance sensitivity to n and p: first row ($n = 500$)

- when $p = 6$ concentrated boxplots and the order is clear, variables 2 and 6 having nearly the same importance
- when p increases, the order of magnitude of importance decreases
- in addition, VI is more unstable for huge values of p
- What is remarkable is that all noisy variables have a zero VI. So one can easily recover variables of interest
- the variability of VI is large for true variables with respect to useless ones

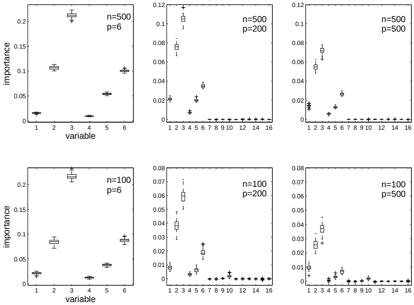


Figure: Variable importance sensitivity to n and p: second row($n = 100$)

- greater instability but variable ranking remains quite the same

- in the difficult situations ($p = 200, 500$) importance of some noisy variables increases

- decreasing behavior of VI with p growing, coming from the fact that when $p = 500$ the algorithm randomly choose only 22 variables at each split (with the *mtry* default value) and the probability of choosing one of the 6 true variables is really small

- variability of VI is large for true variables with respect to useless ones

Behavior

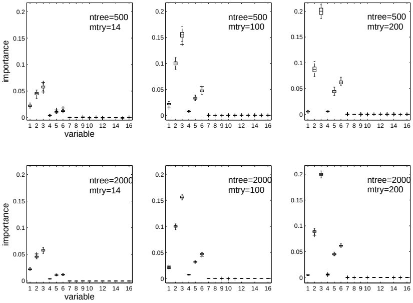


Figure: Sensitivity to `mtry` and `ntree`

- effect of *ntree*: VI is more than doubled starting from $mtry = 14$ to $mtry = 100$, and it again increases with $mtry = 200$

- effect of *ntree*: less visible, but $ntree = 2000$ leads to better stability

- *same order for all true variables in every run of the procedure*

- top left: the mean OOB error rate is about 5% and in the bottom right one it is 3%. *The gain in error may not be considered as large, but what we get in VI is interesting*

Behavior

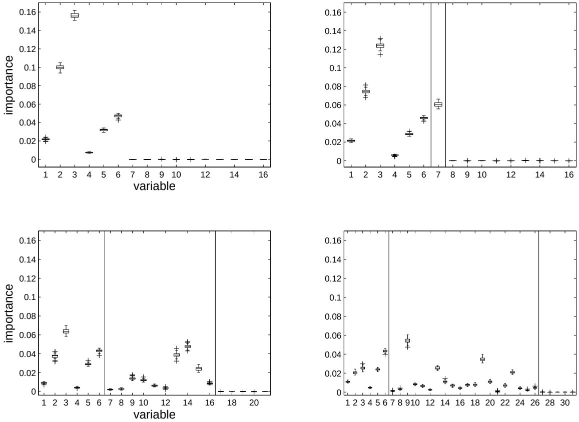


Figure: *Variable importance of a group of correlated variables*

Basic model: previous context with $n = 100$, $p = 200$, $ntree = 2000$ and $mtry = 100$

Replications (plotted between the two vertical lines): we simulate 1, 10 and 20 (resp.) variables with a correlation of 0.9 with variable 3 (the most important one)

Behavior

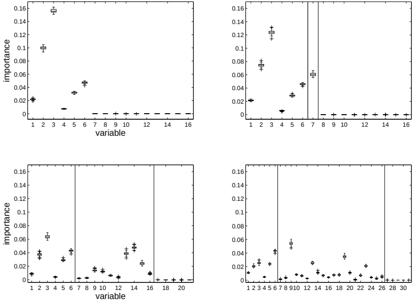


Figure: *Variable importance of a group of correlated variables*

- the magnitude of importance of the group $1, 2, 3$ is steadily decreasing when adding more replications of variable 3. On the other hand, the importance of the group $4, 5, 6$ is unchanged

- VI is not divided by the number of replications. Even with 20 replications the maximum importance of the group containing variable 3 is only three times lower than the initial importance of variable 3

- even if some variables in this group have small importance, they cannot be confused with noise

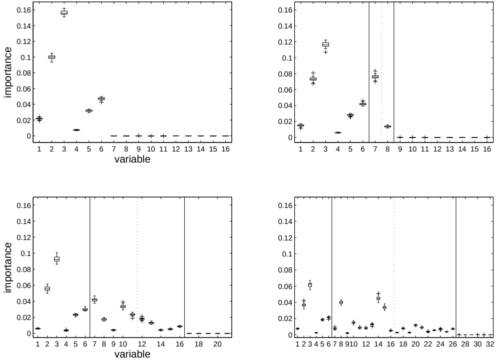


Figure: VI of two groups of correlated variables - replications are plotted between the two vertical lines, V3 replicates then V6 ones
- the magnitude of importance of each group (1, 2, 3 and 4, 5, 6 respectively) is steadily decreasing when adding more replications
- the relative importance between the two groups is preserved. And the relative importance between the two groups of replications is of the same order than the one between the two initial groups

Behavior

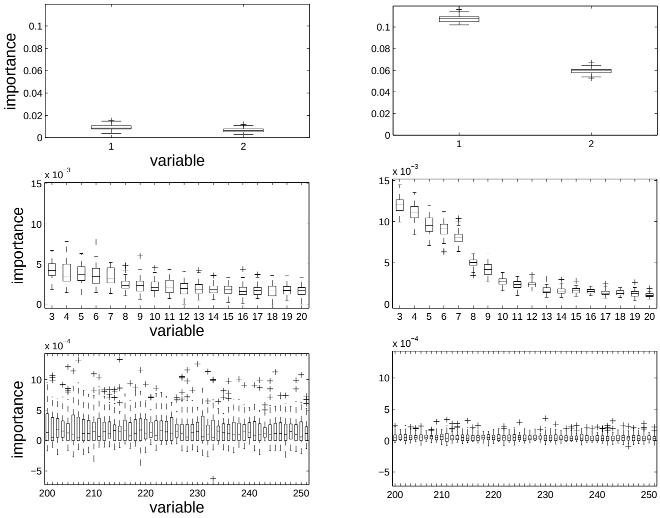


Figure: Variable importance for Prostate data (using $ntree = 2000$ and $mtry = p/3$, on the right and using default values on the left)

1 Introduction
- CART
- Bagging
- Random Forests
- Data Sets

2 Method parameters
- Regression
- Classification

3 Variable Importance
- Definition
- Behavior

4 Variable Selection
- Two objectives
- Procedure
- Examples

Two objectives

- Variable selection usually based on the cooperation of variable importance for ranking and model estimation to evaluate and compare a family of models
- Three types of methods (Kohavi *et al.* (1997)):
  - "filter": VI score does not depend on model design method;
  - "wrapper": prediction performance included in VI score
  - "embedded" relates closely variable selection and model estimation
- *Nonparametric methods for classification*:
  - CART Breiman *et al.* (1984), RF Breiman (2001)
  - Embedded methods: Poggi, Tuleau (2006), SVM-RFE Guyon *et al.* (2002), Rakotomamonjy (2003), Ben Ishak, Ghattas (2008) for a stepwise variant, Park *et al.* (2007) "LARS" type strategy
- Mention mixed strategy for the case $n << p$: descending first to reach a classical situation $n \sim p$, and then ascending, see Fan, Lv (2008)

Two objectives

We distinguish two different objectives:

1. to magnify all the important variables, even with high redundancy, for interpretation purpose
2. to find a sufficient parsimonious set of important variables for prediction

Two earlier works must be cited:

- Díaz-Uriarte, Alvarez de Andrés (2006)
- Ben Ishak, Ghattas (2008)

- *Díaz-Uriarte, Alvarez de Andrés (2006): a strategy based on recursive elimination of variables*
  - compute RF VI
  - at each step, eliminate the 20% of the variables having the smallest importance and build a new forest
  - finally select the set of variables leading to the smallest OOB error rate
  - *Drawback: the proportion of variables to eliminate is arbitrary and does not depend on the data*
- *Ben Ishak, Ghattas (2008): ascendant strategy based on a sequential introduction of variables*
  - compute some SVM-based VI
  - build a sequence of SVM models invoking at the beginning the k most important variables, by step of 1 (for too large k, additional variables are invoked by packet
  - select the set of variables leading to the model of smallest error rate

Procedure

1. Preliminary elimination and ranking:
   - Compute the RF scores of importance, cancel the variables of small importance
   - Order the m remaining variables in decreasing order of importance
2. Variable selection:
   - For *interpretation*:
     - Construct the nested collection of RF models involving the k first variables, for $k = 1$ to m
     - Select the variables involved in the model leading to the smallest OOB error
   - For *prediction* (conservative version):
     - Starting from the ordered variables retained for interpretation, construct an ascending sequence of RF models, by invoking and testing the variables stepwise
     - The variables of the last model are selected

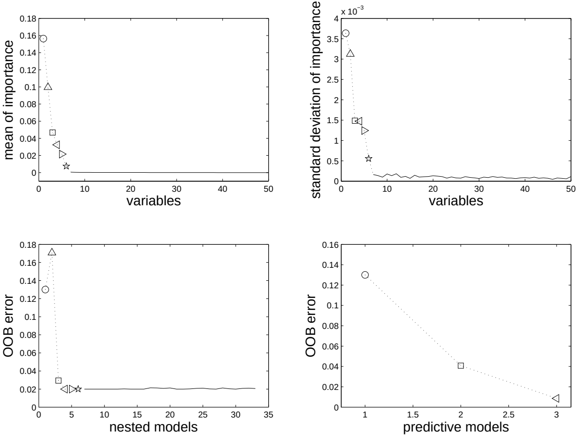


Figure: Variable selection procedure for interpretation and prediction: toys data $n = 100$, $p = 200$

- True variables (1 to 6) represented by $(\triangleright, \triangle, \circ, \star, \triangleleft, \square)$

- VI based on 50 forests with $ntree = 2000$, $mtry = 100$

Procedure

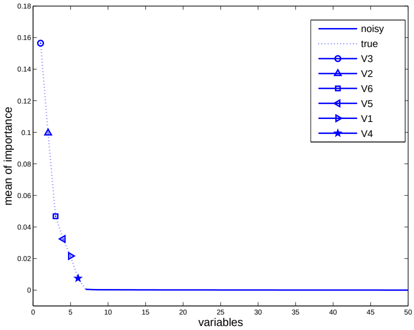


Figure: Variable selection procedure: Ranking

Ranking by sorting the VI in descending order

- Graph for the 50 most important variables (the other noisy variables having an importance very close to zero too)
- True variables are significantly more important than the noisy ones

Procedure

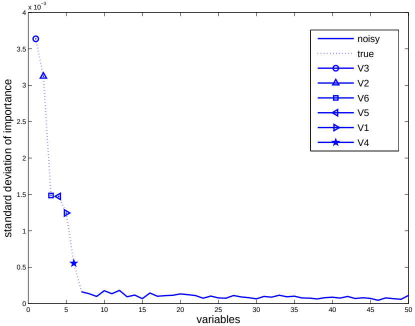


Figure: Variable selection procedure: Elimination

Consider corresponding standard deviations of VI to estimate a threshold and keep variables of importance exceeding this level

- Threshold = argmin of the prediction value given by a CART model fitting this curve (conservative in general)

- True variables standard deviation large w.r.t. the noisy variables one, which is close to zero

- The selected threshold leads to retain 33 variables

Procedure

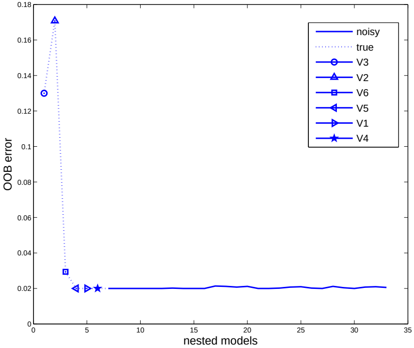


Figure: Variable selection procedure for interpretation

Compute OOB error rates of RF for the nested models and select the variables of the model leading to the smallest OOB error

- Error decreases quickly and reaches its minimum when the first 4 true variables are included in the model, then it remains *almost* constant

- The model containing 4 of the 6 true variables is selected. In fact, the actual minimum is reached for 24 variables but we use a rule similar to the 1 SE rule of Breiman *et al.* (1984) used for cost-complexity selection

Procedure

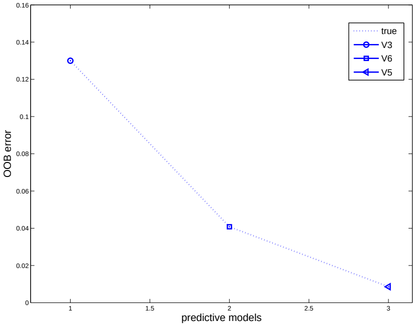


Figure: Variable selection procedure for prediction

Sequential variable introduction with testing

- A variable is added only if the error gain exceeds a threshold since the error decrease must be significantly greater than the average variation obtained by adding noisy variables

- Final prediction model involves only variables 3, 6 and 5

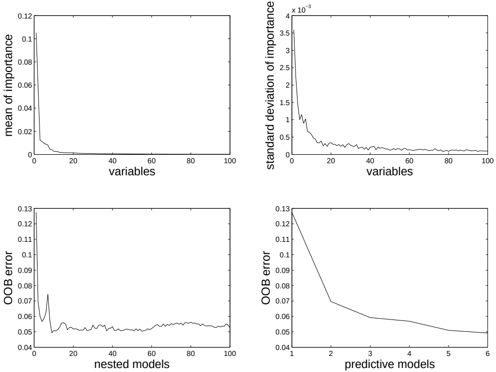


Figure: Variable selection procedures for Prostate data, *ntree* $= 2000$, *mtry* $= p/3$

- same picture as previously, except for the *OOB rate* along the nested models which is *less regular*

- Key point: it selects 9 variables for interpretation, and 6 variables for prediction (both very much smaller than $p = 6033$)

Dataset	*interpretation*	*prediction*	*original*
Colon	0.16 (35)	0.20 (8)	0.14
Leukemia	0 (1)	0 (1)	0.02
Lymphoma	0.08 (77)	0.09 (12)	0.10
Prostate	0.085 (33)	0.075 (8)	0.07

Table: Variable selection for four high dimensional real datasets.
CV-error rate calculated using the same partition in 5 parts and with $ntree = 2000$ and $mtry = p/3$. Into brackets, average number of selected variables

- *Number of interpretation variables* is hugely smaller than p: at most tens to be compared to thousands
- *Number of prediction variables* is very small (always smaller than 12) and the additional reduction can be very important
- Errors for the two variable selection procedures are of the same order of magnitude as the original error (but a little bit larger)
- *Error rates* are comparable with the results reported by Ben Ishak, Ghattas (2008) which have compared their method with 5 competitors

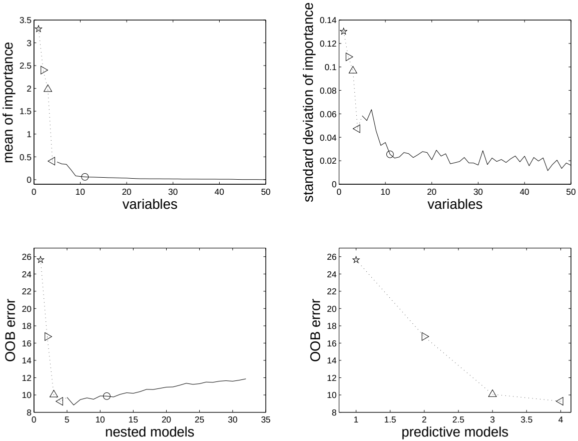


Figure: Variable selection procedures for Friedman1 data, $n = 100$ with $p = 200$ variables. True variables of the model (1 to 5) are respectively represented by $(\triangleright, \triangle, \circ, \star, \triangleleft)$

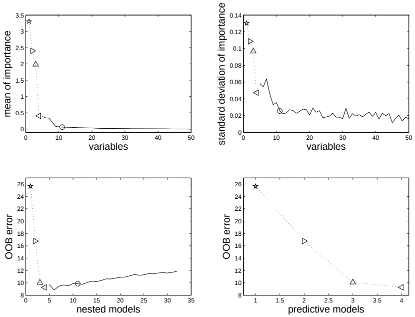


Figure: Friedman1 data

- Interpretation procedure selects the true variables except variable 3 and two noisy variables, and the prediction set of variables contains only the true variables (except variable 3 hardly correlated with the response variable)
- The whole procedure is stable across several runs
- In addition, the test mean squared error with all variables is about 19.2, with the 6 interpretation variables 12.6 and the one with the 4 prediction variables 9.8

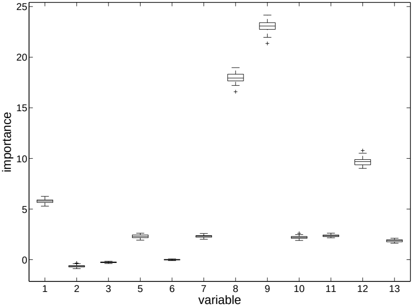


Figure: Variable importance for Ozone data using *mtry* $= p/3 = 4$ and *ntree* $= 2000$.

$n = 366$ observations of the daily maximum one-hour-average ozone together with $p = 12$ meteorologic explanatory variables

From the left to the right: 1-Month, 2-Day of month, 3-Day of week, 5-Pressure height, 6-Wind speed, 7-Humidity, 8-Temperature (Sandburg), 9-Temperature (El Monte), 10-Inversion base height, 11-Pressure gradient, 12-Inversion base temperature, 13-Visibility

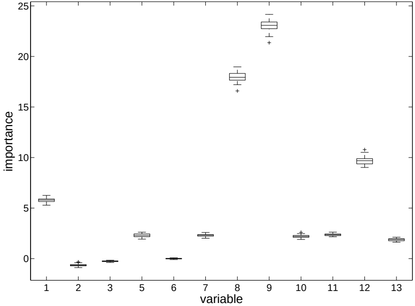


Figure: Three groups of variables for Ozone data.
1- Best ozone predictors: the two temperatures (8 and 9), the inversion base temperature (12) and the month (1)
2- pressure height (5), humidity (7), inversion base height (10), pressure gradient (11) and visibility(13)
3- day of month (2), day of week (3) of course and more surprisingly wind speed (6): wind enter in the model only when ozone pollution arises(see Cheze *et al.* (2003))

Short bibliography

- **Genuer R., Poggi J.-M. and Tuleau C.** *Random Forests: some methodological insights*. Rapport de Recherche INRIA, 2008, http://hal.inria.fr/inria-00340725/fr/
- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Chapman & Hall.* (1984)
- Breiman, L. *Machine Learning* (1996)
- Breiman, L. *Machine Learning* (2001)
- Biau G., Devroye L., and Lugosi G. *Journal of Machine Learning Research* (2007)
- Bühlmann, P. and Yu, B. *Annals of Statistics* (2002)
- Díaz-Uriarte R., Alvarez de Andrés S. *BMC Bioinformatics* (2006)