Advanced Machine Learning

Paper: Variable selection using random forests [Genuer et al., 2010]

Antonio Sutera
a.sutera@uliege.be

February 28, 2019

Variable selection using random forests

Supervised learning

Principle

X_1	X_2	$\dots$	X_p	Y
x_1^1	x_2^1	$\dots$	x_p^1	y^1
x_1^2	x_2^2	$\dots$	x_p^2	y^2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^n	x_2^n	$\dots$	x_p^n	y^n

$\xRightarrow{\text{Learning}} \hat{Y} = f(X_1, X_2, \dots, X_p)$

Goal: From a learning set of n samples, find a function f of the inputs $V = \{X_1, X_2, \dots, X_p\}$ that approximates at best the output Y.

Supervised learning

High-dimensionality

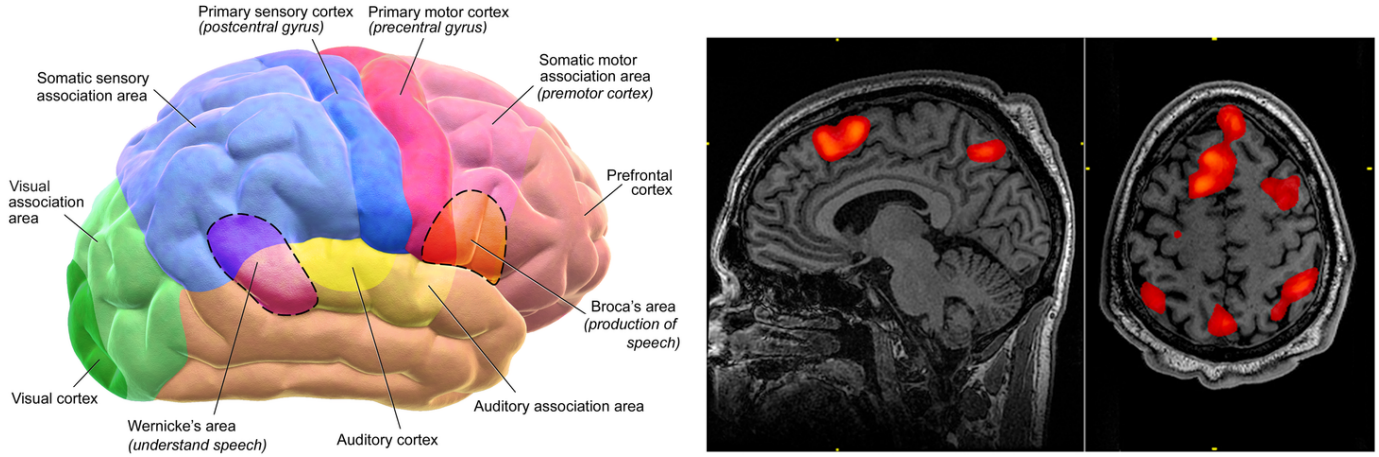


Figure: Brain regions [Blausen, 2014] and fMRI [Graner et al., 2013]

- High-dimensional classification problem : $\mathrm{n} \ll \mathrm{p}$
- A lot of useless variables
- There exists unknown groups of highly correlated variables corresponding to brain regions

Variable selection

Motivation

- Improving interpretability
  - Data understanding : features involved in the underlying mechanism
  - Data visualisation
- Increasing performances
  - Dimensionality reduction
  - Avoid overfitting
  - Reduce storage and computation requirements

Variable selection

Relevance

A variable $X_m \in V$ is **relevant** with respect to the output Y iff there exists a subset $B \subseteq V^{-m}$ such that $X_m \not\perp\!\!\!\perp Y|B$. A variable is **irrelevant** if it is not relevant.

Variable selection

Two objectives

- **Prediction**: The **minimal-optimal feature selection problem** consists in finding a subset of V of minimal size that minimises the generalisation error of a given learning algorithm.
- **Interpretation**: The **all-relevant feature selection problem** consists in finding all relevant features.

Variable selection

Approaches (see iML)

- **Filter** (e.g., statistical test): a priori selection of the variables (ie, independently of the supervised learning algorithm);
- **Embedded** (e.g., decision tree node splitting): feature selection embedded in the learning algorithm;
- **Wrapper**: use CV to find the optimal set of features for a given algorithm.

Variable selection

Recursive elimination of variables [Díaz-Uriarte and De Andres, 2006]

Given a learning model that can rank the features or provide variable importance (e.g., RF);

Iterate (from the full feature set)

- Build a model with the remaining features;
- Compute feature ranking (or variable importance);
- Remove the feature (or more, e.g. 20%) with the smallest ranking (or importance).

The set of features leading to the lowest (CV) error is then **selected**.

Variable selection

Sequential introduction of variables [Ghattas and Ben Ishak, 2008]

Step 1. Feature ranking (e.g., RF variable importance)

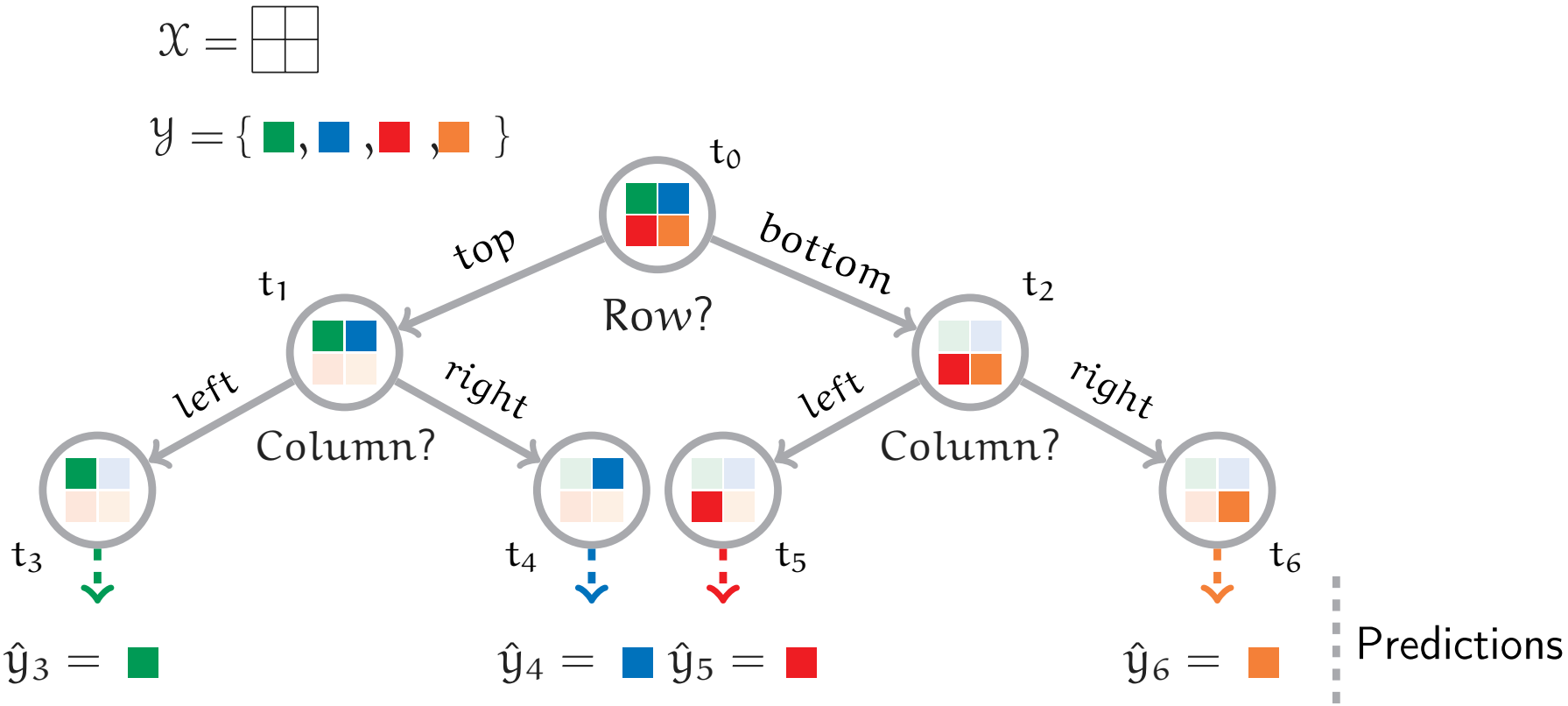
Step 2. Iterate (with k from 1 to p, by step of 1 or more)

- Build a model (e.g., RF) with the k most important feature.

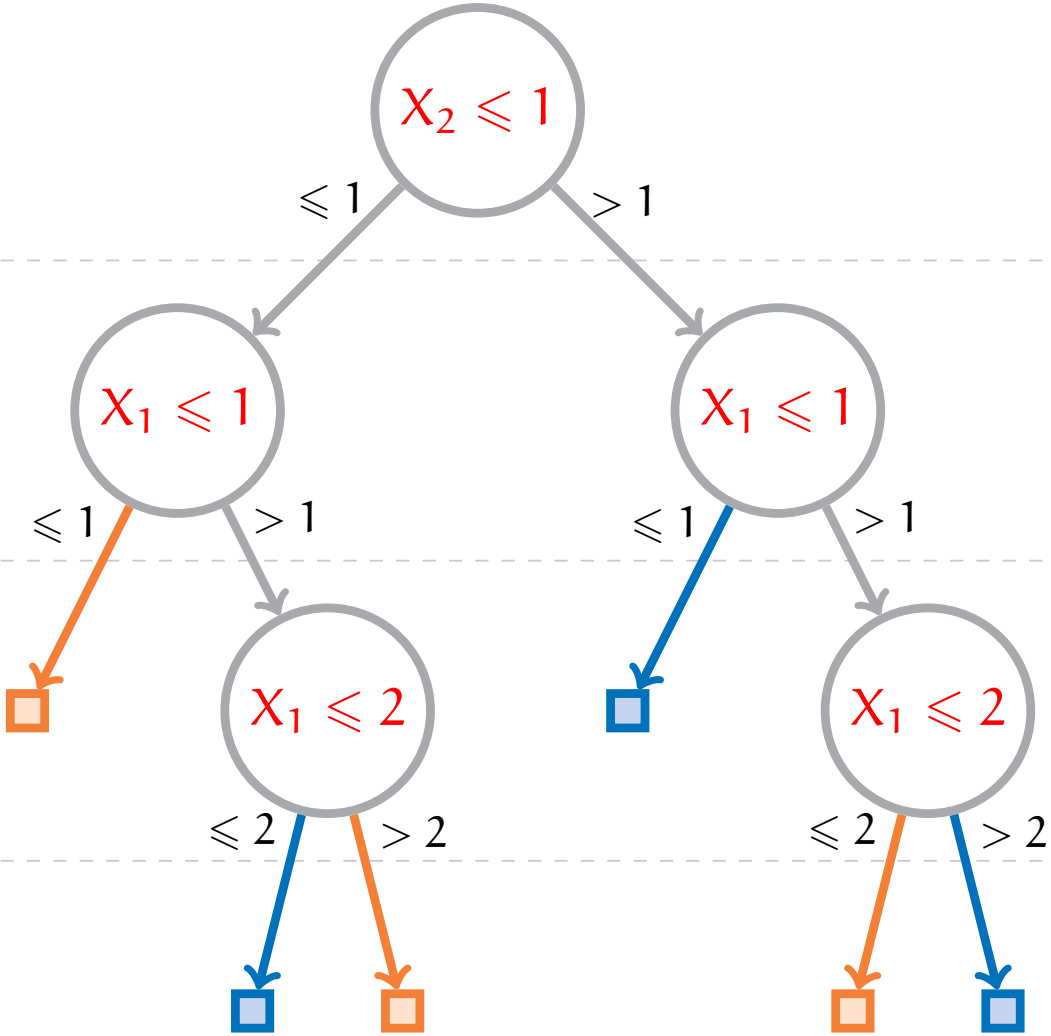
The set of features leading to the lowest (CV) error is then **selected**.

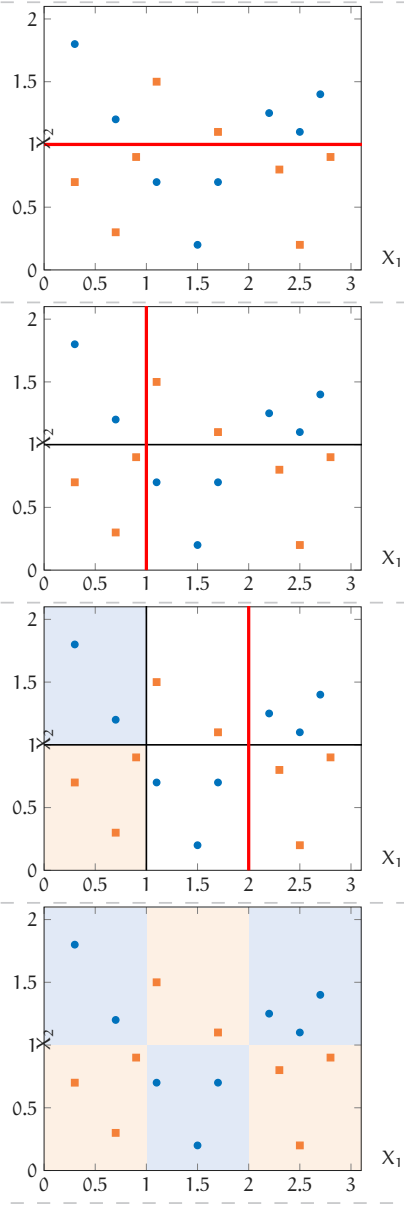
Variable selection using **random forests**

Decision tree



Decision tree





Decision tree

Maximizing the impurity reduction

The impurity reduction is

$$\Delta i(s, t) = i(t) - \frac{N_{t_L}}{N_t} i(t_L) - \frac{N_{t_R}}{N_t} i(t_R)$$

where

- s is a binary split dividing node t into t_L and t_R,
- N_t is the number of samples in node t,
- N_{t_L} and p_{t_L} (respectively, N_{t_R} and p_{t_R}) are the number of samples and the proportion of samples that fall into t_L (resp., t_R),
- $i(\cdot)$ is an **impurity measure**.

Decision tree

Several impurity measures

Three main properties

1. Minimal when the node is pure,
2. Maximal for uniform distribution,
3. Not biased towards some output values.

Suitable impurity measures for classification

With C classes:

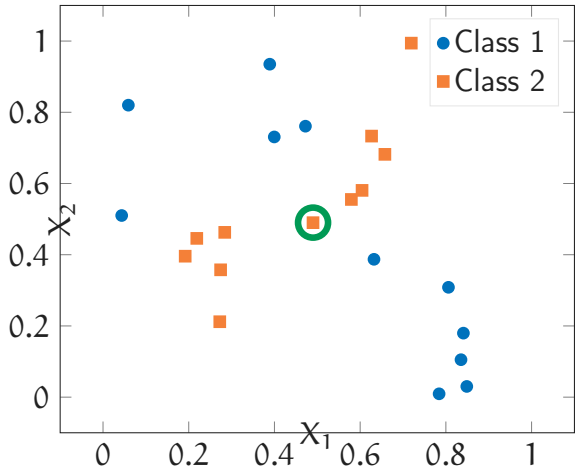
- Shannon entropy : $i_h(t) = -\sum_{j=1}^{C} p(c_j|t) \log_2 p(c_j|t)$
- Gini index: $i_g(t) = \sum_{j=1}^{C} p(c_j|t)(1 - p(c_j|t))$

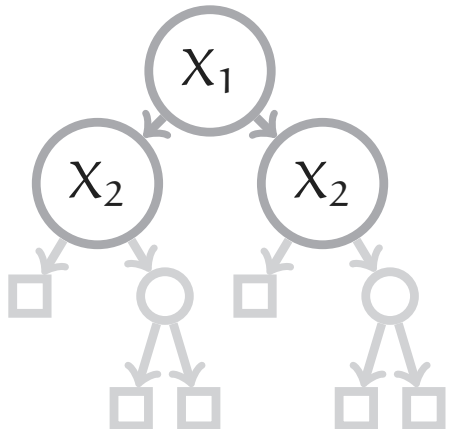
Extension for regression

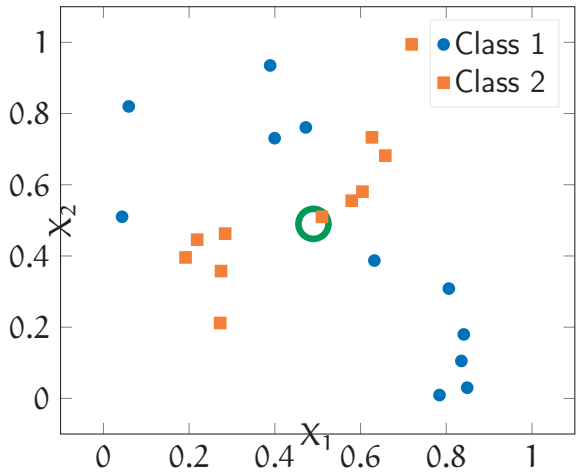
- Variance: $i_v(t) = \frac{1}{N_t} \sum_{y \in y} (y - \bar{y}_t)^2$

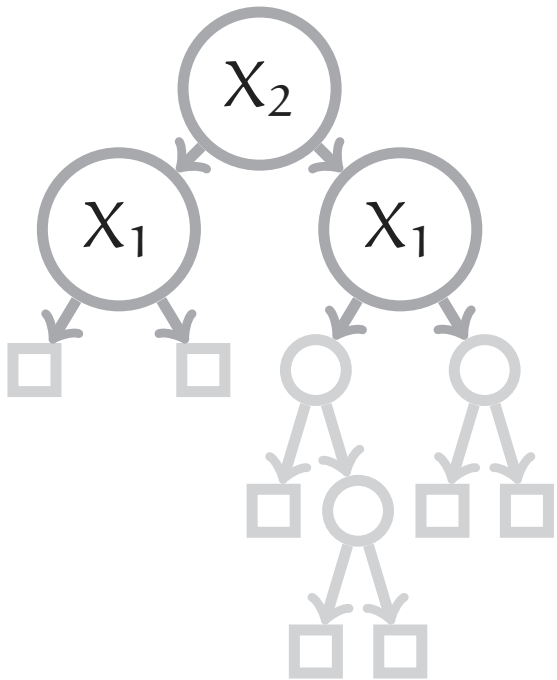
Decision tree

High learning variance

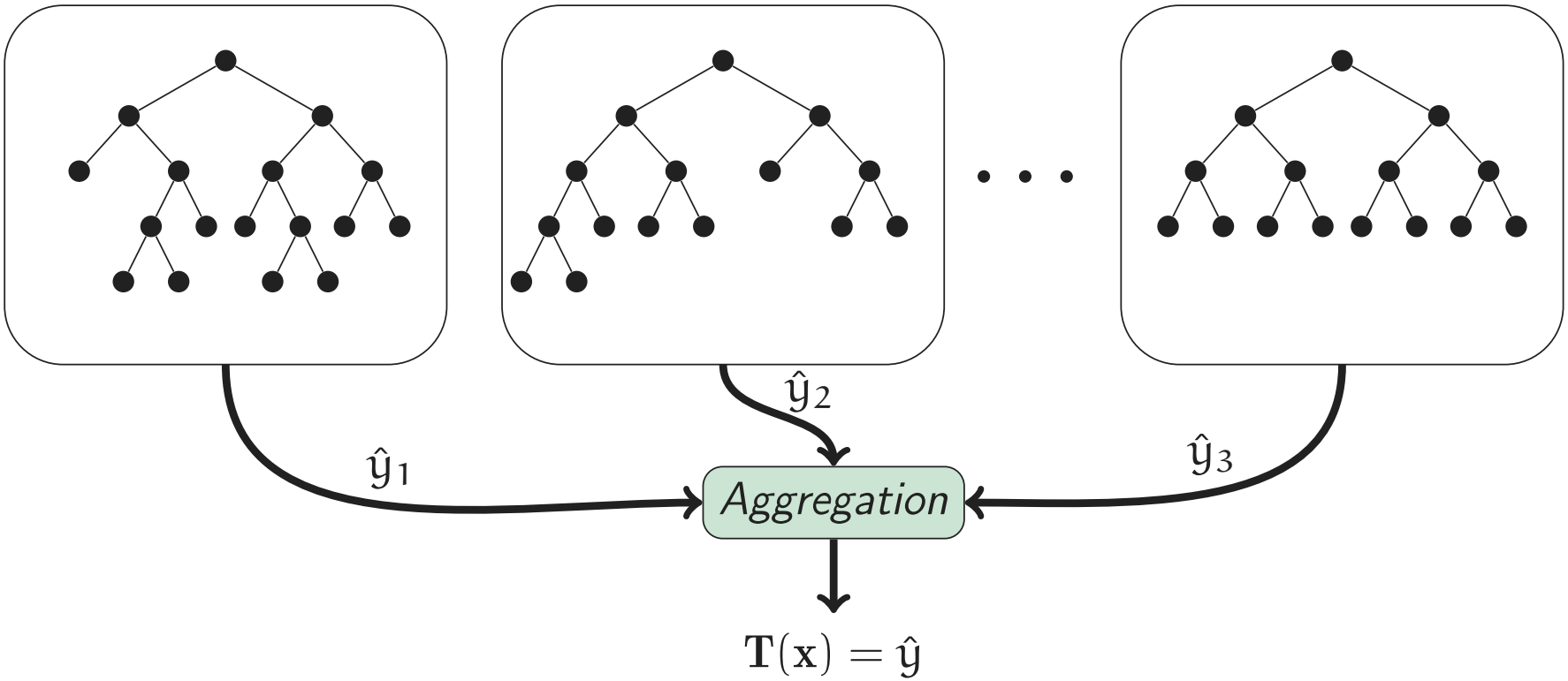








Random forests



Random forests

Techniques to build diverse trees

Ways of introduction randomization in the tree growing procedure

- Tree-wise learning set randomization: e.g. bootstraps or feature (or sample) subspaces
- Node-wise variable randomization
- Node-wise split randomization

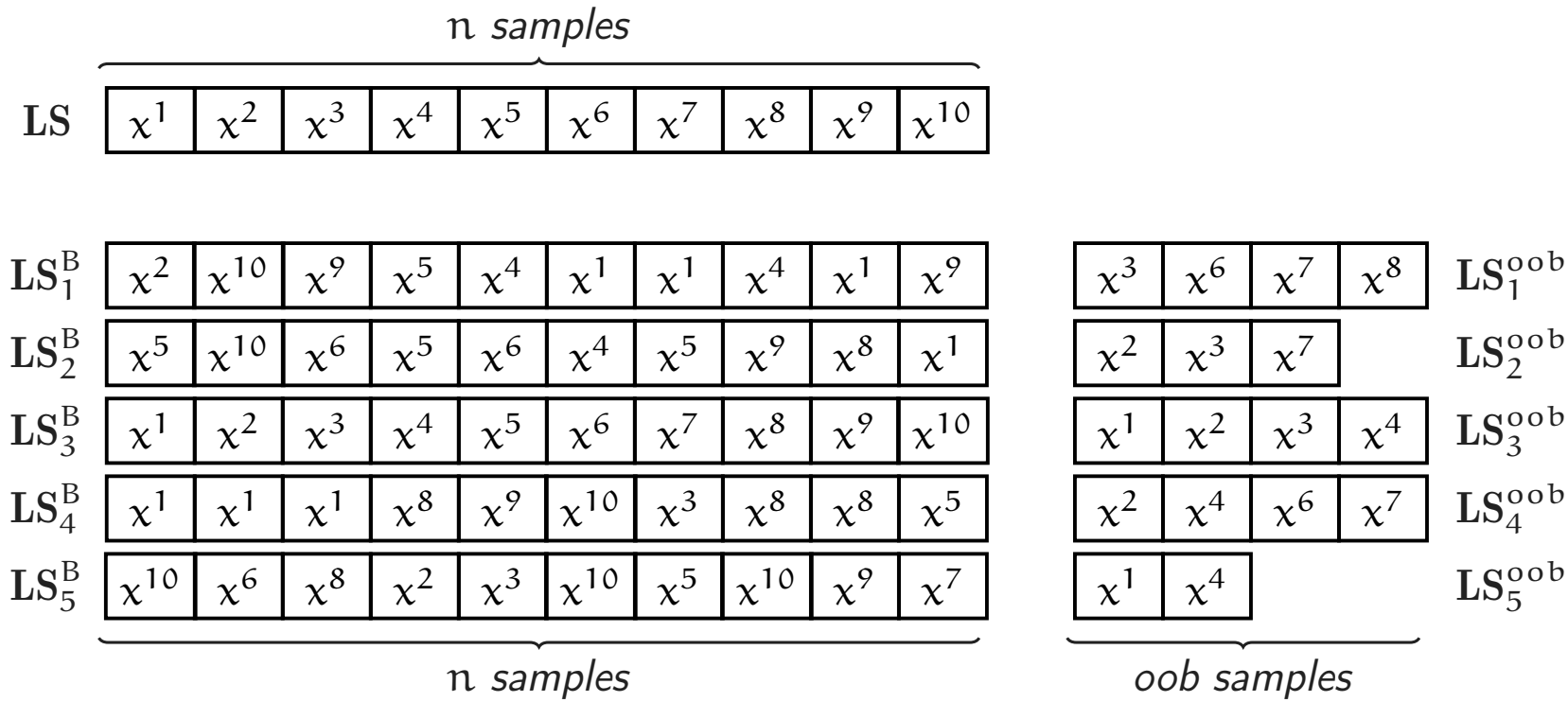
Random forests

Random forest methods vs randomization techniques

	Boot.	Subspace	Feat. rand.	Split rand.
Bagging	✓			
Random Subspace		✓	(✓)	
Random Forest	✓		✓	
Extra-Trees			✓	✓

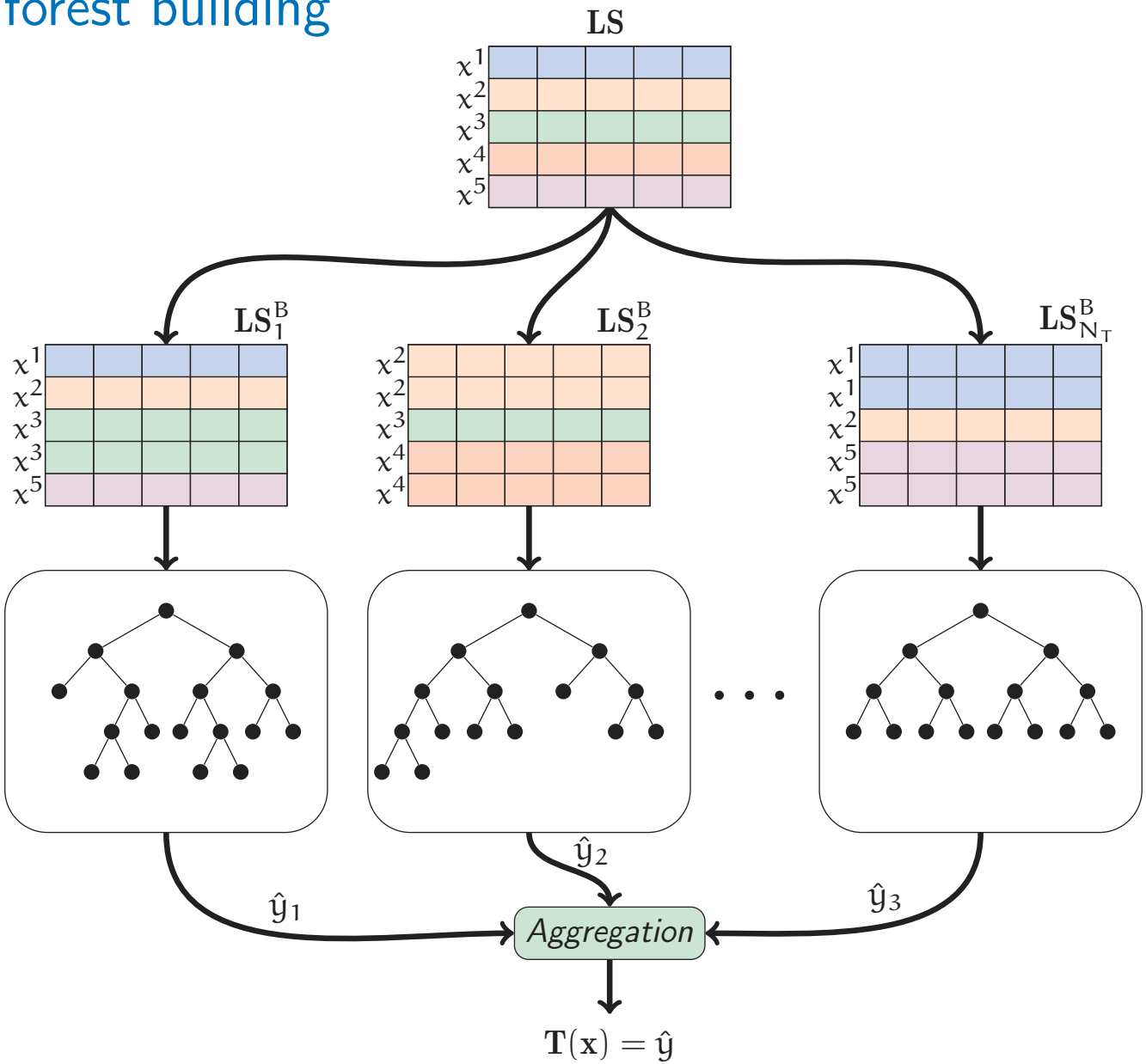
Bagging

How to make bootstrap samples?



Bagging

Random forest building



Variable importances

Two measures

- Mean Decrease of Impurity (MDI)
- Mean Decrease of Accuracy (MDA)

Variable importances

MDI

Mean Decrease of Impurity (MDI)

$$\mathrm{Imp}^{\mathrm{mdi}}(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t^*) = X_m} p(t) \Delta i(s_t^*, t)$$

where

- X_m is an input variable,
- $p(t) = \frac{N_t}{N}$ is the proportion of samples reaching node t,
- N_T is the number of trees;
- s_t^* is the split yielding the largest impurity reduction.

Variable importances

MDA

Mean Decrease of Accuracy (MDA)

$$\mathrm{Imp}^{\mathrm{mda}}(X_m) = \frac{1}{N_T} \sum_{T} (\widetilde{\mathrm{errOOB}}_T^{\,m} - \mathrm{errOOB}_T)$$

where

- N_T is the number of trees,
- errOOB_T is the error (MSE, $0-1$, ...) of a single tree T on its OOB_T sample,
- $\widetilde{\mathrm{errOOB}}_T^{\,m}$ is the error of a single tree T on its OOB_T sample where the values of X_m have been permuted.

Variable selection using random forests

[Genuer et al., 2010]

Experiments

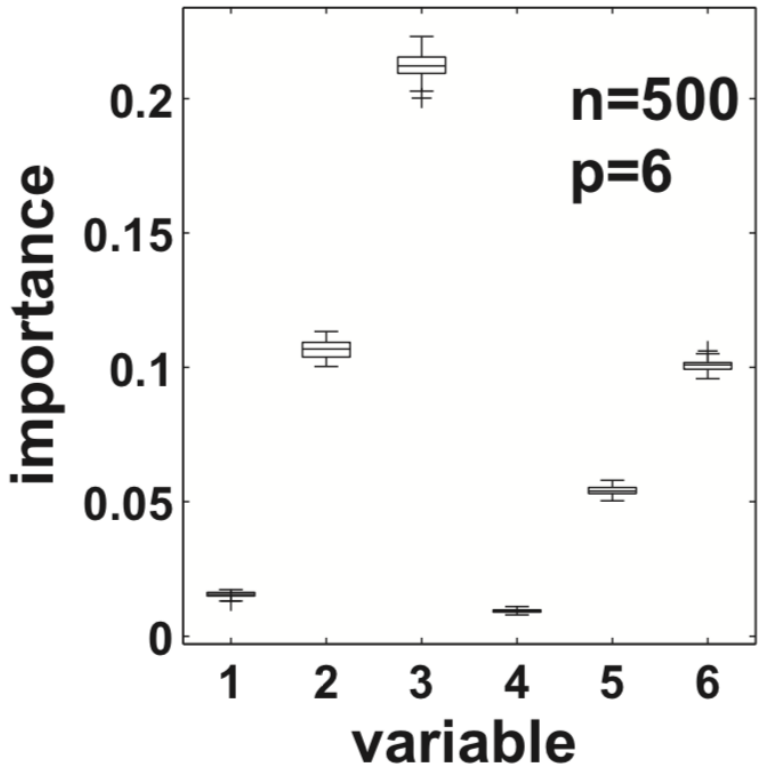
A simulated dataset

A classification problem:

- Two classes : $Y \in \{-1, 1\}$
- Six informative features: $X_1, X_2, \ldots, X_6$ such that
  - There are two independent groups of three relevant features,
  - The first group (X_1, X_2, X_3) is more significant than the second one (X_4, X_5, X_6),
  - Within each group, the third feature (X_3 or X_6) is the most correlated with the output while the first one (X_1 or X_4) is the less correlated.
- All others features are noise: $X_7, \ldots, X_p$

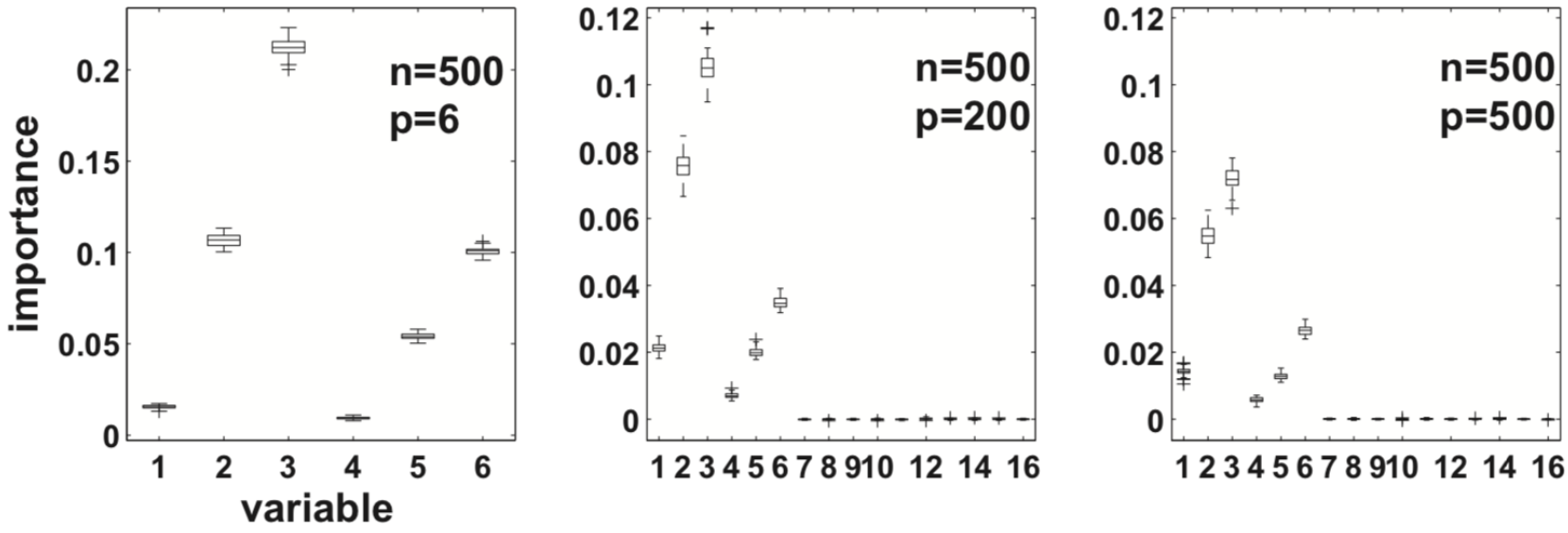
Sensitivity to n and p

Only informative features, $ntree = 500$, $mtry = \sqrt{p}$



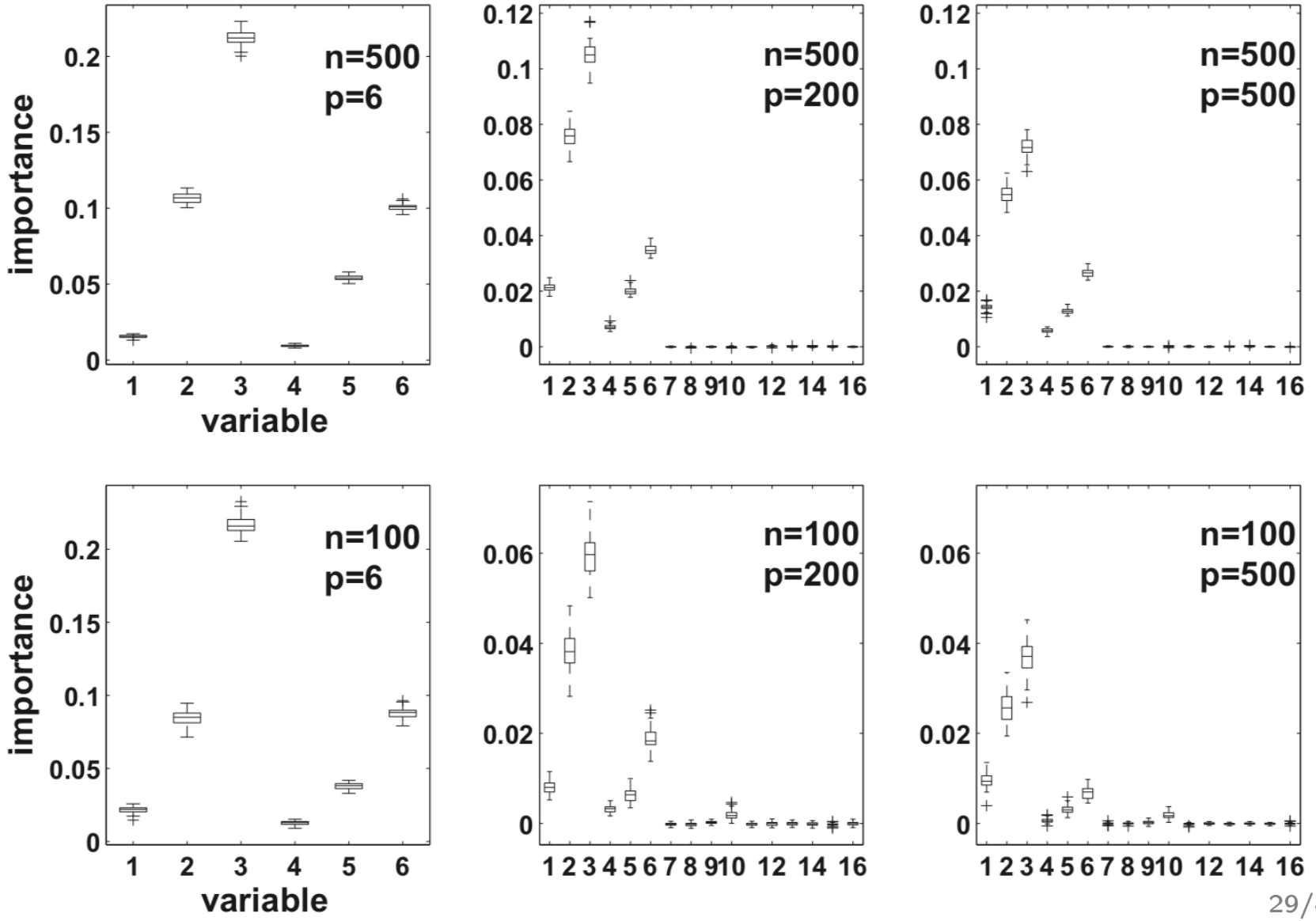
Sensitivity to n and p

Increasing p, $n = 500$, $ntree = 500$, $mtry = \sqrt{p}$



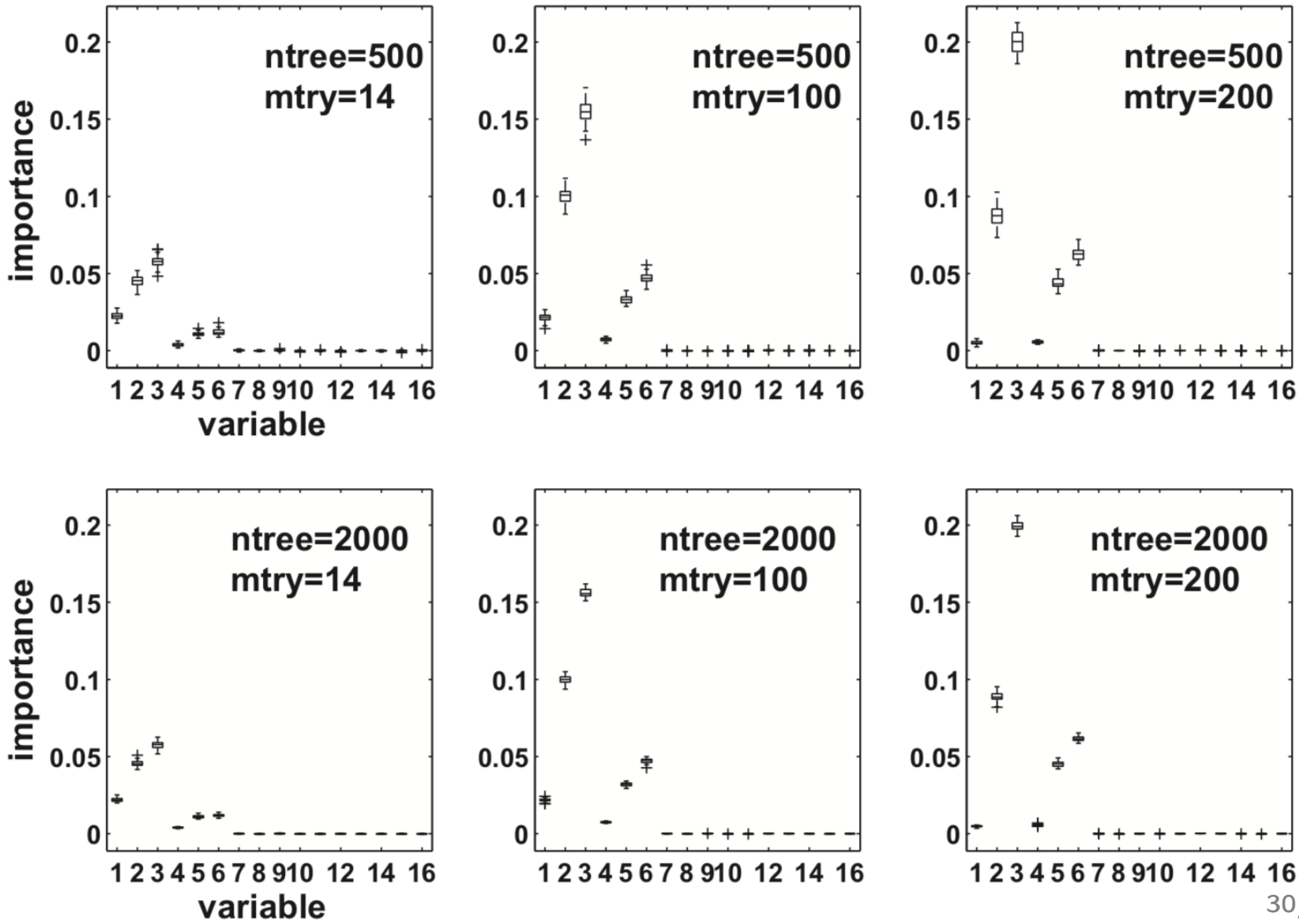
Sensitivity to n and p

Increasing p, $n = \{100, 500\}$, $ntree = 500$, $mtry = \sqrt{p}$



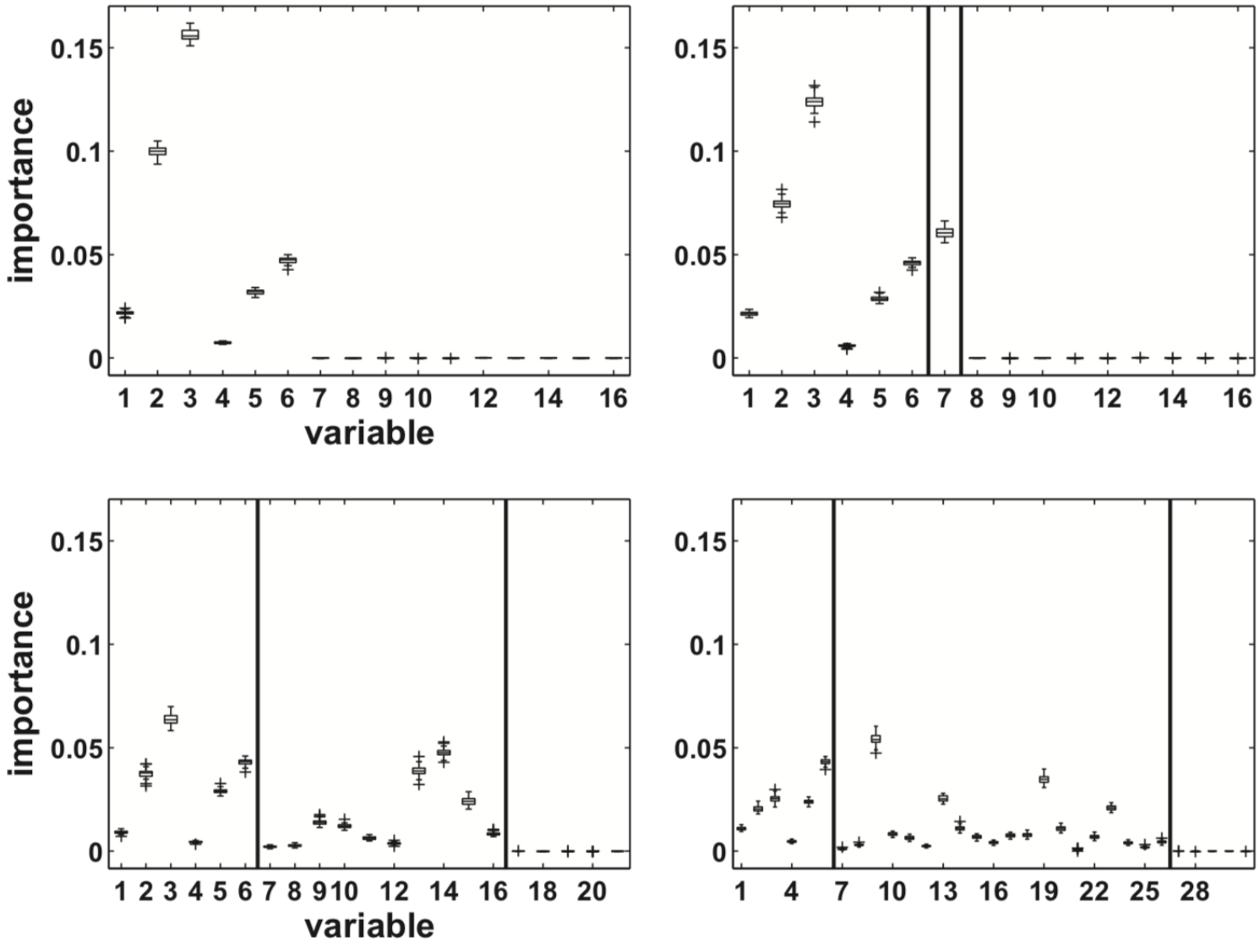
Sensitivity to mtry and ntrees

Increasing mtry, ntree $= \{500, 2000\}$, n $= 100$, p $= 200$



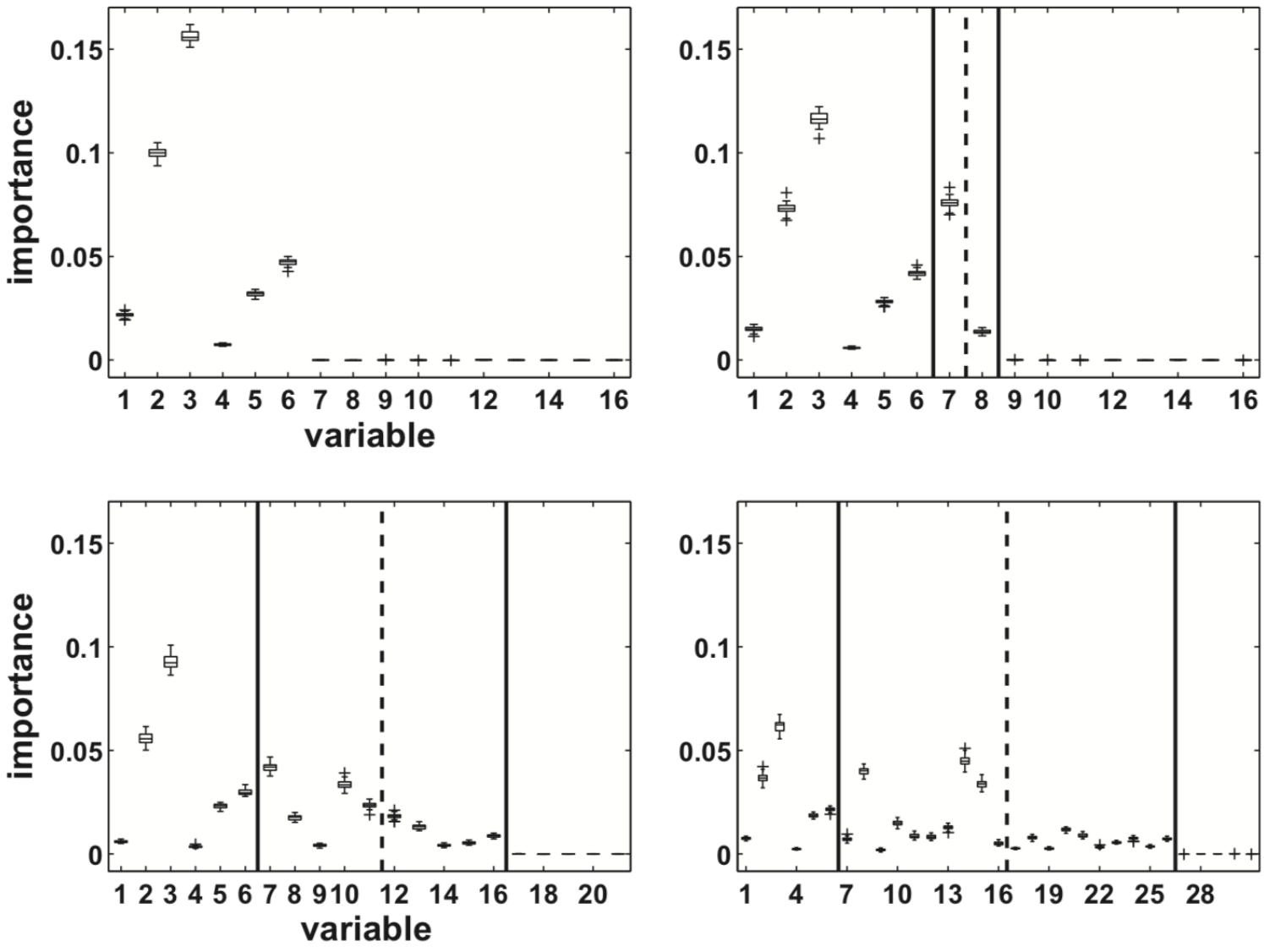
Sensitivity to highly correlated predictors

with X_3, $n = 100$, $p = 200$, $ntree = 2000$, $mtry = 100$



Sensitivity to highly correlated predictors

with X_3 and X_6, $n = 100$, $p = 200$, $ntree = 2000$, $mtry = 100$



Experiment on a real microarray dataset

$n = 102$, $p = 6033$, $ntree = \{500, 2000\}$, $mtry = \{\sqrt{p}, p/3\}$

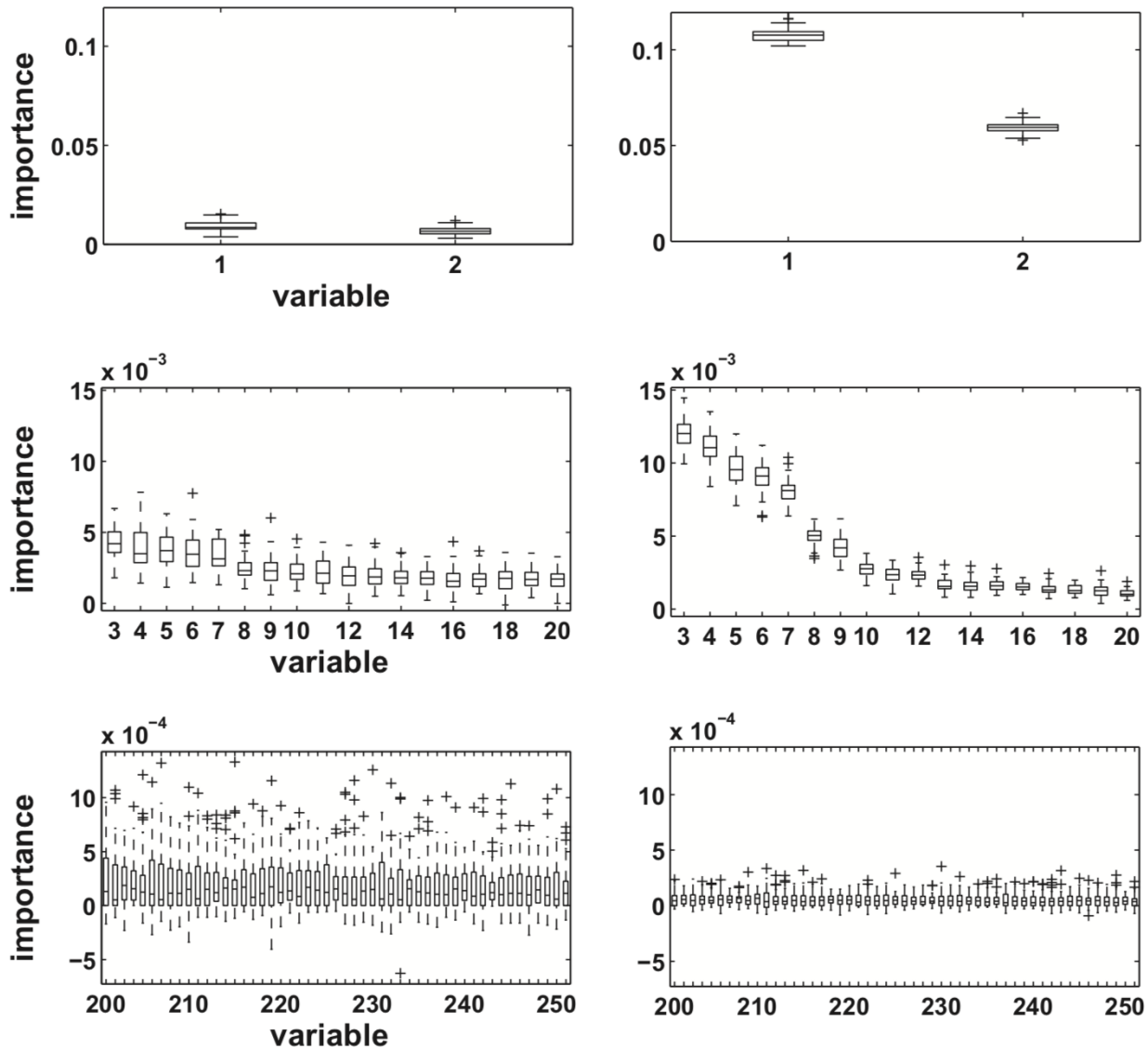


Fig. 5. Variable importance for Prostate data (using *ntree* = 2000 and *mtry* = $p/3$, on the right and using default values on the left).

Variable selection

Two objectives

Their two variable selection objectives:

- To find important variables highly related to the response variable for interpretation purpose;
- To find a small number of variables sufficient to a good parsimonious prediction of the response variable.

A two-steps procedure

Sketch of the algorithm

Step 1. Preliminary elimination and ranking:

- Sort the variables in decreasing order of RF scores of importance.
- Cancel the variables of small importance. Denote by m the number of remaining variables.

Step 2. Variable selection:

- For *interpretation*: construct the nested collection of RF models involving the k first variables, for $\mathrm{k} = 1$ to m, and select the variables involved in the model leading to the smallest OOB error;
- For *prediction*: 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.

Results on the simulated dataset

$n = 100,\ p = 200,\ ntree = 2000,\ mtry = 100$

$X_1, \ldots, X_6$ are respectively represented by ($\triangleright$, $\triangle$, $\circ$, $\star$, $\triangleleft$, $\square$).

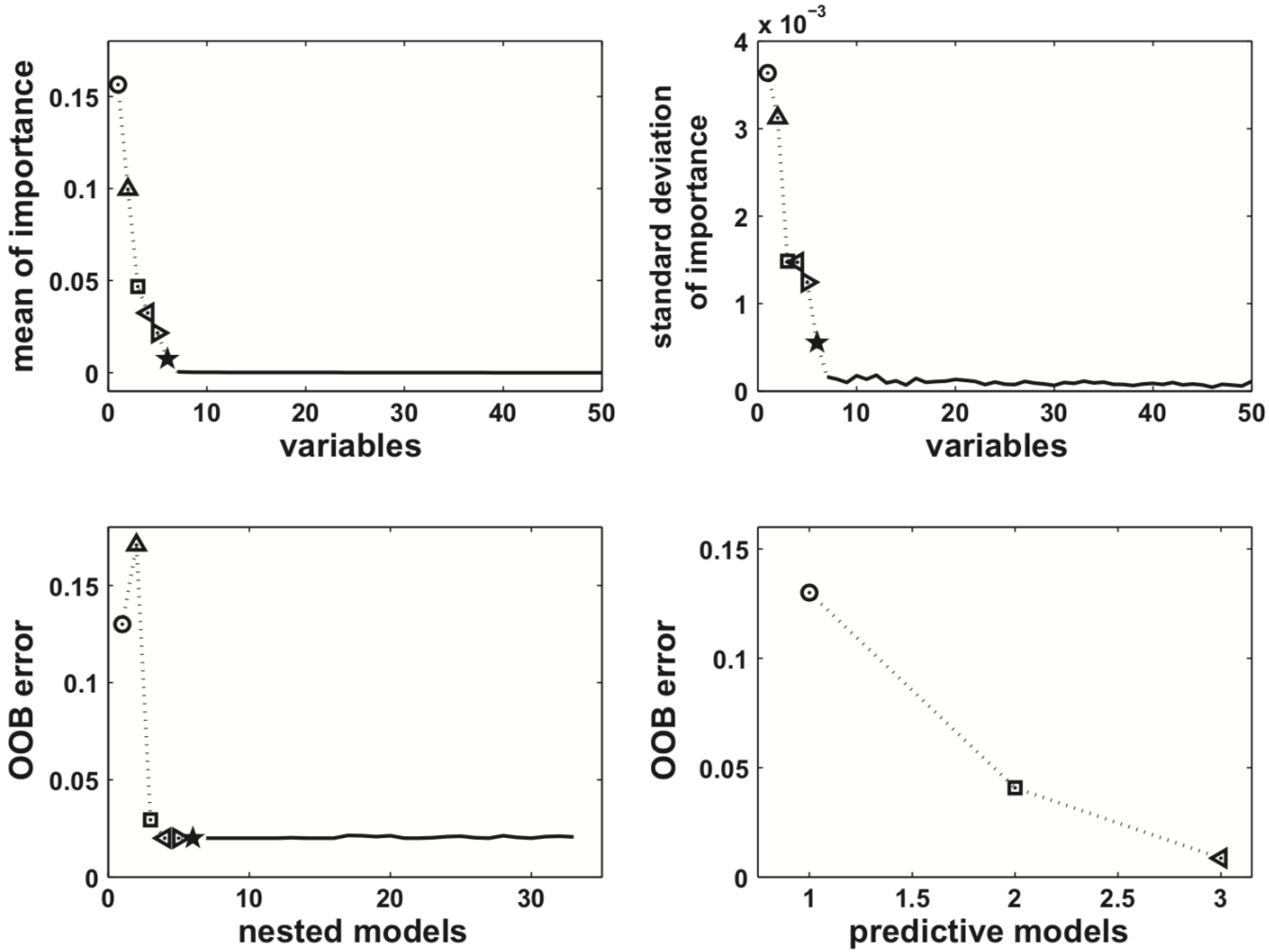


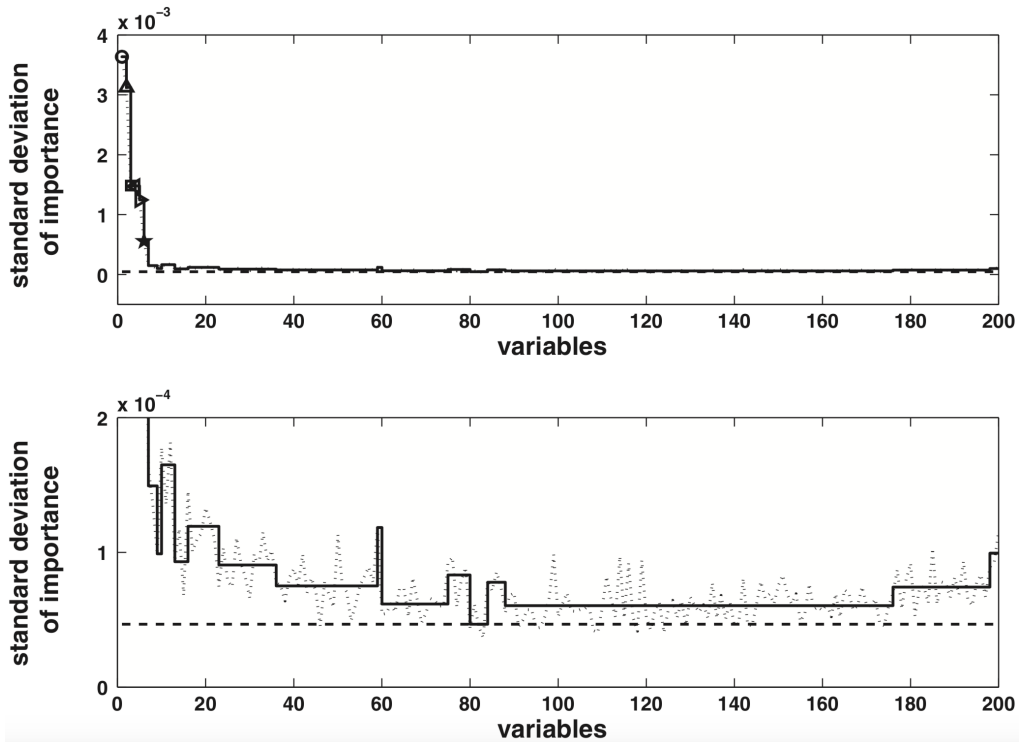
Fig. 6. Variable selection procedures for interpretation and prediction for toys data.

Results on the simulated dataset

Details

Step 1

- Feature ranking (averaged from 50 runs) in descending order;
- Use the standard deviations of importance scores to estimate the threshold* ($p_{elim} = 33$ variables);



* Minimum prediction value given by a CART model fitting this curve.

Results on the simulated dataset

Details

Step 2

- For interpretation: nested models (from only the most important variable, to the one with all 33 kept features);
- For interpretation: they select the smallest model with a sufficiently* small OOB error ($p_{interp} = 4$ variables while minimal OOB error corresponds to 24 variables);
- For prediction: a variable is added only if the error gain exceeds a threshold given by

$$\frac{1}{p_{elim} - p_{interp}} \sum_{j=p_{interp}}^{p_{elim}-1} |errOOB(j+1) - errOOB(j)|$$

where $errOOB(j)$ is the OOB error of the RF built using the j most important variables (gives X_3, X_6, X_5 in that order).

* OOB error less than the minimal OOB error augmented by its empirical standard deviation.

Discussion

Main contributions of their empirical analysis of MDA

- They give some experimental insights about RF importance measure.
- They take into account $\mathtt{mtry}$ and $\mathtt{ntree}$ parameters simultaneously.
- They notice that importance scores of correlated (or duplicated) variables decrease.

Discussion

Main limitations of their empirical analysis of MDA

- They do not consider a (true) high-dimensional setting.
- Only three values of $\texttt{mtry}$, and all are above the number of informative features.

Discussion

Main contributions of their variable selection procedure

They propose:

- a method that addresses both objectives of feature selection;
- a *cheap* way to reduce the number of features in a preliminary step;
- a practical way to find a feature ranking threshold.

Discussion

Main limitations of their variable selection procedure

They note:

- that the threshold value is based on standard deviations while the effective thresholding is performed on importance mean;
- that the threshold estimation strategy is sensible when there exist irrelevant variables;
- that an error evaluation on a test set or using a cross-validation scheme should be preferred.

Discussion

Main limitations of their variable selection procedure

- Importance of their preliminary step.
- Some relevant features may be missed.
- Selection bias.

Conclusion

In their conclusion

it remains unclear. The second remark is about the random feature selection step. The most widely used version of RF selects randomly *mtry* input variables according to the discrete uniform distribution. Two variants can be suggested: the first is to select random inputs according to a distribution coming from a preliminary ranking given by a pilot estimator; the second one is to adaptively update this distribution taking profit of the ranking based on the current forest which is then more and more accurate.

Conclusion

In their conclusion

Finally, let us mention an application for fMRI brain activity classification (see Genuer et al., 2010). This is a typical situation where $n << p$, with a lot of highly correlated variables and where the two objectives have to be addressed: find the most activated (whole) regions of the brain, and build a predictive model involving only a few voxels of the brain. An interesting aspect for us will be the feedback given by specialists, needed to interpret the set of variables found by our algorithm. In addition a lot of well known methods have already been used for these data, so fair comparisons will be easy and fruitful.

Appendices

Experiments

A simulated dataset [Weston et al., 2003]

Linear problem We start with an artificial problem, where six dimensions out of 100 were relevant. The probability of $y = 1$ or -1 was equal. The first three features $\{x_1, x_2, x_3\}$ were drawn[8] as $x_i = yN(i, 1)$, and the second three features $\{x_4, x_5, x_6\}$ were drawn as $x_i = N(0, 1)$ with a probability of 0.7, otherwise the first three were drawn as $x_i = N(0, 1)$ and the second three as $x_i = yN(i-3, 1)$. The remaining features are noise $x_i = N(0, 20)$, $i = 7, \ldots, 100$. The inputs are then scaled to have mean zero and standard deviation one.

In this problem the first six features have redundancy and the rest of the features are irrelevant.

References I

Blausen, S. (2014). com staff" medical gallery of blausen medical 2014. *WikiJournal of Medicine*, 1(2):10.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Ghattas, B. and Ben Ishak, A. (2008). Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. *Journal de la société française de statistique*, 149(3):43–66.

References II

Graner, J. L., Oakes, T. R., French, L. M., and Riedy, G. (2013). Functional mri in the investigation of blast-related traumatic brain injury. *Frontiers in neurology*, 4:16.

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of machine learning research*, 3(Mar):1439–1461.