Forêts aléatoires
Importance et sélection de variables

Jean-Michel Poggi
U. Paris Descartes et LMO, Orsay, U. Paris Saclay

Séminaire nantais inter-établissements en Science des Données
Université de Nantes, Campus Lombarderie

13 décembre 2018

Remerciements et références

- Une référence librement accessible :
  Robin Genuer, Jean-Michel Poggi
  Arbres CART et Forêts aléatoires Importance et sélection de variables
  45 pages, 2017
  https ://hal-descartes.archives-ouvertes.fr/hal-01387654v2

- Texte écrit en collaboration avec Robin Genuer (ISPED, Bordeaux)

- Remerciements à S. Arlot, S. Gey, C. Tuleau-Malot et N. Villa-Vialaneix

Plan

Leo Breiman

- De CART aux RF : 20 ans d'une trajectoire scientifique
- Olshen, Breiman (2001) et Cutler (2010)
- D'abord, en probabilités sous un angle très proche des mathématiques pures, il a ensuite marqué de son empreinte la statistique appliquée et l'apprentissage
- Série de papiers dans les *Annals of Statistics* et dans *Machine Learning*

Cadre

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ v.a. i.i.d. de même loi que (X, Y).

$X \in \mathbb{R}^p$ (variables explicatives) ; on peut aussi avoir $X \in \mathbb{R}^{p'} \otimes \mathcal{Q}$ mixte. $Y \in \mathcal{Y}$ (réponse) :

- $\mathcal{Y} = \mathbb{R}$: régression
- $\mathcal{Y} = \{1, \ldots, L\}$: classification

But : construire un prédicteur $\widehat{h} : \mathbb{R}^p \to \mathcal{Y}$

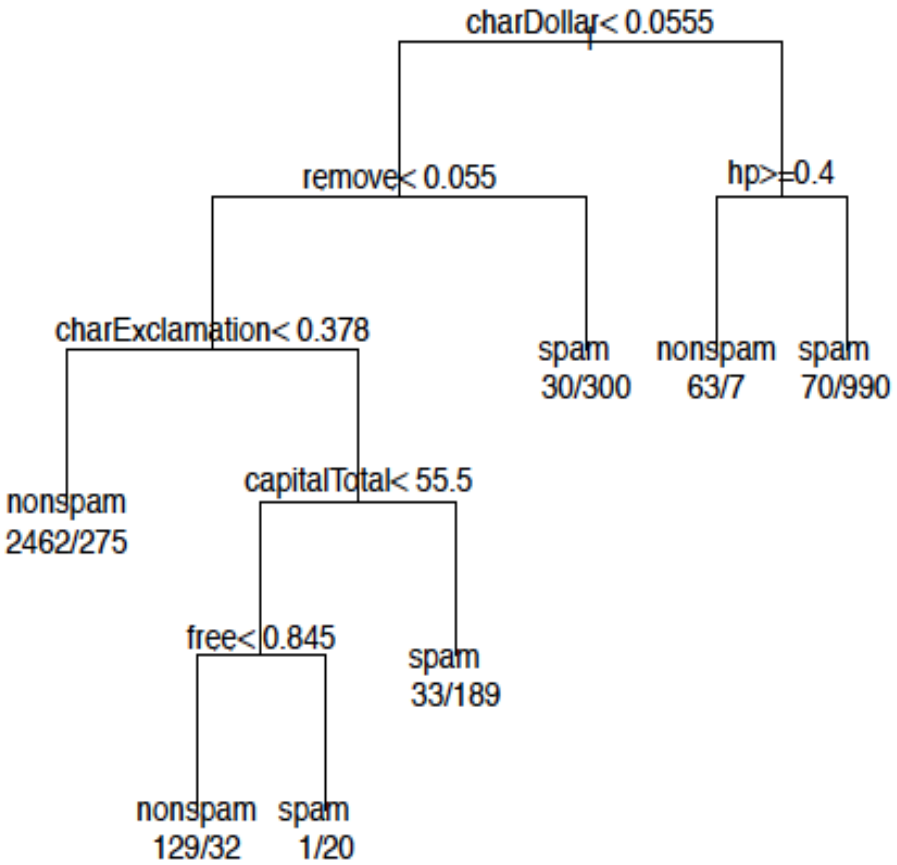
Arbres CART Breiman et al. (1984)

- famille des méthodes d'arbres de décision
- algorithme qui est la base de méthodes très efficaces

Forêts aléatoires Breiman (2001)

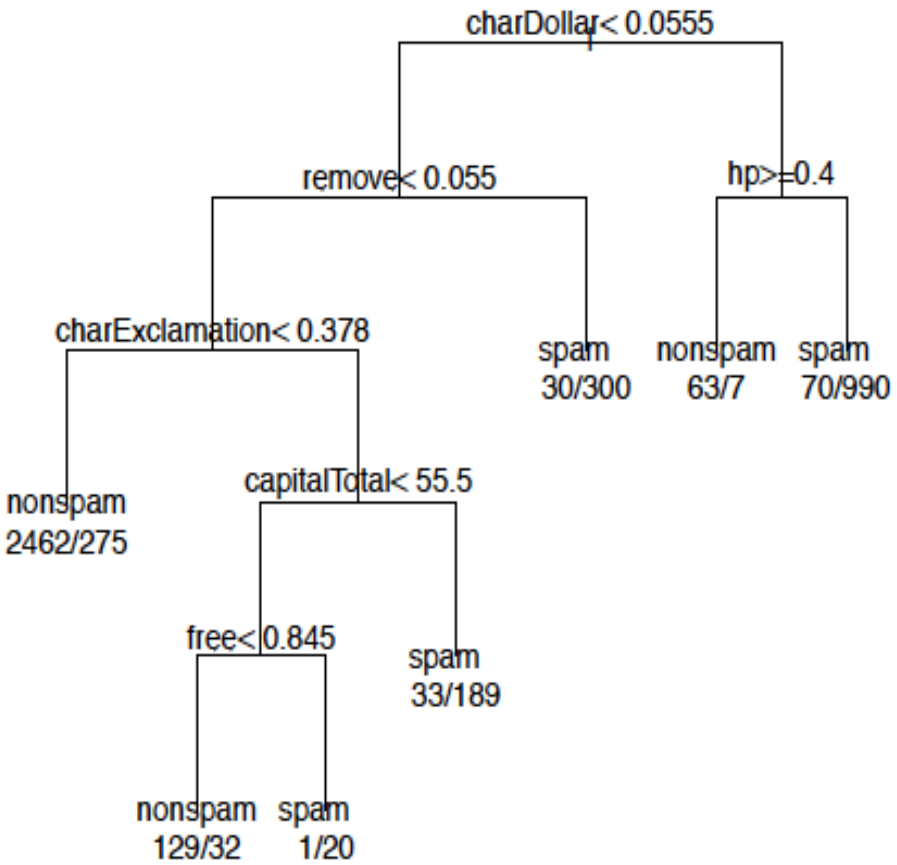
- famille des méthodes d'ensemble
- algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression

Le jeu de données *spam*



- Construire un détecteur automatique de spams et déterminer les variables importantes
- n=4601 emails (1813 spams, 40%)
- p=57 prédicteurs :
  - 54 sont des % de mots ou de caractères donnés comme "$", "!", "remove", "free"
  - 2 liées aux longueurs de suites de majuscules (moyenne, maximum) et enfin le nombre de majuscules

Un arbre CART pour les données *spam*



- Structure de l'arbre : 5 noeuds internes et 7 feuilles ; splits basés sur *charDollar*, *remove*, *hp*, *free*, *charExclamation*, et *capitalTotal*
- Prédiction par l'arbre : les feuilles donnent les prédictions de Y (*spam* ou *nonspam*) et sa distribution
- Interprétation : chemin racine - la feuille la plus à droite : si beaucoup de $ et peu de *hp* alors presque toujours spam

CART *vs* arbres de décision

- Parfois introduites avant CART, d'autres méthodes pour construire des arbres de décision sont disponibles :
  - CHAID par Kass (1980)
  - *C*4.5 par Quinlan (1993)
- La méthode des arbres de décision souffrait de fortes critiques justifiées et CART leur offre un cadre conceptuel de type sélection de modèles, qui leur confère ainsi à la fois une large applicabilité, une facilité d'interprétation et des garanties théoriques
- L'actualité des arbres de décision perceptible dans deux synthèses récentes :
  - Patil et Bichkar (2012) en informatique
  - Loh (2014) en statistique

Construction

Arbre : prédicteur constant par morceaux, obtenu par partitionnement récursif binaire de $\mathbb{R}^p$

Restriction : coupures parallèles aux axes

Classiquement, à chaque étape du partitionnement binaire, on vise à séparer "au mieux" les données du noeud courant, en recherchant la coupure qui conduit à la plus forte décroissance de l'hétérogénéité des deux noeuds fils

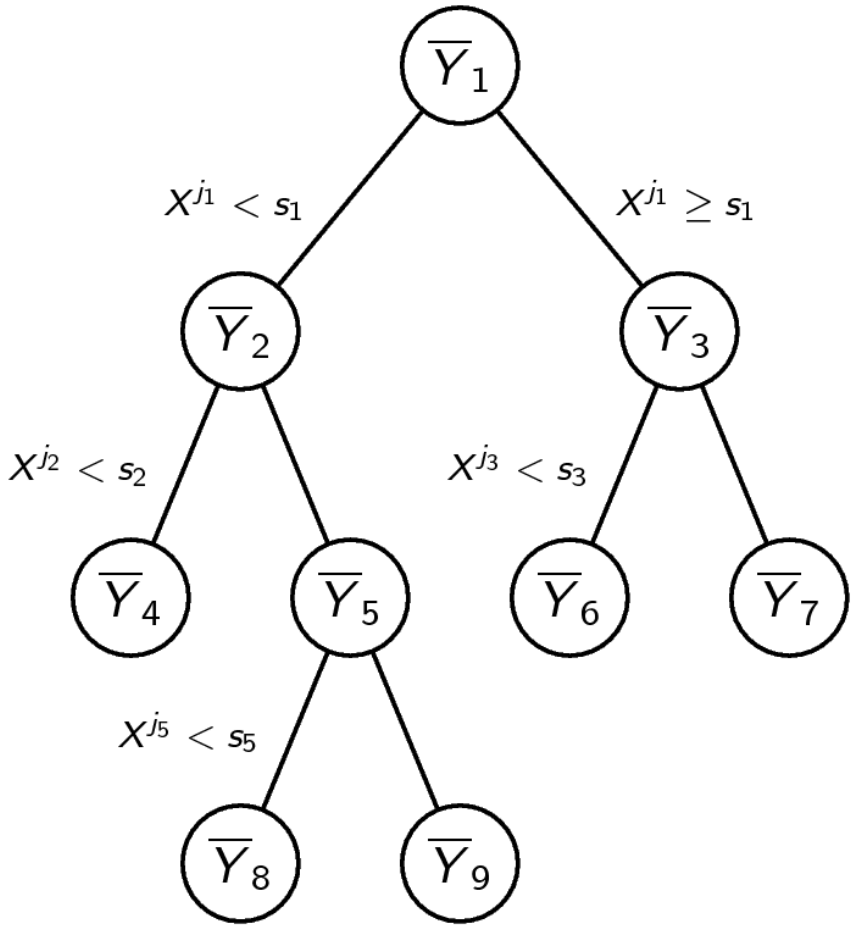
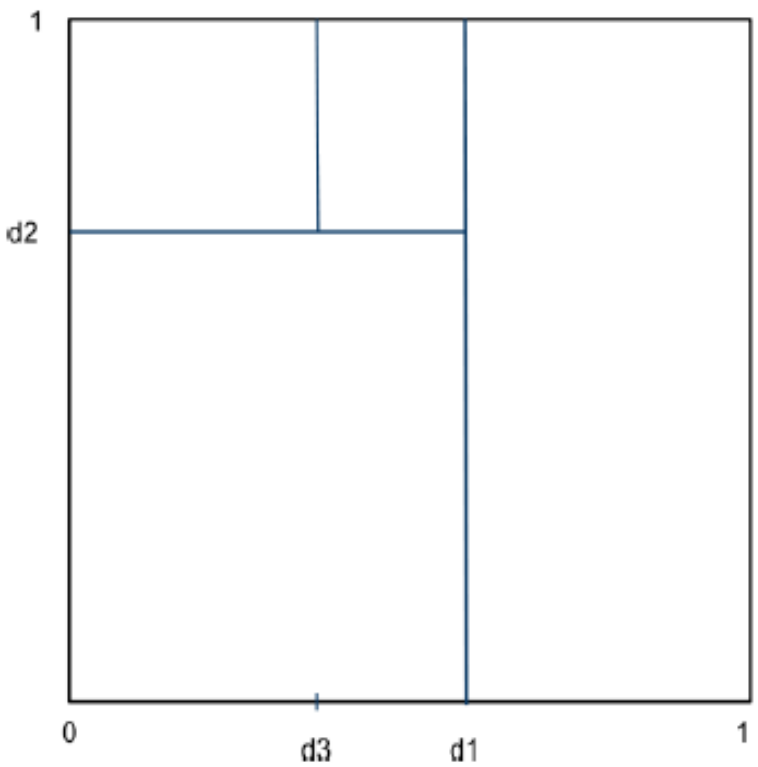
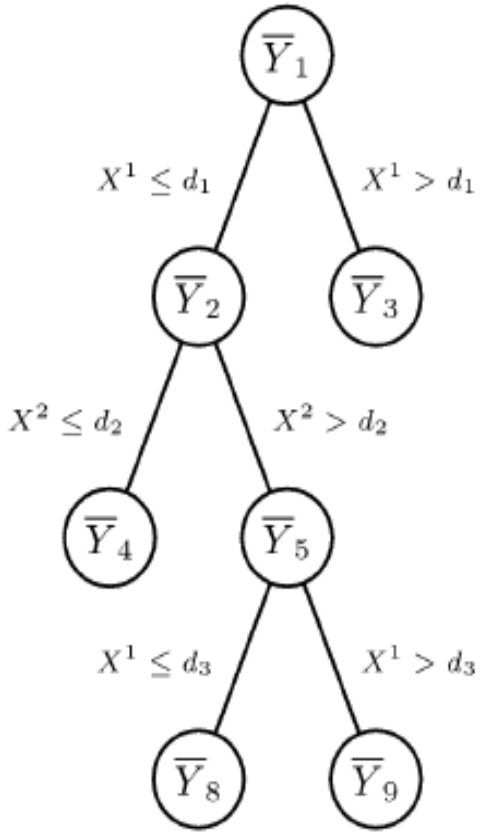


FIGURE : Arbre de régression

Arbre CART et fonction constante par morceaux





Arbre de régression *vs* de classification

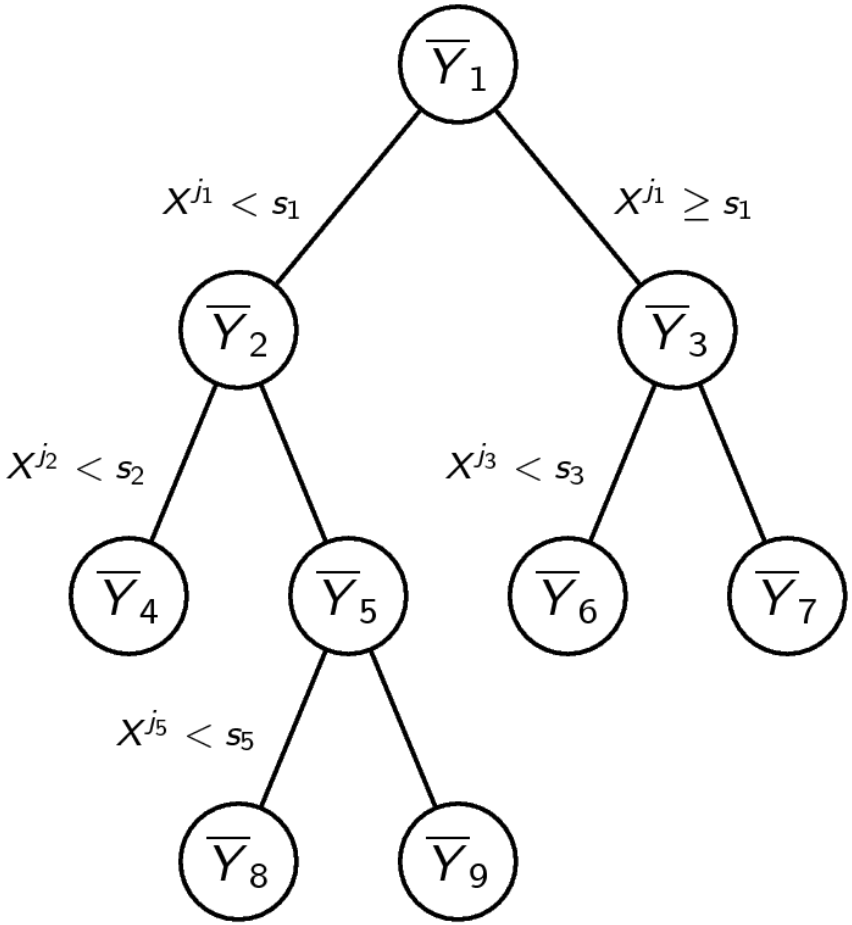


FIGURE : Arbre de régression

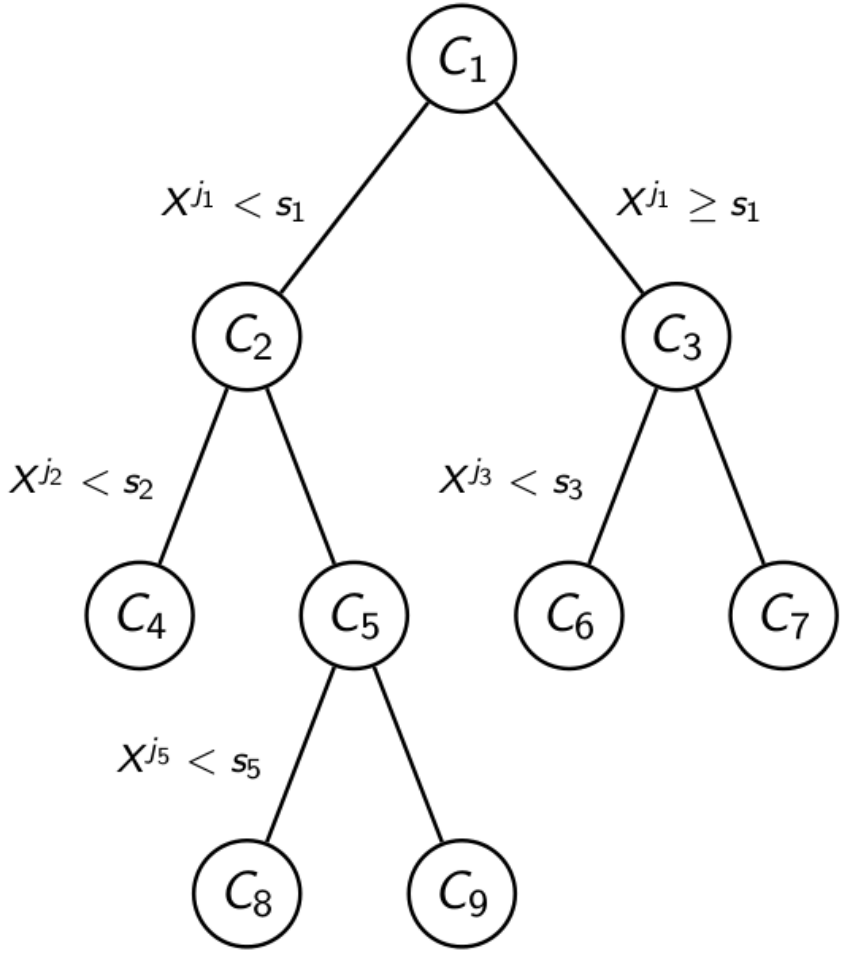
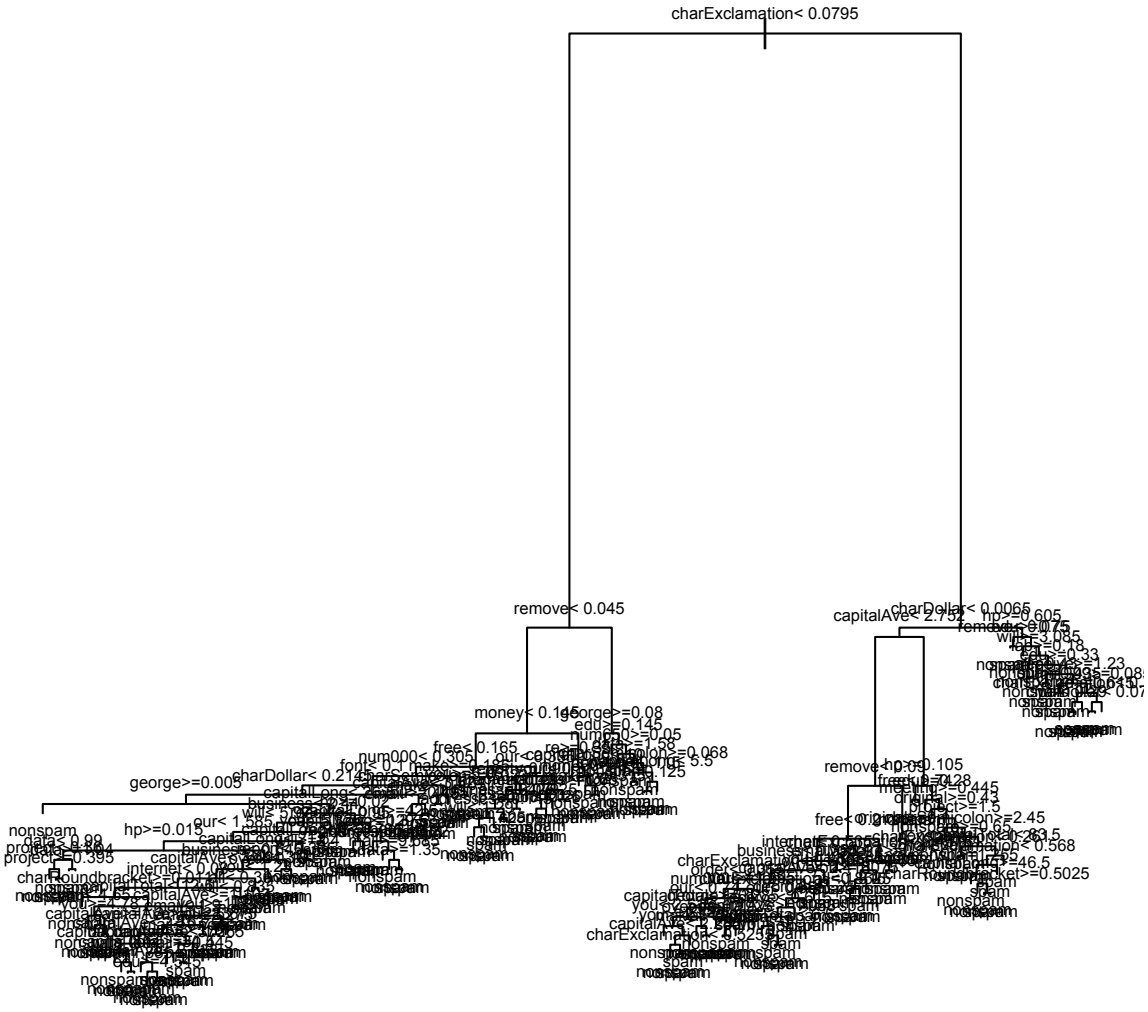
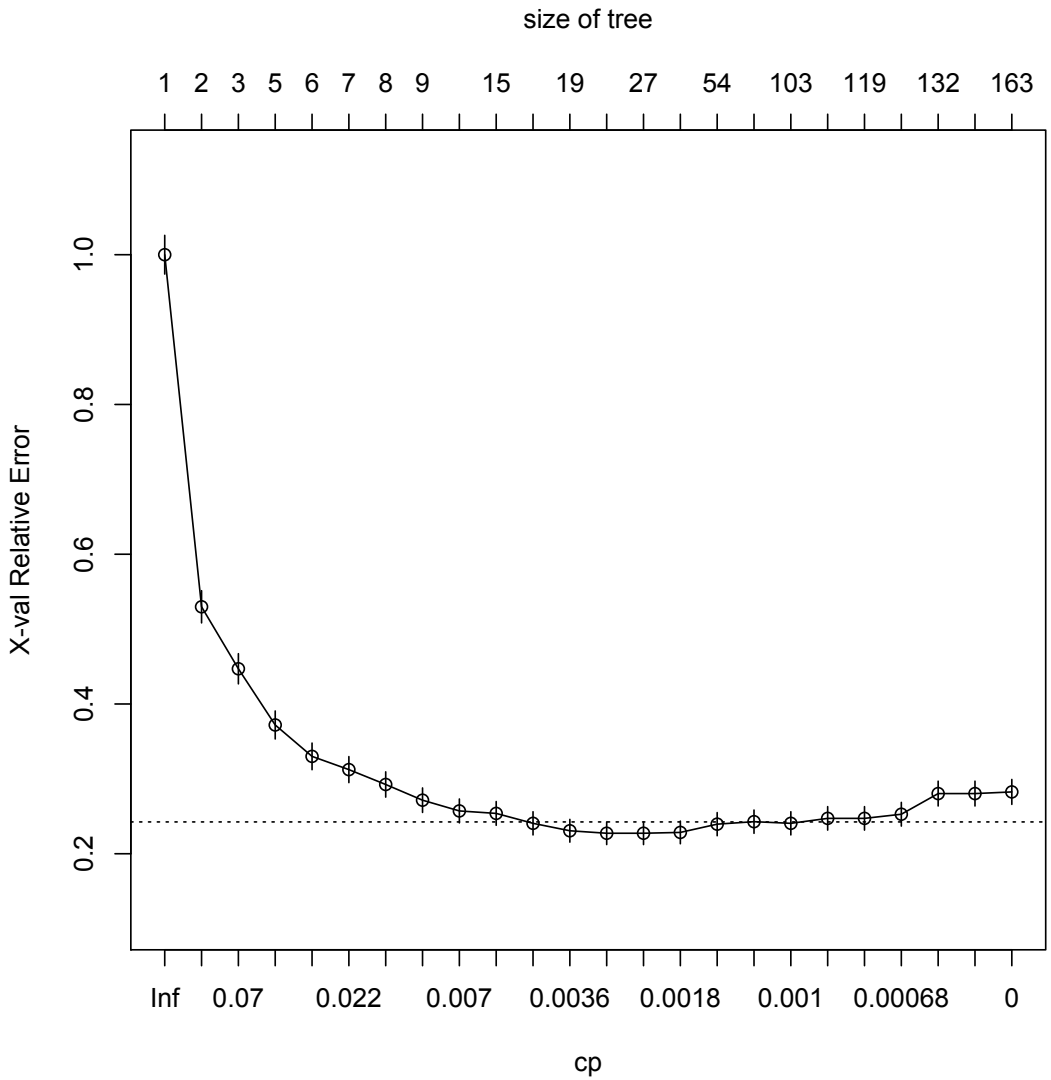


FIGURE : Arbre de classification

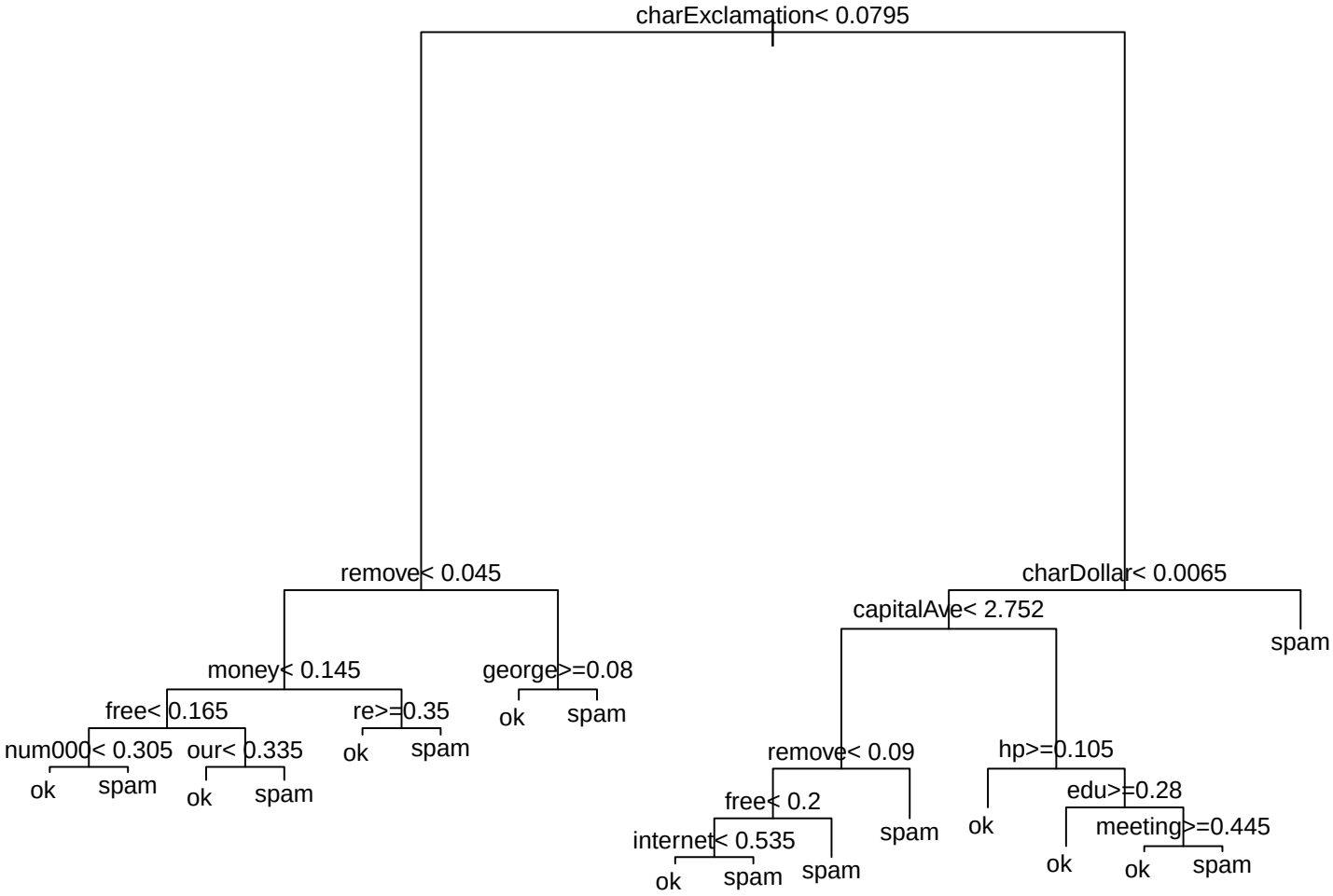
Données *spam* : arbre maximal



Données *spam* : suite de sous-arbres élagués



Données *spam* : arbre optimal à 1 SE près



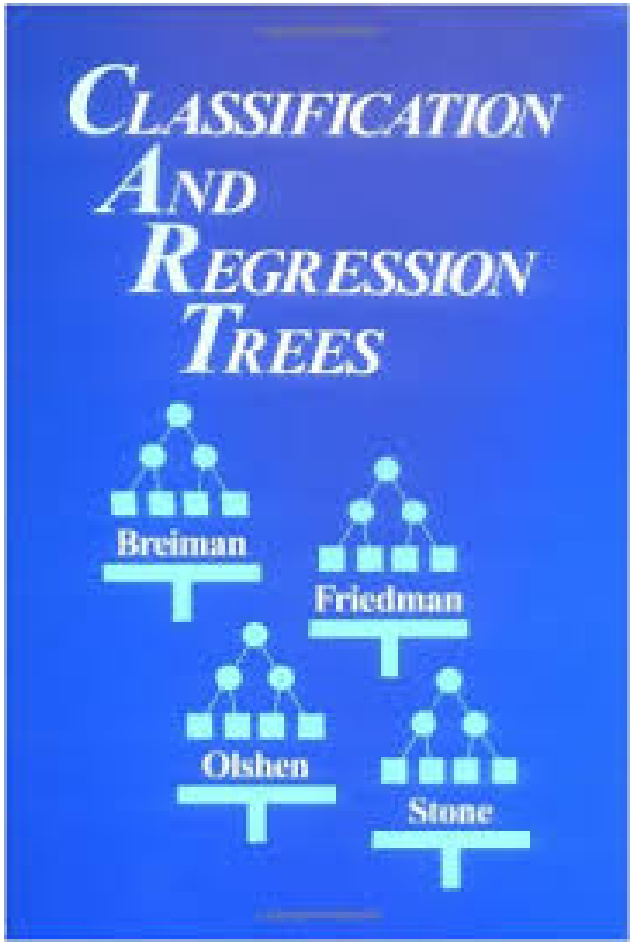
Données *spam* : arbre optimal à 1 SE près

- Le meilleur sous-arbre élagué de l'arbre maximal (à 1 SE près)
  - 17 feuilles
  - seules 14 variables (parmi les 57 initiales) figurent dans les découpes des 16 nœuds internes : `charExclamation`, `charDollar`, `remove`, `capitalAve`, `money`, `george`, `hp`, `free`, `re`, `num000`, `our`, `edu`, `internet` `meeting`
- Deux chemins interprétés :
  - de la racine à la feuille la plus à droite : un mail qui contient beaucoup de $ et de ! est presque toujours un spam
  - de la racine à la cinquième feuille la plus à droite : un mail contenant beaucoup de !, de lettres capitales et de `hp` mais peu de $ n'est presque jamais un spam

Arbre	2 feuilles	1 s.e.	maximal	optimal
Erreur empirique	0.208	0.073	0.000	0.062
Erreur test	0.209	0.096	0.096	0.086

TABLE : Erreurs (empirique et test) des 4 arbres

La référence



- CART Classification And Regression Trees, Breiman et al. (1984)
- Une introduction compacte et claire de la méthode CART en régression se trouve dans le chapitre 2 de la thèse de S. Gey (2002)
- voir Zhang, Singer (2010) et bien entendu le livre Hastie, Tibshirani, Friedman (2009)

Pour et ... contre

- Modele non paramétrique + partition des données
- Un cadre unique pour la régression et la classification binaire or multi-classes
- Modèles faciles à interpréter
- Predicteurs numériques mélangés à des catégoriels
- Découpes compétitives : développement manuel de l'arbre maximal
- Traitement élégant des valeurs manquantes en prédiction : découpes de substitution

- Principal inconvénient : manque de stabilité
- Prédicteur de base pour : bagging, boosting, random forests

Forêts aléatoires

- Introduites par Breiman (2001), elles font partie de la famille des méthodes d'ensemble, Dietterich (1999,2000), on peut citer *Bagging, Boosting, Randomizing Outputs, Random Subspace*
- Algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression. De plus en plus utilisées pour traiter de nombreuses données réelles dans des domaines d'application variés :
  - biopuces Díaz-Uriarte et Alvarez De Andres (2006)
  - l'écologie Prasad et al. (2006)
  - la prévision de la pollution Ghattas (1999)
  - la génomique Goldstein et al. (2010) et Boulesteix et al. (2012)
  - et pour une revue plus large, voir Verikas et al. (2011)
- "Couronnées" dans Fernández-Delgado et al. (2014), elles étaient absentes de Wu et al. (2008) qui mentionne CART

Forêts aléatoires

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ v.a. i.i.d. de même loi que (X, Y).
$X \in \mathbb{R}^p$ (variables explicatives), $Y \in \mathcal{Y}$ (variable réponse) $\mathcal{Y} = \mathbb{R}$ en régression et $\mathcal{Y} = \{1, \ldots, L\}$ en classification
But : construire un prédicteur $\widehat{h} : \mathbb{R}^p \to \mathcal{Y}$

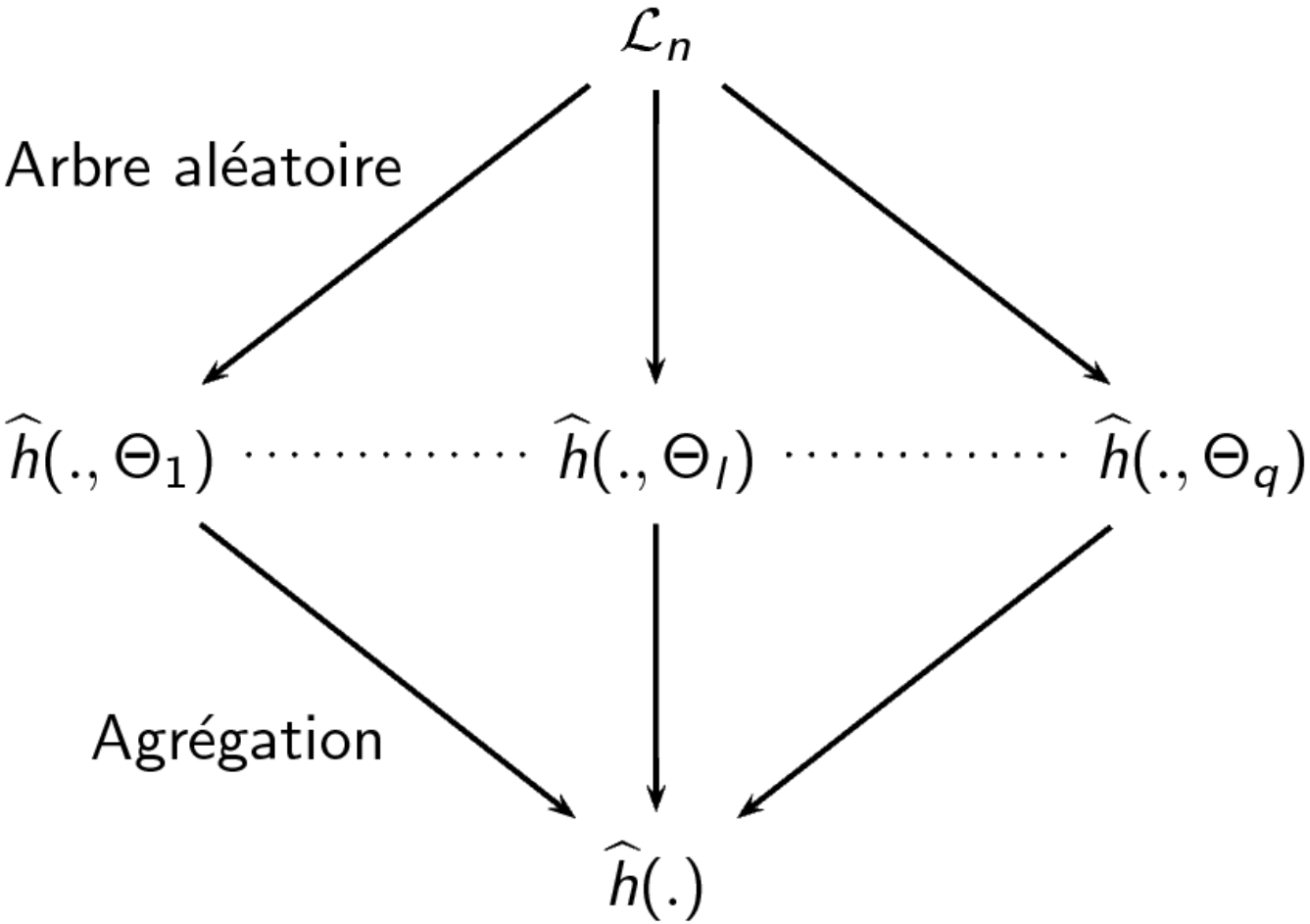
Définition : Forêts aléatoires (Breiman 2001)

$\left\{\widehat{h}(., \Theta_\ell), 1 \leq \ell \leq q\right\}$ collection de prédicteurs par arbre,
$(\Theta_\ell)_{1 \leq \ell \leq q}$ v.a. i.i.d. indépendantes de $\mathcal{L}_n$.
Prédicteur des forêts aléatoires $\widehat{h}$ obtenu en agrégeant la collection d'arbres.

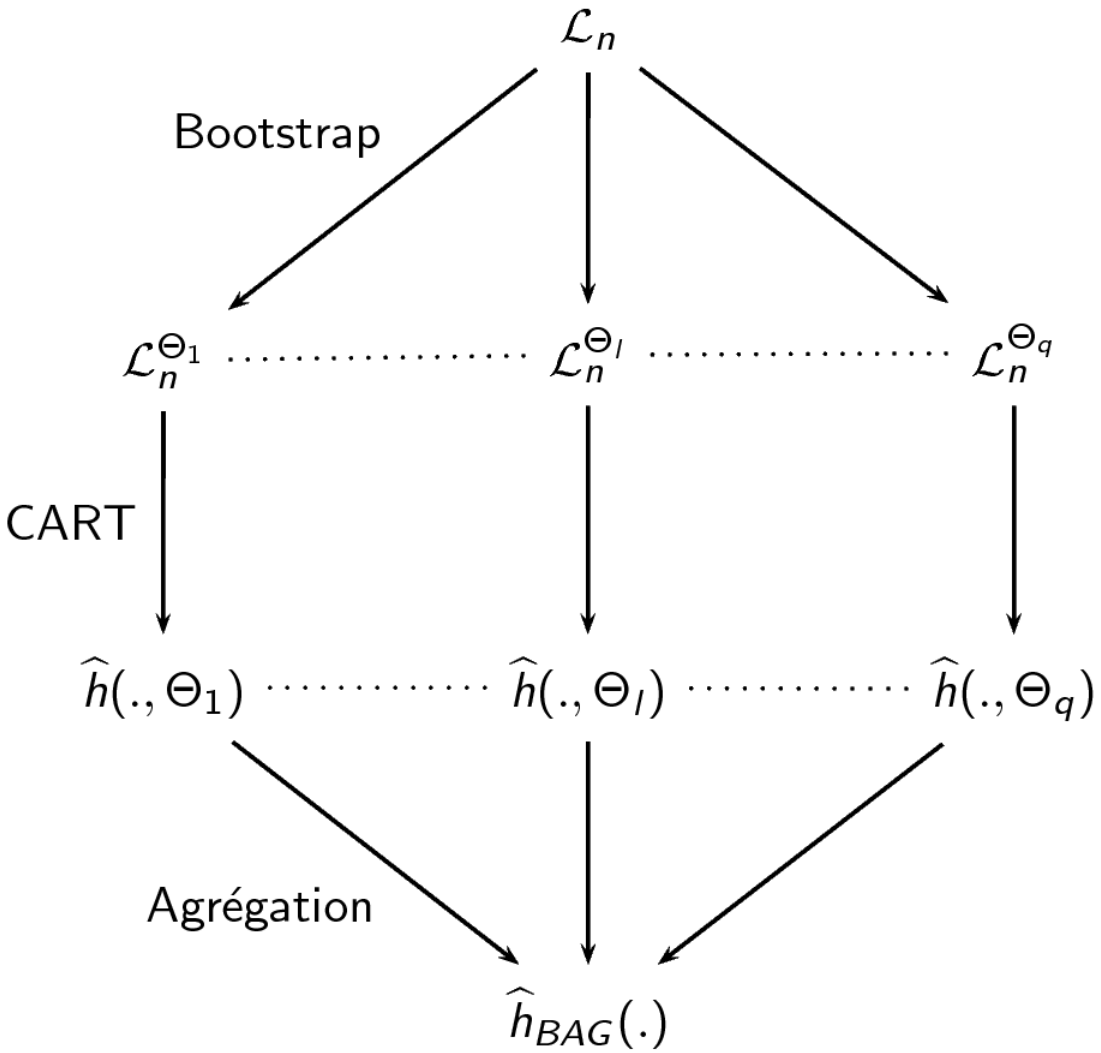
Agrégation :

- $\widehat{h}(x) = \dfrac{1}{q} \displaystyle\sum_{\ell=1}^{q} \widehat{h}(x, \Theta_\ell)$ en régression
- $\widehat{h}(x) = \underset{1 \leq c \leq L}{\operatorname{argmax}} \displaystyle\sum_{\ell=1}^{q} \mathbb{1}_{\widehat{h}(x, \Theta_\ell) = c}$ en classification

Forêts aléatoires



Bagging (Breiman 1996)



Instabilité de CART $\Rightarrow$ amélioration des performances

Random Forests-Random Inputs (Breiman 2001)

Définition : Arbre RI

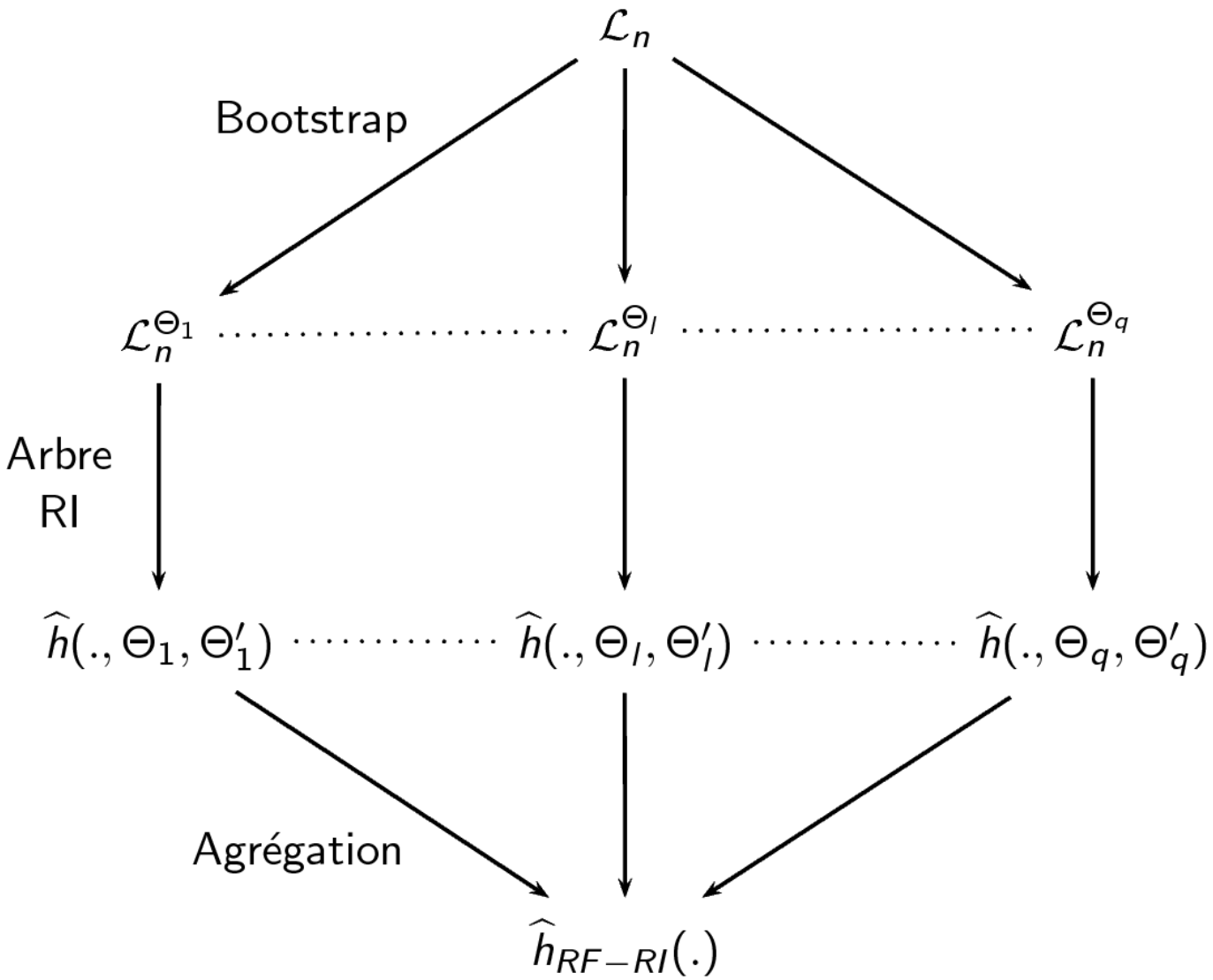
Un arbre RI consiste à tirer aléatoirement, à chaque noeud `mtry` variables, puis à chercher la meilleure coupure uniquement parmi les variables sélectionnées.

`mtry` est le même pour tous les noeuds de tous les arbres de la forêt mais, bien sûr, les variables considérées en chaque noeud pour le choix de la meilleure découpe changent aléatoirement

Définition : Random Forests-RI

Une forêt Random Forests-RI est obtenue en effectuant du Bagging avec des arbres RI.

Random Forests-RI



Aléa supplémentaire $\Rightarrow$ amélioration des performances

Données *spam* : arbre optimal, Bagging, RF

Prédicteur	arbre optimal	bagging	forêt aléatoire
Erreur test	0.086	0.060	0.052

TABLE : Erreurs test du bagging et des forêts aléatoires, comparées à celles de l'arbre optimal pour les données `spam`

- Bagging en utilisant aussi le package `randomForest` et en construisant un prédicteur Bagging avec comme règle de base un arbre CART non-élagué (le package ne permet pas d'élaguer les arbres d'une forêt)
- Forêt aléatoire construite à l'aide du package `randomForest` avec les paramètres par défaut

Famille de forêts aléatoires

Exemples d'aléas supplémentaires :

- rééchantillonnage préalable à la construction de l'arbre,
- choix aléatoire de la variable de coupure à chaque noeud,
- choix aléatoire du point de coupure à chaque noeud.

Deux grandes familles de forêts aléatoires :

- Classiques : partition optimisée sur les données d'apprentissage $\mathcal{L}_n$
- Purement aléatoires : partition tirée aléatoirement, indépendamment de $\mathcal{L}_n$

Des variantes et ... des résultats théoriques

Définition : Forêts purement aléatoires (PRF)

Une PRF est une agrégation d'arbres purement aléatoires, si la partition associée à chacun de ces arbres est tirée aléatoirement indépendamment de $\mathcal{L}_n$

- PRF en théorie :
  - Breiman (2000, Biau et al. (2008), Zhu et al. (2015), Ishwaran, Kogalur (2010), Denil et al. (2014) : résultats de consistance
  - Genuer (2012) : résultat de réduction de variance et vitesse de convergence en dim. 1 puis Arlot, Genuer (2014) en dim. d
  - Biau (2012) : résultat de réduction de variance et de biais dans un contexte de réduction de dimension
  - Mentch, Hooker (2014), Wager (2014) : normalité asympt.
  - Scornet, Biau, Vert (2015) : consistance pour les RF de Breiman, pour les modèles additifs
- PRF en pratique :
  - Cutler, Zhao (2001), Geurts et al. (2006), Duroux et al. (2016)

Des variantes et ... des résultats théoriques

- Récent papier de revue Biau, Scornet (2016) : excellente synthèse des travaux théoriques + discussion

- Dans celle-ci, Arlot, Genuer (2016) étudient l'apport des ingrédients des RF, théoriquement pour une variante simple de RF et par simulation pour une variante proche des RF-RI

  - c'est la randomisation des partitions (qu'elle soit obtenue grâce au bootstrap, au tirage des m variables à chaque nœud ou au tirage du point de coupure) qui serait la plus cruciale

  - Voici pourquoi le Bagging (qui ne randomise pas sur la recherche de la coupure) et Extra-Trees de Geurts et al. (2006) (qui n'utilise pas de bootstrap) donnent des résultats très satisfaisants en pratique alors bien que très différentes dans le choix de l'aléa supplémentaire Θ

Erreur OOB et Estimation de l'erreur de prédiction

Erreur OOB, Out Of Bag (≈ "En dehors du Bootstrap")

Pour prédire Y_i, on agrège uniquement les prédicteurs $\widehat{h}(.,\Theta_\ell)$ construits sur des échantillons bootstrap ne contenant pas (X_i, Y_i)

- Erreur OOB $= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ en régression
- Erreur OOB $= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq \widehat{Y}_i}$ en classification

- Estimation semblable aux estimateurs classiques de l'erreur de généralisation (par échantillon test ou par validation croisée)
- Pas de découpage de l'échantillon d'apprentissage, inclus dans la génération des échantillons bootstrap
- Mais attention : c'est bien une sous-forêt différente (en général) qui est utilisée pour calculer chaque $\widehat{Y}_i$

Importance des variables

- Au delà des performances et du caractère automatique des RF, l'un des aspects les plus importants sur le plan appliqué est la quantification de l'importance des variables
- Azen et Budescu (2003) : discussion générale sur cette notion
- Notion relativement peu examinée par les statisticiens et principalement dans le cadre des modèles linéaires, Grömping (2015) ou la récente thèse de Wallard (2015)

- Les RF offrent un cadre idéal alliant
  - une méthode non-paramétrique, ne prescrivant pas de forme particulière à la relation entre Y et les composantes de X
  - le rééchantillonnage bootstrap

  pour disposer d'une définition à la fois efficace et commode de tels indices

Importance des variables

Breiman (2001), Strobl *et al.* (2007, 2008), Ishwaran (2007), Archer *et al.* (2008), Genuer et al. (2010), Gregorutti et al. (2013, 2015), Louppe et al. (2013)

Importance des variables

Soit $j \in \{1, \ldots, p\}$. Pour chaque échantillon OOB, on permute aléatoirement les valeurs de la j-ième variable des données

Importance de la j-ième variable = augmentation moyenne de l'erreur d'un arbre après permutation

Plus l'augmentation d'erreur est forte,
plus la variable est importante

Données *spam* : importance des variables

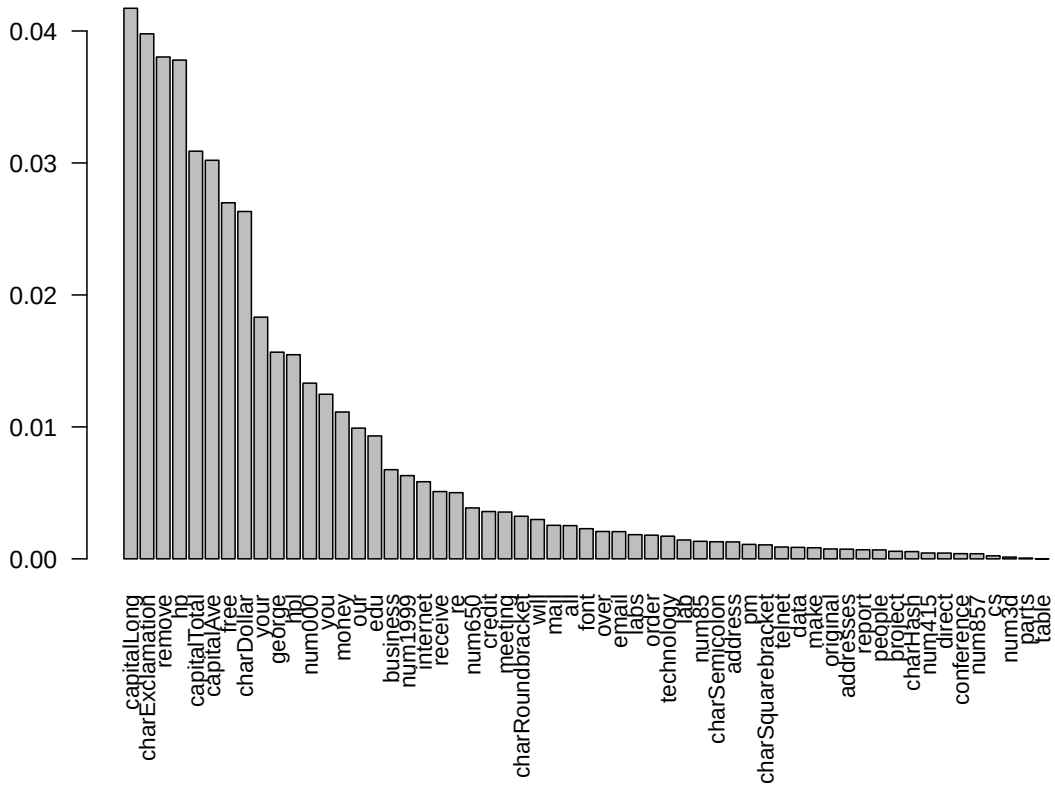


FIGURE : Les 8 plus importantes : les proportions d'occurrences des mots ou caractères *remove*, *hp*, *$*, *!*, *free* ainsi que les 3 variables liées aux longueurs des suites de lettres majuscules

Données *spam* : importance des variables

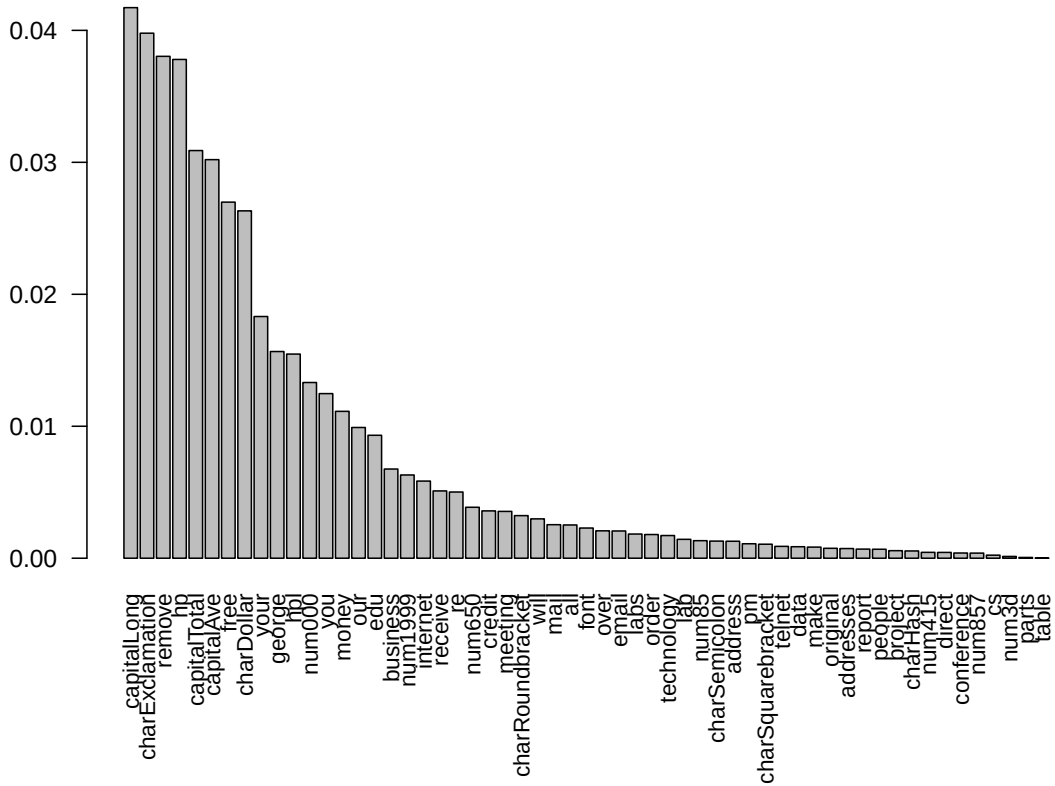


FIGURE : Les variables des 1ères découpes de l'arbre CART optimal ne sont pas en tête et la plus importante : *capitalLong* n'y figure pas

Procédure(s) de sélection

Genuer, Poggi, Tuleau (2010), PRL et (2015), R Journal

Deux objectifs différents de sélection de variables :

1. sélectionner toutes les variables importantes, même si elles sont redondantes, dans un but d'interprétation
2. trouver un ensemble parcimonieux de variables importantes suffisant pour la prédiction

Notre but est de proposer une procédure automatique qui vise ces deux objectifs

Citons simplement deux travaux antérieurs qui ont inspiré notre proposition :

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

Une application

Genuer, Michel, Eger, Thirion (2010)

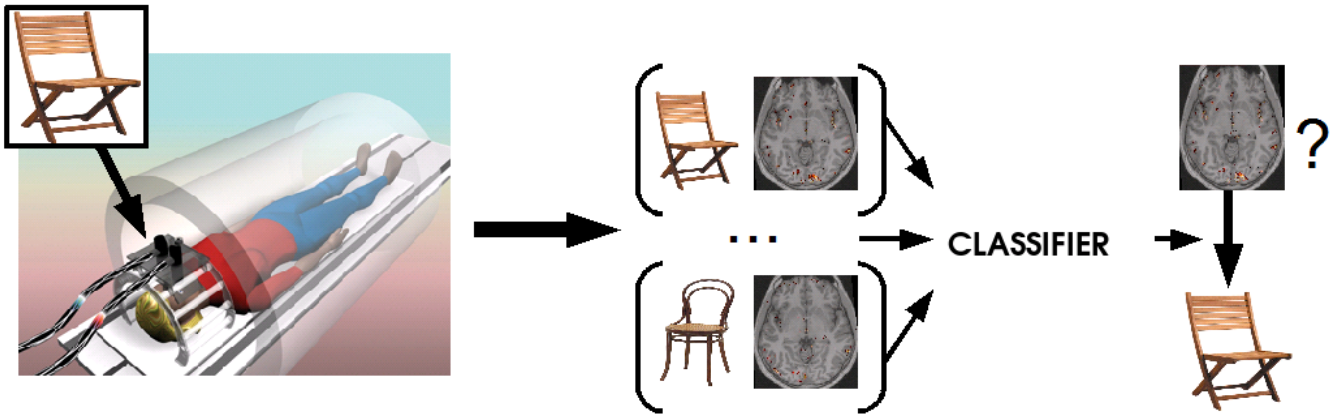


FIGURE : Expérience, IRMf

12 sujets : 4 types de chaises (4 classes), 100 000 voxels, 72 observations.
Etape préliminaire : réduction à 1000 parcelles (et donc 1000 variables) par un algorithme de Ward.

Classification $n = 72$ $p = 1000$ $L = 4$

Procédure de sélection pour un sujet

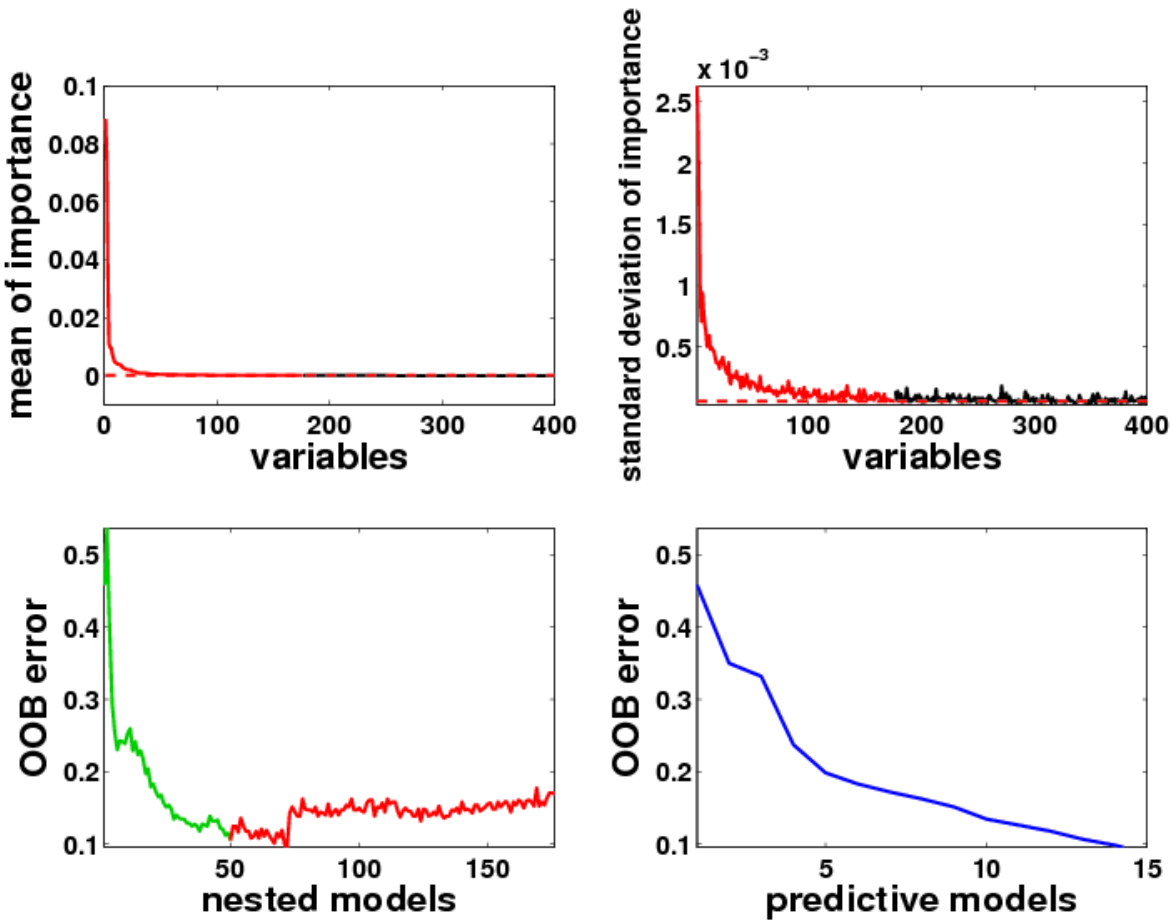


FIGURE : Procédure de sélection de variables pour un sujet ($ntree = 2000$, $mtry = p/3$)
Elimination : 176 variables, Interprétation : 50, Prédiction : 15

Les résultats : vue globale

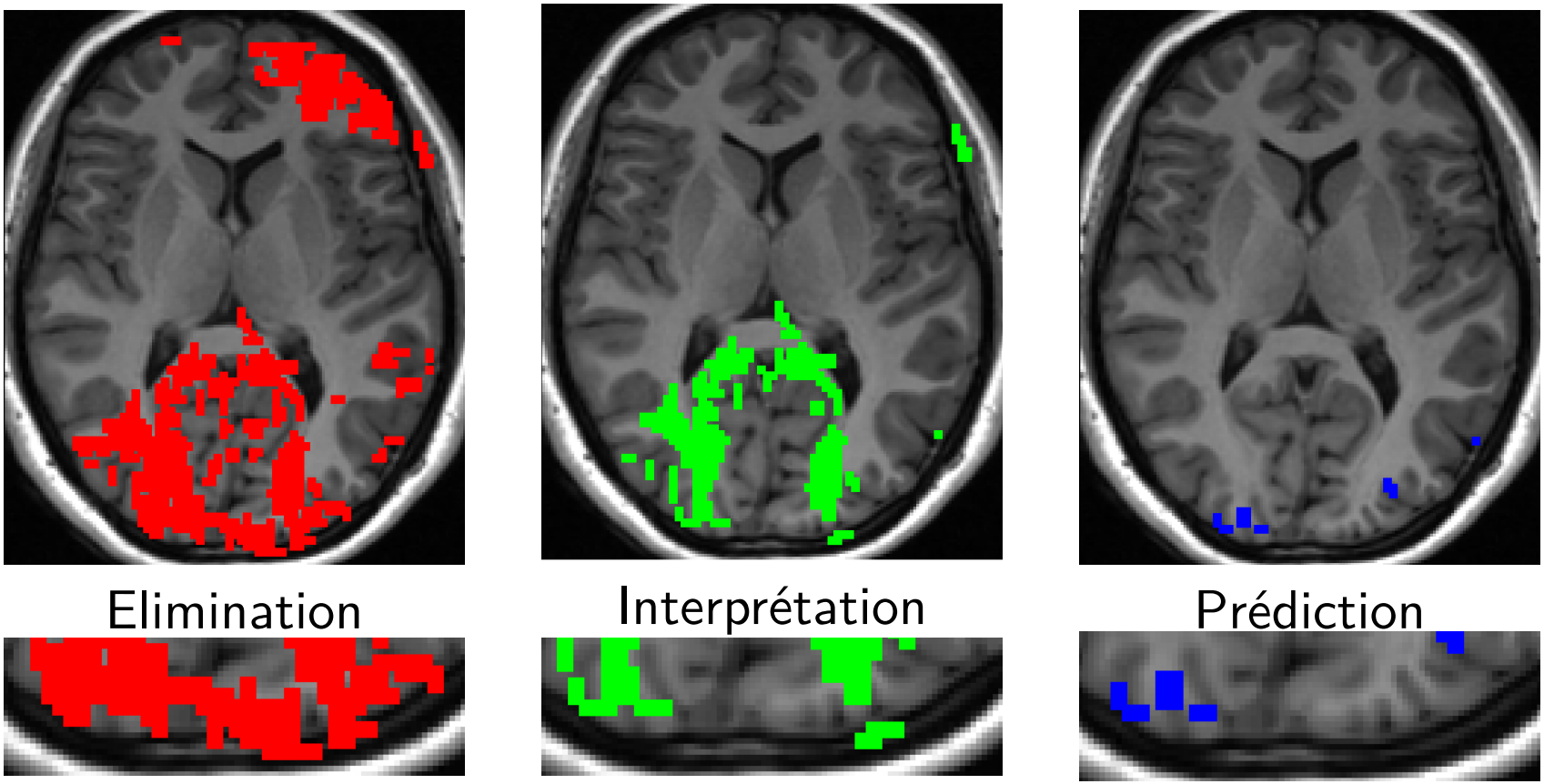


FIGURE : Variables sélectionnées aux différentes étapes de la procédure

Une comparaison

	Initiale	Elim.	Interp.	Préd.	Référence
Erreur	34 %	29 %	27 %	30 %	31 %
Nombre var.	1000	146	23	8	350

FIGURE : Résultats sur les 12 sujets de l'étude

- Méthode de référence : SVM linéaire (F-test + validation croisée)
- Taux d'erreurs comparables
- Beaucoup moins de variables

Pour en savoir plus :

- Une référence librement accessible, ainsi que les références incluses :
  Robin Genuer, Jean-Michel Poggi
  Arbres CART et Forêts aléatoires, Importance et sélection de variables
  45 pages, 2017
  https ://hal-descartes.archives-ouvertes.fr/hal-01387654v2
- Un chapitre de livre, à la suite des JES 2016, dans :
  Apprentissage Statistique et Données Massives,
  Maumy-Bertrand M., Saporta G. et Thomas Agnan C. (eds),
  Technip, 295-342, 2018
- Un livre à venir :
  R. Genuer, J.-M. Poggi, *Les forêts aléatoires avec R*, Editions PUR, coll. Pratique de la statistique, 120 pages, à paraître, 2019