Promenade en forêts aléatoires

Erwan Scornet

Les progrès informatiques des trois dernières décennies permettent désormais d' amasser une quantité considérable de données dans des domaines aussi divers que la finance, l'économie, la biologie ou le marketing. La création de base de données toujours plus massives continue de poser de nombreux problèmes. L'exploitation et le traitement des données est également un enjeu majeur : la collecte des données ne trouve son intérêt que dans l'information qui peut en être extraite. Afin de tirer parti de ces grands volumes de données, de nombreux algorithmes d'apprentissage statistique existent : c'est sur l'algorithme des forêts aléatoires que nous nous concentrerons dans la suite de cet article. Cependant, avant de se perdre dans les méandres des forêts aléatoires, arrêtons-nous tout d'abord devant l'un de ses éléments essentiels : l'arbre.

Comme leur nom l'indique, les arbres de décision sont des algorithmes représentant un processus simple de décision. Concrètement, imaginons qu'un médecin veuille établir un diagnostic. Pour ce faire, il va poser au malade une série de questions afin d'établir une liste des maladies potentielles dont il souffre. Dans un monde idéal, le médecin finit par déterminer la cause exacte du mal-être du patient. Cependant, pour des raisons budgétaires (le diagnostic peut nécessiter des techniques d'analyse de pointes), temporelles (le médecin ne peut obtenir l'attention complète du patient que pendant un court laps de temps) voire métaphysiques, le médecin peut être amené à hésiter entre plusieurs maladies. Face à cette cruelle incertitude, le médecin, sommé par le patient de statuer sur son état, peut alors utiliser son expérience et diagnostiquer la maladie qu'il aura le plus souvent observée parmi les patients présentant les mêmes symptômes. Cet exemple a le mérite d'expliquer simplement le fonctionnement des algorithmes d'arbre de décision : un arbre de décision crée différentes catégories de population (en posant des questions aux individus) et effectue une prédiction pour un nouvel individu en fonction de la catégorie à laquelle il appartient.

Les arbres de décision sont très utilisés en pratique car ils sont facilement interprétables : la prédiction fournie par l'algorithme est le résultat d'un cheminement logique facilement identifiable. Différents cas de figure peuvent se présenter : les variables caractérisant les individus peuvent être quantitatives discrètes ("Classez votre douleur au bras de 1 à 10"), quantitatives continues ("Combien mesure votre bras ?") ou catégorielles ("Avez-vous moins d'un bras ou plus de deux bras ?"). La variable à prédire peut, elle aussi, être discrète (le type de la maladie) ou continue (la probabilité de développer une maladie spécifique). Dans la suite, nous noterons $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)}$ les d variables décrivant les individus et Y la variable que l'on cherche à prédire. Pour simplifier, nous supposerons que toutes ces variables sont continues avec $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)} \in [0,1]$ et $Y \in \mathbb{R}$.

Afin de construire un arbre, on se donne un jeu de données $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ où $\mathbf{X}_i = (\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_1^{(d)})$ correspond aux variables caractérisant le $i$ème individu et Y_i est la valeur associée à cet individu. On suppose que les couples $(\mathbf{X}_i, Y_i) \in$

$[0,1]^d \times \mathbb{R}$ sont indépendants et identiquement distribués, de même loi que le couple $(\mathbf{X}, Y)$. On fait également l'hypothèse que le lien entre les variables d'entrée $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)})$ et la variable à estimer Y est donnée par la fonction de régression m qui vérifie :

$$Y = m(\mathbf{X}) + \varepsilon,$$

où ϵ est un bruit centré. Notre but est alors de construire à partir des données $\mathcal{D}_n$ un estimateur $m_n : [0,1]^d \to \mathbb{R}$ de la fonction de régression m, permettant d'estimer, pour un nouveau point $\mathbf{x} \in [0,1]^d$ la valeur $m(\mathbf{x})$.

Afin de construire un tel estimateur, les algorithmes d'arbre de décision partitionnent de manière récursive l'espace des variables d'entrée $[0,1]^d$. Un exemple d'arbre est décrit en Figure 1 pour deux variables d'entrée ($d = 2$). La première question de cet arbre est : "la variable $\mathbf{X}^{(1)}$ est-elle inférieure ou égale à 0.45 ?". Cette question est posée à chacune des données de l'ensemble $\mathcal{D}_n$ (une donnée correspondant à un point du carré $[0,1]^2$ avec la valeur qui lui est associée) et a pour effet de séparer le carré $[0,1]^2$ en deux cellules distinctes : la cellule $\boxed{1} \cup \boxed{2}$ à gauche et la cellule $\boxed{3} \cup \boxed{4}$ à droite, la séparation s'opérant à l'emplacement $\mathbf{X}^{(1)} = 0.45$. L'algorithme coupe ensuite chacune des deux cellules résultantes et poursuit jusqu'à ce qu'un critère d'arrêt soit vérifié, ce qui conduit dans notre cas à quatre nœuds terminaux $\boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}$. Finalement, pour estimer la valeur associée à un nouveau point $\mathbf{x} \in [0,1]^2$, il suffit de le propager dans l'arbre pour identifier la cellule à laquelle il appartient. La prédiction est alors donnée par la valeur moyenne des points tombant dans cette cellule. Par exemple, le point $\mathbf{x} = (0.6, 0.9)$ appartient à la cellule $\boxed{3}$; la valeur estimée $m_n(\mathbf{x})$ vaut donc 85.

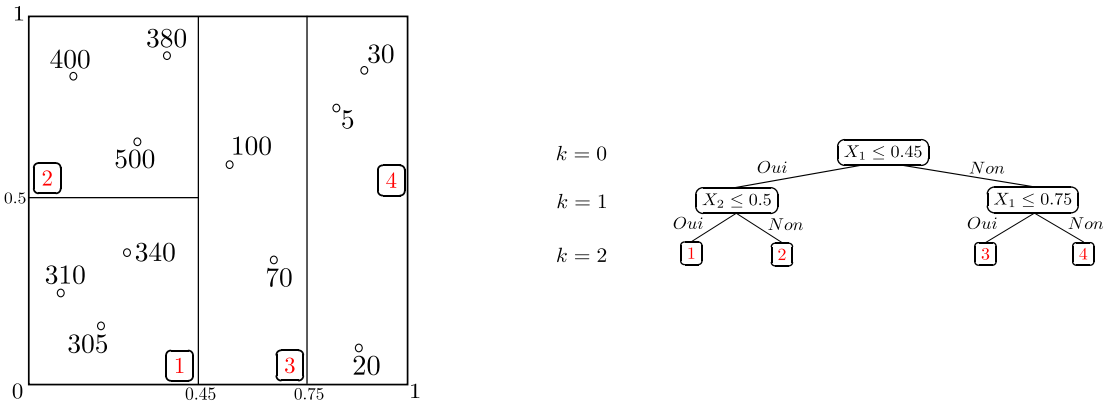

FIGURE 1 – Un arbre de décision de profondeur $k = 2$ en dimension $d = 2$.

Tout algorithme d'arbre de décision est défini par un *critère de coupure* et un *critère d'arrêt*. L'algorithme CART (Classification And Regression Tree) représenté dans la Figure 1 est un des algorithmes d'arbre de décision les plus employés et constituera l'une des pierres angulaires de cet article.

Le *critère de coupure CART* est défini comme suit.

1. Dans chaque cellule, pour chaque coupure possible, on construit l'estimateur constant par morceaux égal en tout point à la moyenne des observations situées dans la cellule contenant ce point.
2. L'erreur quadratique empirique de cet estimateur est ensuite calculée.

3. Le critère de coupure CART consiste alors à choisir la coupure qui correspond à l'estimateur dont l'erreur quadratique est minimale.

En effet, dans la Figure 1, on remarque que la première coupure en $\mathbf{X}^{(1)} = 0.45$ sépare les grandes valeurs (à gauche) des petites valeurs (à droite). Ce critère de coupure crée donc des cellules contenant des valeurs similaires [voir Breiman et al., 1984, pour plus de détails].

Différents *critères d'arrêt CART* sont souvent envisagés : le critère d'arrêt peut porter sur la forme de l'arbre (la construction s'arrête lorsque l'arbre est de la profondeur voulue, ici $k = 2$) ou sur le nombre de points dans les nœuds terminaux (arrêt lorsque chaque nœud terminal contient moins de `nodesize` $= 3$ observations). Compte-tenu de l'arbre obtenu en Figure 1, l'un ou l'autre de ces critères a pu être employé. Une procédure d'élagage peut également être utilisée : l'arbre est développé jusqu'à ce que chaque cellule contienne exactement une donnée, puis les cellules adjacentes sont progressivement fusionnées (i.e., les branches les plus profondes de l'arbre sont supprimées, d'où le nom d'élagage) jusqu'à ce qu'un compromis entre le nombre de nœuds terminaux et la capacité prédictive de l'arbre soit trouvé. Un arbre possédant un faible nombre de coupures (et donc de nœuds terminaux) sera facile à interpréter mais ses performances statistiques en seront amoindries.

1 Vers les forêts aléatoires

Dans la section précédente, nous avons détaillé la construction des arbres CART. Pour construire une forêt aléatoire, il va maintenant falloir construire plusieurs arbres CART distincts. C'est dans cette optique que Breiman [2001] proposa d'introduire de l'aléatoire à la fois dans le jeu de données (étape de sous-échantillonnage) et dans la construction des arbres :

1. Tout d'abord, pour chaque arbre, un sous-échantillon de taille a_n est créé en sélectionnant avec remise a_n observations parmi les n observations initiales. Seules ces a_n observations seront utilisées pour construire l'arbre et effectuer la prédiction.
2. À chaque étape, au lieu de sélectionner la coupure minimisant le critère CART (i.e., l'erreur quadratique de l'estimateur associé) selon les d variables, un petit nombre de variables m_{try} est sélectionné uniformément parmi les d variables disponibles. Le critère CART est alors minimisé seulement sur les m_{try} variables précédemment sélectionnées : la coupure ne peut donc s'effectuer que sur l'une de ces m_{try} directions.

L'algorithme complet est décrit en détail ci-après.

D'un point de vue mathématique, l'aléatoire introduit dans la construction d'un arbre générique sera noté Θ. L'aléatoire utilisé pour construire le $j$ème arbre sera ainsi noté Θ_j et correspondra à la fois aux observations choisies pour construire l'arbre ainsi qu'aux choix des variables pré-sélectionnées dans chaque cellule. Les variables $\Theta_1, \ldots, \Theta_M$ seront supposées indépendantes et identiquement distribuées selon la loi de Θ. On notera $m_n(\mathbf{x}, \Theta)$ l'estimateur associé à un arbre construit à partir de l'échantillon $\mathcal{D}_n$ et avec l'aléatoire Θ, évalué en un point $\mathbf{x}$. Les arbres sont ensuite combinés pour

Algorithme 1 : Construction des forêts de Breiman.

1. Construire M arbres comme suit :
   (a) Pour le $j$ème arbre, tirer uniformément avec remise a_n observations parmi $\mathcal{D}_n$. Seulement ces observations seront utilisées pour construire le $j$ème arbre.
   (b) Considérer la cellule $[0,1]^d$.
   (c) Sélectionner uniformément sans remise m_{try} coordonnées parmi $\{1, \ldots, d\}$.
   (d) Sélectionner la coupure qui minimise le critère de coupure CART par -mi les m_{try} directions pré-sélectionnées.
   (e) Découper la cellule selon la coupure précédente.
   (f) Répéter $(c) - (e)$ pour chacune des deux cellules engendrées jusqu'à ce que chaque cellule de l'arbre contienne moins de `nodesize` données.
   (g) Pour un point $\mathbf{x}$, la prédiction du $j$ème arbre est donnée par la moyen -ne des Y_i tombant dans la cellule contenant $\mathbf{x}$.
2. Pour un point $\mathbf{x}$, la prédiction de la forêt de Breiman est donnée par la moyenne des prédictions de chacun des M arbres au point $\mathbf{x}$.

former l'estimateur de la forêt aléatoire finie

$$m_{M,n}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M} m_n(\mathbf{x}, \Theta_j). \tag{1}$$

D'après la loi des grands nombres, pour tout $\mathbf{x} \in [0,1]^d$, presque sûrement, l'estimateur de la forêt finie tend vers celui de la forêt infinie

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}, \Theta) \right], \tag{2}$$

où $\mathbb{E}_\Theta$ représente l'espérance sous Θ, conditionnellement à $\mathcal{D}_n$.

2 Littérature

Les forêts aléatoires, créées par Breiman [2001] font partie des algorithmes d'apprentissage qui restent efficaces, tant d'un point de vue computationnel que prédictif, lorsqu'ils sont appliqués à des grands jeux de données. Leur construction repose sur les travaux fondateurs de Amit and Geman [1997], Ho [1998] et Dietterich [2000] et s'appuie sur le principe de *diviser pour régner* : la forêt est composée de plusieurs arbres qui sont chacun construits avec une partie du jeu de données. La prédiction de la forêt est alors obtenue simplement en moyennant les prédictions des arbres.

Le fait que les forêts puissent être employées pour résoudre un grand nombre de problèmes d'apprentissage a fortement contribué à leur popularité. Outre leur simplicité d'utilisation (voir l'implémentation du package `R`, `randomForest`), les forêts sont également connues pour leur précision et leur capacité à traiter des jeux de données composés de peu d'observations et de nombreuses variables. Étant par ailleurs facilement parallélisables, elles font partie des méthodes permettant de traiter de grands systèmes de données réelles.

Les bons résultats des forêts dans divers domaines appliqués sont légion : dans l'environnement [voir `http://www.kaggle.com/c/dsg-hackathon` et Prasad et al., 2006, Cutler et al., 2007], en chimio-informatique [Svetnik et al., 2003], dans l'identification d'objets tridimensionnels [Shotton et al., 2011], ou encore en bioinformatique [Díaz-Uriarte and de Andrés, 2006]. J. Howard (Kaggle) et M. Bowles (Biomatica) vont même jusqu'à affirmer dans Howard and Bowles [2012] que *"ensembles of decision trees—often known as "random forests"—have been the most successful general-purpose algorithm in modern times"*.

Néanmoins, le florilège de résultats appliqués contraste avec le peu de résultats théoriques sur les forêts : bien qu'utilisées dans toute une variété de domaines, certaines de leurs propriétés mathématiques demeurent mal comprises. Parmi les résultats théoriques les plus célèbres figure celui de Breiman [2001] qui consiste en une borne supérieure sur le risque quadratique des forêts, montrant que le risque des forêts est minimal lorsque les arbres sont de bons estimateurs faiblement corrélés. Lin and Jeon [2006] ont ensuite établi un lien entre les forêts et les estimateurs du type plus proche voisin, étudié plus en détail par Biau and Devroye [2010]. Plusieurs articles théoriques [e.g., Biau et al., 2008, Ishwaran and Kogalur, 2010, Biau, 2012, Genuer, 2012, Zhu et al., 2012, Arlot and Genuer, 2014] ont également porté sur des versions simplifiées des forêts de Breiman. Plus récemment, certains auteurs se sont concentrés sur des forêts très proches de l'algorithme originel de Breiman. Denil et al. [2013] ont prouvé le premier résultat de consistance pour les forêts en ligne. Mentch and Hooker [2014] et Wager [2014] ont étudié la distribution limite des forêts aléatoires, lorsque le nombre d'observations et le nombre d'arbres tendent vers l'infini, permettant ainsi d'établir des intervalles de confiance.

L'algorithme des forêts aléatoires est souvent considéré comme une véritable *boîte noire* qui combine de manière complexe plusieurs mécanismes difficiles à appréhender, comme le sous-échantillonnage du jeu de données et le critère de coupure des arbres CART [Classification And Regression Trees Breiman et al., 1984]. Cela explique pourquoi la majorité des travaux théoriques ont eu pour principaux objets des versions simplifiées de l'algorithme original, supprimant notamment l'étape de sous-échantillonnage des données et/ou remplaçant le critère de coupure CART et le critère d'arrêt (arrêt de l'algorithme lorsque chacun des nœuds terminaux contient un faible nombre d'observations) par des procédures plus élémentaires.

Le reste de l'article décrit les contributions de ma thèse. Le lecteur avide de précisions pourra se référer aux articles associés pour de plus amples détails ainsi qu'à Biau and Scornet [2016] pour une revue des propriétés théoriques des forêts aléatoires. La plupart des travaux théoriques se concentrant sur les propriétés des forêts infinies (2), les résultats de la Section 3 établissent un lien entre les forêts finies (utilisées en pratique) et les forêts infinies. En particulier, nous calculons le nombre d'arbres nécessaires pour que les risques des forêts finies et infinies soient proches [Première partie de Scornet, 2016a].

La Section 4 a pour objet l'étude de l'expression explicite de l'estimateur des forêts aléatoires infinies. En modifiant légèrement l'algorithme des forêts, on montre que l'estimateur résultant est un estimateur à noyau facilement interprétable. Ce résultat permet de faire le lien avec ces estimateurs largement étudiés et permet ainsi de s'affranchir de l'appellation "boîte noire" des forêts aléatoires [Scornet, 2016b].

La Section 5 [Deuxième partie de Scornet, 2016a] met en lumière l'intérêt de la procédure d'agrégation des arbres. Nous montrons ainsi qu'une forêt peut avoir un

bon pouvoir prédictif même lorsque les arbres agrégés sont individuellement mauvais.

La Section 6 [Scornet et al., 2015] contient, à mon avis, le résultat le plus intéressant de cette thèse qui est la preuve de consistance des forêts aléatoires de Breiman, c'est-à-dire le fait que les prédictions soient exactes lorsque le nombre d'observations tend vers l'infini. Ce résultat, simple en apparence et fondamental pour tout algorithme d'apprentissage, nécessite d'étudier de manière approfondie le critère de coupure CART pour en déduire des propriétés géométriques des cellules de chacun des arbres.

3 Forêt finie et infinie

Avant de nous intéresser aux performances prédictives des forêts aléatoires, il convient d'étudier le lien entre les forêts finies (utilisées en pratique, voir équation 1) et les forêts infinies (analysées en théorie, voir équation 2). Pour $\mathbf{x} \in [0,1]^d$ fixé, la Loi des Grands Nombres montre que l'estimateur des forêts finies évalué en $\mathbf{x}$, $m_{M,n}(\mathbf{x})$ converge presque sûrement vers l'estimateur des forêts infinies évalué en $\mathbf{x}$, $m_{\infty,n}(\mathbf{x})$, conditionnellement au jeu de données. De plus, le Théorème Central Limite fournit la vitesse de convergence en $\sqrt{M}$. Nous avons prouvé (voir Théorème 1) que ces résultats de convergence restent vrais pour les estimateurs des forêts finies et infinies, vus en tant que fonction : on étend ainsi la convergence ponctuelle précédente (à $\mathbf{x}$ fixé) à une convergence fonctionnelle.

Théorème 1. *Soit $m_{M,n}$ (resp., $m_{\infty,n}$) l'estimateur des forêts finies (resp., infinies) de Breiman. Alors, conditionnellement à $\mathcal{D}_n$,*

$$\sqrt{M}\left(m_{M,n}(\bullet) - m_{\infty,n}(\bullet)\right)$$

converge en loi vers un processus Gaussien.

Le Théorème 1 montre la convergence de l'estimateur $m_{M,n}$, vu en tant que processus. Ce résultat n'est pas une simple extension du Théorème Central Limite (l'espace des variables d'entrées $[0,1]^p$ n'étant pas dénombrable) et nécessite l'utilisation de la théorie des processus empiriques [voir van der Vaart and Wellner, 1996].

Lorsque M croît, l'estimation donnée par les forêts finies s'approche de celle des forêts infinies. On pourrait alors utiliser le Théorème 1 pour comparer les risques des forêts finies et infinies. Cependant, une analyse plus simple ne requérant pas la théorie des processus conduit au Théorème 2.

Théorème 2. *Supposons que $Y = m(\mathbf{X}) + \varepsilon$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ est indépendant de $\mathbf{X}$, et où $\|m\|_\infty < \infty$. Alors, pour tout $M, n \in \mathbb{N}^\star$,*

$$0 \leq R(m_{M,n}) - R(m_{\infty,n}) \leq \frac{8}{M} \times \Big(\|m\|_\infty^2 + \sigma^2(1 + 4\log n) \Big),$$

où, pour tout estimateur m_n, son risque quadratique $R(m_n)$ est défini par

$$R(m_n) = \mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2.$$

Le Théorème 2 établit que le risque quadratique des forêts finies est arbitrairement proche de celui des forêts infinies lorsque le nombre d'arbres est grand.

Lorsque le risque d'un estimateur tend vers zero quand la taille du jeu de données tend vers l'infini, on dit que l'estimateur est consistant. La consistance d'un estimateur

est une propriété statistique élémentaire : elle permet d'assurer qu'asymtotiquement la valeur prédite par l'estimateur est proche de la valeur que l'on cherche à estimer. En d'autres termes, un estimateur consistant est un estimateur qui remplit sa fonction. Le Théorème 2 montre que, si les forêts infinies sont consistantes (i.e., $R(m_{\infty,n})$ tend vers 0 lorsque n tend vers l'infini) alors les forêts finies sont consistantes (i.e., $R(m_{M_n,n})$ tend vers 0 lorsque n tend vers l'infini) en choisissant le nombre d'arbres M_n de sorte que $M_n/\log n$ tende vers l'infini. Sous cette hypothèse, les résultats de consistance obtenus pour de nombreux modèles de forêts infinies peuvent s'étendre aux forêts finies correspondantes, permettant ainsi d'appliquer les résultats théoriques aux forêts utilisées en pratique. Par conséquent, lorsqu'on s'intéresse aux propriétés de consistance des forêts aléatoires, il suffit d'étudier le comportement des forêts infinies.

4 Forêts aléatoires et estimateurs à noyau

L'estimateur associé à un arbre de régression construit avec l'aléatoire Θ fournit une prédiction $m_n(\mathbf{x}, \Theta)$ en un point $\mathbf{x}$ en calculant la moyenne des observations appartenant à la cellule contenant $\mathbf{x}$, notée $A_n(\mathbf{x}, \Theta)$. Mathématiquement, cet estimateur s'écrit

$$m_n(\mathbf{x}, \Theta) = \frac{1}{N_n(\mathbf{x}, \Theta)} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)},$$

où $N_n(\mathbf{x}, \Theta)$ est le nombre d'observations dans la cellule contenant $\mathbf{x}$. Par convention, on pose également $0/0 = 0$ dans la formule précédente (l'estimateur prédit 0 lorsque la cellule ne contient aucune observation). Les arbres de régression effectuent ainsi une moyenne des observations se situant dans un voisinage de $\mathbf{x}$, ce voisinage étant défini comme la cellule de l'arbre contenant $\mathbf{x}$. La forêt, qui agrège plusieurs arbres, opère également en calculant une moyenne pondérée des observations dans un voisinage de $\mathbf{x}$. Cependant, dans le cas des forêts, ce voisinage résulte de la superposition des voisinages de chacun des arbres, et a donc une forme plus complexe. La complexité de ce voisinage est probablement l'une des raisons qui explique les bonnes performances des forêts aléatoires.

Les estimateurs à noyau [Nadaraya, 1964, Watson, 1964] sont des estimateurs classiques basés sur une fonction mesurant la distance entre les observations et le point $\mathbf{x}$. Tout estimateur à noyau prédit donc une valeur en $\mathbf{x}$ en calculant une moyenne pondérée des Y_i en fonction de la distance entre $\mathbf{x}$ et $\mathbf{X}_i$ grâce à la fonction K :

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K(\mathbf{X}_i, \mathbf{x})}{\sum_{j=1}^{n} K(\mathbf{X}_j, \mathbf{x})}. \tag{3}$$

Afin de trouver une forme explicite facilement interprétable de l'estimateur des forêts aléatoires, nous avons établi un lien entre les forêts et les méthodes à noyau. En effet, en modifiant légèrement la procédure d'agrégation, l'estimateur des forêts peut se réécrire comme un estimateur à noyau de la forme

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K(\mathbf{X}_i, \mathbf{x})}{\sum_{j=1}^{n} K(\mathbf{X}_j, \mathbf{x})}, \tag{4}$$

où $K(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta \left[\mathbf{z} \in A_n(\mathbf{x}, \Theta)\right]$ est la probabilité de connexion entre $\mathbf{x}$ et $\mathbf{z}$ dans la forêt, c'est-à-dire la probabilité que $\mathbf{x}$ et $\mathbf{z}$ soient dans la même cellule d'un arbre

aléatoire de la forêt. Le noyau K correspond donc à une mesure de proximité particulière, intrinsèque à la forêt considérée. Ce résultat est particulièrement intéressant car il est non asymptotique et permet de voir les forêts comme des estimateurs à noyau dont le noyau est directement lié à la structure des arbres qui composent la forêt. De plus, les estimateurs de la forme (4) ont des performances similaires (que ce soit en précision ou en temps de calcul) à celles des forêts de Breiman, tout en étant plus facilement interprétables.

5 Agrégation d'arbres et arbre unique

Avant d'aborder la question difficile de la consistance des forêts aléatoires de Breiman, nous avons établi un théorème général portant sur les forêts dont la construction est indépendante des données et qui sont ainsi plus simples à analyser.

On appelle forêt aléatoire de niveau k, toute forêt dont chaque nœud terminal de chaque arbre a été obtenu en effectuant k coupures (les arbres associés sont alors binaires, complets, de niveau k). On rappelle également que $A_n(\mathbf{X}, \Theta)$ est la cellule de l'arbre aléatoire construit avec le paramètre aléatoire Θ et contenant $\mathbf{X}$.

Théorème 3. *Considérons une forêt aléatoire de niveau k dont le mécanisme de coupure ne dépend pas des données $\mathcal{D}_n$. Supposons de plus que le diamètre des cellules tend vers zero en probabilité, i.e., pour tout $\varepsilon > 0$,*

$$\mathbb{P}\Big[\mathit{diam}(A_n(\mathbf{X}, \Theta)) > \varepsilon\Big] \to 0, \quad \textit{lorsque } n \to \infty.$$

Alors, si $k \to \infty$ et $2^k/n \to 0$, l'estimateur $m_{\infty,n}$ de la forêt aléatoire infinie précédente est consistant, i.e.,

$$\lim_{n\to\infty} \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2 = 0.$$

Ce théorème montre ainsi que les forêts construites indépendamment des données sont consistantes dès lors que le diamètre de leurs cellules tend vers 0 et que la profondeur des arbres est correctement choisie. Cependant, la consistance de ces forêts résulte de la consistance des arbres individuels qui les composent : au regard du Théorème 3, la procédure d'agrégation des arbres n'a donc aucun intérêt.

Afin de mettre en exergue certaines propriétés propres aux forêts aléatoires, nous considérons les forêts médianes dont les arbres sont construits de la manière suivante :

1. On tire sans remise a_n points parmi les n observations initiales.
2. À chaque nœud, on sélectionne uniformément une direction de coupure. On coupe ensuite à la médiane empirique des $\mathbf{X}_i$ selon la direction précédemment choisie.
3. La procédure de coupure s'arrête lorsque chaque cellule contient exactement une observation.

Les forêts médianes sont construites grâce aux données $\mathbf{X}_i$. Elles offrent donc un bon compromis entre les forêts analysées dans le Théorème 3 dont la construction est indépendante des données et les forêts de Breiman, dont la construction dépend à la fois des positions $\mathbf{X}_i$ et des valeurs observées Y_i. Elles sont également proches des forêts de Breiman en ce sens que chaque cellule ne contient qu'un petit nombre d'observations. Pour ces forêts, le Théorème 4 montre que le sous-échantillonnage est crucial pour

assurer la consistance des forêts médianes. En effet, les arbres de la forêt médiane ne sont pas consistants (car leurs nœuds terminaux ne contiennent qu'un seul point) mais la forêt médiane l'est grâce au sous-échantillonnage.

Théorème 4. *Supposons que* $Y = m(\mathbf{X}) + \varepsilon$*, où* ε *est un bruit centré vérifiant* $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2$*, avec* $\sigma^2 < \infty$ *une constante. Supposons de plus que* $\mathbf{X}$ *admette une densité sur* $[0,1]^d$ *et que* m *est continue. Alors, si* $a_n \to +\infty$ *et* $a_n/n \to 0$*, la forêt infinie médiane est consistante.*

De précédents résultats [Genuer, 2012] montrent que la variance de certaines forêts aléatoires est réduite d'un facteur 3/4 par rapport à la variance des arbres individuels. Le Théorème 4 est le premier théorème portant sur des forêts complètement développées (contenant un seul point par cellule) et prouvant l'intérêt asymptotique de la procédure d'agrégation : alors que la variance de chaque arbre est constante lorsque $n \to +\infty$, la variance de la forêt tend vers 0 grâce au sous-échantillonnage. La procédure d'agrégation (et de sous-échantillonnage) permet de rendre consistant un ensemble d'estimateurs (arbres) non consistants.

Afin d'illustrer ce phénomène, les performances prédictives des forêts de Breiman et d'un arbre CART (qui compose ces forêts) ont été comparées dans la Figure 2 pour deux modèles de régression choisis arbitrairement :

- Modèle 1 ($d = 100$) : $Y = 10(X^{(1)})^2 + 2X^{(2)} + \mathcal{N}(0, 0.05^2)$;
- Modèle 2 ($d = 100$) : $Y = -X^{(3)} \times \sin(2X^{(1)}) + \exp(-X^{(4)}) \times (X^{(2)})^2 + \mathcal{N}(0, 0.05^2)$.

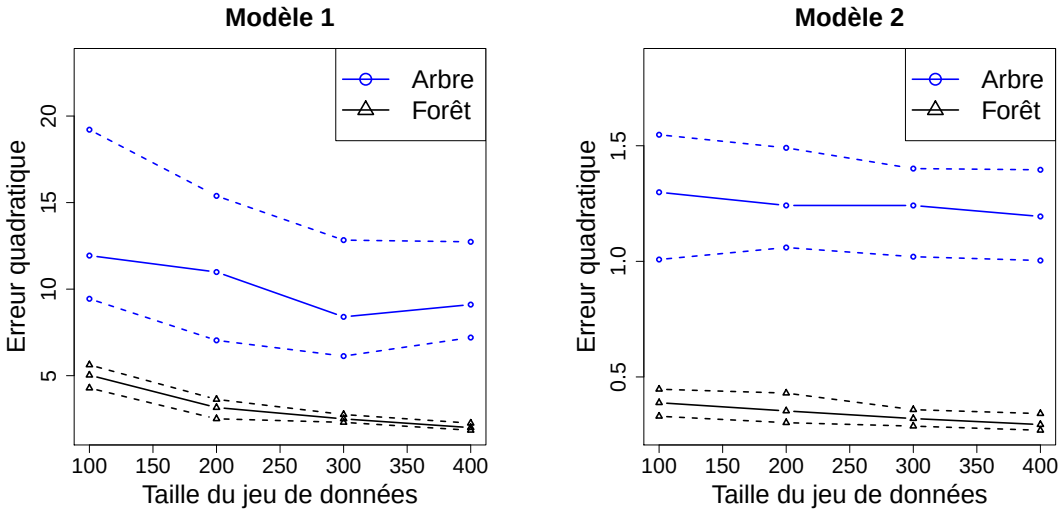

FIGURE 2 – Performances d'une forêt de Breiman (en noir) et d'un arbre CART seul (en bleu). Pour chaque estimateur (forêt ou arbre), l'expérience a été répétée 50 fois. Le trait plein correspond à la médiane des performances et les traits pointillés correspondent au premier et troisième quartile des performances.

Le risque des forêts est très largement inférieur à celui d'un arbre seul. De plus, la variabilité des performances entre différentes expériences est plus faible pour les forêts que pour les arbres : les prédictions des forêts sont plus stables que celles des arbres seuls, ce qui est bien ce pour quoi elles ont été créées.

6 Consistance

Comme mentionné précédemment, la consistance des forêts de Breiman est un problème difficile à résoudre car cette procédure comprend plusieurs mécanismes (sous-échantillonnage, critère de coupure CART) complexes qui nécessitent chacun une étude approfondie. Par conséquent, bien que la consistance soit une propriété fondamentale et bien souvent facile à démontrer pour nombre d'estimateurs, il n'en va pas de même pour les forêts aléatoires de Breiman.

Nous énonçons ici deux théorèmes sur la consistance de ces forêts dans le cadre d'un modèle de régression additif vérifiant les hypothèses suivantes :

(H6) *La réponse Y vérifie*

$$Y = \sum_{j=1}^{p} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

où $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ est uniformément distribué sur $[0,1]^p$, ε est un bruit Gaussien indépendant de variance $0 < \sigma^2 < +\infty$, et chaque composante m_j est continue.

Les modèles additifs (popularisés par Stone [1985] et Hastie and Tibshirani [1986]) sont une extension des modèles linéaires, où la fonction de régression m s'écrit comme une somme de fonctions univariées m_j, non nécessairement linéaires. Ils sont flexibles, faciles à interpréter et fournissent un bon compromis entre la complexité du modèle et le temps de calcul. Ils ont de ce fait été étudiés en détail durant les trente dernières années. Bien que les procédures de forêts aléatoires ne nécessitent pas d'hypothèses sur le modèle de régression pour être opérationnelles, leur analyse est grandement facilitée dans le cadre des modèles additifs.

Le Théorème 5 porte sur des forêts aléatoires contenant des arbres non complètement développés (dont les nœuds terminaux contiennent un grand nombre d'observations) et repose sur le fait que les arbres individuels sont consistants, quel que soit le taux de sous-échantillonnage utilisé pour construire la forêt. Plus précisément, si on suppose que chaque arbre de la forêt de Breiman est construit à partir de a_n observations (sous-échantillonnage) et contient au maximum t_n nœuds terminaux (arrêt précoce de la construction des arbres) alors le Théorème 5 assure que la forêt correspondante est consistante.

Théorème 5. *Supposons que* **(H6)** *est vérifiée. Si $a_n, t_n \to +\infty$, et si $t_n(\log a_n)^9/a_n \to 0$, alors l'estimateur $m_{\infty,n}$ des forêts de Breiman est consistant, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2 = 0.$$

Remarquons que le Théorème 5 est encore vrai lorsque $a_n = n$. Dans ce cas, l'étape de sous-échantillonnage ne joue aucun rôle dans la consistance de la procédure. En effet, contrôler le niveau des arbres grâce au paramètre t_n est suffisant pour borner l'erreur de la forêt. En réalité, le Théorème 5 montre également la consistance de chaque arbre CART qui compose la forêt, ce qui constitue le premier résultat de consistance pour ces arbres [Breiman et al., 1984].

Le Théorème 6 quant à lui concerne les forêts de Breiman complètement développées et suppose un taux de sous-échantillonnage bien choisi (comme pour les forêts médianes).

Théorème 6. *Supposons que* **(H6)** *est vérifiée. Sous une hypothèse supplémentaire, si* $a_n \to +\infty$ *et* $a_n \log n/n \to 0$*, alors l'estimateur* $m_{\infty,n}$ *des forêts de Breiman complètement développées (i.e.,* $t_n = a_n$*) est consistant, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2 = 0.$$

Les Théorèmes 5 et 6 sont les premiers résultats de consistance pour l'algorithme original de Breiman [2001] et utilisent respectivement des résultats de Stone [1977] et de Nobel [1996].

Ils reposent sur la Proposition 1 ci-dessous qui établit une importante caractéristique du mécanisme des forêts aléatoires. Elle montre que la variation de la fonction de régression m à l'intérieur d'une cellule d'un arbre aléatoire est petite, à condition que n soit assez grand. On définit au préalable pour toute cellule A, la variation de m à l'intérieur de A par

$$\Delta(m, A) = \sup_{\mathbf{x},\mathbf{x}'\in A} |m(\mathbf{x}) - m(\mathbf{x}')|.$$

Proposition 1. *Supposons que* **(H6)** *est vérifiée. Alors, pour tout* $\varepsilon > 0$*,*

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) > \varepsilon\right] \to 0, \quad \textit{lorsque } n \to +\infty.$$

Remarquons que lorsqu'on étudie des arbres construits indépendamment des Y_i, l'erreur d'approximation de ces estimateurs est contrôlée en s'assurant que le diamètre des cellules tend vers zéro en probabilité. En lieu et place de cette hypothèse géométrique, la Proposition 1 assure que la variation de m à l'intérieur d'une cellule est petite, forçant ainsi l'erreur d'approximation de la forêt à tendre vers zéro.

Tandis que la Proposition 1 offre un bon contrôle de l'erreur d'approximation de la forêt dans les deux régimes, une analyse séparée est requise pour l'erreur d'estimation. Dans le premier régime (Théorème 5), le paramètre t_n permet de contrôler la structure de l'arbre. Ceci est en accord avec les preuves de consistances classiques pour les arbres de régression [voir, par exemple, Devroye et al., 1996, Chapitre 20]. Les choses sont différentes pour le second régime (Théorème 6), dans lequel les arbres individuels sont complètement développés. Dans ce cas, l'erreur d'estimation est contrôlée en s'assurant que le taux de sous-échantillonnage a_n/n est choisi pour être un $o(1/\log n)$, ce qui est inhabituel et mérite en conséquence quelques explications.

Notons tout d'abord que le terme en $\log n$ dans le Théorème 6 est utilisé pour contrôler le bruit Gaussien ε. Donc, si le bruit est supposé borné, le terme en $\log n$ disparaît et la condition se réduit à $a_n/n \to 0$. La condition $a_n \log n/n \to 0$ garantie que chaque observation $(\mathbf{X}_i, Y_i)$ est utilisée dans la construction d'un arbre avec une probabilité qui devient petite lorsque n grandit. Cela implique également qu'aucun point $\mathbf{x}$ n'est connecté avec la même observation dans une grande proportion des arbres. Dans le cas contraire, la valeur prédite en $\mathbf{x}$ serait trop influencée par une seule observation $(\mathbf{X}_i, Y_i)$, rendant alors la forêt inconsistante. En réalité, une étude de la preuve du Théorème 6 révèle que l'erreur d'estimation de la forêt est petite dès que la probabilité de connexion maximale entre le point $\mathbf{x}$ et toutes les observations est petite. Par conséquent, l'hypothèse portant sur le taux de sous-échantillonnage est simplement un moyen pratique de contrôler ces probabilités, en s'assurant que les partitions sont assez diversifiées (i.e., que $\mathbf{x}$ est connecté à beaucoup d'observations au sein de la forêt). Cette notion de diversité parmi les arbres a été introduite par Breiman [2001], mais

est généralement difficile à analyser. Dans notre approche, le sous-échantillonnage est la composante clé permettant d'imposer la diversité des arbres.

Malheureusement, le Théorème 6 vient au prix de d'une hypothèse additionnelle non détaillée ici dont on ne sait pas si elle est vraie en toute généralité. Cependant, la forêt étudiée dans ce théorème correspond presque parfaitement à l'algorithme utilisé en pratique : c'est donc une étape importante dans la compréhension des forêts de Breiman. Contrairement à la plupart des précédents travaux, le Théorème 6 suppose qu'il n'y a qu'une seule observation dans chaque feuille de chaque arbre. Cela implique que les arbres individuels ne sont pas consistants, puisque les conditions classiques pour la consistance des arbres imposent que le nombre d'observations dans les nœuds terminaux tend vers l'infini lorsque n tend vers l'infini [voir, e.g., Devroye et al., 1996, Györfi et al., 2002]. Par conséquent, sous les hypothèses du Théorème 6, l'algorithme des forêts aléatoires agrège des arbres individuels non consistant en un estimateur consistant. Ce résultat est le pendant du Théorème 4 pour les forêts de Breiman.

Notre analyse permet également de mieux comprendre les bonnes performances des forêts aléatoires dans des contextes de parcimonie. À cet effet, considérons un modèle parcimonieux, i.e., il existe $S < d$ tel que

$$Y = \sum_{j=1}^{S} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

où S représente la véritable dimension du problème, S étant inconnu. Par conséquent, parmi les d variables, nous supposons que seules les S premières sont informatives. Dans ce contexte de réduction de la dimension, la dimension ambiante d du problème peut être très grande mais nous supposons que le signal est parcimonieux : l'information n'est portée que par un petit nombre S de variables. La Proposition 2 ci-dessous montre que les forêts aléatoires s'adaptent au contexte précédent car les coupures des forêts sont effectuées asymptotiquement et avec grande probabilité selon les S premières variables.

Dans cette proposition, nous fixons $m_{\text{try}} = d$ (i.e., on choisit la meilleure coupure selon les d directions possibles), et pour tout k, nous posons $j_{1,n}(\mathbf{X}), \ldots, j_{k,n}(\mathbf{X})$ les k premières directions de coupures utilisées pour construire la cellule contenant $\mathbf{X}$ avec la convention $j_{q,n}(\mathbf{X}) = \infty$ si la cellule a été coupée strictement moins de k fois.

Proposition 2. *Supposons que* **(H6)** *est vérifiée. Soit $k \in \mathbb{N}^\star$. Supposons que pour tout intervalle $[a, b]$ et tout $j \in \{1, \ldots, S\}$, m_j ne soit pas constante sur $[a, b]$. Alors, pour tout $1 \leq q \leq k$,*

$$\lim_{n\to\infty} \mathbb{P}\Big[j_{q,n}(\mathbf{X}) \in \{1, \ldots, S\}\Big] = 1.$$

Références

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9 :1545–1588, 1997.

S. Arlot and R. Genuer. Analysis of purely random forests bias. arXiv :1407.3939, 2014.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13 : 1063–1095, 2012.

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101 :2499–2518, 2010.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25 :197–227, 2016.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9 :2015–2033, 2008.

L. Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

D.R. Cutler, T.C. Edwards Jr, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler. Random forests for classification in ecology. *Ecology*, 88 :2783–2792, 2007.

M. Denil, D. Matheson, and N. de Freitas. *Consistency of online random forests*, 2013. arXiv :1302.4853.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7 :1–13, 2006.

T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees : bagging, boosting, and randomization. *Machine Learning*, 40 :139–157, 2000.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24 :543–562, 2012.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1 :297–310, 1986.

T. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence*, 20 :832–844, 1998.

J. Howard and M. Bowles. The two most important algorithms in predictive modeling today. In *Strata Conference : Santa Clara*. `http://strataconf.com/strata2012/public/schedule/detail/22658`, 2012.

H. Ishwaran and U.B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80 :1056–1064, 2010.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101 :578–590, 2006.

L. Mentch and G. Hooker. Ensemble trees and clts : Statistical inference for supervised learning. arXiv :1404.6473, 2014.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9 : 141–142, 1964.

A. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24 :1084–1105, 1996.

A.M. Prasad, L.R. Iverson, and A. Liaw. Newer classification and regression tree techniques : Bagging and random forests for ecological prediction. *Ecosystems*, 9 :181–199, 2006.

E. Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146 : 72–83, 2016a.

E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62 :1485–1500, 2016b.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43 :1716–1741, 2015.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5 :595–645, 1977.

C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985.

V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random forest : A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43 :1947–1958, 2003.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer, New York, 1996.

S. Wager. Asymptotic theory for random forests. arXiv :1405.0352, 2014.

G. S. Watson. Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

R. Zhu, D. Zeng, and M.R. Kosorok. *Reinforcement learning trees*. Technical Report, University of North Carolina, Chapel Hill, 2012.