SORBONNE UNIVERSITÉ

Les forêts aléatoires & leurs estimateurs à noyau

Laure Ferraris, Paul Liautaud

11 février 2022

"Si quelqu'un vous demande quelle est la probabilité que la Lune tombe sur la Terre, vous n'avez qu'à simplement lui répondre que vous êtes fréquentiste !"
Gérard Biau, 2021.

Table des matières

1 Introduction

Les forêts aléatoires introduites par Leo BREIMAN au début des années 2000 [7] sont une méthode de classification et de régression par apprentissage supervisé. L'approche repose sur un principe simple mais puissant, "diviser pour mieux régner" : faire plusieurs sous-échantillonnages des données, construire un arbre de décision pour chaque sous-ensemble selon un paramètre aléatoire, agréger les réponses pour obtenir la prédiction finale. Cette stratégie affiche d'excellents résultats dans divers domaines appliqués, pour en nommer quelques-uns : bio-informatique, économétrie ou encore reconnaissance d'objets 3D. La robustesse des forêts aléatoires dans des problèmes de très grande dimension associée à leur simplicité pratique (peu de paramètres sont à ajuster) en ont fait une méthode populaire. Ce succès contraste avec le peu de résultats théoriques présents dans la littérature. Les forêts demeurent une question mathématique ouverte [15], ce sont des objets complexes et de ce fait difficiles à analyser. Dans le cadre d'un problème de régression, notre présent objectif sera d'établir le lien entre les forêts aléatoires et des estimateurs à noyau obtenus après une légère modification de leur définition. Cette approche permet d'obtenir de nouveaux estimateurs nommés KeRF (Kernel Random Forest), qui sont proches des forêts aléatoires sous certaines hypothèses, tout en ouvrant des perspectives d'analyses mathématiques plus profondes.

Dans un premier temps nous introduirons les notations et les définitions qui nous seront utiles. Puis, nous montrerons que, les forêts aléatoires et les estimateurs KeRF peuvent être "proches". Nous étudierons plus particulièrement l'estimateur KeRF dans le modèle des forêts centrées. Nous démontrerons sa consistance et nous obtiendrons une borne supérieure caractérisant sa vitesse de convergence.

L'étude théorique sera enrichie par des exemples et appuyée par des observations empiriques obtenues à partir de simulations. L'ensemble des simulations sont issues d'un travail d'implémentation personnel dans le langage de programmation `Python`, dont on met à disposition une partie sur ce Notebook (cliquable).

2 Notations et définitions préliminaires

Dans tout ce cours, on considérera un jeu de données d'entraînement $\mathcal{D}_n := \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_n, Y_n)\}$ contenant n paires $(\mathbf{X}, Y) \in \mathcal{X} \times \mathbb{R}$, où $\mathcal{X} = [0;1]^d \subset \mathbb{R}^d$ est l'ensemble de définition des observations $(\mathbf{X}_i)$. Les données $\mathcal{D}_n$ seront alors considérées aléatoires, indépendantes et identiquement distribuées (dites i.i.d.). Par ailleurs, on supposera que la variable aléatoire (v.a.) $Y \in \mathbb{R}$ admet un moment d'ordre 2 (i.e. elle est de carré intégrable : $\mathbb{E}\left[Y^2\right] < \infty$).

On établit la convention suivante : on utilisera la désignation **RF** pour évoquer l'*estimateur forêt aléatoire* et **KeRF** pour l'*estimateur forêt aléatoire à noyau.*

But : On s'intéressera, dans le cadre de ce cours, à *estimer* la fonction de régression :

$$m : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R} \\ \mathbf{x} & \longmapsto & \mathbb{E}\left[Y|\mathbf{X} = \mathbf{x}\right] \end{array} \right. \tag{1}$$

Une des façons d'approcher cette fonction m est d'utiliser les **forêts aléatoires**. Une telle estimation de (1) sera alors notée m_∞ : $[0;1]^d \to \mathbb{R}$.
Commençons par introduire les notions qui nous seront essentielles pour la suite de ce cours.

2.1 De l'arbre décisionnel à la forêt aléatoire

Attardons-nous donc un instant devant une composante essentielle des forêts aléatoires : l'arbre. Cet algorithme suit un cheminement logique simple. Il pose une série de questions binaires afin de créer différentes catégories de populations et effectuer une prédiction sur un nouvel individu en fonction du groupe auquel il appartient. Les variables caractérisant les individus peuvent être quantitatives discrètes, quantitatives continues ou catégorielles. La variable à prédire peut être discrète (penser à un problème de classification) ou bien continue (dans le cadre d'un problème de régression).

Les arbres pouvant être vus comme des graphes, on rappelle ici dans Définition 2.1 les composantes de ceux-ci.

Définition 2.1 (Arbres & composantes d'un arbre, [5])

Un **arbre binaire** est un *graphe connexe acyclique* qui illustre un processus de **partitionnement récursif** de l'espace étudié $[0;1]^d$.
Il est composé :
- d'une **racine** : $\mathcal{X} = [0;1]^d$;
- de **nœuds**, ayant au plus **2 fils** ;
- des **feuilles**, qui sont des **nœuds terminaux**, i.e. à 0 enfant.

Pour un premier exemple d'arbre de régression, voir Figure 1.

Pour obtenir une forêt à partir d'une collection d'arbres, nous disposons d'un unique jeu de données $\mathcal{D}_n$. Afin de construire des arbres distincts, nous pouvons alors injecter de l'aléatoire à deux niveaux :

— dans le jeu de données, en construisant chaque arbre uniquement à partir d'un sous-échantillon de $\mathcal{D}_n$ tiré aléatoirement, avec ou sans remise parmi les n observations originelles ;
— dans la procédure de construction de l'arbre, en effectuant des "découpes" sur un nombre l de coordonnées tirées aléatoirement parmi $\{1, ..., d\}$ (où d est la dimension de l'espace d'appartenance des $(\mathbf{X}_i)$). Le critère de découpe, sur chaque coordonnée, peut lui aussi être aléatoire.

L'aléatoire paramétrant la construction est noté Θ. Si l'on collectionne M arbres, nous disposons alors de $\Theta_1, \cdots, \Theta_M$ variables aléatoires, supposées i.i.d.selon la loi de Θ et indépendantes de l'échantillon $\mathcal{D}_n$.
La variété des arbres ainsi que les différentes modalités d'aléa caractérisent les diverses forêts aléatoires que l'on peut obtenir.

Introduisons plus formellement l'arbre et la forêt :

Définition 2.2 (Arbre & forêt aléatoires - [7], [12], [13])

Soient $\mathbf{x} \in [0;1]^d$, $\mathcal{D}_n$ les données d'entraînement introduites précédemment et une v.a. Θ.
Un **arbre** est un *unique* prédicteur $\mathbf{x} \mapsto m(\mathbf{x}, \Theta)$, constant par morceaux et obtenu par partitionnement récursif dyadique, selon la v.a. Θ, de l'espace des données $\mathcal{X}$. En particulier, un **arbre de régression** est un prédicteur admettant la forme suivante :

$$m(\mathbf{x}, \Theta; \mathcal{D}_n) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x}, \Theta)\}}}{N(\mathbf{x}, \Theta)} \tag{2}$$

où $A(\mathbf{x}, \Theta)$ est la cellule contenant $\mathbf{x}$ sous l'aléa Θ, et $N(\mathbf{x}, \Theta) = \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x}, \Theta)\}}$ le nombre d'entités qui y sont contenues. Par convention si $N(\mathbf{x}, \Theta) = 0$, alors $m(\mathbf{x}, \Theta) = 0$.

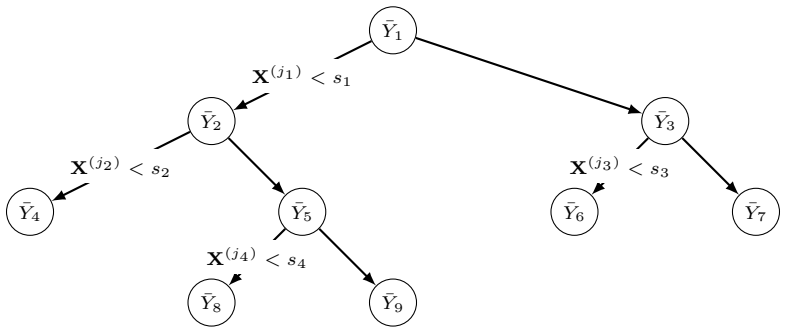
L'arbre décisionnel est un prédicteur de base. En construisant une collection de tels estimateurs simples, nous pouvons en considérer un plus complexe : la forêt aléatoire.

Soient M prédicteurs par arbre $(m(\cdot, \Theta_1), m(\cdot, \Theta_2), \ldots, m(\cdot, \Theta_M))$, où $\Theta_1, \ldots, \Theta_M \underset{\text{i.i.d.}}{\sim} \mathcal{L}(\Theta)$ et indépendantes de $\mathcal{D}_n$.
Une **forêt aléatoire** (RF) est une statistique, notée m_M, construite en agrégeant la collection d'**arbres** $\{m(\cdot, \Theta_j), 1 \leq j \leq M\}$:

$$m_M(\mathbf{x}, \Theta_1, \cdots, \Theta_M; \mathcal{D}_n) := \frac{1}{M} \sum_{j=1}^M m(\mathbf{x}, \Theta_j) \tag{3}$$

Remarque.
— Dans le cadre d'une *classification*, chaque arbre indique la classe la plus vraisemblable à laquelle appartient une donnée $\mathbf{x} \in [0;1]^d$. La forêt aléatoire associée à ces arbres opère alors en une agrégation de ces derniers par *vote majoritaire*. Pour la *régression*, l'agrégation des arbres est faite par *moyenne*.
— Par souci d'allègement de notation, on adoptera l'écriture $m(\mathbf{x}, \Theta) := m(\mathbf{x}, \Theta; \mathcal{D}_n)$ pour mentionner l'arbre estimateur et $m_M(\mathbf{x}) := m(\mathbf{x}, \Theta_1, \ldots, \Theta_M; \mathcal{D}_n)$ pour RF. Cependant, on gardera bien à l'esprit sa dépendance à $\mathcal{D}_n$, et aux v.a. $\Theta_1, \cdots, \Theta_M$ dans le cas de l'estimateur forêt (RF).

Exemple 2.1 (Arbre de régression). Soit $s_1, s_2, s_3, s_4 \in [0;1]$ des seuils de "découpe" associés à des coordonnées $j_1, j_2, j_3, j_4 \in [\![1; d]\!]$. On peut par exemple considérer l'arbre d'ordre $k = 3$ en Figure 1, qui réalise la moyenne empirique des (Y_i) associés à leur $(\mathbf{X}_i)$ correspondants et appartenant à chaque sous-cellule.



où, par exemple, $\bar{Y}_3$ est la moyenne empirique associée aux $\{Y_i, i \in C\}$, avec $C = \{i \in [\![1; n]\!] | \mathbf{X}_i^{(j_1)} \geq s_1\}$.

FIGURE 1 – Exemple d'un premier arbre de régression.

Remarque.
— Le terme de *forêt* fait bien sens : il vient du fait que l'on utilise plusieurs arbres prédicteurs sur $\mathcal{D}_n$. Chaque arbre dépendant d'une v.a. Θ, de là naît alors le caractère *aléatoire* ;

— La Loi Forte des Grands Nombres induit la convergence, pour $\mathbf{x} \in [0;1]^d$,

$$m_M(\mathbf{x}) \underset{M \to +\infty}{\longrightarrow} \mathbb{E}_\Theta[m(\mathbf{x}, \Theta)|\mathcal{D}_n] = m_\infty(\mathbf{x}) \tag{4}$$

où $\mathbb{E}_\Theta$ désigne l'espérance sous la v.a. Θ. La consistance de notre estimateur RF motivera notamment l'étude des forêts infinies (i.e. avec un nombre d'arbres infini) ;

— Dans toute la suite, on écrira $\mathbb{E}_\Theta[\cdot]$ pour $\mathbb{E}_\Theta[\cdot|\mathcal{D}_n]$, en supposant désormais que l'espérance est toujours prise sous Θ et conditionnellement aux données $\mathcal{D}_n$;

— Pour mieux se représenter le rôle de la v.a. Θ (qui peut être tant uni- que multi-dimensionnelle), on peut voir cette dernière comme la *façon* ou la *loi* de partitionnement de l'ensemble $\mathcal{X}$ en cellules (ou partitions) $A_1, A_2, \dots$ telles que $A_1 \sqcup A_2 \sqcup \dots = [0;1]^d$. Dans Figure 4, on présente 2 exemples d'arbres de niveau $k = 2$ sur $[0;1]^2 \subset \mathbb{R}_+^2$ associés à 2 partitions différentes $(A_1^{(1)}, A_2^{(1)}, A_3^{(1)}, A_4^{(1)})$ et $(A_1^{(2)}, A_2^{(2)}, A_3^{(2)}, A_4^{(2)})$ issues respectivement de deux v.a. Θ_1 et Θ_2.

On peut illustrer Définition 2.2 avec le diagramme suivant :

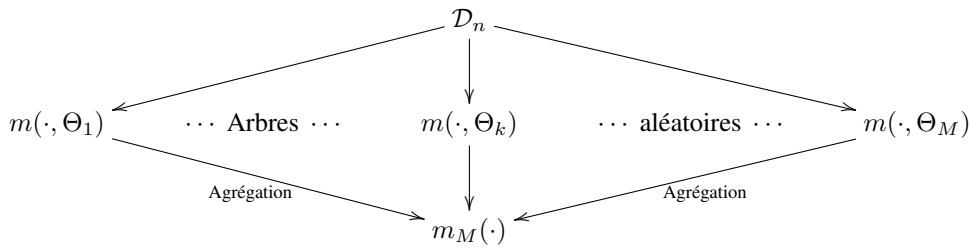


FIGURE 2 – Forêt aléatoire en schéma.

On va maintenant étudier plus spécifiquement l'expression d'un arbre, dans le cas d'une régression comme il est question ici (pour rappel, $Y_i \in \mathbb{R}, 1 \leq i \leq n$).

Pour résumer : après partitionnement, chaque arbre assorti de sa v.a. Θ se place dans la même cellule que $\mathbf{x}$, et effectue une moyenne locale. En d'autres termes, il s'agit de la moyenne empirique des (Y_i) associés aux $(\mathbf{X}_i)$ appartenant à la même cellule que $\mathbf{x}$ (i.e. $A(\mathbf{x}, \Theta)$).

Afin de comprendre le principe de ces prédicteurs, on étudie ci-après la construction de 2 types d'arbres parmi les plus connus, et sur lesquels notre étude se basera postérieurement.

2.2 Exemples d'arbres

2.2.1 Arbre et forêt de BREIMAN : CART

Pour une introduction en détail, on se référera à [9].
L'algorithme **CART** (Classification And Regression Trees) a été développé en 1984 par BREIMAN, FRIEDMAN, OLSHEN et STONE. Il permet de concevoir de façon *simple* et *rapide* des *estimateurs* constants par morceaux (par histogramme) de la fonction cible m de (1). Cette méthode repose sur le *partitionnement récursif* et *dyadique* de l'espace des observations $\mathcal{X} = [0;1]^d$, ce qui se représente par un *arbre* binaire de décision.

Une forêt aléatoire de BREIMAN est construite d'arbres ayant pour racine $\mathcal{X}$ et dont les nœuds sont associés à des cellules hyper[1]-rectangulaires. À chaque étape de construction de l'arbre, un nœud (ou autrement dit sa cellule) est sub-divisé en 2 parties qui seront référencées comme ses 2 fils, pour se conformer à Définition 2.1.

L'algorithme consiste en l'élaboration de $M \geq 2$ arbres aléatoires construits chacun comme suit :

1. **Ré-échantillonage** de notre ensemble $\mathcal{D}_n$ en $\mathcal{D}_n'$ par *boostrap* (avec répétitions possibles) ou *sous-échantillonnage* (sans remises) en effectuant $t \in [\![1; n]\!]$ tirages (hyperparamètre à fixer initialement). On a donc $|\mathcal{D}_n^{'}| = t \leq n$;
2. **Splitting rule** en appliquant le *critère CART* sur l coordonnées tirées uniformément parmi les d possibles (l est un hyperparamètre fixé dans $[\![1; d]\!]$). On appellera $\mathcal{E}_{\text{dir}}$ l'ensemble des coordonnées retenues, tel que $|\mathcal{E}_{\text{dir}}| = l \leq d$;
3. **Stopping rule** pour laquelle on arrête l'algorithme lorsque l'on atteint le nombre minimal d'observations par cellule (hyperparamètre fixé préalablement). En général (en particulier lors d'implémentations sous `R` et `Python`), on demande à ce que chaque feuille de l'arbre contienne entre 1 et 5 observations.

1. Le terme d'"hyper" fait bien référence à la dimension $d \geq 1$ que peut prendre notre espace d'observations à partitionner.

Sans perte de généralité et pour faciliter la compréhension du *critère CART*, on considère, dans les lignes qui suivent, que $\mathcal{D}_n^{'} = \mathcal{D}_n$ (pas de ré-échantillonnage).
Considérons $A \subset \mathcal{X}$ une cellule quelconque (à n'importe quelle itération du processus de partitionnement) qui sera caractérisée par $|A|$, le nombre d'observations $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(d)}) \in \mathcal{X}$ présentes dans A (i.e. son cardinal).
Soit $j \in \{1, \dots, d\}$ une direction de découpe (parmi les l tirées dans la deuxième étape), et $z \in [0; 1]$ une position de découpe sur cette coordonnée j. On pose $\mathcal{E}_{\text{cut}} := \{(j, z) \in [\![1; d]\!] \times [0; 1] | j \in \mathcal{E}_{\text{dir}}\}$ l'ensemble de tous les couples (j, z) de découpes possibles. Le critère de CART est le suivant :

$$L(j, z) = \frac{1}{|A|} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\{\mathbf{X}_i \in A\}} - \frac{1}{|A|} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\left\{X_i^{(j)} < z\right\}} - \bar{Y}_{A_R} \mathbb{1}_{\left\{X_i^{(j)} \geq z\right\}}\right)^2 \tag{5}$$

où $A_L = \{\mathbf{x} \in A | \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A | \mathbf{x}^{(j)} \geq z\} = A \backslash A_L$ et $\bar{Y}_A$ (respectivement $\bar{Y}_{A_L}$ et $\bar{Y}_{A_R}$) étant les moyennes des (Y_i) appartenant à A (respectivement A_L et A_R).

Pour chaque cellule, la meilleure découpe possible $(j^*, z^*) \in \mathcal{E}_{\text{cut}}$ est alors celle *maximisant* le critère (5), i.e. telle que :

$$(j^*, z^*) \in \underset{(j,z) \in \mathcal{E}_{\text{cut}}}{\arg\max}\, L(j, z).$$

Pour résumer, l'arbre de BREIMAN procède à une construction dépendant à la fois des $(\mathbf{X}_i)$ mais également des (Y_i) [2]. Cet arbre se propose ainsi de partitionner nos données $\mathcal{D}_n$ selon le critère CART (5), qui consiste à choisir itérativement les coupures qui correspondent à l'estimateur dont l'*erreur quadratique* est *minimale*.

Exemple 2.2. On clôture cette introduction des arbres de BREIMAN avec 1 exemple de partitionnement CART sur $\mathcal{X} = [0; 1]^2$ où l'on considère les 10 données $(\mathbf{X}_i, Y_i)_{1 \leq i \leq 10} \subset (\mathcal{X} \times \mathbb{R}_+)$ suivantes :

$$((0.08, 0.25), 310), ((0.2, 0.13), 305), ((0.27, 0.4), 340), ((0.31, 0.62), 500), ((0.15, 0.83), 400),$$
$$((0.4, 0.9), 380), ((0.52, 0.6), 100), ((0.68, 0.35), 70), ((0.875, 0.86), 30), ((0.82, 0.74), 5), ((0.87, 0.1), 20).$$

En appliquant l'algorithme CART avec l'arbre de BREIMAN, on peut alors obtenir le partitionnement représenté en Figure 3, où l'ensemble de découpes $\{(1, 0.45), (2, 0.5), (1, 0.75)\} \subset \mathcal{E}_{\text{cut}} = [\![1; 2]\!] \times [0; 1]$ vérifie bien la maximisation du critère (5).

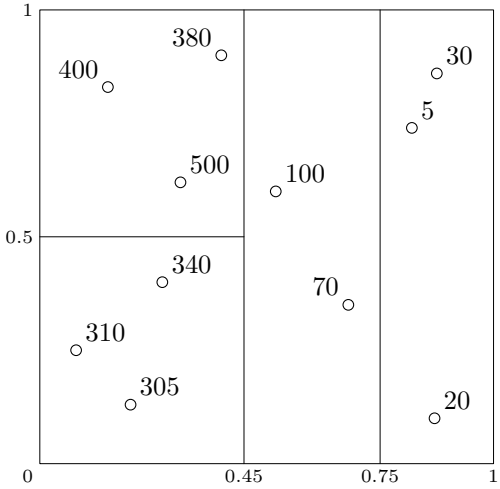


FIGURE 3 – Exemple de représentation d'un arbre de BREIMAN.

2.2.2 Arbre et forêt centrés

Comme exposé dans [2], l'arbre centré est à l'origine d'un modèle de forêt aléatoire intéressant car il est, bien que plus simpliste, proche du modèle des forêts de BREIMAN. L'analyse des forêts centrées est alors motivée par la compréhension des mécanismes statistiques des "vraies" forêts (i.e. les forêts de BREIMAN). Dans la suite, nous étudierons un estimateur dérivé de ce modèle, en gardant à l'esprit ce dessein.

2. Ici l'arbre ne vérifie pas la $\mathbf{X}-$propriété.

Un des éléments essentiels de la forêt centrée est son indépendance de $\mathbf{X}$ et des données $\mathcal{D}_n$. Ceci exclut en particulier une étape de *bootstrapping* ou *re-sampling* du jeu d'entraînement lors de la construction de la collection des arbres.

Chaque arbre suit la procédure suivante itérée k fois (k étant un paramètre à déterminer) :

1. **A chaque nœud** une variable de $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$ est tiré aléatoirement. Chaque j-ième coordonnée ayant une probabilité $p_j \in [0; 1]$ d'être sélectionnée, avec $\sum_j p_j = 1$;
2. **Splitting rule** : la découpe est effectuée au centre de la coordonnée tirée aléatoirement.

La Figure 4 illustre 2 exemples d'arbres centrés de niveau $k = 2$ pour des observations considérées dans $[0; 1]^2$.

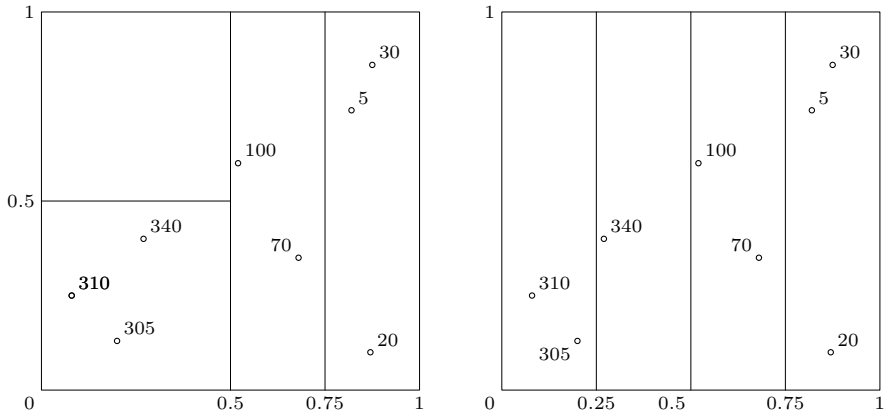


FIGURE 4 – Exemple de représentation de 2 arbres centrés sur $[0; 1]^2$.

Remarque. k peut éventuellement dépendre de la taille de l'échantillon $(\mathbf{X}_i)_{1 \leq i \leq n}$. Ce paramètre est crucial pour contrôler en moyenne le nombre de données dans chaque nœud terminal. Dans l'exemple de la Figure 4, on remarque que le processus de partitionnement centré peut en effet conduire à des cellules dépourvues d'observations. La racine étant $\mathcal{X} = [0; 1]^d$, remarquons alors que chaque feuille est un hyper-rectangle de mesure de LEBESGUE 2^{-k}. Si les $(\mathbf{X}_i)$ sont uniformément distribuées sur $[0; 1]^d$, on peut espérer en moyenne compter $n2^{-k}$ observations par nœud terminal. Le choix de ce paramètre sera discuté en section 4.1.2 dans laquelle on applique le résultat 4.1 au modèle des forêts centrées.

3 L'estimateur KeRF

Cette partie est dédiée à l'introduction d'un nouvel estimateur de (1) : en modifiant légèrement la procédure d'agrégation, l'estimateur des forêts (RF) peut alors se réécrire comme un *estimateur à noyau*, dénommé *KeRF* (pour Kernel Random Forest, en anglais).

3.1 Motivation des estimateurs à noyau

Revenons tout d'abord à notre estimateur de forêt aléatoire (3) à $M \geq 1$ arbres de régressions admettant la forme (2).
La pondération

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

pour les couples de données $(\mathbf{X}_i, Y_i)$ étant *uniforme* au sein de chaque cellule, alors la contribution de chaque Y_i est relative au nombre d'observations qui y sont contenues.

Ceci pose un problème d'estimation dans certaines conditions :

— Considérons par exemple une configuration des données avec une forte disparité [3], i.e. des cellules à forte densité et d'autres faiblement peuplées. La représentativité d'une variable Y_i sera alors d'autant plus forte que la structure de sa cellule d'appartenance est *lacunaire* ;

— Par ailleurs (en particulier dans le cas d'arbres construits indépendemment des données, e.g. les arbres centrés), l'estimation de l'arbre décisionnel est 0 par convention lorsque la cellule est vide. La valeur 0 va alors biaiser l'estimation globale de la forêt.

Exemple 3.1 (pathologique). Pour illustrer cette dernière remarque, considérons l'exemple suivant :
Soit $n = 200, d = 2, k = \lfloor \log_2(200) \rfloor$ et on considère $(\mathbf{X}_i)_{1 \leq i \leq n} \underset{\text{i.i.d.}}{\sim} \mathcal{U}([0, 1]^2)$ (où $\log_2$ est le logarithme de base 2). On construit 4

3. Remarquons que dans le cas où $\forall 1 \leq i \leq d, \mathbf{X}_i \sim \mathcal{U}\left([0, 1]^d\right)$, alors p.s. la configuration de données disparates est évitée.

arbres décisionnels centrés de niveau k, représentés dans Figure 5. On cherche à estimer la fonction de régression en un $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, représenté en rouge. Les points bleus représentent, quant à eux, les données partageant la même cellule que $\mathbf{x}$ et donc celles prises en compte dans l'estimation de l'arbre.

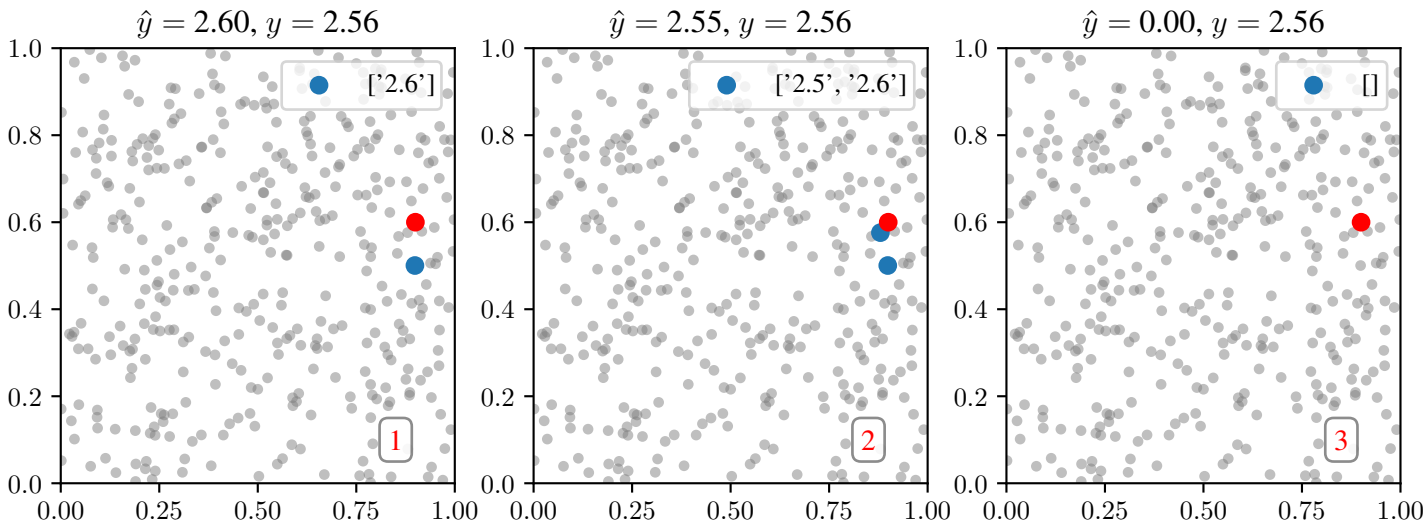


FIGURE 5 – Exemple de cas pathologiques pour les arbres de régression centrés appliqués à des données uniformes sur $\mathcal{X}$.

Dans cet exemple, on introduit $\tilde{\mathbf{X}} \in [-1; 1]^2$ tel que $\tilde{X}_i = 2X_i - 1, 1 \leq i \leq d$ et $Y = 2\tilde{X}_1 + \exp(-\tilde{X}_2^2)$. Pour $\mathbf{x} = (0.9, 0.6)$, la valeur cible est $Y \simeq 2.6$. L'estimateur RF à $M = 3$ arbres (voir 2.2) prédirait :

$$m_3(\mathbf{x}) = \frac{2.6 + \frac{2.5+2.6}{2} + 0}{3} \simeq 1.7$$

En prédisant 0 par défaut, l'arbre numéro 3 a biaisé l'estimation de la forêt. Nous allons voir que l'estimateur KeRF parvient à détourner ce travers.

3.2 L'estimateur KeRF

Pour contourner le problème rencontré dans l'exemple 3.1 ci-dessus, une idée a été d'introduire l'estimateur KeRF suivant, qui découle d'une légère modification de m_M en (3) :

Définition 3.1 (KeRF, [17])

Soit $\mathbf{x} \in [0; 1]^d$, et les v.a. $\Theta_1, \cdots, \Theta_M \underset{\text{i.i.d.}}{\sim} \mathcal{L}(\Theta)$, indépendantes de $\mathcal{D}_n$. On introduit l'estimateur **KeRF** :

$$\tilde{m}_M(\mathbf{x}) := \frac{1}{\sum_{j=1}^M N(\mathbf{x}, \Theta_j)} \sum_{j=1}^M \sum_{i=1}^n Y_i \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x}, \Theta_j)\}} \tag{6}$$

Remarque.

— En remarquant que, pour $\mathbf{x} \in [0; 1]^d$ et $\Theta_1, \cdots, \Theta_M$ comme dans Définition 3.1,

$$\tilde{m}_M(\mathbf{x}) = \frac{\frac{1}{M}\sum_{j=1}^M \sum_{i=1}^n Y_i \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x}, \Theta_j)\}}}{\frac{1}{M}\sum_{j=1}^M N(\mathbf{x}, \Theta_j)} = \frac{\sum_{i=1}^n Y_i \left(\frac{1}{M}\sum_{j=1}^M \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x}, \Theta_j)\}}\right)}{\frac{1}{M}\sum_{j=1}^M N(\mathbf{x}, \Theta_j)} \tag{7}$$

alors (6) devient une moyenne pondérée de l'ensemble des (Y_i) : le poids de chaque Y_i étant le nombre de fois que la donnée $\mathbf{X}_i$ associée est comptée dans le même nœud terminal que $\mathbf{x}$, i.e. parmi $A(\mathbf{x}, \Theta_1), \ldots, A(\mathbf{x}, \Theta_M)$.

— Considérant cet estimateur, la moyenne n'étant plus *intra* mais *inter-cellulaire* (i.e. sur la globalité des (Y_i) qui interviennent), la faible ou forte densité d'une cellule (et en particulier le cas où une cellule est vide) n'influera plus sur l'estimation. Reprenons l'exemple 5, où la valeur cible est 2.6. L'estimateur KeRF prédirait $\tilde{m}_3(\mathbf{x}) = \frac{2.6+2.5+2.6}{3} \simeq 2.6$ ce qui est bien meilleur que l'estimation 1.7 de m_3.

3.3 Noyau KeRF & estimateur associé

Il existe une analogie entre l'estimateur KeRF et les méthodes à noyau existe [6]. Les noyaux sont des fonctions grandement utilisées, par exemple dans des problèmes d'estimation de densité ou de régression.
On introduit maintenant le noyau suivant, qui nous sera utile pour la suite :

Définition 3.2 (Noyau de KeRF, [14], [15], [17])

Considérons une forêt aléatoire à M arbres basés sur les M v.a. $\Theta_1, \ldots, \Theta_M \underset{\text{i.i.d.}}{\sim} \mathcal{L}(\Theta)$, indépendantes de $\mathcal{D}_n$. Soit $(\mathbf{x}, \mathbf{z}) \in \mathcal{X}^2$, on introduit la **fonction noyau** $K_M : [0;1]^d \times [0;1]^d \to [0;1]$ suivante :

$$K_M : (\mathbf{x}, \mathbf{z}) \mapsto \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\{\mathbf{z} \in A(\mathbf{x}, \Theta_j)\}} \tag{8}$$

La forme de KeRF rappelle celle de l'estimateur classique de NADARAYA-WATSON [18]. Cependant, bien qu'un lien puisse être établi entre ces derniers, K_M en (8) et la forme de $\tilde{m}_M$ introduit en (6), l'analyse de l'*estimateur KeRF à noyau* que l'on définira en Proposition 3.1 n'a pourtant rien d'évident. C'est pourquoi la Proposition 3.1 suivante sera admise comme étant l'écriture d'un estimateur à noyau, faisant appelle à des *fonctions de connexion* relatives à notre $M-$forêt (i.e. notre forêt à M arbres).

Proposition 3.1 (Estimateurs à noyau KeRF, [17])

Pour tout $\mathbf{x} \in [0;1]^d$, et $\Theta_1, \ldots, \Theta_M \underset{\text{i.i.d.}}{\sim} \mathcal{L}(\Theta)$, indépendantes de $\mathcal{D}_n$, on a le résultat suivant, presque-sûrement :

$$\tilde{m}_M(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_M(\mathbf{x}, \mathbf{X}_i)}{\sum_{l=1}^{n} K_M(\mathbf{x}, \mathbf{X}_l)}, \tag{9}$$

où K_M, définie comme dans Définition 3.2, est également appelée **fonction de connexion** de la forêt aléatoire finie estimée par m_M.
En particulier, 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.

Remarque. On peut néanmoins établir que les estimateurs à noyau [18] sont des estimateurs classiques basés sur une fonction mesurant la distance entre les observations et le point $\mathbf{x} \in \mathcal{X}$ considéré (on peut justement citer les mesures de similarité et les fonctions de connexion [4] très utilisées en clustering spectral par exemple). Ici l'analyse se veut plus complexe, mais l'intéressé peut se référer à [1], [11] et plus anciennement [6] qui étudient et quantifient la connexion entre *forêts aléatoires* et *estimateurs à noyau*.

Démonstration. Par définition,

$$\begin{aligned}
\tilde{m}_M(\mathbf{x}) &= \frac{\sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\sum_{j=1}^{M} N(\mathbf{x}, \Theta_j)} \\
&= \frac{\frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\frac{1}{M} \sum_{j=1}^{M} N(\mathbf{x}, \Theta_j)} \\
&= \frac{\sum_{i=1}^{n} Y_i \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\sum_{i=1}^{n} \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}} \\
&= \frac{\sum_{i=1}^{n} Y_i K_M(\mathbf{x}, \mathbf{X}_i)}{\sum_{l=1}^{n} K_M(\mathbf{x}, \mathbf{X}_l)}
\end{aligned}$$

□

Remarque. On soulève un intérêt de l'estimateur KeRF par rapport à l'estimateur forêt aléatoire (RF). La forme de KeRF révèle une interprétation sur le partitionnement construit par la forêt aléatoire qui lui est associée. En effet, le noyau de l'estimateur KeRF (8) est aussi la fonction de connexion K_M. Pour $\mathbf{x}, \mathbf{z} \in [0;1]^d$, $K_M(\mathbf{x}, \mathbf{z})$ est la probabilité empirique sous Θ que $\mathbf{x}$ et $\mathbf{z}$ soient dans la même cellule. Ceci caractérise la géométrie des cellules d'une forêt.

Rappelons que par la loi forte des grands nombres l'estimateur m_M converge presque sûrement :

$$m_M(\mathbf{x}) \underset{M \to +\infty}{\longrightarrow} \mathbb{E}_\Theta[m(\mathbf{x}, \Theta) | \mathcal{D}_n] = m_\infty(\mathbf{x}).$$

4. Voir [14] et [10] pour ces notions.

On peut naturellement se demander ce qu'il se passe pour $\tilde{m}_M$ lorsque M devient infiniment grand. Afin de répondre à cette interrogation, nous définissons l'estimateur KeRF infini.

Définition 3.3 (KeRF infini, [17])

Soit $\mathbf{x} \in [0;1]^d$, et $\Theta_1, \cdots, \Theta_M \underset{\text{i.i.d.}}{\sim} \mathcal{L}(\Theta)$, indépendantes de $\mathcal{D}_n$.
On définit l'estimateur **KeRF infini** :

$$\tilde{m}_\infty(\mathbf{x}) := \lim_{M \to +\infty} \tilde{m}_M(\mathbf{x}, \Theta_1, ..., \Theta_M) \tag{10}$$

Remarque (de notation). À partir de maintenant, on désignera par KeRF l'estimateur $\tilde{m}_M$ (9) à noyau basé sur un nombre *fini* d'arbres, et par KeRF infini $\tilde{m}_\infty$ (11) celui construit via une *infinité* d'arbres.

La proposition suivante établie que lorsque $M \to \infty$, la fonction de connexion K_M, évaluée en $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, converge presque-sûrement vers la probabilité sous Θ que $\mathbf{x}$ et $\mathbf{z}$ soient dans la même cellule, conditionnellement aux données $\mathcal{D}_n$. On remarquera notamment que cette convergence a lieu, que la fonction K_M soit discrète ou continue. Cette probabilité sera notée K et devient alors la *fonction de connexion* de la forêt infinie.

Proposition 3.2 (Estimateurs à noyau KeRF infinis, [17])

On considère une forêt aléatoire avec un nombre d'arbres infini. Alors pour tous $\mathbf{x}, \mathbf{z} \in [0,1]^d$, on a la convergence :

$$\lim_{M \to +\infty} K_M(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{z}) \quad p.s.,$$

où

$$K(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta[\mathbf{z} \in A(\mathbf{x}, \Theta)],$$

$\mathbb{P}_\Theta$ désignant la probabilité sous Θ et conditionnellement aux données $\mathcal{D}_n$.

De plus, pour tout $\mathbf{x} \in [0,1]^d$, on a :

$$\tilde{m}_\infty(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n K(\mathbf{x}, \mathbf{X}_i)} \tag{11}$$

On appelle K la **fonction de connexion** de la forêt aléatoire infinie.

Ce résultat montre que les KeRF infinis sont encore des estimateurs à noyaux. L'interprétation que nous avions établie pour KeRF (fini) s'applique de nouveau, K caractérise bien la forme des cellules de la forêt aléatoire.

Démonstration. Idée de preuve : En se plaçant dans le cas de forêts aléatoires continues et en dimension $d = 2$ (le cas discret et d quelconque en découle), la première limite résulte de la loi forte des grands nombres appliquée à K_M en $\mathbf{x}, \mathbf{z} \in [0;1]^2 \cap \mathbb{Q}$ (par un argument de densité, on a le résultat sur $[0,1]^2$).
L'égalité obtenue pour $\tilde{m}_\infty$ en fonction de K est quant à elle travaillée en utilisant l'hypothèse de continuité de K sur $[0;1]^2 \cap \mathbb{Q}$. En utilisant le critère de continuité sur des voisinages des points $\mathbf{x}$ et $\mathbf{z}$, chacun peut alors contrôler la quantité $|K_M(\mathbf{x}, \mathbf{z}) - K(\mathbf{x}, \mathbf{z})|$ et aboutir au résultat souhaité. On pensera à disjoindre le cas $\sum_{j=1}^M K(\mathbf{x}, \mathbf{X}_i) = 0$ pour lequel on admet la convention $\tilde{m}_\infty(\mathbf{x}) = \tilde{m}_M(\mathbf{x}) = 0$. □

Nous verrons par la suite que dans certains cas nous pouvons expliciter un tel noyau K, ce qui offre de bonnes perspectives d'analyse. Avant cela, nous allons étudier dans quels cas KeRF et KeRF infini sont une approximation raisonnable des forêts aléatoires : c'est l'objet de la prochaine section.

4 La relation entre KeRF et les forêts aléatoires

Dans cette partie nous étudierons la proximité entre KeRF et l'estimateur RF. Moralement, ces estimateurs sont proches lorsque le nombre de données par cellule est contrôlé. Ceci est facilement obtenu lorsque les forêts sont adaptatives, c'est-à-dire que le processus de partitionnement dépend des données. Les forêts de BREIMAN vues en partie 2.2.1 en sont un exemple. On peut déterminer exactement le nombre d'observations dans chaque nœud terminal. En revanche cela est moins évident pour des modèles indépendants du jeu d'entraînement. C'est le cas des forêts centrées vues en 2.2.2 dont on approfondit l'étude dans la partie 4.1.2. Cependant, sous l'hypothèse de distribution uniforme des $(\mathbf{X}_i)$ dans $[0,1]^d$ et d'un paramétrage approprié des arbres de décision, on peut tout de même obtenir de bonnes garanties.

Dans toute cette section on supposera $Y \geq 0$ p.s..

4.1 Le cas des forêts finies

Proposition 4.1 (Distance relative de KeRF à RF, [17])

Soit $\mathbf{x} \in \mathcal{X}$. On suppose qu'il existe 2 suites (a_n) et (b_n), telles que,

$$a_n \leq N(\mathbf{x}, \Theta) \leq b_n \quad p.s..$$

Alors,

$$\left| \frac{m_M(\mathbf{x}) - \tilde{m}_M(\mathbf{x})}{\tilde{m}_M(\mathbf{x})} \right| \leq \frac{b_n - a_n}{a_n} \quad p.s.$$

avec la convention que $\frac{0}{0} = 1$.

Démonstration. Prenons $\mathbf{x} \in \mathcal{X}$ et supposons que $Y \geq 0$ p.s. Sous l'hypothèse, on a $\forall 1 \leq j \leq M$, $a_n \leq N(\mathbf{x}, \Theta_j)$ et donc, presque-sûrement :

$$|m_M(\mathbf{x}) - \tilde{m}_M(\mathbf{x})| = \left| \sum_{i=1}^{n} Y_i \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{N(\mathbf{x}, \Theta_j)} \right) - \sum_{i=1}^{n} Y_i \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\frac{1}{M}\sum_{j=1}^{M} N(\mathbf{x}, \Theta_j)} \right) \right| \quad \text{(par (7))}$$

En notant $\bar{N}_M(\mathbf{x}) := \frac{1}{M} \sum_{j=1}^{M} N(\mathbf{x}, \Theta_j)$ le cardinal moyen de la cellule contenant $\mathbf{x}$, on obtient :

$$\begin{aligned}
&= \frac{1}{M} \left| \sum_{i=1}^{n} Y_i \left(\sum_{j=1}^{M} \frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{N(\mathbf{x}, \Theta_j)} - \sum_{j=1}^{M} \frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\bar{N}_M(x)} \right) \right| \\
&\leq \underbrace{\frac{1}{M} \sum_{i=1}^{n} Y_i \sum_{j=1}^{M} \frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x}, \Theta_j)\}}}{\bar{N}_M(\mathbf{x})}}_{=\tilde{m}_M(\mathbf{x})} \left| \frac{\bar{N}_M(\mathbf{x})}{N(\mathbf{x}, \Theta_j)} - 1 \right| && \text{(factorisation et inégalité triangulaire)} \\
&\leq \frac{b_n - a_n}{a_n} \tilde{m}_M(\mathbf{x}) && \text{(majoration du terme en } |\cdot| \text{ par hypothèse)}
\end{aligned}$$

□

On voit immédiatement que si a_n et b_n sont proches alors nos estimateurs aussi. Appliquons ce résultat pour des modèles spécifiques.

4.1.1 Les forêts de Breiman

Si l'on détermine que chaque nœud terminal des arbres CART contient exactement une observation, alors $a_n = b_n$. D'après la proposition 4.1, p.s.,

$$m_M(\mathbf{x}, \Theta_1, \cdots, \Theta_M) = \tilde{m}_M(\mathbf{x}, \Theta_1, \cdots, \Theta_M).$$

Plus généralement le nombre de points dans chaque nœud terminal des arbres CART est paramétré entre 1 et 5. Ainsi toujours par la proposition 4.1, p.s.,

$$\left| \frac{m_M(\mathbf{x}, \Theta_1, ..., \Theta_M)}{\tilde{m}_M(\mathbf{x}, \Theta_1, ..., \Theta_M)} - 1 \right| \leq 4.$$

4.1.2 Les forêts centrées de niveau k

On suppose que $\mathbf{X} \sim \mathcal{U}\left([0,1]^d\right)$, sous cette hypothèse nous avons vu que chaque cellule avait pour mesure de LEBESGUE 2^{-k}. Ainsi le nombre moyen de points $\mathbf{X}_i$ est $\frac{n}{2^k}$. Fixant $\mathbf{x} \in [0,1]^d$, pour n assez grand, on admet que :

$$\left| N(\mathbf{x}, \Theta) - \frac{n}{2^k} \right| \leq \frac{\sqrt{2n \log\log n}}{2}.$$

Ce résultat est une conséquence de la loi du logarithme itéré. Pour une preuve détaillée, on peut se référer à [17].

L'hypothèse de contrôle du nombre d'observations par cellule est satisfaite. En effet,

$$a_n = \frac{n}{2^k} - \frac{\sqrt{2n \log\log n}}{2} \leq N(\mathbf{x}, \Theta) \leq \frac{n}{2^k} + \frac{\sqrt{2n \log\log n}}{2} = b_n.$$

Cela induit, par la proposition 4.1, p.s.,

$$\left| \frac{m_M(\mathbf{x}, \Theta_1, \cdots, \Theta_M)}{\tilde{m}_M(\mathbf{x}, \Theta_1, \cdots, \Theta_M)} - 1 \right| \leq \frac{\sqrt{2n \log\log n}}{n2^{-k} - \sqrt{2n \log\log n}/2}.$$

En choisissant par exemple $k = (\log_2 n)/3$, on a,

$$n2^{-k} = n2^{-(\log_2 n)/3} = nn^{-1/3} = n^{2/3}.$$

Alors,

$$\frac{\sqrt{2n\log\log n}}{n2^{-k} - \sqrt{2n\log\log n}/2} = \frac{\sqrt{2n\log\log n}}{n^{2/3} - \sqrt{2n\log\log n}/2} \underset{n\to+\infty}{\longrightarrow} 0.$$

Les forêts centrées et les estimateurs KeRF sont asymptotiquement équivalents.

Beware my Lord ! Attention ici, c'est le nombre de données qui devient infiniment grand. Nous sommes toujours dans le cas de forêts centrées finies, c'est-à-dire que les statistiques m_M et $\tilde{m}_M$ sont construites en agrégeant un nombre *fini* d'arbres centrés (i.e. $M < \infty$).

Nous allons à présent voir que sous des hypothèses assez proches, ce résultat s'étend pour les forêts infinies.

4.2 Le cas des forêts infinies

Proposition 4.2 (Distance de KeRF à RF infinis, [17])

On suppose qu'il existe des suites (ϵ_n), (a_n), (b_n), telles que, p.s.,

$$1 \leq a_n \leq \mathbb{E}_\Theta[N(\mathbf{x},\Theta)] \leq b_n,$$
$$\mathbb{P}_\Theta\left[a_n \leq N(\mathbf{x},\Theta) \leq b_n\right] \geq 1-\epsilon_n.$$

Alors, p.s.,

$$|m_\infty(\mathbf{x}) - \tilde{m}_\infty(\mathbf{x})| \leq \frac{b_n - a_n}{a_n}\tilde{m}_\infty(\mathbf{x}) + n\epsilon_n \max_{1\leq i\leq n} Y_i.$$

Démonstration. Soit $\mathbf{x} \in [0,1]^d$, on suppose que $Y \geq 0$ p.s. et qu'il existe des suites (ϵ_n), (a_n), (b_n), telles que, p.s.,

$$1 \leq a_n \leq \mathbb{E}_\Theta[N(\mathbf{x},\Theta)] \leq b_n,$$
$$\mathbb{P}_\Theta\left[A\right] \geq 1-\epsilon_n,$$

où A désigne l'événement $\{a_n \leq N(\mathbf{x},\Theta) \leq b_n\}$.
On rappelle que $\tilde{m}_\infty$ et m_∞ admettent respectivement les formes (11) et (4). Alors,

$$\begin{aligned}
|m_\infty(\mathbf{x}) - \tilde{m}_\infty(\mathbf{x})| &= \left|\sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{N(\mathbf{x},\Theta)}\right] - \sum_{i=1}^n Y_i \frac{\mathbb{E}_\Theta\left[\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}\right]}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\right| \\
&= \left|\sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{N(\mathbf{x},\Theta)} - \frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\right]\right| \\
&= \left|\sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)} - 1\right)\right]\right|
\end{aligned}$$

En conditionnant l'équation selon l'évènement A et son complémentaire, on obtient,

$$|m_\infty(\mathbf{x}) - \tilde{m}_\infty(\mathbf{x})| = \left|\sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)} - 1\right)\mathbb{1}_{\{A\}}\right]\right| + \left|\sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)} - 1\right)\mathbb{1}_{\{A^c\}}\right]\right|.$$

Pour le premier terme, nous sommes sous l'événement A, les hypothèses suivantes sont vérifiées p.s. :

$$\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right] \leq b_n, \mathbb{P}_\theta[N(\mathbf{x},\Theta) \geq a_n] \geq 1-\epsilon_n.$$

Alors,

$$\begin{aligned}
\left|\sum_{i=1}^n Y_i \mathbb{E}_\theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)} - 1\right)\mathbb{1}_{\{A\}}\right]\right| &\leq \left|\frac{\mathbb{E}_\theta\left[\sum_{i=1}^n Y_i \mathbb{1}_{\{\mathbf{x}_i\in A(\mathbf{x},\Theta)\}}\right]}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{b_n - a_n}{a_n}\right)\mathbb{1}_{\{A\}}\right| \\
&\leq \tilde{m}_\infty(\mathbf{x})\left(\frac{b_n - a_n}{a_n}\right).
\end{aligned}$$

En ce qui concerne le second terme, on a,

$$\begin{aligned}\left|\sum_{i=1}^{n} Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)}-1\right)\mathbb{1}_{\{A^c\}}\right]\right| &= \left|\sum_{i=1}^{n} Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\{\mathbf{x}_i \in A(\mathbf{x},\Theta)\}}}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\left(\frac{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}{N(\mathbf{x},\Theta)}-1\right)|\mathbb{1}_{\{A^c\}}\right]\mathbb{P}_\Theta(A^c)\right| \\ &\leq \epsilon_n \max_{1\leq i\leq n} Y_i \mathbb{E}_\Theta\left[\left|1-\frac{N(\mathbf{x},\Theta)}{\mathbb{E}_\Theta\left[N(\mathbf{x},\Theta)\right]}\right||A^c\right] \\ &\leq n\epsilon_n \max_{1\leq i\leq n} Y_i.\end{aligned}$$

On en déduit que p.s.,

$$|m_\infty(\mathbf{x}) - \tilde{m}_\infty(\mathbf{x})| \leq \tilde{m}_\infty(\mathbf{x})\left(\frac{b_n - a_n}{a_n}\right) + n\epsilon_n \max_{1\leq i\leq n} Y_i.$$

□

Proposition 4.2 nous montre que si l'on contrôle le nombre d'observations par cellule avec une probabilité assez grande (on voudrait que ϵ_n soit petit), alors les forêts infinies sont proches des KeRF infinis. Par ailleurs, et c'est l'objet de la suite, on aimerait que la géométrie de chacune de ces cellules soit suffisamment simple pour pouvoir expliciter la probabilité K. Obtenir les deux garanties à la fois est cependant loin d'être évident.

En effet, lorsque la construction d'un arbre dépend des données (par exemple l'arbre de BREIMAN), on peut paramétrer l'algorithme pour n'obtenir qu'un point par nœud terminal. Dans ce cas, il est intéressant de remarquer que les estimations KeRF et forêts infinies sont équivalentes. En effet, on a alors que $N(\mathbf{x},\Theta) = 1$ et donc :

$$m_M(\mathbf{x},\Theta_1,...,\Theta_M) = \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{i=1}^{n} Y_i \mathbb{1}_{\{\mathbf{X}_i \in A(\mathbf{x},\Theta_j)\}}\right) = \tilde{m}_M(\mathbf{x},\Theta_1,...,\Theta_M).$$

Et par passage à la limite $M \to \infty$,

$$m_\infty(\mathbf{x}) = \tilde{m}_\infty(\mathbf{x}).$$

Cependant le partitionnement des arbres dépendant fortement du jeu d'entraînement, sa géométrie est ainsi difficilement prédictible, il est alors compliqué d'exprimer la probabilité K établie en proposition 3.2.

Si l'on s'intéresse aux arbres dont la procédure de découpe ne dépend pas des données, la probabilité K peut être explicitée dans certains cas. C'est ce que nous allons faire ci-dessous pour les forêts centrées. En contrepartie, le nombre de points par cellule est difficile à borner, et l'estimation KeRF peut être potentiellement loin de l'estimation forêt aléatoire. Certes nous savons exprimer KeRF, mais nous ne pouvons en déduire une expression explicite pour la forêt associée.

A présent, on s'intéressera à l'étude des KeRF centrés, pour lesquels nous déterminerons la fonction de connexion K avant d'en déduire dans la dernière partie la *consistance de l'estimateur* ainsi qu'une borne supérieure pour sa *vitesse de convergence*.

5 Approfondissement de cas : les KeRF centrés

Nous allons approfondir l'analyse des estimateurs à noyaux en nous concentrant sur les KeRF infinis. On en rappelle la formule générique :

$$\tilde{m}_\infty(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K(\mathbf{x},\mathbf{X}_i)}{\sum_{i=1}^{n} K(\mathbf{x},\mathbf{X}_i)}.$$

Deux raisons justifient ce choix :

- Les KeRF infinis ont l'avantage de ne pas dépendre des arbres particuliers avec lesquels la forêt a été construite. En effet ils sont caractérisés par la fonction de connexion K. En ce sens ils sont plus généraux que les estimateurs KeRF finis et plus aptes à l'analyse mathématique ;
- De plus, ces estimateurs sont connus pour avoir une *meilleure précision* d'estimation que les KeRF (finis) construits à partir d'un nombre *finis* d'arbres.

5.1 La fonction de connexion des KeRF centrés

Les forêts centrées sont un modèle simplifié des forêts de BREIMAN. Si l'on se place dans un modèle de régression linéaire, en supposant les $(\mathbf{X}_i)$ i.i.d. uniformes sur $[0,1]^d$, alors on peut présupposer que les premières découpes de chaque arbre CART seront localisées non loin du centre de la cellule. Dans ce contexte les forêts de BREIMAN et les forêts centrées sont proches l'une de l'autre, ce qui justifie l'intérêt pour les KeRF centrés. Ce modèle suggéré dans [8] par BREIMAN (en 2004) a déjà été l'objet d'une étude dans [2].

Comme nous allons le voir avec Proposition 5.1, il est possible d'en expliciter la fonction de connexion, i.e. la probabilité que deux points soient dans le même nœud terminal d'un arbre centré. La construction d'un tel estimateur est indépendante des données. Nous avons vu en appliquant la proposition 4.1 dans 4.1.2 que le choix du paramètre k en fonction de la taille du jeu d'entraînement est crucial pour garantir une proximité entre KeRF et RF centrés. Nous garderons cette dépendance à l'esprit et noterons K^c le noyau de l'estimateur KeRF centré infini.

Proposition 5.1 (KeRF centré,[17])

Soit $k \in \mathbb{N}$. On considère une forêt aléatoire centrée de niveau k. Alors, pour tout $\mathbf{x} = (x_1, \dots, x_d), \mathbf{z} = (z_1, \dots, z_d) \in [0,1]^d$,

$$K^c(\mathbf{x}, \mathbf{z}) = \sum_{\substack{k_1,\dots,k_d \\ \sum_{l=1}^d k_l = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}^k\right) \prod_{j=1}^d \mathbb{1}_{\{\lceil 2_j^k x_j \rceil = \lceil 2_j^k z_j \rceil\}}. \tag{12}$$

Remarque.
— En pratique, il est plus commode d'utiliser les KeRF finis : le coefficient multinomial apparaissant dans l'écriture de K^c (12) complexifie le calcul machine. Comme le justifie la Proposition 3.2, les KeRF finis peuvent s'interpréter comme des *estimateurs Monte-Carlo* des KeRF infinis ;
— Les nœuds terminaux sont des hypercubes définis par le produit $\prod_{i=1}^d A_i$, où, par convention, les A_i sont des intervalles ouverts à gauche sauf en 0, i.e. $A_i =]a, b], 0 < a < b \leq 1$ ou $A_i = [0, b], 0 < b \leq 1$.

Démonstration. On fixe $\mathbf{x}, \mathbf{z}, \in [0,1]^d$.
Sélectionnons une coordonnées sur chacun de nos vecteurs, disons $x_j, z_j \in [0;1]$, pour un $j \in \{1, \dots, d\}$. On découpe l'axe j en 2^{k_j} sous intervalles égaux, k_j représentant le nombre de fois que l'axe j a été tiré au hasard lors de la construction de l'arbre de niveau k. Les intervalles sont de la forme : $[0; \frac{1}{2^{k_j}}]$ pour celui contenant 0 puis $]\frac{l}{2^{k_j}}; \frac{l+1}{2^{k_j}}], \forall l \in \left\{1, \dots, 2^{k_j}\right\}$.
Supposons que x_j est dans l'intervalle $]\frac{l}{2^{k_j}}; \frac{l+1}{2^{k_j}}]$, alors :

$$\frac{l}{2^{k_j}} < x_j \leq \frac{l+1}{2^{k_j}} \iff l < 2^{k_j} x_j \leq (l+1)2^{k_j} \iff \lceil 2^{k_j} x_j \rceil = l + 1.$$

Sans perte de généralité, supposons que $x_j \leq z_j$, alors l'événement $\{x_j$ et z_j sont dans le même intervalle$\}$ est caractérisé par :

$$\mathbb{1}_{\left\{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil\right\}}. \tag{13}$$

Pour que $\mathbf{x}$ et $\mathbf{z}$ soient dans la même cellule, il suffit de vérifier la condition (13) pour chaque dimension, i.e. pour tous les couples : $(\mathbf{x}_1, \mathbf{z}_1), \dots, (\mathbf{x}_d, \mathbf{z}_d)$.
Reste à déterminer la loi de probabilité du vecteur $(k_1, \dots, k_d)$. A chaque nœud, on tire uniformément et de manière indépendante $j \in \{1, \dots, d\}$, après k étapes de découpe, on a $(k_1, \dots, k_d)$ tirages de chaque variable j. Alors on a $\sum_{j=1}^d k_j = k$, et $0 \leq k_j \leq k$. Ainsi $(k_j)_{1 \leq j \leq d}$ suit une loi multinomiale de paramètres $(\frac{1}{d}, \dots, \frac{1}{d}) \in \mathbb{R}^d$.

Alors :

$$K^c(\mathbf{x}, \mathbf{z}) = \sum_{\substack{k_1,\dots,k_d \\ \sum_{l=1}^d k_l = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}^k\right) \prod_{j=1}^d \mathbb{1}_{\left\{\lceil 2_j^k \mathbf{x}_j \rceil = \lceil 2_j^k \mathbf{z}_j \rceil\right\}}.$$

□

Nous disposons d'une formule explicite pour les KeRF infinis centrés associés à la fonction de connexion K^c (on rappelle que K^c dépend également du paramètre k) :

$$\tilde{m}_\infty^c(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K^c(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n K^c(\mathbf{x}, \mathbf{X}_i)}$$

Nous continuons notre analyse des KeRF infinis centrés en montrant leur *consistance* accompagnée d'une borne supérieure dans un modèle de *régression* avec *bruit Gaussien*.

5.2 Vitesse de convergence des KeRF centrés

Theorem 5.1 (Convergence L^2 de KeRF centré)

On se place dans le modèle suivant :

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

avec $\epsilon \sim \mathcal{N}(0,1)$ indépendante de $\mathbf{X}$, et de variance $\sigma^2 < \infty$.

Alors pour $k \to \infty$ et $n/2^k \to \infty$, il existe une constante $C_1 > 0$ telle que, pour tout $n > 1$, et pour tout $x \in [0,1]^d$,

$$\mathbb{E}[|\tilde{m}^c_\infty(\mathbf{x}) - m(\mathbf{x})|^2] \leq C_1 n^{-1/(3+d\log 2)} (\log n)^2.$$

À titre de comparaison, la vitesse minimax des estimateurs à noyaux pour la classe des fonctions LIPSCHITZ est de l'ordre de $n^{-2/(d+2)}$ et notre estimateur ne l'atteint pas.

La Figure 6 représente le risque quadratique empirique des RF centrées, KeRF centrés et des forêts de BREIMAN. On évalue les performances de ces 3 estimateurs sur deux modèles de données que l'on décrit ci-dessous. Pour chaque expérience, l'échantillon des données est divisé en un jeu d'entraînement (80%) et en un jeu de test (20%). Le temps de calcul pour les forêts centrées finies est raisonnable, $\simeq$ 3 minutes pour le modèle 1, contre $\simeq$ 20 pour le modèle 2 [5]. On rappelle qu'ils représentent des estimateurs MONTE-CARLO des forêts infinies. Le risque empirique (noté L^2) est évalué sur le jeu de test. Les forêts ont été construites sans bootstrap. Les forêts centrées sont implémentées en choisissant un niveau $k = \lfloor \log_2 n \rfloor$. Dans ce cas, chaque nœud terminal contient en moyenne $n/2^k \simeq 1$ observation. Les paramètres des forêts de Breiman sont laissés par défaut mise à part `maxfeatures = 0.333`. La variable `maxfeatures` est la proportion de coordonnées (i.e. les features d'une observation) considérée pour chaque coupure.
On considère des v.a. multivariées $\mathbf{X} = (X_1, ..., X_d)$ uniformément distribuées sur $[0,1]^d$. On définit $\tilde{X}_i = 2X_i - 1, 1 \leq i \leq d$. On introduit ainsi les 2 modèles considérés pour la Figure 6 :

— Modèle 1 : $n = 800$, $d = 2$, $Y = \tilde{X}_1^2 + \exp(-\tilde{X}_2^2)$;
— Modèle 2 : $n = 600$, $d = 100$, $Y = \tilde{X}_1\tilde{X}_2 + \tilde{X}_3^2 - \tilde{X}_4\tilde{X}_7 + \tilde{X}_8\tilde{X}_{10} - \tilde{X}_6^2 + \mathcal{N}(0, 0.5)$.

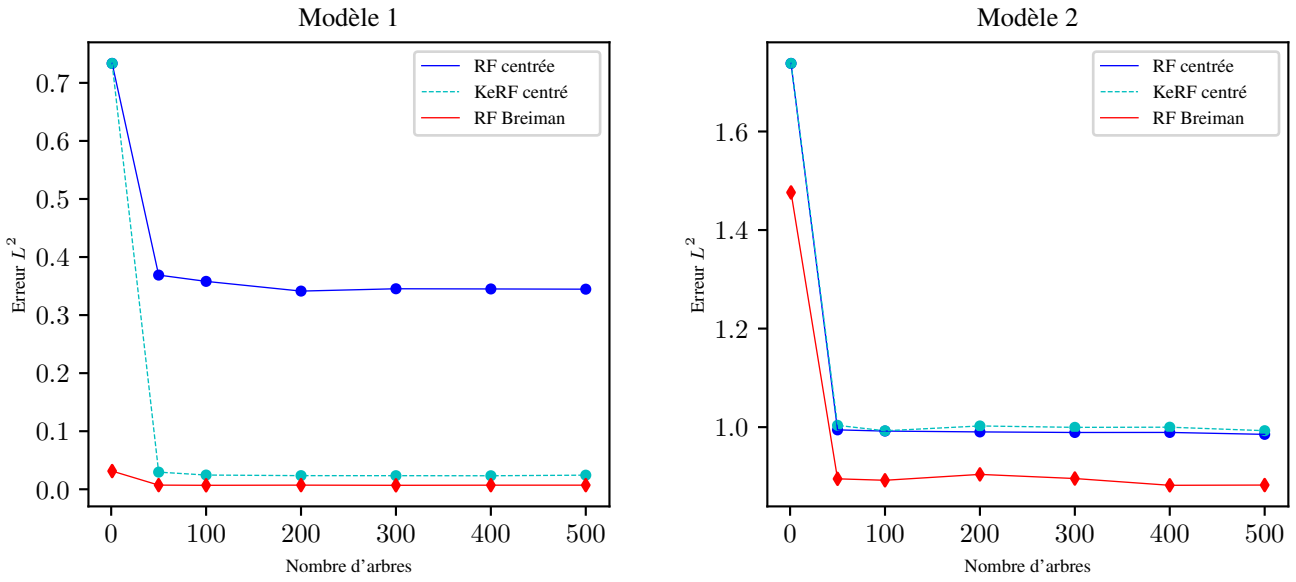


FIGURE 6 – Erreurs quadratiques empiriques de RF centrée, KeRF centré et BREIMAN évaluées sur les modèles 1 et 2

Observons dans un premier temps les performances des RF centrés et KeRF centrés. Dans le modèle 1 on remarque que nos estimateurs ont des résultats différents : KeRF apparaît nettement meilleur. Ceci peut notamment s'expliquer par le fait que les forêts centrées ne sont pas adaptatives. Pour le niveau k choisi, chaque arbre contient potentiellement des cellules vides altérant l'estimation. Ce travers est illustré en Figure 4. Cet effet indésirable disparaît dans l'estimation KeRF qui attribue le même poids à chaque observation comme nous l'avions montré dans l'exemple 5.
Une constatation satisfaisante est la proximité des KeRF et RF centrés avec les forêts de BREIMAN, ceci conforte l'intérêt des résultats théoriques obtenus sur la consistance et la borne supérieure de la vitesse de convergence.
Enfin remarquons que les forêts sont robustes dans le modèle 2 qui est sparse, $d = 100$ et seulement 8 variables sont actives. Ce fait est soulevé et analysé dans l'article [2].

5. Cependant ces résultats sont probablement améliorables, les algorithmes implémentés pouvant être optimisés davantage.

Démonstration. On rappelle que pour $\mathbf{x} \in [0,1]^d$, l'estimateur KeRF centré infini s'écrit :

$$\tilde{m}_\infty^c = \frac{\sum_{i=1}^n Y_i K^c(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n K^c(\mathbf{x}, \mathbf{X}_i)}.$$

Une méthode classique pour contrôler le risque L^2, est la décomposition biais-variance. Cependant le calcul de l'espérance de notre estimateur n'est pas évident. Nous allons détourner cette difficulté par une décomposition un peu différente et une disjonction de cas concernant le comportement du dénominateur.
Dans un premier temps on retranche artificiellement $\mathbb{E}\left[YK^c(\mathbf{x}, \mathbf{X})\right]$, $\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]$ au numérateur et au dénominateur respectivement. Et l'on multiplie de part et d'autre par $\frac{1}{\mathbb{E}[K^c(\mathbf{x}, \mathbf{X})]}$.

$$\tilde{m}_\infty^c = \frac{1}{n}\sum_{i=1}^n \left(\frac{\left(Y_i K^c(\mathbf{x}, \mathbf{X}_i) - \mathbb{E}\left[YK^c(\mathbf{x}, \mathbf{X})\right] + \mathbb{E}\left[YK^c(\mathbf{x}, \mathbf{X})\right]\right) \frac{1}{\mathbb{E}[K^c(\mathbf{x}, \mathbf{X})]}}{\left(K^c(\mathbf{x}, \mathbf{X}_i) - \mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right] + \mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]\right) \frac{1}{\mathbb{E}[K^c(\mathbf{x}, \mathbf{X})]}} \right) = \frac{M_n(\mathbf{x}) + A_n(\mathbf{x})}{1 + B_n(\mathbf{x})}.$$

où,

$$A_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i K^c(\mathbf{x}, \mathbf{X}_i)}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} - \frac{\mathbb{E}\left[YK^c(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} \right), B_n = \frac{1}{n}\sum_{i=1}^n \left(\frac{K^c(\mathbf{x}, \mathbf{X}_i)}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} - 1 \right), M_n = \frac{\mathbb{E}\left[YK^c(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]}.$$

Alors,

$$\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}) = \frac{M_n(\mathbf{x}) - m(\mathbf{x}) + A_n(x) - B_n(\mathbf{x})m(\mathbf{x})}{1 + B_n(\mathbf{x})}.$$

Nous allons à présent considérer deux événements complémentaires pour contrôler le risque selon les comportements de A_n et B_n.
Soit $\alpha \in]0, 1/2]$, on considère l'évènement $\mathcal{C}_\alpha(\mathbf{x}) = \{\{|A_n(\mathbf{x})|\} \leq \alpha \cap \{|B_n(\mathbf{x})|\} \leq \alpha\}$. Sous l'évènement $\mathcal{C}_\alpha(\mathbf{x})$, on a $1 + B_n(\mathbf{x}) \geq 1/2$.
De plus en appliquant l'inégalité de convexité $(a+b)^2 \leq 2a^2 + 2b^2 \ \forall a, b \in \mathbb{R}$, il vient :

$$\left(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x})\right)^2 \leq 8(M_n(\mathbf{x}) - m(\mathbf{x}))^2 + 8(A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x}))^2. \tag{14}$$

On s'intéresse d'abord au premier terme. On rappelle que par hypothèse, $Y = m(\mathbf{X}) + \epsilon$ avec $\mathbf{X} \sim \mathcal{U}([0,1]^d)$ et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ indépendant de $\mathbf{X}$.

$$\begin{aligned}
|M_n(\mathbf{x}) - m(\mathbf{x})| = \left| \frac{\mathbb{E}\left[m(\mathbf{X})K^c(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} + \frac{\mathbb{E}\left[\epsilon K^c(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} - m(\mathbf{x}) \right| &= \left| \frac{\mathbb{E}\left[m(\mathbf{X})K^c(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} - m(\mathbf{x}) \right| \qquad \text{(par indépendance entre } \epsilon \text{ et } \mathbf{X}) \\
&= \left| \frac{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\left(m(\mathbf{X}) - m(\mathbf{x})\right)\right]}{\mathbb{E}\left[K^c(\mathbf{x}, \mathbf{X})\right]} \right| \\
&= \left| \frac{\int_{[0,1]^d} K^c(\mathbf{x}, \mathbf{u})\left(m(\mathbf{u}) - m(\mathbf{x})\right) \mathrm{d}\mathbf{u}}{\int_{[0,1]^d} K^c(\mathbf{x}, \mathbf{u}) \mathrm{d}\mathbf{u}} \right|,
\end{aligned}$$

où l'on a noté par soucis d'allègement d'écriture $\mathrm{d}\mathbf{u} = \mathrm{d}u_1 \ldots \mathrm{d}u_d$. Or par hypothèse m est L-LIPSCHITZ pour la norme ℓ_1, i.e. $|m(\mathbf{u}) - m(\mathbf{x})| \leq \|\mathbf{u} - \mathbf{x}\|_1$.
Ainsi pour le terme au numérateur,

$$\left| \int_{[0,1]^d} K^c(\mathbf{x}, \mathbf{u})\left(m(\mathbf{u}) - m(\mathbf{x})\right) \mathrm{d}\mathbf{u} \right| \leq L \sum_{l=1}^d \int_{[0,1]^d} K^c(\mathbf{x}, \mathbf{u}) \left|u_l - x_l\right| \mathrm{d}\mathbf{u}$$

Par FUBINI-TONELLI et en injectant (12), il vient,

$$\begin{aligned}
&\leq L \sum_{l=1}^d \sum_{\substack{k_1, \ldots, k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \int_0^1 \mathbb{1}_{\left\{\lceil 2^{k_l} x_l \rceil = \lceil 2^{k_l} u_l \rceil\right\}} \left|u_l - x_l\right| \mathrm{d}u_l \\
&\qquad \times \prod_{j \neq l} \int_0^1 \mathbb{1}_{\left\{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} u_j \rceil\right\}} \mathrm{d}u_j
\end{aligned}$$

Or sous l'évènement $\mathbb{1}_{\left\{\lceil 2^{k_l} x_l \rceil = \lceil 2^{k_l} u_l \rceil\right\}}$, nous avons $|u_l - x_l| \leq 2^{-k_l}$,

$$\begin{aligned}
&\leq L \sum_{l=1}^{d} \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}\right)^k 2^{-k_l} \prod_{j=1}^{d} \int_0^1 \mathbb{1}_{\left\{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} u_j \rceil\right\}} \mathrm{d}u_j \\
&\leq L \sum_{l=1}^{d} \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}\right)^k 2^{-k_l} \prod_{j=1}^{d} 2^{-k_j} \\
&\leq L \sum_{l=1}^{d} \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}\right)^k 2^{-k_l} 2^{\sum_j k_j} \\
&\leq L \sum_{l=1}^{d} \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}\right)^k 2^{-(k+k_l)}
\end{aligned}$$

En remarquant par un simple calcul que $\int_0^1 \mathbb{1}_{\left\{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} u_j \rceil\right\}} \mathrm{d}u_j = 2^{-k_j}$ et par conséquent que le dénominateur $\int_{[0,1]^d} K^c(\mathbf{x}, \mathbf{u}) \, \mathrm{d}\mathbf{u} = 2^{-k}$:

$$|M_n(\mathbf{x}) - m(\mathbf{x})| \leq L \sum_{l=1}^{d} \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{2d}\right)^k 2^{-k_l}$$

Regardons le premier terme de la somme ci-dessus :

$$\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{2d}\right)^k 2^{-k_1} = \sum_{k_1=0}^{k} \left(\frac{1}{2d}\right)^{k_1} \left(1 - \frac{1}{d}\right)^{k-k_1} \frac{k_1!}{k_1!(k-k_1)!} \leq \left(1 - \frac{1}{2d}\right)^k.$$

Au final,

$$|M_n(\mathbf{x}) - m(\mathbf{x})| \leq Ld \left(1 - \frac{1}{2d}\right)^k.$$

Sous l'événement $\mathcal{C}_\alpha$, il est alors maintenant aisé de majorer le premier terme dans (14). Le deuxième terme se traite plus facilement, puisque m est bornée et que, sous $\mathcal{C}_\alpha$, $|A_n(\mathbf{x})| \leq \alpha$ et $|B_n(\mathbf{x})| \leq \alpha$, alors :

$$\left(A_n(\mathbf{x}) - B_n(\mathbf{x}) m(\mathbf{x})\right)^2 \leq \alpha^2 (1 + \|m\|_\infty)^2.$$

Par conséquent :

$$\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha\}}\right] \leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2 (1 + \|m\|_\infty)^2 \tag{15}$$

On considère maintenant $\mathcal{C}_\alpha^c(\mathbf{x})$ comme l'évènement complémentaire de $\mathcal{C}_\alpha(\mathbf{x})$. Il nous reste à travailler sur $\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right]$ pour obtenir une borne supérieure du risque L^2 :

$$\begin{aligned}
\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] &\leq \mathbb{E}\left[(\max_{1\leq i\leq n} Y_i + m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] && \text{(car } \tilde{m}_\infty^c \text{ est une moyenne pondérée des } Y_i\text{)} \\
&\leq \mathbb{E}\left[(\max_{1\leq i\leq n} (m(\mathbf{X}_i) + \epsilon_i) + m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] \\
&\leq \mathbb{E}\left[|2\|m\|_\infty + \max_{1\leq i\leq n} \epsilon_i|^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] \\
&\leq \left(\mathbb{E}\left(2\|m\|_\infty + \max_{1\leq i\leq n} \epsilon_i\right)^4 \mathbb{P}[\mathcal{C}_\alpha^c(\mathbf{x})]\right)^{1/2} && \text{(par CAUCHY-SCHWARZ)} \\
&\leq \left(\left(16\|m\|_\infty^4 + 8\mathbb{E}\left[\max_{1\leq i\leq n} \epsilon_i\right]^4\right) \mathbb{P}[\mathcal{C}_\alpha^c(\mathbf{x})]\right)^{1/2} && \text{(par inégalité de convexité).}
\end{aligned}$$

On utilise le fait suivant : il existe une constante $C > 0$ telle que $\forall n$

$$\mathbb{E}\left[\max_{1\leq i\leq n} \epsilon_i\right]^4 \leq C(\log n)^2.$$

On trouvera une preuve de ce fait par exemple dans le cours [5], Lemme 5, applicable aux variables aléatoires gaussiennes et donc en particulier aux variables ϵ_i.

Alors il existe une constante C_2 tel que pour tout $n > 1$,

$$\mathbb{E}[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))]^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}] \leq C_2(\log n)(\mathbb{P}[\mathcal{C}_\alpha^c(\mathbf{x})])^{1/2} \tag{16}$$

On travaille à présent sur la probabilité $\mathbb{P}[\mathcal{C}_\alpha^c(\mathbf{x})]$. On rappelle que $\mathcal{C}_\alpha(\mathbf{x}) = \{|A_n(\mathbf{x})|, |B_n(\mathbf{x})| \leq \alpha\}$, alors $\mathcal{C}_\alpha^c(\mathbf{x}) = \{|A_n(\mathbf{x})| \geq \alpha \text{ ou } |B_n(\mathbf{x})| \geq \alpha\}$.
D'après l'imégalité de Bienaymé Chebyshev :

$$\begin{aligned}
\mathbb{P}[|A_n(\mathbf{x})| > \alpha] = \mathbb{P}[|A_n(\mathbf{x})|^2 > \alpha^2] \leq \frac{\mathbb{V}ar[A_n(\mathbf{x})]}{\alpha^2} &\leq \frac{1}{\alpha^2}\mathbb{V}ar\left[\frac{1}{n}\sum_{i=1}^n\left(\frac{Y_i K^c(\mathbf{x},\mathbf{X}_i)}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]} - \frac{\mathbb{E}\left[YK^c(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]}\right)\right] \\
&\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\left(\frac{Y_i K^c(\mathbf{x},\mathbf{X}_i)}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]} - \frac{\mathbb{E}\left[YK^c(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]}\right)^2\right] \\
&\qquad \text{(en remarquant que } \mathbb{E}[A_n(\mathbf{x})] = 0 \text{ et par indépendance des } (\mathbf{X}_i, Y_i)_{1\leq i\leq n})
\end{aligned}$$

Par hypothèse, $Y \geq 0$, on omet donc le second terme dans l'espérance et on applique l'inégalité de JENSEN pour obtenir :

$$\begin{aligned}
&\leq \frac{1}{n\alpha^2}\frac{1}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2}\mathbb{E}\left[Y^2 K^c(\mathbf{x},\mathbf{X})^2\right] \\
&\leq \frac{1}{n\alpha^2}\frac{1}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2}\mathbb{E}\left[(m(\mathbf{x})+\epsilon)^2 K^c(\mathbf{x},\mathbf{X})^2\right] \\
&\leq \frac{2}{n\alpha^2}\frac{1}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2}\left(\mathbb{E}\left[m(\mathbf{x})^2 K^c(\mathbf{x},\mathbf{X})^2\right] + \mathbb{E}\left[\epsilon^2 K^c(\mathbf{x},\mathbf{X})^2\right]\right) && \text{(par inégalité de convexité)} \\
&\leq \frac{2(\|m\|_\infty^2 + \sigma^2)}{n\alpha^2}\frac{1}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2}\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right] && \text{(car } \sup_{\mathbf{x},\mathbf{z}\in[0,1]^d} K^c(\mathbf{x},\mathbf{z}) \leq 1) \\
&\leq \frac{2M_1^2}{\alpha^2}\frac{2^k}{n} && \text{(car } \int_{[0,1]^d} K^c(\mathbf{x},\mathbf{u})\,\mathrm{d}\mathbf{u} = 2^{-k})
\end{aligned}$$

où $M_1^2 = \|m\|_\infty^2 + \sigma^2$.
On majore la probabilité $\mathbb{P}[|B_n(\mathbf{x})| > \alpha]$ également grâce à l'inégalité de Bienaymé Chebyshev,

$$\begin{aligned}
\mathbb{P}[|B_n(\mathbf{x})| > \alpha] &\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\left(\frac{K^c(\mathbf{x},\mathbf{X})}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]}\right)^2\right] && \text{(en remarquent que } \mathbb{E}[B_n(\mathbf{x})] = 0) \\
&\leq \frac{1}{n\alpha^2}\frac{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})K^c(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2} \\
&\leq \frac{1}{n\alpha^2}\frac{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K^c(\mathbf{x},\mathbf{X})\right]^2} && \text{(car } \sup_{\mathbf{x},\mathbf{z}\in[0,1]^d} K^c(\mathbf{x},\mathbf{z})) \leq 1) \\
&\leq \frac{2^k}{n\alpha^2}. && \text{(car } \int_{[0,1]^d} K^c(\mathbf{x},\mathbf{u})\,\mathrm{d}\mathbf{u} = 2^{-k})
\end{aligned}$$

Ainsi par la borne union,

$$\mathbb{P}[\mathcal{C}_\alpha^c(\mathbf{x})] = \mathbb{P}[\{|A_n(\mathbf{x})|\rangle\,\alpha\} \cup \{|B_n(\mathbf{x})|\rangle\,\alpha\}] \leq \mathbb{P}[|A_n(\mathbf{x})| > \alpha] + \mathbb{P}[|B_n(\mathbf{x})| > \alpha] \leq \frac{2M_1^2}{\alpha^2}\frac{2^k}{n} + \frac{2^k}{n\alpha^2} \leq \frac{2^k(2M_1^2+1)}{n\alpha^2}.$$

En réinjectant cette dernière inégalité dans (16),

$$\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] \leq C_2(\log n)\left(\frac{2^k(2M_1^2+1)}{n\alpha^2}\right)^{1/2} \tag{17}$$

Nous avons à présent tous les éléments pour majorer le risque L^2.

$$\begin{aligned}
\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2\right] &= \mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha(\mathbf{x})\}}\right] + \mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2 \mathbb{1}_{\{\mathcal{C}_\alpha^c(\mathbf{x})\}}\right] \\
&\leq 8C_1^2\left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 + C_2(\log n)\left(\frac{2^k(2M_1^2+1)}{n\alpha^2}\right)^{1/2} && \text{(d'après (15) et (17)).}
\end{aligned}$$

En optimisant le terme de droite en α, il vient,

$$\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2\right] \leq 8C_1^2\left(1 - \frac{1}{2d}\right)^{2k} + C_3\left(\frac{(\log n)^2 2^k}{n}\right)^{1/3},$$

pour une constante $C_3 > 0$. En minimisant à présent en k, pour

$$k = C_4 + \frac{1}{\log 2 + \frac{3}{d}}\log\left(\frac{n}{(\log n)^2}\right),$$

où $C_4 = \left(\frac{1}{d} + \frac{\log 2}{3}\right)^{-1}\log\left(\frac{C_3 \log 2}{24 C_1^2}\right)$. Par conséquent, il existe une constante C_5 telle que, pour tout $n > 1$,

$$\mathbb{E}\left[(\tilde{m}_\infty^c(\mathbf{x}) - m(\mathbf{x}))^2\right] \leq C_5 n^{-\frac{1}{d\log 2 + 3}}(\log n)^2$$

□

6 Conclusion

Les forêts aléatoires sont des méthodes d'ensemble qui font "pousser" des arbres comme apprenants de base et combinent par la suite leurs prédictions en faisant une moyenne. Les forêts aléatoires sont connues pour leurs bonnes performances pratiques, notamment dans des contextes de grande dimension : elles allient rapidité, simplicité et estiment avec précision et ce sans sur-apprentissage. Nous avons également montré de manière empirique que les estimations KeRF se comparent favorablement aux estimations des forêts aléatoires.

Ce cours a donc été l'objet de l'introduction d'un nouvel estimateur. En particulier, en modifiant légèrement la définition des forêts aléatoires, ces dernières pouvant être réécrites comme des méthodes à noyau (KeRF) qui sont plus interprétables et plus faciles à analyser. Du point de vue théorique, plusieurs études ont mis en évidence le lien potentiellement fructueux entre les forêts aléatoires et les méthodes à noyau. Ce point reste tout de même une question ouverte à l'heure actuelle [1] [11], [6].

Le temps de calcul pour l'estimation du KeRF fini est très acceptable et similaire à celui de la forêt aléatoire. Cependant, la situation est différente pour les estimations infinies de KeRF. Les méthodes KeRF infinies doivent être considérées comme des outils théoriques plutôt que comme un substitut pratique aux forêts aléatoires.

Références

[1] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv :1407.3939*, 2014.

[2] Gérard Biau. Analysis of a random forest model. *Journal of machine learning research*, 13 :1063–1095, 2012.

[3] Gérard Biau and Luc Devroye. Cellular tree classifiers. In *International Conference on Algorithmic Learning Theory*, pages 8–17. Springer, 2014.

[4] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2) :197–227, 2016.

[5] Gérard Biau. Cours d'apprentissage statistique. 2021.

[6] Leo Breiman. Some infinity theory for predictor ensembles. Technical report, Citeseer, 2000.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[8] Leo Breiman. Consistency for a simple model of random forests. 2004.

[9] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[10] Marin Ferecatu & Michel Crucianu. Cours sur les méthodes à noyaux. `http://cedric.cnam.fr/vertigo/cours/ml2/coursMethodesNoyaux.html`.

[11] Serdar Demir and Öniz Toktamiş. On the adaptive nadaraya-watson kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, 39(3) :429–437, 2010.

[12] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[13] Robin Genuer and Jean-Michel Poggi. Arbres cart et forêts aléatoires, importance et sélection de variables. 2017.

[14] Maxime Sangnier. Cours d'introduction à l'apprentissage automatique. 2021.

[15] Erwan Scornet. *Learning with random forests*. Theses, Université Pierre et Marie Curie - Paris VI, November 2015.

[16] Erwan Scornet. Promenade en forêts aléatoires. *MATAPLI*, 111, 2016.

[17] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3) :1485–1500, 2016.

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