# Arbres et forêts aléatoires de Fréchet

Louis Capitaine et Robin Genuer
INSERM U1219 Bordeaux Population Health Research,
Université de Bordeaux, INRIA Bordeaux Sud-Ouest, équipe SISTM

21 avril 2019

**Résumé**

Les forêts aléatoires sont une méthode d'apprentissage statistique très largement utilisée dans de très nombreux domaines de recherche scientifique tant pour sa capacité à décrire des relations complexes entre des variables explicatives et une variable réponse que pour sa capacité à traiter des données de très grande dimension. Cependant, avec l'emergence de nouvelles techniques d'acquisition de données, nous avons accès à des données de plus en plus complexes, des images, des formes, des données longitudinales, des courbes et la méthode des forêts aléatoire n'est pas toujours adaptée à ces nouvelles entrées. Dans ce travail nous introduisons la notion de forêts aléatoires de Fréchet, qui permet d'apprendre des relations entre des variables de natures diverses dans des espaces métriques non ordonnés, et ce, même dans un cadre de grande dimension. Nous décrivons une nouvelle manière de découper les noeuds des arbres constituant notre forêt de Fréchet puis nous détaillons la procédure de prédiction pour une variable explicative à valeurs dans un espace non euclidien. Nous utilisons la structure des forêts afin de calculer une l'importance des variables constituant notre échantillon d'apprentissage. Nous terminons avec un exemple d'utilisation de nos Forêts de Fréchet pour la régression entre courbes à partir de leurs formes et nous donnons quelques simulations dans ce cadre.

**Abstract**

Random forests are a statistical learning method widely used in many areas of scientific research both for its ability to describe complex relationships between explanatory variables and a response variable and for its capacity to process very large data. However, with the rise of new data acquisition techniques, we have access to increasingly complex data such as images, shapes, curves, and random forests method is not always adapted to these new entries. In this work, we introduce the notion of Fréchet random forests, which allow to learn relationships between variables of various natures in non-ordered metric

spaces, even in high dimensional settings. We describe a new way of cutting the nodes of the trees constituting our Fréchet forests and then we detail the prediction procedure for an explanatory variable with values in a non-Euclidean space. We use forest structure to calculate the importance of the variables in our learning sample. We end with an example of the use of our Fréchet Forests for regression between curves from their shapes and we give some simulations in this context.

# 1 Arbres CART

Soit un $n$-échantillon $\mathcal{L}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ $i.i.d$ de même loi que $(X, Y)$ avec $X_i \in \mathcal{X}$ et $Y_i \in \mathcal{Y}$. L'algorithme **CART** (**C**lassification **A**nd **R**egression **T**rees) est une méthode d'apprentissage statistique introduite par Breiman en 1984 [2]. Le principe général de **CART** est de partitionner récursivement l'espace des variables explicatives $\mathcal{X} = \mathbb{R}^p$ de façon binaire. A chaque étape du partitionnement, on découpe une partie de l'espace en deux sous-parties en partant du noeud racine ($\mathbb{R}^p$ tout entier) contenant toutes les observations de l'échantillon d'apprentissage $\mathcal{D}_n$. Nous appelons coupure un élément de la forme

$$\{X^{(j)} \leq d\} \cup \{X^{(j)} > d\}$$

avec $j \in \{1, ..., p\}$ et $d \in \mathbb{R}$. La méthode sélectionne alors la meilleure découpe, c'est à dire le couple $(j, d)$ qui minimise une certaine fonction de coût :

- En régression, on cherche à minimiser la variance des noeuds fils. La variance d'un noeud $t$ est définie par $\sum_{i:X_i \in t} \left(Y_i - \bar{Y}_t\right)^2$ où $\bar{Y}_t$ est la moyenne des observations $Y_i$ dans le noeud $t$.
- En classification, où $\mathcal{Y} = \{1, ..., L\}$, on cherche à minimiser l'indice de Gini des noeuds fils. L'indice de Gini d'un noeud $t$ est défini par $\sum_{c=1,...,L} \widehat{p}_t^c(1 - \widehat{p}_t^c)$ où $\widehat{p}_t^c$ est la proportion d'observations de la classe $c$ dans le noeud $t$.

Une fois la racine de l'arbre découpée, on se restreint à chacun de ses noeuds fils et on recherche alors, suivant le même procédé la meilleure façon de le découper en deux nouveaux noeuds et ainsi de suite.
On veut étendre cette méthode au cas où $\mathcal{Y}$ est un espace métrique quelconque et $\mathcal{X}$ est un produit de $p$ espaces métriques quelconques non ordonnés.

## 2 Modèle

Soit un $n$-échantillon $\mathcal{L}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ *i.i.d* de même loi que $(X, Y)$.
On suppose que $Y \in (\mathcal{Y}, d_\mathcal{Y})$ un espace métrique et $X \in \mathcal{X} = (\mathcal{X}_1, d_1) \times .... \times (\mathcal{X}_p, d_p)$ un espace produit de $p$ espaces métriques non ordonnés *i.e.* $\forall x \in \mathcal{X},\ x = \left(x^{(1)}, ..., x^{(p)}\right)$ avec $x^{(j)} \in (\mathcal{X}_j, d_j)$ pour tout $j = 1, ..., p$. On dit que l'entier $p > 0$ est la dimension du problème. La dimension du problème ne coïncide pas forcément avec la dimension de l'espace des variables explicatives $\mathcal{X}$. Par exemple, dans le cas $p = 1$ où la seule variable explicative est à valeurs dans un espace de fonctions ($L^2([0, 1], \mathbb{R})$ par exemple) alors $\mathcal{X}$ est de dimension infinie. Un cas intéressant est celui où l'on s'intéresse à l'évolution dans le temps d'un phénomène en fonction d'autres variables évoluants elles aussi au cours du temps. Dans ce cas, on considèrera chaque $\mathcal{X}_j$ (ainsi que $\mathcal{Y}$) comme un espace de courbes paramétrées par le temps. On détaillera ce cas dans la section 5.

On suppose le modèle de régression non paramétrique suivant :

$$Y = f(X) + \epsilon \tag{1}$$

avec $f : \mathcal{X} \longrightarrow \mathcal{Y}$ une fonction inconnue et $\epsilon$ une variable aléatoire à valeurs dans $\mathcal{Y}$ centrée conditionnellement à $X$.
A la manière des arbres **CART**, on cherche à estimer la fonction $f$ en partitionnant récursivement l'espace des variables explicatives $\mathcal{X}$ de façon binaire. Cependant, nos espaces $\mathcal{X}_i$ ne disposent pas de notion de relation d'ordre mais seulement d'une métrique $d_i$ pour tout $i = 1, ..., p$. Ainsi, le critère de découpe des arbres **CART** est inutilisable dans notre cas car on ne peut pas décider si un élément est plus grand qu'un autre. Il nous faut donc changer le critère de découpe pour l'adapter à notre problème.

## 3 Fonction de split

La partie essentielle dans la construction d'un arbre **CART** est le découpage de ses noeuds. Découper un noeud $t$ d'un arbre selon une certaine variable $X^{(j)}$ revient à trouver une manière de regrouper les observations de ce noeud en deux sous-groupes en fonction des $X_i^{(j)}$ se trouvant dans le noeud $t$. Si la $j$ème variable $X^{(j)}$ explique fortement les variations de la variable de sortie $Y$ on s'attend alors à ce que si deux observations de la variable $X^{(j)}$ sont "proches" dans l'espace $(\mathcal{X}_j, d_j)$ (relativement aux autres observations pour la $j$ème variable) alors les deux observations de la variable réponse associée à ces mesures seront proches dans l'espace réponse $(\mathcal{Y}, d_\mathcal{Y})$ (relativement aux autres observations de la variable réponse du noeud $t$). Il nous faut alors pour chaque espace $\mathcal{X}_i$ trouver un moyen

de regrouper un ensemble de points de cet espace en deux sous-groupes les plus homogènes possibles, pour ce faire nous introduisons la notion de fonction de split.

**Définition** : Soit $(\mathcal{H}, d)$ un espace métrique. Une fonction de split $q$ est une application mesurable $q : \mathcal{H} \longrightarrow \mathcal{C} \subset \mathcal{H}$ telle que $Card\,(\mathcal{C}) = 2$.

Ainsi une fonction de split $q$ est entièrement définie par un alphabet $\mathcal{C} = \{c_1, c_2\}$ et une partition $\mathcal{A} = \{A_1, A_2\}$ et pour tout $x \in \mathcal{H}$, il existe $l \in \{1, 2\}$ tel que

$$q(x) = c_l \Leftrightarrow x \in A_l$$

La question qui se pose alors est "qu'est ce qu'une bonne fonction de split ?" On a besoin d'une mesure de la qualité d'une fonction de split $q$ et c'est ce qu'on appelle la distorsion.

**Définition (distorsion)** : La distorsion d'une fonction de split $q$ dans un espace métrique probabilisé $((\mathcal{H}, d), \mathcal{B}, \mathbb{P})$ est définie par :

$$D(\mathbb{P}, q) = \int_{\mathcal{H}} d^2\,(x, q(x))\,\mathbb{P}\,(dx)$$

où $\mathcal{B}$ est la tribu engendrée par la topologie de l'espace métrique $(\mathcal{H}, d)$.

La distorsion d'une fonction de split $q$ nous donne donc la qualité du découpage d'un espace métrique $(H, d)$ par $q$ en deux sous-parties. On sait que les fonctions de split qui minimisent la distortion sont celles de type plus proches voisins, c'est-à-dire les fonctions de split telles que le partitionnement en deux sous ensembles est une partition de Voronoï *i.e.*

$$\forall i \in \{1, 2\}, \forall x \in A_i \quad d\,(x, c_i) \leq d\,(x, c_j) \;\; j \neq i$$

**Découpage d'un noeud**

Soit $A$ une partie de $\mathcal{X}$ et $q_j : A_j \longrightarrow \mathcal{C}_j = \{c_{j,l}, c_{j,r}\}$ une fonction de split sur $(A_j, d_j)$ avec $A_j = \mathcal{X}_{j|A}$ pour tout $j = 1, ..., p$. On définit les noeuds fils droit et gauche issus du découpage sur la $j$ème variable par la fonction de split $q_j$ par

$$A_{j,r} = \{x \in A;\; q_j(x^{(j)}) = c_{j,r}\} \quad \text{et} \quad A_{j,l} = \{x \in A;\; q_j(x^{(j)}) = c_{j,l}\}$$

On définit la qualité du découpage de la $j$ème variable par la fonction de split de type plus proches voisins (**PPV**) $q_j$ par la fonction $L_n$ qui mesure la diminution de la variance de Fréchet pour ce découpage :

$$L_n\,(j) = \frac{1}{N_n(A)} \sum_{i:X_i \in A} d^2_{\mathcal{Y}}\left(Y_i, \bar{Y}_A\right) - \frac{1}{N_n(A)} \sum_{i:q_j(X_i^{(j)}) = c_{j,r}} d^2_{\mathcal{Y}}\left(Y_i, \bar{Y}_{A_{j,r}}\right) - \frac{1}{N_n(A)} \sum_{i:q_j(X_i^{(j)}) = c_{j,l}} d^2_{\mathcal{Y}}\left(Y_i, \bar{Y}_{A_{j,l}}\right)$$

avec $\bar{Y}_A$, $\bar{Y}_{A_{j,l}}$ et $\bar{Y}_{A_{j,r}}$ les moyennes de Fréchet des observations dans les noeuds $A, A_{j,l}$ et $A_{j,r}$

$$\bar{Y}_A = \underset{z\in\mathcal{Y}}{\arg\min} \sum_{i:X_i\in A} d^2_{\mathcal{Y}}(z, Y_i) \quad \bar{Y}_{A_{j,l}} = \underset{z\in\mathcal{Y}}{\arg\min} \sum_{i:q(X_i^{(j)})=c_{j,l}} d^2_{\mathcal{Y}}(z, Y_i) \quad (\text{resp } A_{j,r})$$

et

$$N_n(A) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i\in A\}} \quad \forall A \subseteq \mathcal{X}$$

La variable de découpe retenue est celle qui maximise la fonction $L_n$

$$j_n^* = \underset{j\in\{1,\dots,p\}}{\arg\max}\ L_n(j)$$

Le découpage optimal de la partie $A \subset \mathcal{X}$ retenu est donné par les sous parties $A_{j_n^*,l}$ et $A_{j_n^*,r}$. On applique récursivement cette méthode partant de $\mathcal{X}$ tout entier (le noeud racine) afin de construire une partition de l'espace $\mathcal{X}$.

# 4 Prédiction

Soit $(\Omega_i)_{i=1}^K$ la partition en $K$ parties obtenue en utilisant de manière récursive la méthode de découpage par des fonctions de split de type **PPV** donnée dans la section précédente. L'estimateur de la fonction $f$ est alors donné par

$$\widehat{f}_n = \sum_{j=1}^{K} \widehat{\beta}_j \mathbb{1}_{\Omega_j}$$

avec

$$\widehat{\beta}_j = \underset{z\in\mathcal{Y}}{\arg\min} \sum_{i:(X_i,Y_i)\in\Omega_j} d^2_{\mathcal{Y}}(z, Y_i)$$

la moyenne empirique de Fréchet (lorsqu'elle existe) des observations $Y_i$ tombant dans la partie $A_j$. Pour toute nouvelle observation $x \in \mathcal{X}$, on prédit la variable réponse associée à $x$ par

$$\widehat{y} = \widehat{f}_n(x)$$

## 5 Forêts aléatoires de Fréchet

On appelle forêt aléatoire de Fréchet l'aggrégation de prédicteurs par arbres de Fréchet randomisés. Le principe des forêts aléatoires de Fréchet reste le même que celui des forêts de Léo Breiman [1]. D'abord on génère $k$ échantillons de bootstrap $\mathcal{L}_n^{\Theta_1}, ..., \mathcal{L}_n^{\Theta_k}$ où la variable aléatoire $\Theta_l$ représente le $l$ème tirage. Ensuite, une variante des arbres de Fréchet est appliquée sur chaque sous échantillon $\mathcal{L}_n^{\Theta_l}$. Plus précisément, on construit un arbre de la manière suivante : pour découper un noeud, on tire aléatoirement $m$ variables et on cherche la meilleure coupure selon ces $m$ variables tirées. On note $\Theta'_l$ la variable aléatoire qui représente le tirage de ces $m$ variables pour l'arbre $l$.
Soient $\widehat{f}_n(., \Theta_1, \Theta'_1), ..., \widehat{f}_n(., \Theta_k, \Theta'_k)$ nos $k$ arbres de Fréchet randomisés, on définit le prédicteur par forêt aléatoire de Fréchet par

$$\widehat{f}_n^{RF}(x) = \underset{z \in \mathcal{Y}}{\arg\min} \sum_{i=1}^{k} d_{\mathcal{Y}}^2 \left( z, \widehat{f}_n(x, \Theta_i, \Theta'_i) \right) \quad \forall x \in \mathcal{X}$$

## 6 Régression sur courbes paramétrées

On suppose le modèle (1) où $Y$ est une courbe paramétrée de $\mathcal{I} \subset \mathbb{R} \longrightarrow \mathbb{R}$ ; $X$ est une courbe paramétrée de $\mathcal{J} = \mathcal{J}_1 \times ... \times \mathcal{J}_p \longrightarrow \mathbb{R}^p$ où les $\mathcal{J}_i \subset \mathbb{R}$ peuvent être disjoints et $\epsilon$ est un bruit blanc Gaussien. Dans ce cas, la fonction inconnue $f$ que l'on souhaite estimer part de $\mathcal{CP}(\mathcal{I}, \mathbb{R})$ et est à valeurs dans $\mathcal{CP}(\mathcal{J}_1 \times ... \times \mathcal{J}_p, \mathbb{R})$ où $\mathcal{CP}(A, B)$ désigne l'espace des courbes paramétrées de $A$ dans $B$.

Dans ce cas, une métrique intéressante à considérer est la distance de Fréchet. On utilisera alors comme fonction de split la méthode des 2-means de Genolini et al [3] pour des courbes paramétrées. Nous axerons la présentation autour de cette problématique et présenterons une analyse de simulation de nos arbres de Fréchet pour l'analyse de courbes paramétrées.

## Références


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

[2] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[3] Christophe Genolini, René Ecochard, Mamoun Benghezal, Tarak Driss, Sandrine Andrieu, and Fabien Subtil. kmlshape : An efficient method to cluster longitudinal data (time-series) according to their shapes. *Plos one*, 11(6) :e0150738, 2016.