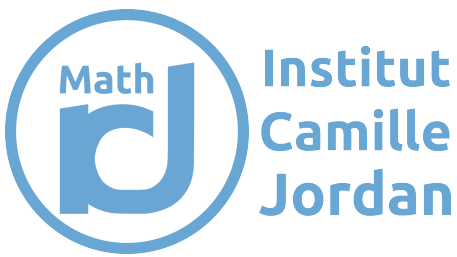
Goal oriented random forest (GORF).

Véronique Maume-Deschamps
joint work with
Kevin Elie-Dit-Cosaque (SCOR SE), Bérénice-Alexia Jocteur (Natixis, ICJ), Clémentine Prieur (LJK - UGA), Pierre Ribereau (ICJ - UCBL), Ri Wang (ICJ).

17 novembre 2023



Plan

Plan

Why considering GORF?

Random Forests are mainly designed for Regression or Classification purposes. In these cases, the target is observed. Other aims may be pursued, e.g.:

- conditional quantiles
- conditional average treatment effects (CATE)
- other conditional risk measure such as expectiles
- ...

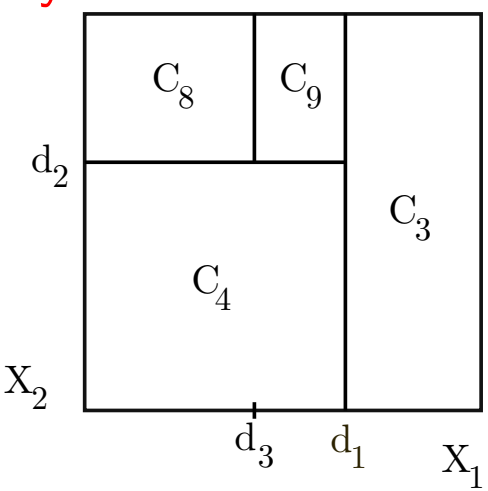
$\Longrightarrow$ Distributional Random Forest[1] vs Goal Oriented Random Forest.

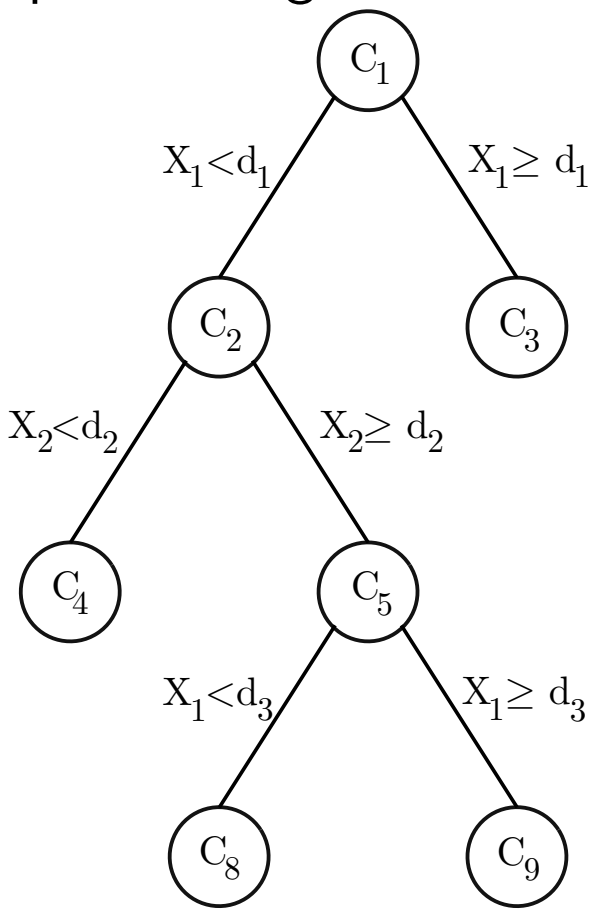
[1] Domagoj Ćevid et al. (2022). In: *Journal of Machine Learning Research*
Qiming Du et al. (2021). In: *International Conference on Artificial Intelligence and Statistics*

Recall CART

Classification And Regression Tree[2]. Input variables: $\mathbf{X} = (X_1, \ldots, X_d)$, Output variable: Y. Tree: constant piecewise predictor, obtained by binary recursive partitioning.

Separate the data from the current node, by looking for the split reducing the most the heterogeneity of Y at the two child nodes.





[2] Leo Breiman (2001). In: *Machine learning*

Loss function

In the regression context, $Y = m(\mathbf{X}) + \varepsilon$, the goal is to estimate $\mathbb{E}(Y|\mathbf{X}) = m(\mathbf{X})$. Consider a random sample $\mathcal{D}_n = (\mathbf{X}^i, Y^i), i = 1, \ldots, n$, the heterogeneity of Y is measured by the intra-groups variance, so that we shall maximise:

$$\mathcal{L}_C^n(j,z) = \frac{1}{\#C} \sum_{i=1}^{n} \left(Y^i - \overline{Y}_C\right)^2 \mathbf{1}_{\{\mathbf{X}^i \in C\}} - \left[\frac{1}{\#C} \sum_{i=1}^{n} \left(Y^i - \overline{Y}_{C_L}\right)^2 \mathbf{1}_{\{\in C_L\}} + \frac{1}{\#C} \sum_{i=1}^{n} \left(Y_i - \overline{Y}_{C_R}\right)^2 \mathbf{1}_{\{\mathbf{X}_i \in C_R\}}\right],$$

where C is the current cell, ie an hyper-rectangle $\prod_{j=1}^{d} [a_j, b_j]$, $\#C$ is the number of elements of $\mathcal{D}_n$ for which $\mathbf{X}^i$, $i = 1, \ldots, n$ belongs to C; $C_L = C \cap \{x_j \leq z\}$; $C_R = C \cap \{x_j > z\}$.

Random Forests

Agregate several CART's to reduce the estimation variance $\Longrightarrow$
Bootstrap aggregating

- Training sample: $\mathcal{D}_n = (\mathbf{X}^i, Y^i)\,, i = 1, \ldots, n$
- $\Theta_\ell, \ell = 1, \ldots, k$ are independent random variables, following $\Theta = (\Theta^1, \Theta^2)$' law: Θ^1 provides the bootstrap indices on $\mathcal{D}_n$ and Θ^2 gives which mtry variables are considered for the splits of each node. Θ_ℓ is assumed to be independent of $\mathcal{D}_n$.
- $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$: the leaf that is obtained when dropping $\mathbf{x}$ down the tree.
- $N_n(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$: the number of points which are in $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$.
- $N_n^b(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$: the number of points of the bootstrapped sample, which are in $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$.

CART estimation

In the regression framework, $\mathbb{E}(Y|\mathbf{X}) = m(\mathbf{X})$ is estimated on the random forest by. Let $B_j(\Theta_\ell, \mathcal{D}_n)$ be the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been drawn from the original dataset for the ℓ-th tree construction. Consider the weights:

$$\omega_{n,i}(\mathbf{x}, \Theta) = \frac{1}{k} \sum_{j=1}^{k} \frac{\mathbf{1}_{\mathbf{X}^i \in A_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)}}{N_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)},$$

$$\omega_{n,i}^b(\mathbf{x}, \Theta) = \frac{1}{k} \sum_{\ell=1}^{k} \frac{B_i(\Theta_\ell, \mathcal{D}_n)\, \mathbf{1}_{\mathbf{X}^i \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)}}{N_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)},$$

and the corresponding estimations of $m(\mathbf{x})$:

$$\hat{m}_n^b(\mathbf{x}) = \sum_{i=1}^{n} \omega_{n,i}^b(\mathbf{x})\, Y^i.$$

Focus on causality and quantile estimation

- Conditional quantile estimation : $\alpha \in]0,1[$, the conditional quantile $q_\alpha(Y|\mathbf{X})$ could be estimated by inverting the estimated conditional distribution function: and the corresponding estimations of $F(y|\mathbf{X}=\mathbf{x})$ (introduced in Meinshausen 2006 and a.s. consistency proved in Elie-Dit-Cosaque and Maume-Deschamps 2022):

$$\hat{F}_n^b(y|\mathbf{X}=\mathbf{x}) = \sum_{i=1}^{n} \omega_{n,i}^b(\mathbf{x})\, \mathbf{1}_{\{Y^i \leqslant y\}}.$$

- Conditional Average Treatement Effect

Focus on causality and quantile estimation

- Conditional quantile estimation
- Conditional Average Treatement Effect:
  $CATE(\mathbf{X}) = \mathbb{E}(Y(1) - Y(0)|\mathbf{X})$ where $Y(W)$ is a target variable (eg some biological quantity) in presence ($Y(1)$) or absence ($Y(0)$) of a treatement W. Some causal models write: $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$, in this case, $CATE(\mathbf{X}) = \tau(\mathbf{X})$. Modifications on the CART construction has to be done in order to estimate $CATE$, since for each individual, we observe either $Y^i(1)$ or $Y^i(0)$.

Other forests

- Generalized Random Forests[3] consider θ the target (eg conditional quantiles or *CATE* and use the following loss function:

$$\frac{\#C_L \#C_R}{\#C^2} \left[\hat{\theta}_{C_L} - \hat{\theta}_{C_R}\right]^2$$

  where $\hat{\theta}_C$ is an estimator of the target θ on the cell C.
- Distributional Random Forests[4] use Maximal Mean Discrepancy as loss function or the Wasserstein distance [5] for the construction of the split criterion.
- Trees designed for Extreme Value Analysis[6].

[3] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

[4] Domagoj Ćevid et al. (2022). In: *Journal of Machine Learning Research*

[5] Qiming Du et al. (2021). In: *International Conference on Artificial Intelligence and Statistics*

[6] Sébastien Farkas et al. (2021). In: *arXiv preprint arXiv:2112.10409*

Plan

Goal oriented loss function

We propose loss functions specifically designed for

- *CATE* estimation,
- Conditional quantile estimation,

with a common proof scheme for a.s. consistency.

HTERF

Heterogeneous Treatment Effect Random Forest[7]

Consider the loss function:

$$\mathcal{L}_C(j,z) = \frac{\#C_L \#C_R}{\#C^2}\left(\left(\overline{Y}_{C_{L1}} - \overline{Y}_{C_{L0}}\right) - \left(\overline{Y}_{C_{R1}} - \overline{Y}_{C_{R0}}\right)\right)^2, \quad (1)$$

where $C_{L1} = \{\mathbf{X}^i \in C_L, W^i = 1\}$, $C_{L0} = \{\mathbf{X}^i \in C_L, W^i = 0\}$,
$C_{R1} = \{\mathbf{X}^i \in C_R, W^i = 1\}$, $C_{R0} = \{\mathbf{X}^i \in C_R, W^i = 0\}$,

[7] Bérénice-Alexia Jocteur, Véronique Maume-Deschamps, and Pierre Ribereau (2023). In: *https://hal.science/hal-04112079*

HTERF estimation

For the estimation of *CATE* we use:

$$\widehat{CATE}_{(}\mathbf{x}) = \sum_{i:W^i=1} \omega_{n,i}\left(\mathbf{x}, \Theta\right) Y^i - \sum_{i:W^i=0} \omega'_{n,i}\left(\mathbf{x}, \Theta\right) Y^i, \qquad (2)$$

where ω (resp. ω') are the weights associated to observations such as $W^i = 1$ (resp. $W^i = 0$).

Pin-ball loss

Consider the pin-ball function[8]: $\psi_\alpha(y,\theta) = (y-\theta)(\alpha - \mathbf{1}_{\{y\leq\theta\}})$
Recall that the α-quantile is given by:
$q^\alpha(Y) = \arg\min_\theta \mathbb{E}[\psi_\alpha(Y,\theta)]$.
Consider the loss function[9]:

$$\mathcal{L}_C(j,z) = \sum_{i=1}^{n} \psi_\tau(Y^i, \hat{\theta}_C)\mathbf{1}_{\mathbf{X}^i\in C} - \left[\sum_{i=1}^{n} \psi_\tau(Y^i, \hat{\theta}_{C_L})\mathbf{1}_{\mathbf{X}^i\in C_L} + \sum_{i=1}^{n} \psi_\tau(Y^i, \hat{\theta}_{C_R})\mathbf{1}_{\mathbf{X}^i\in C_R}\right],$$

where $\hat{\theta}_C$ is an estimator of the α-quantile in C.
Estimate the conditional distribution function then the quantile as before.

[8] It is also named check function, quantile loss.

[9] Harish S Bhat, Nitesh Kumar, and Garnet J Vaz (2015). In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE

Plan

Consistency of random forests

Results by Scornet, Biau, Vert (2015) in a linear model context:

$$Y = m(X) + \varepsilon \text{ with } \varepsilon \rightsquigarrow \mathcal{N}(0, \sigma^2) \text{ and } m(X) = \sum_{j=1}^{d} m_j(X_j).$$

Under various assumptions including tree size wrt n and a forest correlation control, for $\mathbf{X} \rightsquigarrow \mathcal{U}[0,1]^d$,

$$\mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] \longrightarrow 0, \text{ with } m_n = \mathbb{E}_\Theta(\hat{m}_n).$$

- The bootstrap is not taken into account in Θ
- No results for $m(\mathbf{x})$
- Results for fully grown trees and for limited grown trees.

Consistency of GRF

Asymptotic normal laws obtained[10] under

- regularity asumptions of the target function,
- constraints on the tree construction

[10] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

Consistency of GRF

Asymptotic normal laws obtained[10] under

- regularity asumptions of the target function, which write in the *CATE* estimation case:
  $\mathbf{x} \mapsto \mathbb{E}[Y(u)|\mathbf{X} = \mathbf{x}]$ and $\mathbf{x} \mapsto \mathbb{E}[Y(u)^2|\mathbf{X} = \mathbf{x}]$ are Lipschitz-continuous, $Var[Y(u)|\mathbf{X} = \mathbf{x}] > 0$ and $\mathbb{E}[|Y(u) - \mathbb{E}[Y(u)|\mathbf{X} = \mathbf{x}]|^{2+\delta}|\mathbf{X} = \mathbf{x}] \leq M$ for some constants $\delta, M > 0$ uniformly over all $\mathbf{x} \in [0, 1]^d$.
- constraints on the tree construction

[10] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

Consistency of GRF

Asymptotic normal laws obtained[10] under

- regularity asumptions of the target function,
- constraints on the tree construction:
  - at every step of the tree building procedure, the probability that the next split is done along the $j-th$ feature is bounded below by π/d for some $0 < \pi \leq 1$ for all $j = 1, \ldots, d$ random split hypothesis.
  - for some fixed γ, each split leaves at least a fraction γ of the available training sample on each side of the split, γ-regularity hypothesis
  - for some fixed p, the leaf containing $\mathbf{x}$ has at least p observations for each treatment group and the leaf containing $\mathbf{x}$ has either less than $2p-1$ observations with $W^i = 0$ or $2p-1$ observations with $W^i = 1$.

[10] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

Consistency of GRF

Asymptotic normal laws obtained[10] under

- regularity asumptions of the target function,
- constraints on the tree construction

The bootstrap is not taken into account, the proof is done for honest forests (ie independance of the sample used for the tree construction and the estimation).

[10] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

General scheme for the a.s. consistency

Conditions

Relations between k (number of trees) and $N_n^b(\mathbf{x}; \Theta, \mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$.
2. $\forall \mathbf{x}, \quad N_n^b(\mathbf{x}; \Theta, \mathcal{D}_n) = \Omega\left(\sqrt{n}(\ln(n))^\beta\right)$, with $\beta > \frac{5}{2}$, *a.s.*[a] *or an asumption on mean and variance of* $N_n^b(\mathbf{x}; \Theta, \mathcal{D}_n)$

The variation of the target function θ is small on the trees' leaves: for any $\mathbf{x}$,

$$\sup_{\mathbf{z},\mathbf{z}' \in A_n(\mathbf{x}, \Theta_j)} \left|\theta(\mathbf{z}) - \theta(\mathbf{z}')\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

[a] $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

Consistency: result

Theorem

Assume the conditions above are verified then

$$|\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0$$

Proof for conditional distributions[11] and *CATE*[12] but the proof scheme applies more generally.

[11] Kevin Elie-Dit-Cosaque and Véronique Maume-Deschamps (2022). In: *Electronic Journal of Statistics*

[12] Bérénice-Alexia Jocteur, Véronique Maume-Deschamps, and Pierre Ribereau (2023). In: *https://hal.science/hal-04112079*

Remark on the variation hypothesis

In order to get the variation hypothesis we need the continuity of θ and either

- the random split and γ-regularity hypothesis or,
- the convergence of the empirical loss function to the theoretical one (with some uniformity on the cells C) and the fact that the theoretical loss function is 0 on a cell C of a theoretical tree[13] implies that θ is zero on C. This last condition is true e.g. in the *CATE* estimation setting for a large class of functions (including sums, products, dense classes).

[13] introduced in Scornet, Biau, and Vert 2015 and deeply used in Elie-Dit-Cosaque and Maume-Deschamps 2022

Theroetical trees

A theoretical tree is grown following the same rules as an empirical tree, except that the theoretical equivalent of the empirical split criterion on a node C is used to choose the best split. E.g. for the CART-tree, the theoretical split criterion is:

$$\begin{aligned}\mathcal{L}_C^\star(j,z) =& \operatorname{Var}\left(Y|\mathbf{X}\in C\right)\\ &-\mathbb{P}\left(\mathbf{X}\in C_L|\mathbf{X}\in C\right)\operatorname{Var}\left(Y|\mathbf{X}\in C_L\right)\\ &-\mathbb{P}\left(\mathbf{X}\in C_R|\mathbf{X}\in C\right)\operatorname{Var}\left(Y|\mathbf{X}\in C_R\right)\ .\end{aligned}$$

Hence, a theoretical tree is obtained thanks to the best consecutive cuts $(j^\star, z^\star)$, among $j\in\mathcal{M}_{\text{try}}$, $z\in C^j$ optimizing the previous criterion $\mathcal{L}_C^\star(\cdot,\cdot)$.

Empirical vs theoretical trees

Consider the model $Y = m(\mathbf{X}) + \varepsilon$ with $m(\cdot)$ in the ♠-class[14] and ε with ligth tails.

Proposition

For any $h \in \mathbb{N}$ fixed, for any empirical tree with node sizes greater than $C\sqrt{n}(\ln n)^{\beta}$, $\beta > \frac{5}{2}$, consider a node at height h in the theoretical tree (resp. empirical tree) and $\mathcal{T}_h$ the set of theoretical trees of height h, then

$$\text{let } A = \prod_{j=1}^{d}[a_j, b_j] \text{ and } A^n = \prod_{j=1}^{d}[a_j^n, b_j^n]. \text{ We have:}$$

$$\inf_{\mathcal{T}_h} \max_{j=1,\dots,d} \max\left(|a_j - a_j^n|, |b_j - b_j^n|\right) \longrightarrow 0 \text{ a.s. as } n \to \infty.$$

[14] Functions in this class satisfy that if $\mathcal{L}_C^{\star}(j,z) = 0$ for all j, z the m is constant on C.

Empirical vs theoretical trees

Proposition

For any $h \in \mathbb{N}$ fixed, for any empirical tree with node sizes greater than $C\sqrt{n}(\ln n)^{\beta}$, $\beta > \frac{5}{2}$, consider a node at height h in the theoretical tree (resp. empirical tree) and $\mathcal{T}_h$ the set of theoretical trees of height h, then

$$\text{let } A = \prod_{j=1}^{d}[a_j, b_j] \text{ and } A^n = \prod_{j=1}^{d}[a_j^n, b_j^n]. \text{ We have:}$$

$$\inf_{\mathcal{T}_h} \max_{j=1,\dots,d} \max\left(|a_j - a_j^n|, |b_j - b_j^n|\right) \longrightarrow 0 \text{ a.s. as } n \to \infty.$$

Proved for CART trees[14] and for causal forests[15].

[14] Kevin Elie-Dit-Cosaque and Véronique Maume-Deschamps (2022). In: *Electronic Journal of Statistics*

[15] Bérénice-Alexia Jocteur, Véronique Maume-Deschamps, and Pierre Ribereau (2023). In: *https://hal.science/hal-04112079*

The two samples method.

One of the main idea to prove the constistency of random forests is to use an auxiliary sample: let $(\mathbf{X}^{i\diamond}, Y^{i\diamond}, i = 1, \ldots n)$ be a second sample, independent from $(\mathbf{X}^i, Y^i, i = 1, \ldots, n)$ and consider the weights

$$\omega_{n,i}^{\diamond}(\mathbf{x}, \Theta) = \frac{1}{k} \sum_{j=1}^{k} \frac{\mathbf{1}_{\mathbf{X}^{i\diamond} \in A_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)}}{N_n^{\diamond}(\mathbf{x}, \Theta_j, \mathcal{D}_n)}$$

and the corresponding estimator $\theta_n^{\diamond}$ of θ. We prove:

1. $\left|\hat{\theta}_n(\mathbf{x}) - \theta_n^{\diamond}(\mathbf{x})\right| \xrightarrow[n\to\infty]{a.s.} 0$, uses a Hoeffding like inequality + Vapnik-Chervonenkis classes[16] (proximity of $N^{\diamond}$ and N^b),
2. $|\theta_n^{\diamond}(\mathbf{x}) - \theta(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0$, uses Vapnik-Chervonenkis classes again and the variation hypothesis.

[16] V. N. Vapnik and A. Ya. Chervonenkis (1971). In: *Theory of Probability and its Applications*

Plan

CATE estimation: a first example

We consider simulated data close to causal frameworks previously studied [17]. $\mathbf{X} \sim U([0,1]^p)$, $W \sim Bern(0.5)$ and $Y = \tau(\mathbf{X})W + \beta\gamma(\mathbf{X})$, $p = 10, \tau(\mathbf{x}) = \sin(x_1)$ and $\gamma(\mathbf{x}) = \cos(2x_2 + 3x_3)$. The scalar β allows to consider the impact of the magnitude of τ relative to γ.

β	GRF	HTERF
5	0.276	0.117
1	0.122	0.012
0.2	0.079	0.004

Table: Mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect, with 500 tree forests.

[17] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

CATE estimation: interpretability

	GRF				HTERF			
β	dep.3	dep.5	dep.10	imp.	dep.3	dep.5	dep.10	imp.
5	0.870	0.378	0.150	0.852	1	0.498	0.175	0.985
1	0.874	0.526	0.174	0.866	1	0.995	0.282	1
0.2	0.875	0.627	0.2	0.866	1	1	0.603	1

Table: Frequencies of splitting on X_1 at depths 3, 5 and 10 and importance of X_1.

CATE estimation: a non linear framework

Let $\mathbf{X} \sim U([0,1]^p)$, $W \sim Bern(0.5)$ and $Y = \sin(X_1)(W+2)^3 + \cos(X_2)$, where $p = 3$. Hence we have CATE that satisfies: $\tau(\mathbf{x}) = 19\sin(x_1)$.

Method	RMSE	importance
GRF	0.321	0.777
HTERF	0.209	1

Table: Root mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect, with 500 tree forests. We also consider the importance of X_1.

Conditional quantiles estimation: two examples[18]

Example 1:
$Y|\mathbf{X} \sim \mathcal{N}(0, 1 + \mathbf{1}(X_1 > -0.5))$, $p = 10$, $X_j, j = 1, 2, ..., 10$ are independent draws from uniform distribution $\mathcal{U}(-1, 1)$,
$\mathbf{x}_i = (x_i, 0, 0, ..., 0)$, and x_i is taken from regular grid over $[-1, 1]$.

Example 2: $Y = X_1 - X_2 + \varepsilon$, $X_j, j = 1, 2, ..., 10$ are independent draws from $X_j \sim Exp(1)$, $\varepsilon \sim \mathcal{N}(0, 1)$,
$\mathbf{x}_i = (x_{i1}, x_{i2}, 1, 1, ..., 1)$, and (x_{i1}, x_{i2}) are taken from regular grid over $[0, 5]^2$.

[18] Véronique Maume-Deschamps, Clémentine Prieur, and Ri Wang.

Conditional quantiles estimation: Example 1

α	n	GRF	Quantile06	DRF	Pin-ball
0.1	500	0.197	0.219	0.107	**0.085**
	2000	0.038	0.140	**0.027**	0.028
	4000	0.025	0.110	0.019	**0.016**
0.5	500	**0.015**	0.016	0.017	0.035
	2000	**0.007**	**0.007**	0.009	0.010
	4000	**0.006**	**0.006**	0.010	0.010
0.9	500	0.205	0.221	0.112	**0.083**
	2000	0.037	0.137	0.025	**0.022**
	4000	0.026	0.119	0.017	**0.014**

Table: MSE[19] for Example 1: $Y|\mathbf{X} \sim \mathcal{N}(0, 1 + \mathbf{1}(X_1 > -0.5))$

[19] 100 repetitions, number of trees $B = 200$, `min.node.size` $= 20$, `mtry` $= p$. Other parameters use the default setting.

Conditional quantiles estimation: Example 2

τ	n	GRF	Quantile06	DRF	Pin-ball
0.1	500	5.700	2.679	3.074	**1.629**
	2000	4.754	1.770	1.949	**0.763**
	4000	3.959	1.133	1.373	**0.474**
0.5	500	1.848	1.166	1.713	**0.817**
	2000	0.761	0.405	0.644	**0.383**
	4000	0.411	0.243	0.341	**0.235**
0.9	500	5.869	2.833	3.022	**1.812**
	2000	4.842	1.717	1.809	**0.847**
	4000	4.093	1.201	1.332	**0.497**

Table: MSE[20] for Example 2: $Y = X_1 - X_2 + \varepsilon$

[20] 100 repetitions, number of trees $B = 200$, `min.node.size` $= 20$, `mtry` $= p$. Other parameters use the default setting.

Plan

Conclusion.

- Use Goal Oriented Random Forest for specific purposes:
  - better estimation (in general)
  - better interpretability.
- Results developped for *CATE* and conditional quantile estimations but other quantities are reachable.
- General scheme for a.s. consistency results.
- Theoretical random forest = a powerfull tool to better understand the random forests methods.

References I

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Bhat, Harish S, Nitesh Kumar, and Garnet J Vaz (2015). "Towards scalable quantile regression trees". In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 53–60.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Ćevid, Domagoj et al. (2022). "Distributional random forests: Heterogeneity adjustment and multivariate distributional regression". In: *Journal of Machine Learning Research* 23.333, pp. 1–79.

Du, Qiming et al. (2021). "Wasserstein random forests and applications in heterogeneous treatment effects". In: *International Conference on Artificial Intelligence and Statistics*, pp. 1729–1737.

Elie-Dit-Cosaque, Kevin and Véronique Maume-Deschamps (2022). "Random forest estimation of conditional distribution functions and conditional quantiles". In: *Electronic Journal of Statistics* 16, 6553–6583.

Farkas, Sébastien et al. (2021). "Generalized Pareto Regression Trees for extreme events analysis". In: *arXiv preprint arXiv:2112.10409*.

Jocteur, Bérénice-Alexia, Véronique Maume-Deschamps, and Pierre Ribereau (2023). "Heterogeneous Treatment Effect based Random Forest: HTERF". In: *https://hal.science/hal-04112079*.

Maume-Deschamps, Véronique, Clémentine Prieur, and Ri Wang. "Work in progress".

Meinshausen, Nicolai (2006). "Quantile regression forests". In: *Journal of Machine Learning Research* 7.Jun, pp. 983–999.

References II

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert (2015). "Consistency of random forests". In: *The Annals of Statistics*.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities". In: *Theory of Probability and its Applications* 16.2, pp. 264–280.

Merci

Thanks for your attention.

Merci pour votre attention.