Comparing CART trees using subsampling

Pierre-Yves Boëlle[1], **Felix Cheysson**[2],
Olivier Lopez[2]

[1] Institut Pierre-Louis d'Epidémiologie et de Santé Publique
[2] Sorbonne Université, LPSM

EcoDep Conference 2022
June 24th 2022

Outline

Modelling motivation

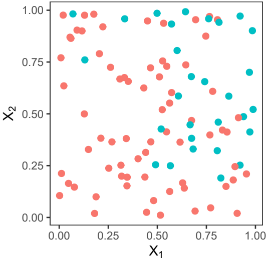
Covid-19 dataset from EDS database: all hospitalised patients in AP-HP hospitals with a diagnosis of Covid-19 (PCR analysis or lung X-ray).

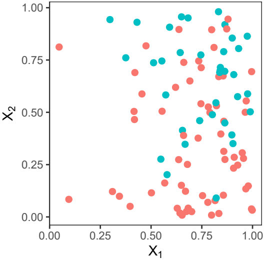
dt.first	dt.last	outcome	sex	age	diabetes
2020-03-17	2020-04-05	alive	F	45	no
2020-03-14	2020-03-25	alive	F	29	no
2020-03-18	2020-03-29	death	H	80	no
2020-03-11	2020-03-15	death	H	62	no
2020-03-04	2020-03-09	death	F	72	yes
2020-03-16	2020-03-20	death	H	92	no

- **Motivation**: Identify the risk factors of the disease, the groups at risk, and determine whether they evolve through the pandemic.
- **Objective**: Improve patient management and care when changes in the vulnerability of groups at risks are detected.

Statistical framework: supervised learning

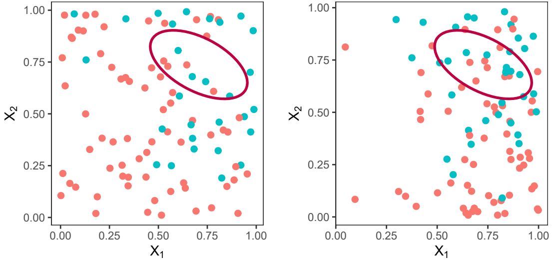
- Independent learning sets $\mathbb{X} = \{X_i\}_{1 \leq i \leq m}$ and $\mathbb{Y} = \{Y_j\}_{1 \leq j \leq n}$, where X_i and Y_j are doublets $(u, v) \in \mathcal{U} \times \{0, 1\}$ with d.f. P_X and P_Y.





Statistical framework: supervised learning

- Independent learning sets $\mathbb{X} = \{X_i\}_{1 \leq i \leq m}$ and $\mathbb{Y} = \{Y_j\}_{1 \leq j \leq n}$, where X_i and Y_j are doublets $(u, v) \in \mathcal{U} \times \{0, 1\}$ with d.f. P_X and P_Y.

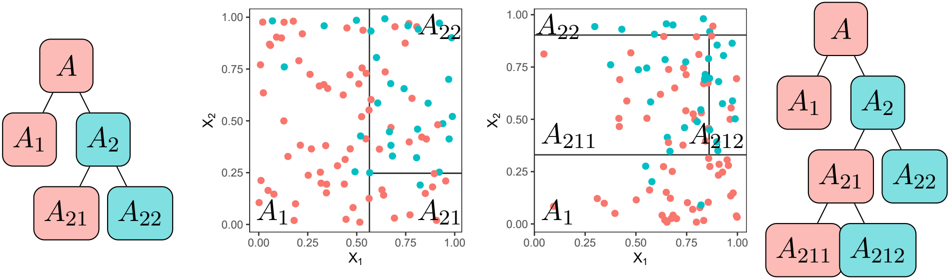


Statistical motivation

Compare $\mathbb{E}_{P_X}[V \mid U = u]$ and $\mathbb{E}_{P_Y}[V \mid U = u]$ for arbitrary $u \in \mathcal{U}$.

Statistical framework: supervised learning

- Independent learning sets $\mathbb{X} = \{X_i\}_{1 \le i \le m}$ and $\mathbb{Y} = \{Y_j\}_{1 \le j \le n}$, where X_i and Y_j are doublets $(u, v) \in \mathcal{U} \times \{0, 1\}$ with d.f. P_X and P_Y.
- Decision tree $T_{\mathbb{X}} = T(X_1, \ldots, X_m)$ generated from sample $\mathbb{X}$.
- $T(u) \in [0, 1]$ denotes the prediction of the tree T at $u \in \mathcal{U}$.



Statistical motivation

Compare $\mathbb{E}_{P_X}[V \mid U = u]$ and $\mathbb{E}_{P_Y}[V \mid U = u]$ for arbitrary $u \in \mathcal{U}$.

Classification and Regression Trees (CART)

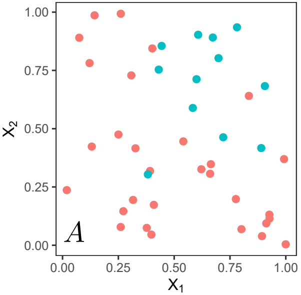
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{U}$ along one of the covariate dimensions:
  - Find the node A, the dimension d and the value z such that the split (A, d, z) maximises the decrease in impurity:

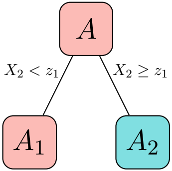
$$\Delta i(A, d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

Extensions include randomised ensembles: random forests, bagging, etc.



Classification and Regression Trees (CART)



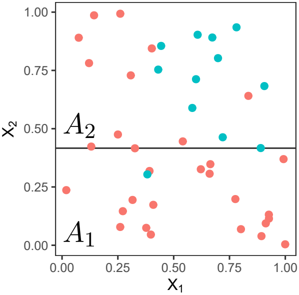
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{U}$ along one of the covariate dimensions:
  - Find the node A, the dimension d and the value z such that the split (A, d, z) maximises the decrease in impurity:

$$\Delta i(A, d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

Extensions include randomised ensembles: random forests, bagging, etc.



Classification and Regression Trees (CART)

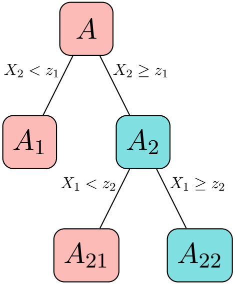
Introduced by Breiman et al., 1984.

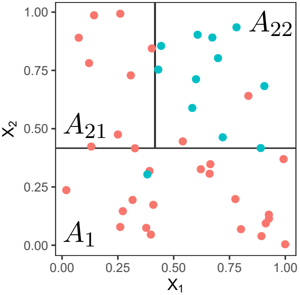
- Construct binary tree by recursively splitting the sample space $\mathcal{U}$ along one of the covariate dimensions:
  - Find the node A, the dimension d and the value z such that the split (A, d, z) maximises the decrease in impurity:

    $$\Delta i(A, d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

Extensions include randomised ensembles: random forests, bagging, etc.



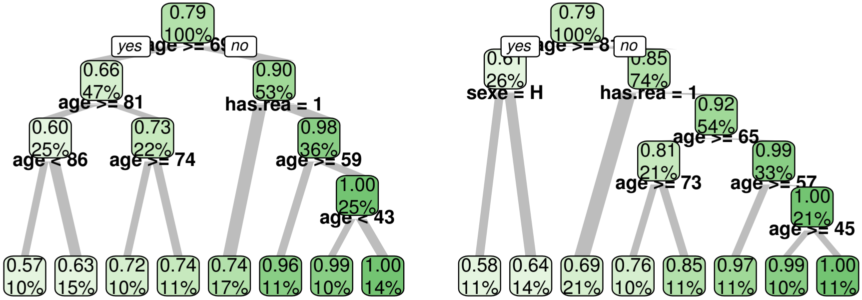


About CART

- Handles missing data, interpretable (particularly for MDs).
- Theoretical properties:
  - Breiman et al., 1984: $\mathbb{L}^2$-consistency of tree structured regression and classification, though not pointwise.
  - Gey and Nedelec, 2005; Gey, 2012: Non asymptotic risk for pruned procedure.
- Empirical results (variance of prediction):
  - Bar-Hen, Gey, and Poggi, 2015: influence functions derived from robust estimation theory.
  - Wager, Hastie, and Efron, 2014: variance of bagged predictors.

CART and learning set

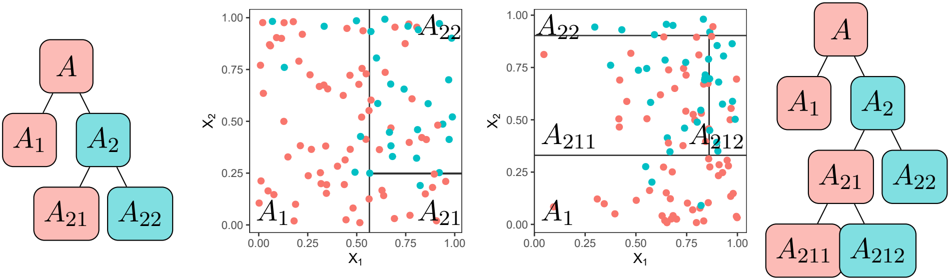
Problem: CART are sensitive to perturbations in the learning set.



- Study predictions rather than structure of the tree: what is the variance associated with the sampling of the learning set?

Hypothesis test for the comparison of trees

- Independent learning sets $\mathbb{X} = \{X_i\}_{1 \leq i \leq m}$ and $\mathbb{Y} = \{Y_j\}_{1 \leq j \leq n}$, where X_i and Y_j are doublets $(u, v) \in \mathcal{U} \times \{0, 1\}$ with d.f. P_X and P_Y.
- Decision tree $T_{\mathbb{X}} = T(X_1, \ldots, X_m)$ generated from sample $\mathbb{X}$.
- $T(u) \in [0, 1]$ denotes the prediction of the tree T at $u \in \mathcal{U}$.



Hypothesis test for the comparison of trees

- Independent learning sets $\mathbb{X} = \{X_i\}_{1 \leq i \leq m}$ and $\mathbb{Y} = \{Y_j\}_{1 \leq j \leq n}$, where X_i and Y_j are doublets $(u, v) \in \mathcal{U} \times \{0, 1\}$ with d.f. P_X and P_Y.
- Decision tree $T_{\mathbb{X}} = T(X_1, \ldots, X_m)$ generated from sample $\mathbb{X}$.
- $T(u) \in [0, 1]$ denotes the prediction of the tree T at $u \in \mathcal{U}$.
- (Dis)similarity between $T_{\mathbb{X}}$ and $T_{\mathbb{Y}}$ at a collection of test points $(u_1, \ldots, u_t)$ via the kernel h

$$h(\mathbb{X}; \mathbb{Y}) = \sum_{i=1}^{t} d\big(T_{\mathbb{X}}(u_i), T_{\mathbb{Y}}(u_i)\big).$$

- Null hypothesis $\mathcal{H}_0 : \forall u \in \mathcal{U},\ \mathbb{E}_{P_X}[V \mid U = u] = \mathbb{E}_{P_Y}[V \mid U = u]$.
- **Question**: What is the d.f. of $h(\mathbb{X}; \mathbb{Y})$ under $\mathcal{H}_0$?

U-statistics for CART

- Mentch and Hooker, 2016; Peng, Coleman, and Mentch, 2019.
- Base learners $h(X_1, \ldots, X_r; Y_1, \ldots, Y_s)$ on subsamples of size r and s.
- Bagging predictions from ensemble method yields a U-statistic

$$U_{m,n,r,s} = \binom{m}{r}^{-1} \binom{n}{s}^{-1} \sum_{(m,r)} \sum_{(n,s)} h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}).$$

- Extension to incomplete U-Statistics

$$U_{m,n,r,s,N} = N^{-1} \sum_{(m,r)} \sum_{(n,s)} \rho_{ij}\, h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}),$$

with (ρ_{ij}) a multinomial with N trials and probabilities $1/\binom{m}{r}\binom{n}{s}$.

A gentle reminder on U-statistics

- Introduced by Halmos, 1946 and Hoeffding, 1948.
- Generalisation of the mean to sum of dependent variables.
- Suppose we are interested in the expected value of a kernel h which is permutation symmetric in its r arguments:

$$\theta = \mathbb{E}h(X_1, \dots, X_r).$$

- For an *i.i.d.* sample $(X_1, \dots, X_n)$, define the *U-statistic with kernel* h:

$$U_n = \binom{n}{r}^{-1} \sum_{(n,r)} h(X_{i_1}, \dots, X_{i_r}).$$

- Examples of U-statistics: sample mean and variance, signed rank statistic, Mann-Whitney statistic ($\mathbb{P}(X < Y)$), ...

A gentle reminder on U-statistics (cont'd)

By projecting U_n on the space $\mathcal{S}_1$ (*Hájek projection*),

$$\mathcal{S}_1 = \left\{ \sum_{i=1}^{n} g_i(X_i) : \mathbb{E} g_i^2(X_i) < \infty \right\},$$

it can be shown that:

Theorem

If $\mathbb{E} h^2(X_1, \ldots, X_r) < \infty$, then

$$\sqrt{n}(U_n - \theta) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, r^2 \zeta_1),$$

where

$$\begin{aligned}\zeta_1 &= \mathrm{Var}\big(\mathbb{E}[h(X_1, X_2, \ldots, X_r) \mid X_1] - \theta\big) \\ &= \mathbb{E}\big[h(X_1, X_2, \ldots, X_r) h(X_1, X_2', \ldots, X_r')\big] - \theta^2.\end{aligned}$$

U-statistics for CART

- Mentch and Hooker, 2016; Peng, Coleman, and Mentch, 2019.
- Base learners $h(X_1, \ldots, X_r; Y_1, \ldots, Y_s)$ on subsamples of size r and s.
- Bagging predictions from ensemble method yields a U-statistic

$$U_{m,n,r,s} = \binom{m}{r}^{-1} \binom{n}{s}^{-1} \sum_{(m,r)} \sum_{(n,s)} h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}).$$

- Extension to incomplete U-Statistics

$$U_{m,n,r,s,N} = N^{-1} \sum_{(m,r)} \sum_{(n,s)} \rho_{ij}\, h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}),$$

with (ρ_{ij}) a multinomial with N trials and probabilities $1/\binom{m}{r}\binom{n}{s}$.

U-statistics for CART

Theorem

Suppose

- $m/(m+n) \to \lambda \in [0,1]$ (relative proportion of samples),
- $r/m \sim s/n \to 0$ (size of subsamples),
- $N = o(n/s)$ (number of subsamples).

Let $\theta_{r,s} = \mathbb{E}h$, $\zeta_{r,s} = \mathrm{Var}(h)$ and assume $\mathbb{E}h^6 < \infty$. Then

$$\frac{U_{m,n,r,s,N} - \theta_{r,s}}{\sqrt{\zeta_{r,s}/N}} \xrightarrow{d} \mathcal{N}(0,1).$$

- Stronger rate of convergence possible with some conditions on the conditional moments of h (Hoeffding decomposition).
- Possible to establish CLTs for the sample p-quantiles of h.

Bootstrap approximation

Under $\mathcal{H}_0$, the parameters $\theta_{r,s}$ and $\zeta_{r,s}$ are unknown.

Idea: Generate a bootstrap approximation $h(\mathbb{X}^*; \mathbb{Y}^*)$ to the d.f. of $h(\mathbb{X}; \mathbb{Y})$ under $\mathcal{H}_0$:

- Build average predictions $\bar{T}$ under the null:

$$\bar{T}(u) = \frac{m}{m+n} T_{\mathbb{X}}(u) + \frac{n}{m+n} T_{\mathbb{Y}}(u);$$

- Generate bootstrapped trees $T^*_{\mathbb{X}}$ and $T^*_{\mathbb{Y}}$ of sizes r and s resp.:
  - Subsample the inputs: u^*;
  - Draw $v^* \sim B(\bar{T}(u^*))$;
  - Build the tree T^* on $\{(u^*, v^*)\}$, using the same control parameters as the subsampled trees.
- Estimate $\theta_{r,s}$ and $\zeta_{r,s}$ with the $\{h(\mathbb{X}^*, \mathbb{Y}^*)\}$.

Numerical experiments

Generative model $\mathcal{M}$ for both $\mathbb{X}$ and $\mathbb{Y}$:

- Continuous variable (age):
  $U_1 = p_a U_a + (1 - p_a) U_b$, where
  $U_a \sim \mathcal{N}(\mu_a, \sigma_a)$ and $U_b \sim \mathcal{N}(\mu_b, \sigma_b)$;
- Discrete variable (gender): $U_2 \sim B(p_f)$;
- Binary outcome (death): $V \mid U_1, U_2 \sim B(p_d)$,

$$\mathrm{logit}(p_d) = \beta_0 + \beta_1 U_1 + \beta_2 U_2.$$

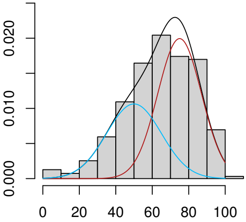
Scenarii, $n = m = 1,000$, $r = s = 50$, $N = 50$:

$\mathcal{H}_0$ Test points $(u_1, \dots, u_t)$ are generated from $\mathcal{M}$;

$\mathcal{H}_0'$ Test points $(u_1, \dots, u_t)$ are generated from $\mathcal{M}$ with $p_a = .85$;

$\mathcal{S}_1$ As for $\mathcal{H}_0'$, with $\beta_1 = 0.06$ (higher risk for old);

$\mathcal{S}_2$ As for $\mathcal{H}_0'$, with $\beta_2 = 0.7$ (higher risk for male).



Param.	Value
μ_a	80
σ_a	13
μ_b	55
σ_b	16
p_a	0.45
p_f	0.5
β_0	-4.85
β_1	0.05
β_2	0.55

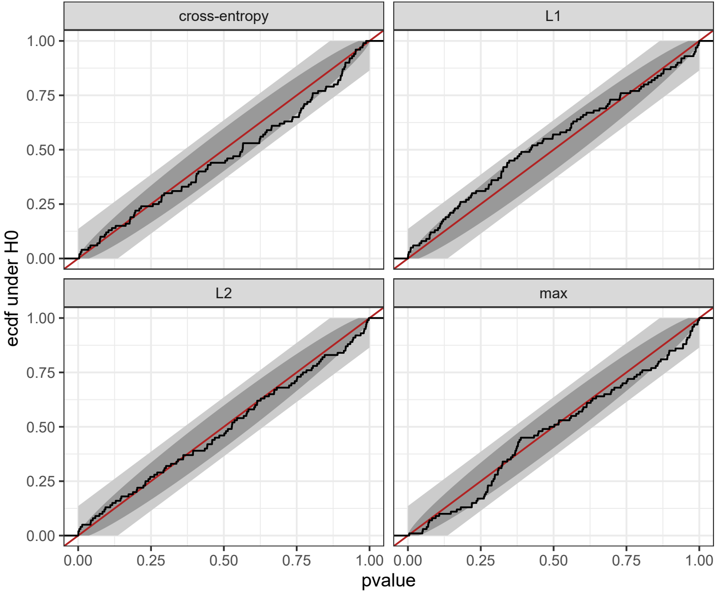
Choice of function d

(Dis)similarity between $T_{\mathbb{X}}$ and $T_{\mathbb{Y}}$ at a collection of test points $(u_1, \ldots, u_t)$ via the kernel (d a distance)

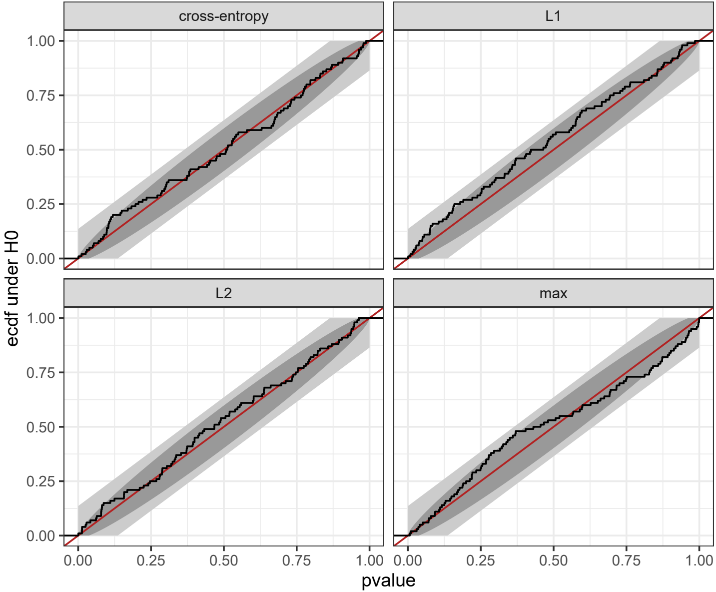
$$h(\mathbb{X}; \mathbb{Y}) = \sum_{i=1}^{t} d\big(T_{\mathbb{X}}(u_i), T_{\mathbb{Y}}(u_i)\big).$$

- Based on $\mathbb{L}^2$-consistency:
  - L2: $d(p, q) = (p - q)^2$;
  - L1: $d(p, q) = |p - q|$;
- cross: $d(p, q) = -p \log q - q \log p$;
- max: $h(\mathbb{X}; \mathbb{Y}) = \sup_{1 \leq i \leq t} |T_{\mathbb{X}}(u_i) - T_{\mathbb{Y}}(u_i)|$.

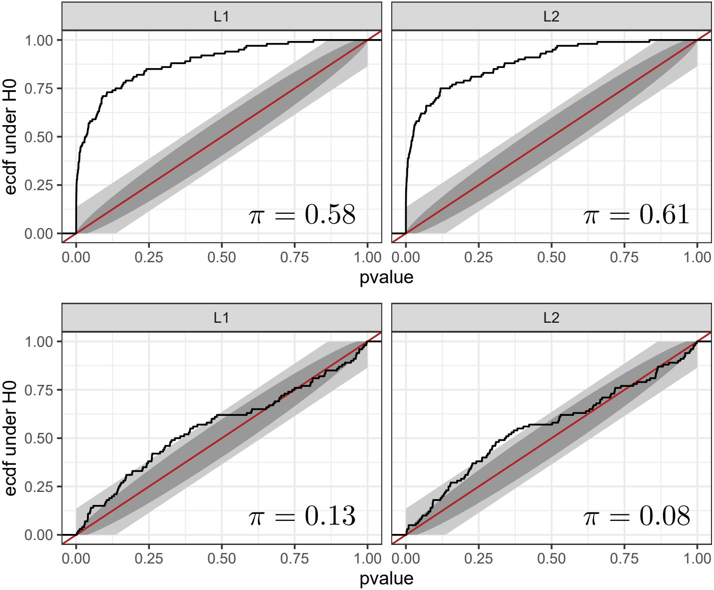
p-values for 100 simulations under $\mathcal{H}_0$



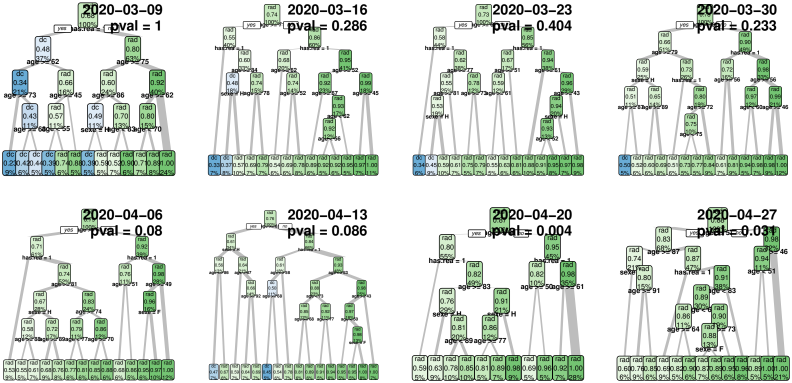
p-values for 100 simulations under $\mathcal{H}_0'$

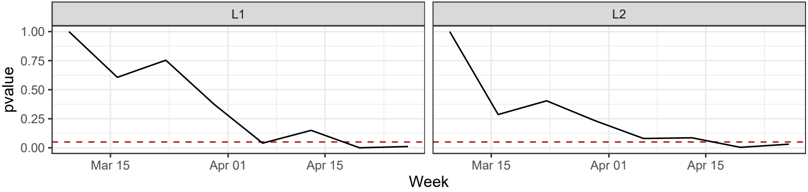


p-values for 100 simulations for scenarii $\mathcal{S}_1$ and $\mathcal{S}_2$



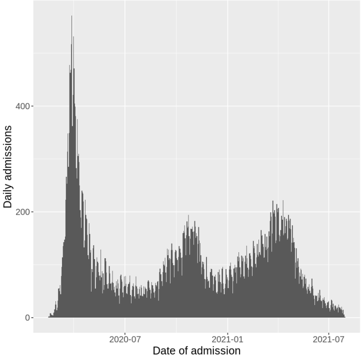
Applications to death rates during first wave





Comparing death rates for the first three waves

- Data from AP-HP's EDS (*Entrepôt de Santé*), covering 39 hospitals.
- Pandemic waves occurring:
  - From mid-March to end of June 2020;
  - From early-Sept. to end of Nov. 2020;
  - From early-Feb. to end of May 2021.



	Healthy < 50 y.o.		Elderly > 60 y.o.	
	Rate	p-value	Rate	p-value
1st wave	0.029	—	0.214	—
2nd wave	0.019	0.34	0.184	< 0.01
3rd wave	0.015	0.61	0.216	0.03

Conclusion and perspectives

Some methodological innovations

- Two-sample test of conditional expectations,
- Adapted to decision trees and ensemble methods,
- And distributional results for the test statistic under $\mathcal{H}_0$.

Some crucial perspectives

- Convergence of the bootstrap approximation for the d.f. of the test statistic under $\mathcal{H}_0$.
- Under $\mathcal{H}_1$: find the test points $(u_1, \dots, u_t)$ so that the test is most powerful.
- Full application to complex Covid-19 data, including more explanatory covariates.

Thank you for your attention.

For Further Reading I

- Bar-Hen, Avner, Servane Gey, and Jean-Michel Poggi (2015). "Influence Measures for CART Classification Trees". In: *J. Classif.* 32.1, pp. 21–45. ISSN: 0176-4268. DOI: 10.1007/s00357-015-9172-4. arXiv: 1610.08203.
- Biau, Gérard and Erwan Scornet (2016). "A random forest guided tour". In: *TEST* 25.2, pp. 197–227. ISSN: 1133-0686. DOI: 10.1007/s11749-016-0481-7. arXiv: 1511.05741.
- Breiman, Leo et al. (1984). *Classification and regression trees*. The Wadsworth statistics / probability series. CRC, p. 366. ISBN: 0-412-04841-8.
- Fernholz, Luisa Turrin (1983). *von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics. New York: Springer-Verlag, p. 133. ISBN: 9788578110796. arXiv: arXiv:1011.1669v3.

For Further Reading II

- Gey, Servane (2012). "Risk bounds for CART classifiers under a margin condition". In: *Pattern Recognit.* 45.9, pp. 3523–3534. ISSN: 00313203. DOI: 10.1016/j.patcog.2012.02.021. arXiv: 0902.3130.
- Gey, Servane and Elodie Nedelec (2005). "Model Selection for CART Regression Trees". In: *IEEE Trans. Inf. Theory* 51.2, pp. 658–670. ISSN: 0018-9448. DOI: 10.1109/TIT.2004.840903.
- Halmos, Paul R. (1946). "The Theory of Unbiased Estimation". In: *Ann. Math. Stat.* 17.1, pp. 34–43. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731020.
- Hoeffding, Wassily (1948). "A Class of Statistics with Asymptotically Normal Distribution". In: *Ann. Math. Stat.* 19.3, pp. 293–325. ISSN: 0003-4851. DOI: 10.1214/aoms/1177730196.

For Further Reading III

- Lopes, Miles E., Suofei Wu, and Thomas C. M. Lee (2020). "Measuring the Algorithmic Convergence of Randomized Ensembles: The Regression Setting". In: *SIAM J. Math. Data Sci.* 2.4, pp. 921–943. DOI: 10.1137/20m1343300. arXiv: 1908.01251.
- Mayer, Michael (2009). "U-Quantile-Statistics". In: pp. 1–9. arXiv: 0906.1266.
- Mentch, Lucas and Giles Hooker (2016). "Quantifying uncertainty in random forests via confidence intervals and hypothesis tests". In: *J. Mach. Learn. Res.* 17, pp. 1–41. ISSN: 15337928. arXiv: 1404.6473.
- Peng, Wei, Tim Coleman, and Lucas Mentch (2019). "Asymptotic Distributions and Rates of Convergence for Random Forests via Generalized U-statistics". In: DOI: 10.1214/19-EJS1643. arXiv: 1905.10651.

For Further Reading IV

- Scornet, Erwan, Gerard Biau, and Jean Philippe Vert (2015). "Consistency of random forests". In: *Ann. Stat.* 43.4, pp. 1716–1741. ISSN: 00905364. DOI: 10.1214/15-AOS1321. arXiv: 1405.2881.
- Vaart, A. W. van der (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, p. 460.
- Wager, Stefan (2014). "Asymptotic Theory for Random Forests". In: pp. 1–17. arXiv: 1405.0352.
- Wager, Stefan, Trevor Hastie, and Bradley Efron (2014). "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife.". In: *J. Mach. Learn. Res.* 15.1, pp. 1625–1651. ISSN: 1532-4435.

For Further Reading V

Wolfson, Julian and Ashwini Venkatasubramaniam (2018). "Branching Out: Use of Decision Trees in Epidemiology". In: *Curr. Epidemiol. Reports* 5.3, pp. 221–229. ISSN: 2196-2995. DOI: 10.1007/s40471-018-0163-y.

U-statistic for the hypothesis test (cont'd)

Proof follows Peng, Coleman, and Mentch, 2019:

- *Hoeffding decomposition*: study the variance by projecting $U_{m,n,r,s}$ on the pairwise orthogonal spaces $S_{i,j}$ of square-integrable functions, of the form

$$\mathcal{S}_{i,j} = \left\{ \sum_{(m,i)} \sum_{(n,j)} g_{i,j}(X_{\alpha_1}, \dots, X_{\alpha_i}; Y_{\beta_1}, \dots, Y_{\beta_j}) \right\}.$$

- We have that $r\zeta_{1,0} \leq \zeta_{r,s}$, similarly for $\zeta_{0,1}$: r and s must be chosen such that the assumption is valid.
- Example: for the one-sample OLS estimator, $(r\zeta_1)^{-1}\zeta_s \to 1$ (Peng, Coleman, and Mentch, 2019).