Evolution of groups at high risk of death from Covid-19 using hospital data

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

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

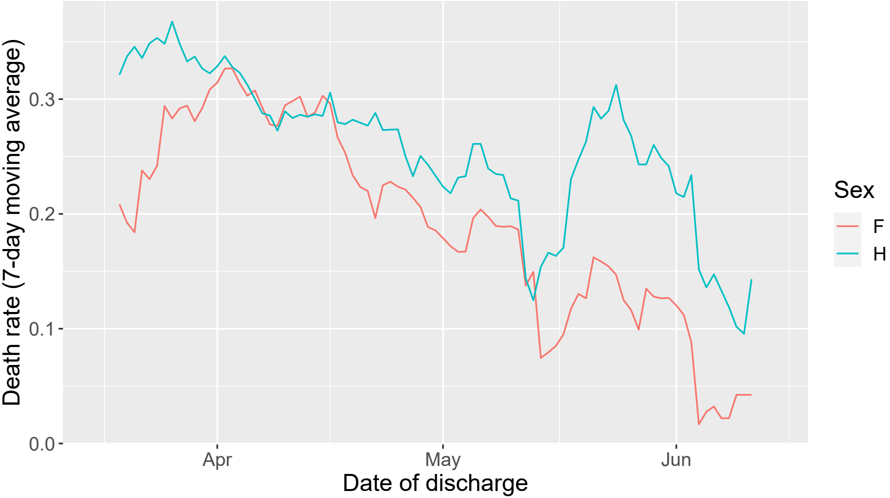
Séminaire MIA Paris
December 6th 2021

Outline

1. Motivation
2. Comparing CART trees
   - Bootstrap based hypothesis test
   - Numerical experiments
3. Some theoretical insight
   - U-statistics
   - Some results

Covid-19 death rates during first wave, Ile-de-France

SI-VIC database, AP-HP hospitals.



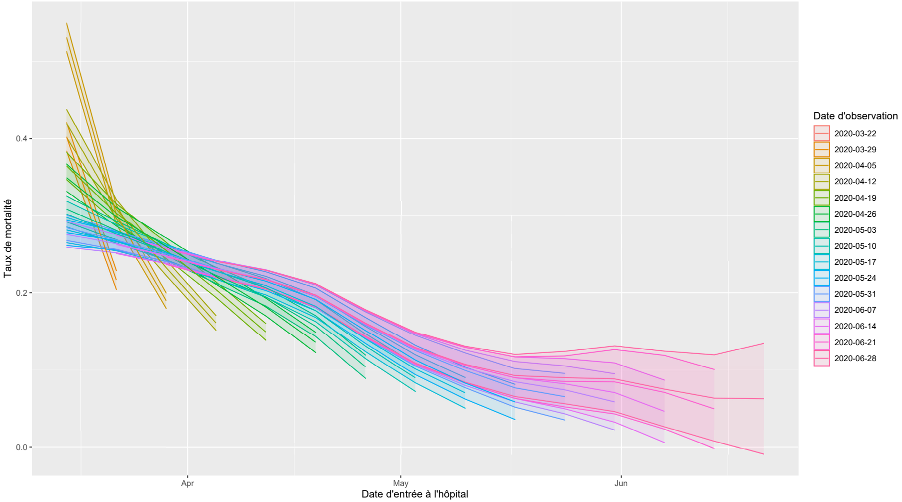
Modelling motivation

Covid-19 Dataset from SI-VIC database: all hospitalisation for Covid-19 patients in AP-HP hospitals.

dt.first	dt.last	outcome	sex	age	hospital
2020-03-17	2020-04-05	rad	F	45	Robert Debré
2020-03-14	2020-03-25	rad	F	29	Robert Debré
2020-03-18	2020-03-29	dc	H	80	St Antoine
2020-03-11	2020-03-15	dc	H	62	St Louis
2020-03-04	2020-03-09	dc	F	72	Pitié Salpétrière
2020-03-16	2020-03-20	dc	H	92	Raymond Poincaré

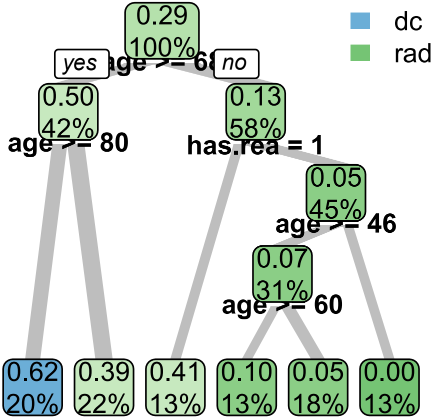
- **Motivation**: We wish to model the risk of death of a patient hospitalised for Covid-19 with respect to covariates.
- **Objective**: Adapt care of patients when changes in the vulnerability of groups at risks are detected.

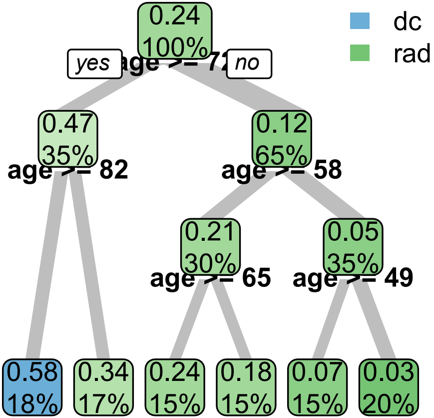
Nadaraya-Watson estimator, corrected for censure



Estimating groups at risk using classification trees

- Classification and Regression Trees (Breiman et al., 1984)
- Build one classification tree per week.
- Study the evolution of mortality in groups at risk.





Classification and Regression Trees (CART)

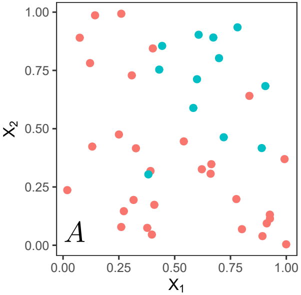
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ 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:

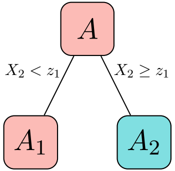
$$\Delta i(A, d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - 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)



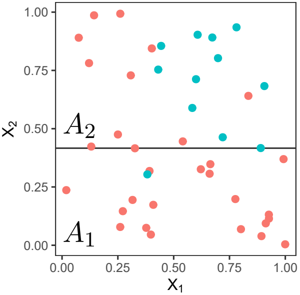
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ 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)

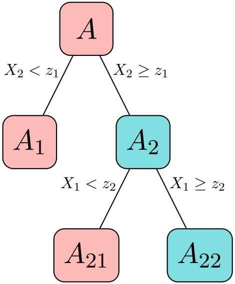
Introduced by Breiman et al., 1984.

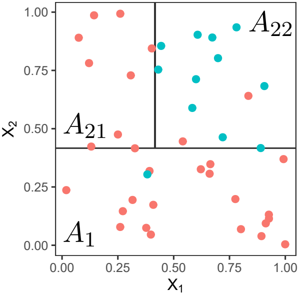
- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ along one of the covariate dimensions:
    - Find the node A, the dimension d and the value z such that the split (A, d, z) maximises the decrease in impurity:

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

    - 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.



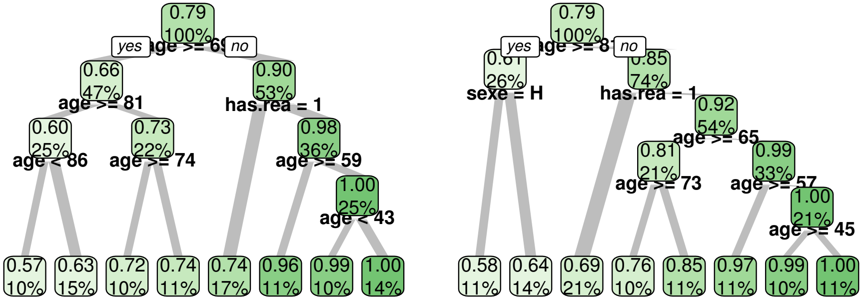


About CART

- Interpretable, handles missing data.
- Not widely used in epidemiology (Wolfson and Venkatasubramaniam, 2018).
- Theoretical properties:
    - Breiman et al., 1984: Consistency of tree structured regression and classification:
      $$\lim_{n\to\infty} \mathbb{E}[p_{T_n}(\mathbf{X}) - p(\mathbf{X})]^2 = 0.$$
      though not pointwise.
    - Gey and Nedelec, 2005; Gey, 2012: Non asymptotic risk for pruned procedure.
- Ensemble methods often preferred (Scornet, Biau, and Vert, 2015; Biau and Scornet, 2016; Lopes, Wu, and Lee, 2020)

CART and learning set

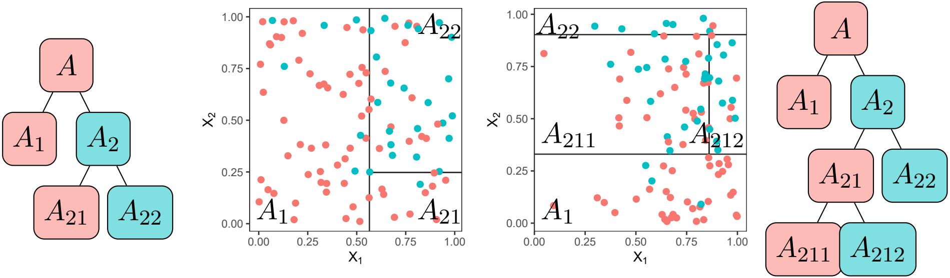
Problem: CART are sensitive to perturbations in the learning set.



- Study predictions rather than structure of the tree: what is the variance associated with the sampling of the learning set?
- Bar-Hen, Gey, and Poggi, 2015: influence functions derived from robust estimation theory.
- Wager, Hastie, and Efron, 2014: variance of bagged predictors.

Hypothesis test for the comparison of trees

- Learning set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $Y_i \in \{0, 1\}$ and $\mathbf{X}_i = (X_i^1, \dots, X_i^p) \in \mathcal{X}$.
- Tree $T_n = T(\mathcal{D}_n)$ generated by CART.
- Predicted probability $p_{T_n}(\mathbf{x})$ of $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ for the tree T_n.
- Hypothesis test for the comparison of two trees T_n and T'_m with respect to a d.f. F defined on a subset $B \subseteq \mathcal{X}$ of the input space:
  - Null hypothesis $\mathcal{H}_0 : \quad \forall \mathbf{x} \in B,\ p_{T_n}(\mathbf{x}) = p_{T'_m}(\mathbf{x})$.



Hypothesis test for the comparison of trees

- Learning set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $Y_i \in \{0,1\}$ and $\mathbf{X}_i = (X_i^1, \ldots, X_i^p) \in \mathcal{X}$.
- Tree $T_n = T(\mathcal{D}_n)$ generated by CART.
- Predicted probability $p_{T_n}(\mathbf{x})$ of $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ for the tree T_n.
- Hypothesis test for the comparison of two trees T_n and T'_m with respect to a d.f. F defined on a subset $B \subseteq \mathcal{X}$ of the input space:
  - Null hypothesis $\mathcal{H}_0$: $\quad \forall \mathbf{x} \in B,\ p_{T_n}(\mathbf{x}) = p_{T'_m}(\mathbf{x})$.
  - Test statistic:

$$I(T_n, T'_m, F) = \int d\big(p_{T_n}(\mathbf{x}), p_{T'_m}(\mathbf{x})\big) \mathrm{d}F(\mathbf{x}),$$

    where $d(p,q)$ is one of $(p-q)^2$, $|p-q|$, or $-p\log q - q\log p$, or

$$I(T_n, T'_m, F) = \sup_{\mathbf{x} \in B} |p_{T_n}(\mathbf{x}) - p_{T'_m}(\mathbf{x})|.$$

  - **Question**: What is the d.f. of $I(T_n, T'_m, F)$?

Bootstrap approximation

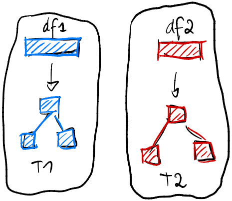
Idea: Generate a bootstrap approximation of the d.f. of $I(T_n, T'_m, F)$ under $\mathcal{H}_0$:

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

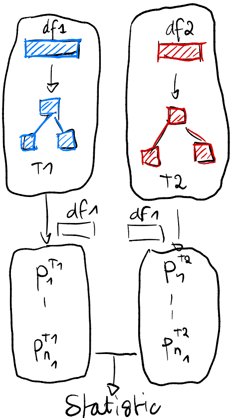
$$\bar{p}(\mathbf{x}) = \frac{n}{n+m} p_{T_n}(\mathbf{x}) + \frac{m}{n+m} p_{T'_m}(\mathbf{x});$$

- Generate bootstrapped trees T_n^* and T'^*_m, where probabilities for Y_i^* are given by $\bar{T}$;
  - Sample with replacement n inputs from $\mathcal{D}_n$: $\mathbf{X}_1^*, \dots, \mathbf{X}_n^*$;
  - For each $\mathbf{X}_i^*$, draw $Y_i^* \sim B(p(\mathbf{X}_i^*; \mathcal{D}_n))$;
  - Build the tree T_n^* on $\mathcal{D}_n^* = \{(\mathbf{X}_i^*, Y_i^*)\}_{i=1}^n$, using the same control parameters as the original tree.
- Compare $I(T_n, T'_m, F)$ to $I(T_n^*, T'^*_m, F)$ to determine a unilateral p-value.

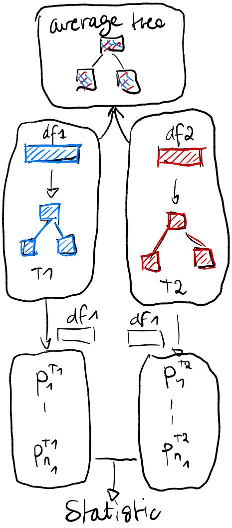
Bootstrap based hypothesis test



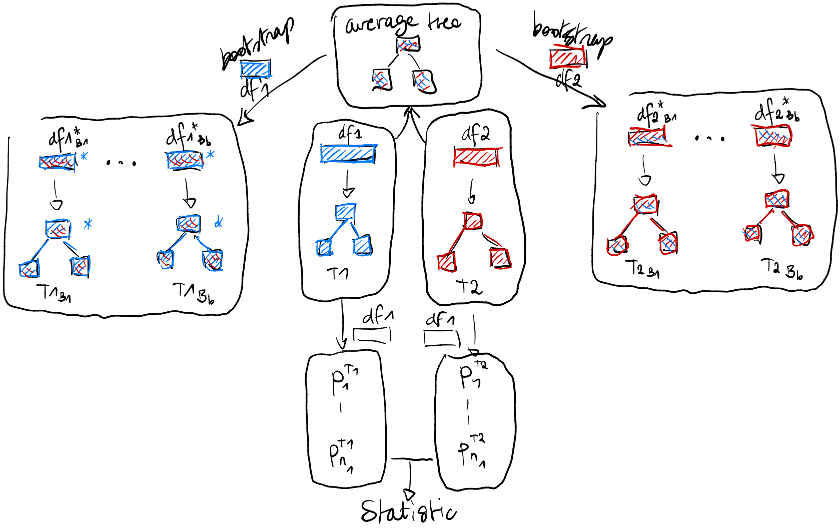
Bootstrap based hypothesis test



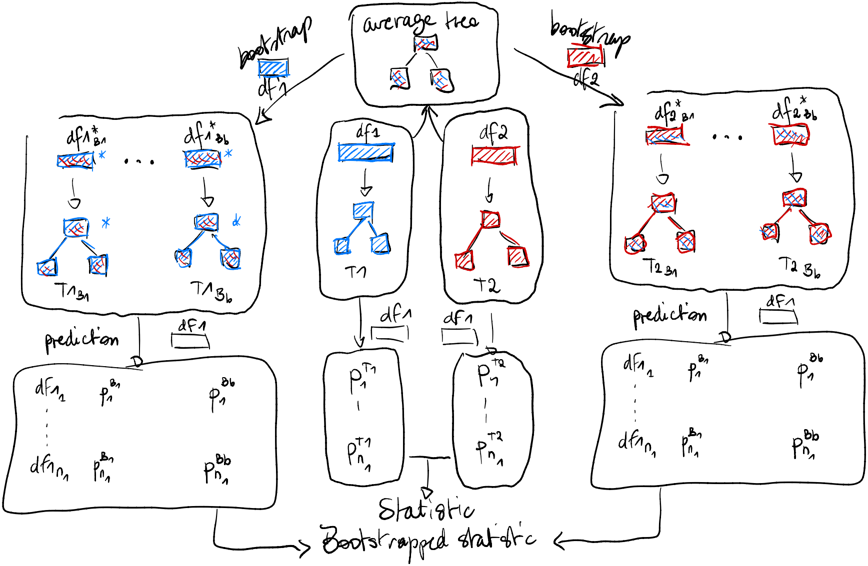
Bootstrap based hypothesis test



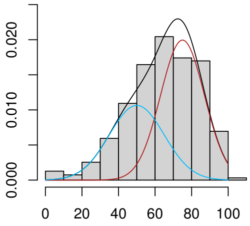
Bootstrap based hypothesis test



Bootstrap based hypothesis test



Numerical experiments



Generative model $\mathcal{M}$ for both $\mathcal{D}_n$ and $\mathcal{D}'_m$:

- Continuous variable (age):
  $X_1 = p_e X_e + (1 - p_e) X_y$, where
  $X_e \sim \mathcal{N}(\mu_e, \sigma_e)$ and $X_y \sim \mathcal{N}(\mu_y, \sigma_y)$;
- Discrete variable (gender): $X_2 \sim B(p_f)$;
- Binary outcome (death): $Y \mid X_1, X_2 \sim B(p_d)$,

$$\mathrm{logit}(p_d) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Param.	Value
μ_e	75
σ_e	12
μ_y	50
σ_y	15
p_e	0.6
p_f	0.5
β_0	-5
β_1	0.05
β_2	0.35

Scenarii, $n = m = 1,000$:

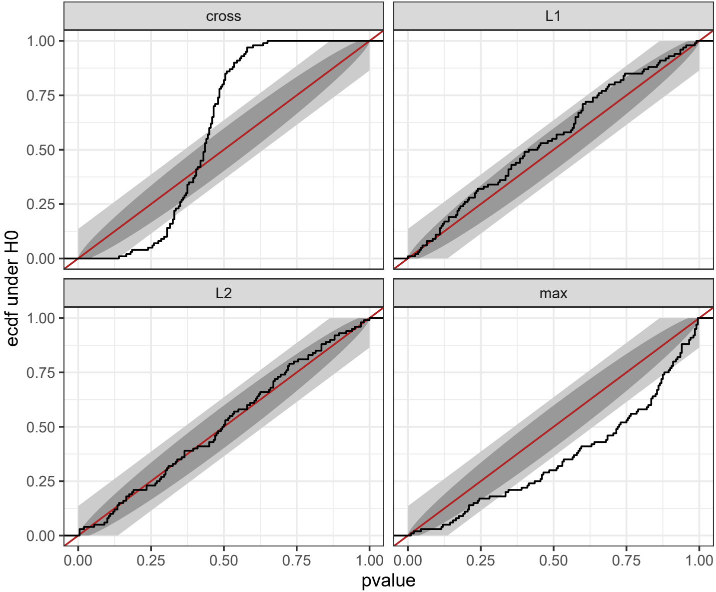
$\mathcal{H}_0$ Testing d.f. are $(\mathbf{X}_i)$ from $\mathcal{D}_n$;

$\mathcal{H}'_0$ Testing d.f. is generated from $\mathcal{M}$ with $p_e = 1$;

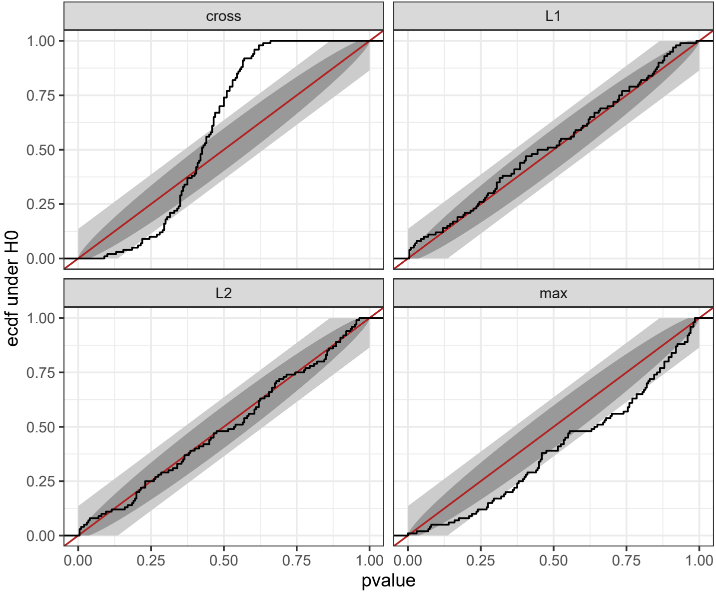
$\mathcal{S}_1$ As for $\mathcal{S}'_0$, with $\beta_1 = 0.06$;

$\mathcal{S}_2$ As for $\mathcal{S}'_0$, with $\beta_2 = 0.7$.

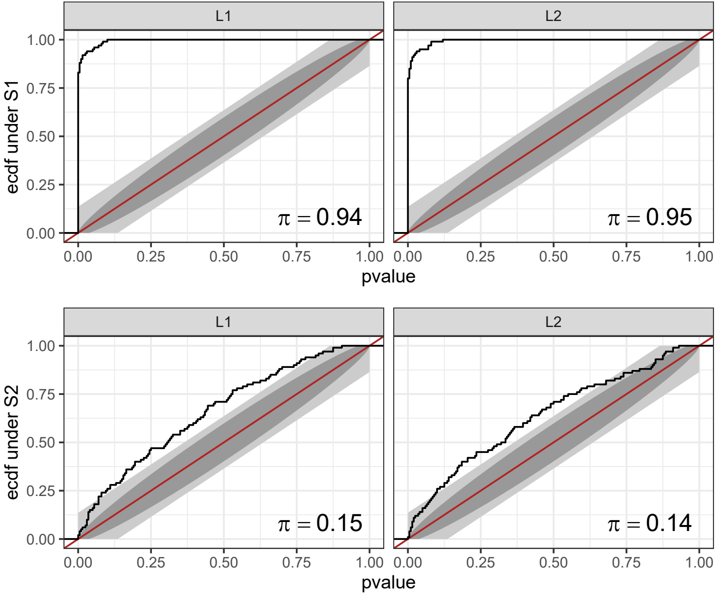
p-values for 100 simulations under $\mathcal{H}_0$



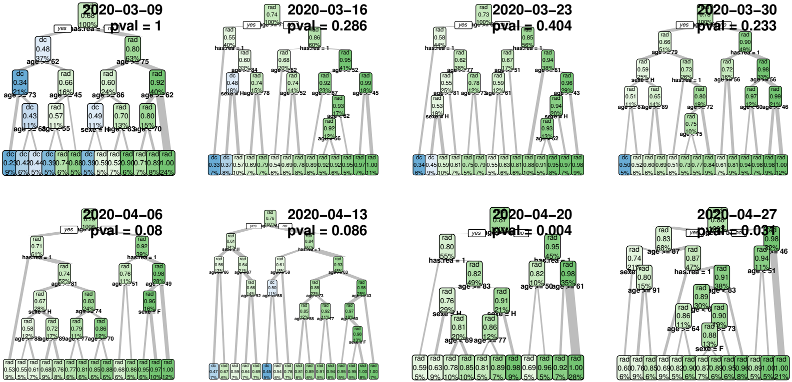
p-values for 100 simulations under $\mathcal{H}_0'$, for

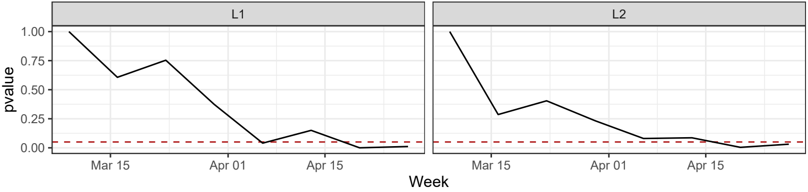


p-values for 100 simulations for scenarii $\mathcal{S}_1$ and $\mathcal{S}_2$



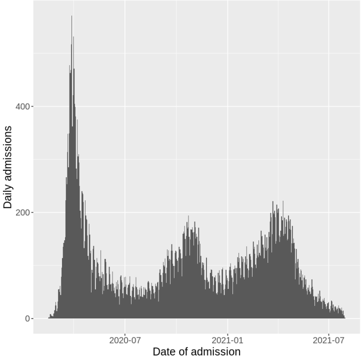
Applications to death rates during first wave





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

Theoretical development

Motivation: Is the bootstrap approximation valid?
Our first approach: Using von Mises calculus (Fernholz, 1983).

- Let $\mathbb{P}_n$ and $\mathbb{P}'_m$ denote the empirical d.f. of $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and $\{(\mathbf{X}'_j, Y'_j)\}_{j=1}^m$, and $\mathbb{P}$ and $\mathbb{P}'$ their true d.f.
- If $I(T_n, T'_m, F) = I(\mathbb{P}_n, \mathbb{P}'_m)$ can be shown to be Hadamard-differentiable at $\mathbb{P}$ and $\mathbb{P}'$, then
  - $I(T_n, T'_m, F)$ is asymptotically normal;
  - $I(T^*_n, T'^*_m, F)$ converges in distribution to the same limit.
- Problem: differentiability of the split point (A, d, z) at $\mathbb{P}$ cannot be shown to hold for CART trees.

Second attempt: Using distributional results on ensemble methods (Wager, 2014; Mentch and Hooker, 2016; Lopes, Wu, and Lee, 2020).

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:

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

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

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

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

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

where

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

U-statistics for CART

- Mentch and Hooker, 2016; Peng, Coleman, and Mentch, 2019:
  - Base learners $T(\mathbf{X}_1, \ldots, \mathbf{X}_r)$ trained on subsamples of size r.
  - Bagging predictions at $\mathbf{x}^*$ from this ensemble method yields

$$U_{n,k}(\mathbf{x}^*) = \binom{n}{k}^{-1} \sum_{(n,k)} T_{\mathbf{x}^*}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_r}).$$

  - Extension to Incomplete Infinite Order U-Statistics:

$$U_{n,r_n,N}(\mathbf{x}^*) = \frac{1}{N} \sum_{(i)} T_{\mathbf{x}^*}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{r_n}}).$$

- Mayer, 2009: U-quantile-statistic = sample p-th quantile of the set

$$\{h(X_1, \ldots, X_r : 1 \leq i_1 \leq \ldots \leq i_r \leq n\}.$$

Idea: Under the null, find the asymptotic properties of the $(1-\alpha)$-th quantile of the test statistic $I(T_n, T'_m, F)$ written as an incomplete infinite order U-quantile-statistic.

U-statistic for the hypothesis test

- Consider two learning sets $\{(X_i, U_i)\}_{i=1}^m$ and $\{(Y_j, V_j)\}_{j=1}^n$.
- Define the kernel h by

$$h(X_1, \dots, X_r; Y_1, \dots, Y_s) = \int d\big(T_x(X_1, \dots, X_r), T_x(Y_1, \dots, Y_s)\big) \mathrm{d}F(x),$$

  with expected value $\theta_{r,s} = \mathbb{E}h$.
- Then we consider

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

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

Define $\zeta_{r,s} = \mathrm{Var}(h)$, $\zeta_{1,0} = \mathrm{Var}\big(\mathbb{E}[h \mid X_1]\big)$ and $\zeta_{0,1} = \mathrm{Var}\big(\mathbb{E}[h \mid Y_1]\big)$.

Proposition

Denote $N = m + n$. Suppose $m/N \to \lambda$, $n/N \to (1-\lambda)$, and $r/m \sim s/n \sim S/N$. Assume that $\mathbb{E}h^2 < \infty$, and

$$\frac{S}{N} \frac{\zeta_{r,s}}{r\zeta_{1,0} + s\zeta_{0,1}} \to 0.$$

Then

$$\sqrt{N} \frac{U_{m,n,r,s} - \theta_{r,s}}{\sqrt{\frac{r^2}{\lambda}\zeta_{1,0} + \frac{s^2}{1-\lambda}\zeta_{0,1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

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

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

- 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).

Perspectives

Applications to more complex data from Covid-19 pandemic

- Extend methodology to include censored data.
  - Weigh observations according to the inverse of the survival function.
- Include more explanatory covariates in the learning set.
  - Biological data, comorbidities, hospital pathways, etc.
- Develop and share a R package.

Theoretical properties

- Finish proofs for incomplete U-statistics and U-quantile-statistics.

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.
- Courtejoie, Noémie and Claire-Lise Dubost (2020). *Parcours hospitalier des patients atteints de la Covid-19 lors de la première vague de l'épidémie*. Tech. rep. 67. Les dossiers de la DREES.

For Further Reading II

- 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.
- 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.

Bootstrap approximation

Parametric bootstrap $T_n^* = T(\mathcal{D}_n^*)$ of tree $T(\mathcal{D}_n)$:

- Sample with replacement n inputs from $\mathcal{D}_n$: $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$;
- For each $\mathbf{X}_i^*$, draw $Y_i^* \sim B(p(\mathbf{X}_i^*; \mathcal{D}_n))$;
- Build the tree T_n^* on $\mathcal{D}_n^* = \{(\mathbf{X}_i^*, Y_i^*)\}_{i=1}^n$, using the same control parameters as the original tree.

back

Linear statistical functionals

Consider a *linear statistical functional* $T : G \mapsto T(G) = \int \phi(x)\mathrm{d}G(x)$ with ϕ a real-valued function. Then, for the empirical d.f. F_n of F:

$$\begin{aligned}\sqrt{n}(T(F_n) - T(F)) &= \sqrt{n}\left\{\int \phi(x)\mathrm{d}F_n(x) - \int \phi(x)\mathrm{d}F(x)\right\}\\ &= \sqrt{n}\left\{n^{-1}\sum \phi(X_i) - \mathbb{E}_F[\phi(X)]\right\}\\ &= \sqrt{n}\left\{n^{-1}\sum \Big(\phi(X_i) - \mathbb{E}_F[\phi(X)]\Big)\right\}\\ &\xrightarrow{\text{CLT}} \mathcal{N}\big(0, \sigma^2 = \mathrm{Var}_F\ \phi(X)\big).\end{aligned}$$

von Mises calculus for statistical functionals

- **von Mises differentiation** generalises this to non-linear functionals:

$$T(F_n) = T(F) + T'_F(F_n - F) + \mathrm{Rem}(F_n - F),$$

where $T'_F(\cdot - F) : G \mapsto \int \phi_F(x)\mathrm{d}G(x)$ is a linear mapping with

$$\phi_F(x) = \frac{\mathrm{d}}{\mathrm{d}t}\Big(T(F + t(\delta_x - F))\Big)\Big|_{t=0}$$

often denoted $\mathrm{IC}(x; F, T)$ the *influence curve* of T at F.

- Existence of
  - the Von Mises derivative $T'_F(\cdot - F)$
  - and convergence of $\sqrt{n}\,\mathrm{Rem}(F_n - F)$ to zero in probability,

and thus validity of the Taylor expansion, can both be ensured by the Hadamard differentiability of the functional T at F (Fernholz, 1983).

Hadamard differentiability (Fernholz, 1983)

A function $T : A \in V \to W$ is *Hadamard-differentiable* at $F \in A$ if there exists $T'_F \in L(V, W)$ such that, for any $K \subset V$ compact,

$$\lim_{t \to 0} \frac{T(F + tH) - T(F) - T'_F(tH)}{t} = 0$$

uniformly for $H \in K$. The linear function T'_F is called the Hadamard-derivative of T at F.

Robustness of CART

- Idea: quantify sensitivity through influence functions $I(\cdot)$ derived from robust estimation theory (Bar-Hen, Gey, and Poggi, 2015).
- If I is Hadamard-differentiable, then:

$$
\begin{aligned}
\sqrt{n}(I(F_n) - I(F)) &\simeq \sqrt{n} \int \mathrm{IC}_{I,F}(x) \mathrm{d}F_n(x) \\
&= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \mathrm{IC}_{I,F}(X_i) \\
&\xrightarrow{\text{TCL}} \mathcal{N}\left(0, \sigma^2 = \int \mathrm{IC}_{I,F}^2(x) \mathrm{d}F(x)\right).
\end{aligned}
$$

- Estimation via Jackknifing:

$$
\mathrm{IC}_{I,F_n}(x_i) \simeq I^*_{n,i} - I(F_n) = nI(F_n) - (n-1)I(F_{n-1}^{(-i)}),
$$

where $I^*_{n,i}$ represents the n-th jackknife pseudo-value.

Perspectives

Study theoretical properties of the test

$$\sqrt{n}(I(F_n) - I(F)) \simeq \sqrt{n} \int \mathrm{IC}_{I,F}(x) \mathrm{d}F_n(x)$$

- What is the induced functional $I(T, T', F)$?
  - $\mathbb{L}^2$-consistency for regression trees and random forests (Scornet, Biau, and Vert, 2015):
    $$\lim_{n \to \infty} \mathbb{E}[(m_{T_n}(\mathbf{X}) - m(\mathbf{X}))^2] = 0,$$
    with $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ and $m_{T_n}(\mathbf{x})$ its prediction.
- Asymptotic properties of the bootstrap statistic $I(T_n^*, T_m'^*, F)$?
  - Vaart, 2000: Conditionally on $X_1, \ldots, X_n$, the sequence $\sqrt{n}(\phi(F_n^*) - \phi(F_n))$ converges in distribution to the same limit as $\sqrt{n}(\phi(F_n) - \phi(F))$, for every Hadamard-differentiable function ϕ.