Asymptotic Property for Generalized Random Forests

Hiroshi Shiraishi[1] Tomoshige Nakamura[2] Ryuta Suzuki[1]

[1]Keio University, Japan

[2]Jyuntendo University, Japan

BOSTON-KEIO-TSINGHUA WORKSHOP
Probability and Statistics
June 26-30, 2023

Contents

Literature

- Conditional Mean (RF)
  - Breiman (2001), Biau et al. (2008), Biau (2012), Scornet et al. (2015), Davis and Nielsen (2020)
- Causal Inference
  - Wager and Walther (2015), Athey and Imbens (2016), Wager and Athey (2018)
- Conditional Quantile
  - Meinshausen (2006), S-N-Shibuki
- Survival Function
  - Ishwaran and Kogalur (2010), Cui et al. (2023)
- Local Estimating Equation (GRF)
  - Athey,P., Tibshirani,J., and Wager,S. (2019)

Outline of This study

Athey et al. (2019)'s result

- Although they showed the asymptotic nomarilty, they did not show the rate of convergence and (closed form of) the asymptotic variance.

Our Contribution

- We show the "Rate of Convergence" and "Asymptotic Variance" in closed form.

Key Idea

- We approximate the "Forest Weight" as a "Kernel Function". Then, the asymptotic theory is reduced to that of "Nadaraya-Watson type estimator".

GRF model

Definition 1. (GRF model)

Suppose that a sequence of i.i.d. random vector $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1,\dots,n}$ satisfies

$$\mathbb{E}\left[\psi_{\theta(x)}(Y_i) | X_i = x\right] = 0 \quad \text{for all } x \in \mathcal{X} \tag{1}$$

where

- $\theta \in \Theta = \{\theta : \mathcal{X} \to \mathbb{R}\}$: parameter of interest
- $\psi : \Theta \times \mathcal{Y} \to \mathbb{R}$: some scoring function

(Note) ψ depends on the parameter for example

- mean : $\psi_{\theta(x)}(y) = y - \theta(x)$
- quantile : $\psi_{\theta(x)}(y) = \tau - \mathbf{1}_{\{y \leq \theta(x)\}}$ for some $\tau \in (0, 1)$
- likelihood : $\psi_{\theta(x)}(y) = \nabla \log(f_{\theta(x)}(y))$ for some (localized) p.d.f f

GRF estimator

Definition 2. (GRF estimator)

Given a data $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1,\dots,n}$ satisfying (1), an estimator of $\theta = (\theta(x))_{x\in\mathcal{X}} \in \Theta$ (defined in Def 1) is defined by

$$\hat{\theta}(x) \in \arg\min_{e\in\mathbb{R}} \left\{ \left| \sum_{i=1}^{n} \alpha_i(x)\psi_e(Y_i) \right| \right\} \quad \text{for all } x \in \mathcal{X}$$

where

- $\alpha_i(x) \in [0,1]$: weight function based on Random Forests

$$\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x), \quad \alpha_{bi}(x) = \frac{\mathbf{1}_{\{X_i \in L_b(x)\}}}{|L_b(x)|}$$

- B : number of trees
- $L_b(x)$: "leaf" of b-th tree containning the test point $x \in \mathcal{X}$
- $|L_b(x)|$: subsample size falling in the leaf $L_b(x)$

Image of Weights

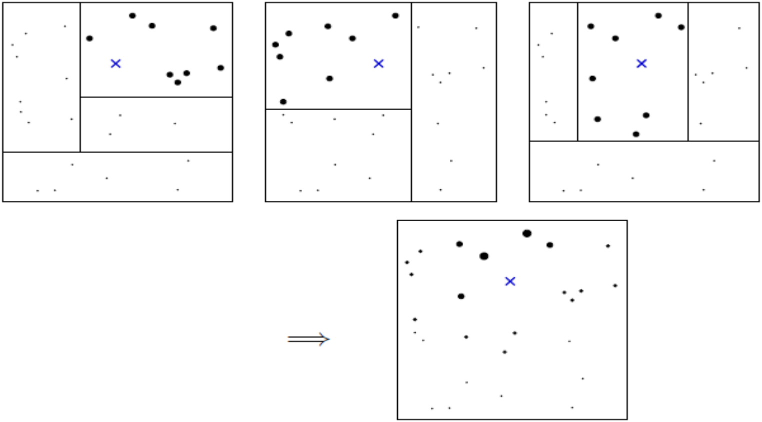


figure 1: Illustration of the random forest weighting functon (Athey et al. 2019)

Double Sample

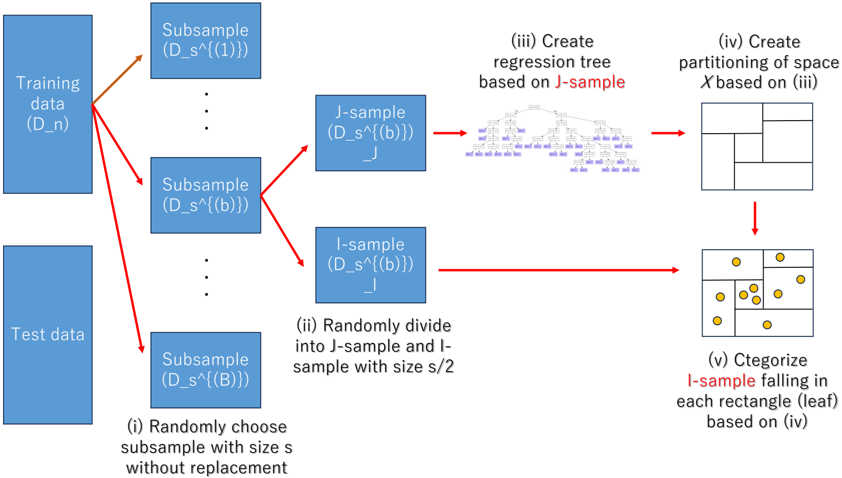


figure 2: Procedure of double sample

Athey et al.'s Result

Theorem 5 of Athey et al. (2019)

Suppose Assumptions 1-6 and a forest trained according to Specification 1 with trees are grown on subsamples of size $s = n^{\beta}$
satisfying (13). Finally, suppose that $\mathrm{Var}[\rho_i^*(x)|X_i = x] > 0$. Then, there is a sequence $\sigma_n(x)$ for which $\left(\hat{\theta}_n(x) - \theta(x)\right)/\sigma_n(x) \xrightarrow{d} N(0,1)$ and $\sigma_n^2(x) = \mathrm{polylog}(n/s)^{-1} s/n$, where $\mathrm{polylog}(n/s)$ is a function that is bounded away from 0 and increases at most polynomially with the log-inverse sampling ratio $\log(n/s)$.

Problem

From the above result, we have for any $x \in \mathcal{X}$

$$r_n \left(\hat{\theta}_n(x) - \theta(x)\right) \xrightarrow{d} N(0, \sigma^2(x)).$$

Two quantities are missing.

- rate of convergence (r_n)
- asymptotic varinace ($\sigma^2(x)$)

Assumption

We impose the following assumptions.

Assumption

(A.1) There exists 2nd order moment, and strictly positive, continuous p.d.f. of (X, Y) on $\mathcal{X} \times \mathcal{Y}$.

(A.2) (Lipschitz x-signal) $M_e(x) := \mathbb{E}\left[\psi_e(Y)|X = x\right]$ is Lipschitz continuous on $\mathcal{X}$.

(A.3) (Smooth identification) M_e is twice continuously differentiable at $e = \theta$ with a uniformly bounded second derivative, and that $\dot{M}(x) := \partial_e M_e(x)|_{e=\theta(x)}$ is invertible for all $x \in \mathcal{X}$.

(A.4) (Lipschitz (θ)-variogram) $\sup_{x\in\mathcal{X}} \left\{|\mathrm{Var}\left(\psi_e(Y) - \psi_{e'}(Y)|X = x\right)|\right\} \leq L|e - e'|$.

(A.5) (Regularity of ψ) $\psi_e = \lambda_e + \zeta_e$ where λ_e is Lipschitz continuous in e, ζ_e is monotone and bounded function.

(A.6) (Existence of solutions) There exists $\hat{\theta}_n(x)$ in Definition 2 and $|\sum_{i=1}^{n} \alpha_i(x)\psi_{\hat{\theta}_n(x)}(Y_i)| \leq C \max\{\alpha_i(x)\}$ for some constant $C \geq 0$.

(A.7) (Convexity) The score function ψ_e is a negative sub-gradient of a convex function, and the expected score M_e is the negative gradient of a strong convex function.

Approximate Kernel Function

Definition 3. (Forest score and Approximate forest score)

Denote the forest score as $\Psi(\theta(x)) = \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta(x)}(Y_i)$ for any $\theta = (\theta(x))_{x\in\mathcal{X}} \in \Theta$. We introduce the approximate forest score by

$$\Psi^{\mathrm{Ker}}(\theta(x)) = \sum_{i=1}^{n} \alpha_i^{\mathrm{Ker}}(x)\psi_{\theta(x)}(Y_i), \quad \alpha_i^{\mathrm{Ker}}(x) = \frac{K\left((x - X_i)/a_n\right)}{\sum_{j=1}^{n} K\left((x - X_j)/a_n\right)}$$

where a_n is a bandwidth and K is a Gaussian Kernel given by

$$K\left(u\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

(Note) $\Psi^{\mathrm{Ker}}(\theta(x))$ is a class of Nadaraya-Watson regression estimatiors and K is a kernel function satisfied with

$$\int_{-\infty}^{\infty} K(u)du = 1, \ \int_{-\infty}^{\infty} uK(u)du = 0, \ \int_{-\infty}^{\infty} u^2K(u)du = \frac{1}{2\sqrt{\pi}} < \infty$$

Fitting Kernel Functions

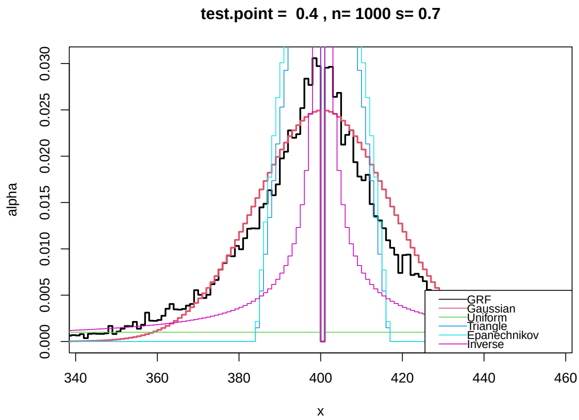


figure 3: Result for fitting some kernel functions such as Gaussian, Uniform, Tiangle, Epanechinikov and Inverse when $Y_i = X_i + N(0,1),\ X_i = i,\ n = 10^4,\ s = n^{0.7},\ a_n = 2\left(\log n / \log s\right)^2 \times (s/n/2)^{1/2}$.

Fitting Kernel Functions (effect for some test points)

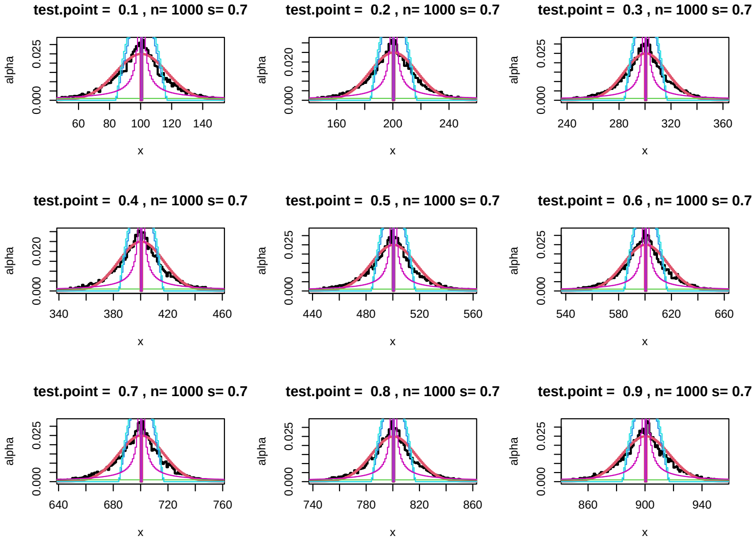


figure 4: Result for test points as 0.1-quantile, 0.2-quantile, $\ldots$,0.8-quantile 0.9-quantile.

Fitting Kernel Functions (effect for sample size)

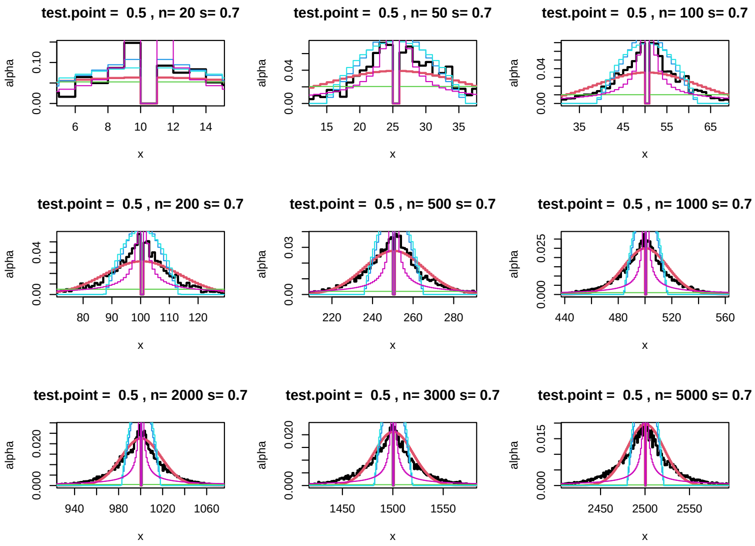


figure 5: Result for sample size (n) as $20, 50, 100, 200, 500, 1000, 2000, 3000, 5000$.

Fitting Kernel Functions (effect for subsample size)

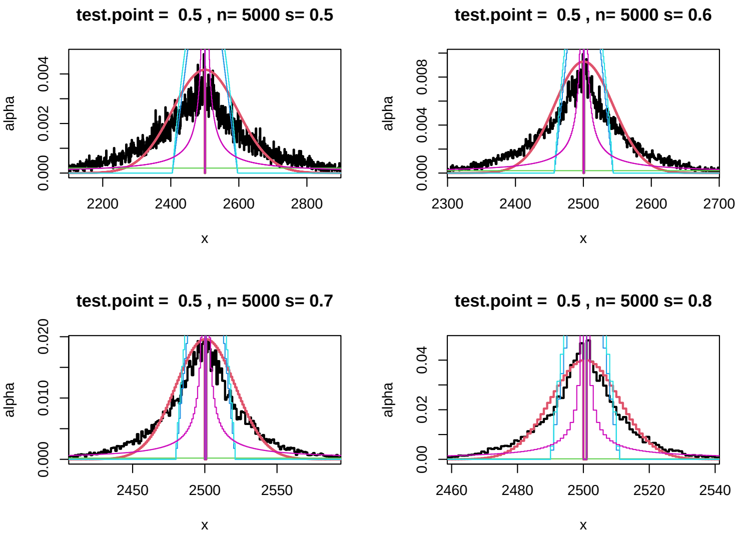


figure 6: Result for subsample size ($s = n^{\beta}$) as $5000^{0.5}, 5000^{0.6}, 5000^{0.7}, 5000^{0.8}$.

Asymptotic Normality

Lemma 1.

Let $s \equiv s(n) = n^{\beta}$ with $\beta \in (\beta_{\min}, 1)$. Under Assumption 1, Specification 1 in Athey et al. (2019), and $a_n = C_{\beta}(s/n)^{1/2}$, we have

$$\max_{i \in \{1,\ldots,n\}} \sup_{x \in \mathcal{X}^{\circ}} \left|\alpha_i(x) - \alpha_i^{\mathrm{Ker}}(x)\right| = o_p\left((na_n)^{-1/2}\right)$$

(Note1) "Specification 1" is the splitting rule for trees with (i) ω-regular (ii) random split (iii) PNN (potential nearest neighbor) k-set in Athey.et.al. (2019).

(Note2) Athey et al. (2019) define the lower bound of β by

$$\beta_{\min} := 1 - \left(1 + \pi^{-1}\left(\log\left(\omega^{-1}\right)\right) / \left(\log\left((1-\omega)^{-1}\right)\right)\right)^{-1}$$

where (i) ω and (ii) π are defined in "Specification 1" (in which the size of rectangle m is assumed to satisfy $\lfloor s/2 \rfloor \omega^m \in [k, 2k-1]$).

(Note3) $C_{\beta} > 0$ is a constant value depending on β. In the simulation study,
$C_{\beta} = 2^{1/2}(\log n / \log s)^2 = 2^{1/2}\beta^{-2}$.

(Note4) $\mathcal{X}^{\circ} \subset \mathcal{X}$ is a compact set.

Asymptotic Normality (cont.)

By some modification of Schuster (1972) or Stute (1984), we have the followings.

Lemma 2.

Under $\beta \in (1/3, 3/5)$, for any fixed $x \in \mathcal{X}^\circ$ and $M_{\theta(x)}(x) = \mathbb{E}\left[\psi_{\theta(x)}(Y)|X = x\right]$

$$\sqrt{na_n}\left\{\Psi^{\mathrm{Ker}}(\theta(x)) - M_{\theta(x)}(x)\right\} \xrightarrow{d} N\left(0, V(x)\right)$$

where

$$V(x) = \int u^2 K(u) du \mathrm{Var}\left(\psi_{\theta(x)}(Y)|X = x\right) = \frac{1}{2\sqrt{\pi}}\mathrm{Var}\left(\psi_{\theta(x)}(Y)|X = x\right)$$

(Note1) Schuster (1972) or Stute (1984) require the condition $na_n^3 \to \infty$ and $na_n^5 \to 0$ in order to vanish the asymptotic bias. In our case, we can see from $\beta \in (1/3, 3/5)$ that

$$na_n^3 = n\left(C_\beta (n/s)^{-1/2}\right)^3 = C_\beta^3 n\left((n^{1-\beta})^{-1/2}\right)^3 = C_\beta^3 n^{(3\beta-1)/2} \to \infty$$

$$na_n^5 = n\left(C_\beta (n/s)^{-1/2}\right)^5 = C_\beta^5 n\left((n^{1-\beta})^{-1/2}\right)^5 = C_\beta^5 n^{(5\beta-3)/2} \to 0.$$

(Note2) However, when $\omega = 0.2, \pi = 0.1$, it follows

$$\beta_{\min} := 1 - \left(1 + \pi^{-1}\left(\log\left(\omega^{-1}\right)\right) / \left(\log\left(\left(1-\omega\right)^{-1}\right)\right)\right)^{-1} = 0.9863249 \notin (1/3, 3/5).$$

Asymptotic Normality (cont.)

If both of Lemma 1 and Lemma 2 are satisfied under some condition, we have the following result.

Theorem.

Under some condition satisfied with Lemmas 1 and 2, we have for any fixed $x \in \mathcal{X}^\circ$

$$\sqrt{na_n}\left\{\hat{\theta}_n(x) - \theta(x)\right\} \xrightarrow{d} N\left(0, \frac{V(x)}{\dot{M}^2(x)}\right)$$

where $\dot{M}(x) = \partial_e M_e(x)|_{e=\theta(x)}$.

Remark.

The Cramér-Wold device may be applied to show that $\sqrt{na_n}\left\{\hat{\theta}_n(x) - \theta(x)\right\}$ converges jointly in distribution at finitely many points $x_1, \ldots, x_k$ with $\hat{\theta}_n(x_1), \ldots, \hat{\theta}_n(x_k)$ being asymptotically independent.

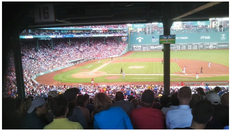
Conclusion and Future Work

- Conclusion
  - We considered the statistical estimation of functions defined by solutions of local estimating equations by using Generalized Random Forests (GRF) for i.i.d. data.
  - We fund that the asymptotic theory of Nadaraya-Watson type estimator is not directly applicable to that of GRF estimator and some additional arguments are needed.
- Future Work
  - Consider the asymptotic theory for the case of $\beta \in (\beta_{\min}, 1) \subset (3/5, 1)$
  - Numerical Result
  - Extend the model from i.i.d. to dependent

Referrence

- Athey, P., Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113, 7353-7360.
- Athey, P., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47, 1148-1178.
- Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research (JMLR)*, 13, 1063-1095.
- Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifires. *Journal of Machine Learning Research (JMLR)*, 9, 2015-2033.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S. and Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forest. *Journal of the Royal Statistical Society Series B*, 85(2), 179-21.
- Davis, R. A. and Nielsen, M. S. (2020). Modeling of time series using random forests: theoretical developments. *Electronic Journal of Statistics*, 14, 3644-3671.
- Iswaran, H. and Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & Probability Letters*, 80(13), 1056-1064.
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research (JMLR)*, 7, 983-999.
- Schuster, E., F. (1972). Joint Asymptotic Distribution of the Estimated Regression Function at a finite number of distinct points. *The Annals of Mathematical Statistics*, 43(1), 84-88.
- Scornet, E., Biau, G., and Vert, J, P. (2015). Consistency of random forests. *Annals of Statistics*, 43, 1716-1741.
- Stute, W. (1984). Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *Annals of Statistics*, 12(3), 917-926.
- Wager, S. and Athey, P. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228-1242.
- Wager, S. and Walther, G. (2015). Adaptive Concentration of Regression Trees, with Application to Random Forests. *arXiv preprint arXiv:1503.06388.*

Thank you very much!





Double Sample

Suppose that we can obtain a data,

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}.$$

Based on $\mathcal{D}_n$, we generate sub-sample $\{\mathcal{I}_s, \mathcal{J}_s\} = \{\mathcal{D}_{A^{\mathcal{I}}}, \mathcal{D}_{A^{\mathcal{J}}}\}$ as follows:

Definition.

Let $s = s(n)$ be a sub-sample size with $s < n$. Let

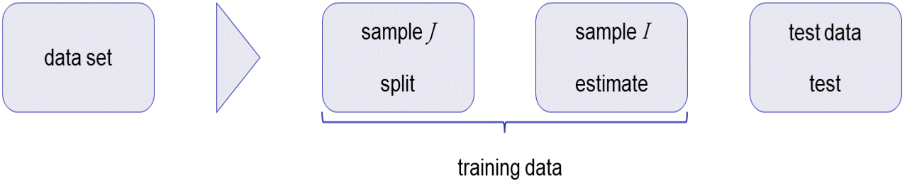
$$\mathcal{A}_s := \left\{ A = \{A^{\mathcal{I}}, A^{\mathcal{J}}\},\ A^{\mathcal{I}}, A^{\mathcal{J}} \subset \{1, 2, \dots, n\} \middle| A^{\mathcal{I}} \cap A^{\mathcal{J}} = \emptyset, \right.$$
$$\left. \left|A^{\mathcal{I}}\right| = \left\lfloor \frac{s}{2} \right\rfloor,\ \left|A^{\mathcal{J}}\right| = \left\lceil \frac{s}{2} \right\rceil \right\}$$

For any $A = \{A^{\mathcal{I}}, A^{\mathcal{J}}\} \in \mathcal{A}_s$, we define two sub-samples $\mathcal{I}_s$ and $\mathcal{J}_s$ by $\mathcal{I}_s = \mathcal{D}_{A^{\mathcal{I}}}, \mathcal{J}_s = \mathcal{D}_{A^{\mathcal{J}}}$ where $\mathcal{D}_{A^\cdot} = \{(X_i, Y_i)\}_{i \in A^\cdot}$.

Why Double Sample ?

The tree score $\mathcal{T}$ is constructed based on sub-samples $\{\mathcal{I}_s, \mathcal{J}_s\} \subset \mathcal{D}_n$ called the Double-sampling (s: sub-sample size).

- $\mathcal{J}_s$-sample : To place the splits (i.e., partitioning of the covariate space)
- $\mathcal{I}_s$-sample : To do within-leaf estimation (i.e., estimation of the interest quantities based on the elements of the $\mathcal{I}_s$-sample within the leaf)



- Thanks to the Double-sampling, regression tree (which is based on the $\mathcal{I}_s$-sample) do not depend on the covariate space partitioning (which is based on the $\mathcal{J}_s$-sample)!
- However, the sample size to be able to estimate becomes half.

Splitting Rule

By using $\mathcal{J}_s$-sample, we consider the partitioning of the covariate space $\mathbb{R}^p$.

Definition.

Given $\mathcal{J}_s = \mathcal{J}_s(A)$, we define a sequence of partitions $\mathcal{P}_0, \mathcal{P}_1, \ldots$ by starting form $\mathcal{P}_0 = \{\mathcal{X}\}$ and then, for each $\ell \geq 1$, construct $\mathcal{P}_\ell$ from $\mathcal{P}_{\ell-1}$ by replacing one set (parent node) $P \in \mathcal{P}_{\ell-1}$ by (child node)

$$C_1 := \{x = (x_1, \ldots, x_p) \in P \subset \mathcal{X} : x_\xi \leq \zeta\}$$
$$C_2 := \{x = (x_1, \ldots, x_p) \in P \subset \mathcal{X} : x_\xi > \zeta\}$$

where

- the split direction $\xi \in \{1, \ldots, p\}$: randomly chosen (i.e., random split) [a]
- the split position $\zeta = \zeta(\xi) \in \{x_\xi \in \mathcal{X}_j \cap P\}$: chosen by optimizing a criterion $\Delta(C_1, C_2)$

[a] In case of Athey et al. (2019), ξ is defined by $\xi \sim \min\{\max\{\text{Poisson}(m), 1\}, p\}$ at each step, where $m > 0$ is a turning parameter.

Criterion for Split Position

Athey et al. (2019) introduces the criterion for split position, which is an approximation of

$$\mathrm{err}(C_1, C_2) = \sum_{j=1}^{2} \mathbb{P}[\boldsymbol{X}_t \in C_j | \boldsymbol{X}_t \in P] \mathbb{E}[(\|\hat{\boldsymbol{\theta}}_{C_j}(\mathcal{J}_s) - \boldsymbol{\theta}_0(\boldsymbol{X}_t)\|^2 | \boldsymbol{X}_t \in C_j]$$

where P is a parent node, C_1, C_2 are children, $\hat{\boldsymbol{\theta}}_{C_j}(\mathcal{J}_s)$ is an estimator based on $C_j \subset \mathcal{J}_s$ and $\boldsymbol{\theta}_0$ is the target function (true parameter).

Proposition 1,2 (Athey et al., 2019)

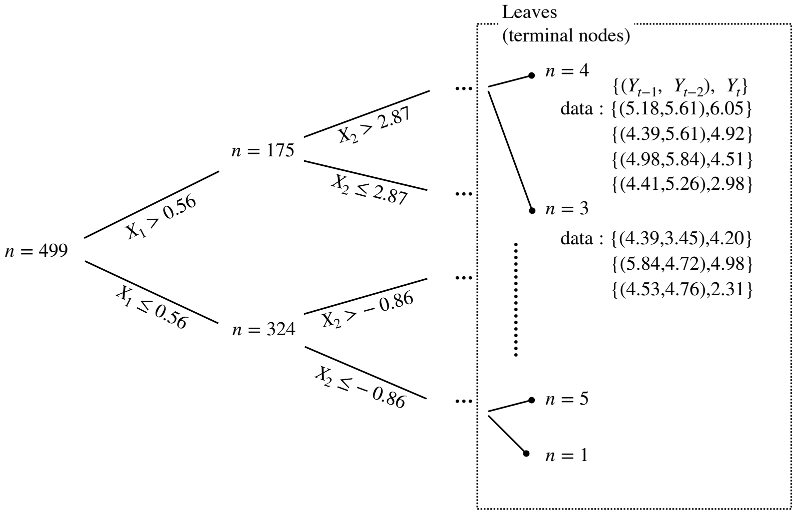
$$\Delta_{\mathrm{I}}(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \hat{\theta}_{C_1}(\mathcal{J}_s) - \hat{\theta}_{C_2}(\mathcal{J}_s) \right\|^2$$

$$\Delta_{\mathrm{II}}(C_1, C_2) := \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left\| \sum_{i : X_i \in C_j} \rho_i \right\|^2$$

Example for Splitting

One of the tree in quantile regression forest via GRF
(s = 499, $\mathcal{J}$-sample = 250, honest splitting)

AR(2) : $Y_t = 0.5Y_{t-1} + 0.4Y_{t-2} + N(0,1)$

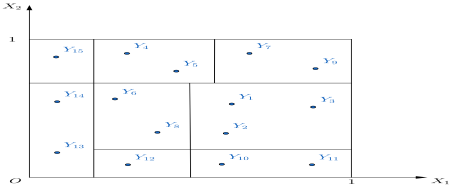


Splitting Rule (cont.)

For the above splitting rule, we impose the following assumption.

Assumption.

- (ω-regular) $\min(n_{C_1}, n_{C_2}) \geq \omega \times n_P$ where n_P, n_{C_1}, n_{C_2} are sample size of P, C_1, C_2, respectively.
- (random split) $\mathbb{P}(\xi = j) \geq \pi$ for all $j \in \{1, \dots, p\}$.
- (PNN (potential nearest neighbor) k-set) Let $L \in \mathcal{P}_\ell$ be a leaf of the tree and let $\sharp L := |\{X_t : X_t \in L\}|$ be a sub-sample size falling in L. Then $\sharp L$ satisfies $k \leq \sharp L \leq 2k - 1$ for some $k \in \mathbb{N}$.



Result for Test Data

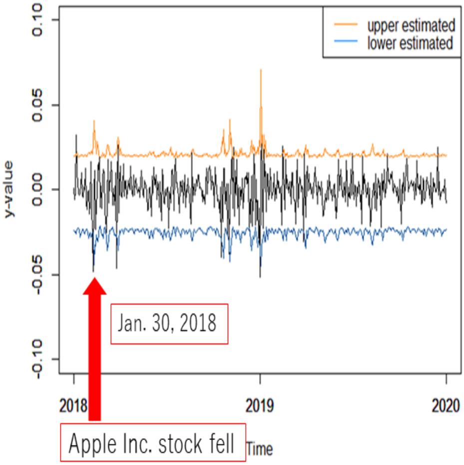


figure 8: WNW predictor by Cai (2002)

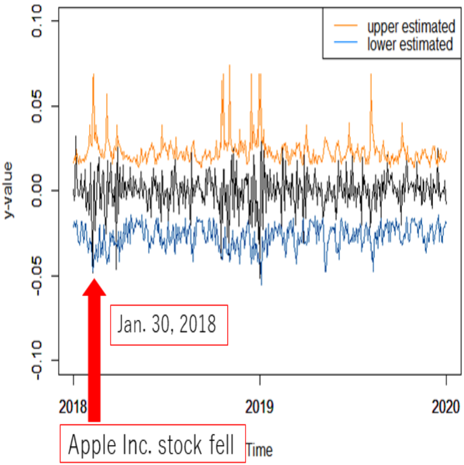


figure 9: tsQRF predictor