FACT: High-Dimensional Random Forests Inference

Authors: Chien-Ming Chi, Yingying Fan, Jinchi Lv

https://arxiv.org/abs/2207.01678

Presenter: Beining Wu

Department of Statistics and Finance,
University of Science and Technology of China

August 25, 2022

Outline

Preliminaries

FACT inference method

- Vanilla version
- Bias issue and debiasing techniques
- Ensemble test and general version

Theoretical guarantees

- Technical conditions
- Null behavior
- Power analysis

Reference

High-dimensional random forests

Notations

- Underlying probability space: $(\Omega, \mathcal{F}, \mathbb{P})$.
- Feature $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in [0,1]^p$.
- Scalar response: $Y \in \mathbb{R}$.
- Inference samples: $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ where $(\boldsymbol{X}_i, Y_i) \overset{\text{d.}}{=} (\boldsymbol{X}, Y)$.
- Fitting samples: $\mathcal{X}_0 = \{(\boldsymbol{U}_i, V_i)\}_{i=1}^N$ where $(\boldsymbol{U}_i, V_i) \overset{\text{d.}}{=} (\boldsymbol{X}, Y)$.
- Leave one coordinate out vector: $\boldsymbol{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$.
- The i-th leave-one-out feature: $\boldsymbol{X}_{-ij} = (X_{i1}, \ldots, X_{i,j-1}, X_{i,j+1}, \ldots, X_{ip})$.

High-dimensional random forests

Let $\widehat{Y}(\boldsymbol{X}_{-j})$ be the random forests estimate using the samples $\{(\boldsymbol{U}_{-ij}, V_i)\}_{i=1}^{N}$.

Condition 1.

Assume that $\mathbb{E}\{\mathbb{E}(Y|\boldsymbol{X}_{-j}) - \widehat{Y}(\boldsymbol{X}_{-j})\}^2 \leq B_1$ for some small $B_1 > 0$.

- Here $B_1 = o(1)$ implicitly depends on N.
- [?] obtained the consistency rates $B_1 = O(N^{-c})$ under SID condition.
- It which allows for $p = O(N^{K_0})$ for some $c, K_0 > 0$.

High-dimensional random forests

Sufficient impurity decrease

Condition 2.

Assume that there exists some $\alpha_1 \geq 1$ such that for each cell $\boldsymbol{t} \subset [0,1]^p$, we have

$$\alpha_1^{-1} \mathsf{Var}(m(\boldsymbol{X})|\boldsymbol{X} \in \boldsymbol{t}) \leq \mathsf{Var}(m(\boldsymbol{X})|\boldsymbol{X} \in \boldsymbol{t}) -$$

$$\inf_{j,x} \Big\{ \mathbb{P}(\boldsymbol{X} \in \boldsymbol{t}_1|\boldsymbol{X} \in \boldsymbol{t}) \mathsf{Var}(m(\boldsymbol{X})|\boldsymbol{X} \in \boldsymbol{t}_1) + \mathbb{P}(\boldsymbol{X} \in \boldsymbol{t}_2|\boldsymbol{X} \in \boldsymbol{t}) \mathsf{Var}(m(\boldsymbol{X})|\boldsymbol{X} \in \boldsymbol{t}_2) \Big\}.$$

where $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ represent the two daughter cells of $\boldsymbol{t}$ after the split (j, x) along feature X_j and feature value $x \in \mathbb{R}$, and the infimum is taken over all possible feature and split value combinations (j, x).

Outline

Preliminaries

FACT inference method

- Vanilla version
- Bias issue and debiasing techniques
- Ensemble test and general version

Theoretical guarantees

- Technical conditions
- Null behavior
- Power analysis

Reference

Overview of FACT

Goal we're to test the null hypothesis

$$H_0 : X_j \perp Y | \boldsymbol{X}_{-j}$$

Intuition Under null, for any user specified transformation $g : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}\big\{[Y - \mathbb{E}(Y|\boldsymbol{X}_{-j})][g(X_j) - \mathbb{E}g(X_j)]\big\} = 0$$

Define $\sigma_{j0}^2 = \text{Var}\{[Y - \mathbb{E}(Y|\boldsymbol{X}_{-j})][g(X_j) - \mathbb{E}g(X_j)]\}$. Under null hypothesis, we can use the following population level quantity to test the null.

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\left[Y_i - m(\boldsymbol{X}_{-ij})\right]\left[g(X_{ij}) - \mathbb{E}g(X_j)\right]}{\sigma_{j0}}$$

The basic FACT

- Under null hypothesis, last equation should behaves like standard normal distribution.
- If X_j is relevant, then this quantity should approach ∞.
- The power should depends on the choice of g.

Last quantity provides an ideal test, but is not available. So we replace the population quantity to the sample version.

The basic FACT

Sample version

- Let $d_{i1} = [Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})][g(X_{ij}) - n^{-1}\sum_{i=1}^{n} g(X_{ij})]$.
- Variance estimate $\widehat{\sigma}_{j0}^2 = n^{-1}\sum_{i=1}^{n}\left(d_{i1} - n^{-1}\sum_{i=1}^{n} d_{i1}\right)^2$.
- Forests estimate $\widehat{Y}(\boldsymbol{X}_{-j})$.

And our sample version test statistic is

$$\mathrm{FACT}_{j,0} := n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\left[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\right]\left[g(X_{ij}) - n^{-1}\sum_{i=1}^{n} g(X_{ij})\right]}{\widehat{\sigma}_{j0}}. \qquad (1)$$

The basic FACT

Rejection criterion

- For each threshold $t > 0$, we reject H_0 if $|\mathsf{FACT}_{j,0} > t|$.
- Let $t = -\Phi(\alpha/2)$, then the test has asymptotic size α.
- A conservative p-value for given realization is $2\Phi(-|\mathsf{FACT}_{j,0}|)$.

Bias issue

The bias of the basic FACT test is

$$\begin{aligned}\text{Bias}_1(N) &:= \mathbb{E}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\big[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\big]\big[g(X_{ij}) - \mathbb{E}g(X_j)\big]\Big\} \\ &= \sqrt{n}\mathbb{E}\big\{\big[\mathbb{E}(Y|\boldsymbol{X}_{-j}) - \widehat{Y}(\boldsymbol{X}_{-j})\big]\big[g(X_j) - \mathbb{E}g(X_j)\big]\big\}\end{aligned}$$

- When $X_j \perp \boldsymbol{X}_{-j}$, then $\mathbb{E}\big\{\big[\mathbb{E}(Y|\boldsymbol{X}_{-j}) - \widehat{Y}(\boldsymbol{X}_{-j})\big]\big[g(X_j) - \mathbb{E}g(X_j)\big]\big\} = 0$
- When X_j and $\boldsymbol{X}_{-j}$ are correlated, then our test statistic is biased because $\mathbb{E}[g(X_{ij})] \neq \mathbb{E}[g(X_{ij})|\boldsymbol{X}_{-ij}]$.

Bias issue

We can bound the bias as

$$\text{Bias}_1(N) \leq 2 \sup_{x\in[0,1]} |g(x)| \sqrt{B_1 n},$$

- If $N \sim n$ are of the same order, then the bias is not asymptotically negligible because B_1 is much slower than n^{-1}.
- If we have access to the conditional $\mathbb{E}[g(X_j)|\boldsymbol{X}_{-j}]$, then the bias issue can also be resolved.

These two observations motivate us with two debiasing techniques introduced below.

FACT test with imbalancing

Intuition If $N \gg n$, then B_1 can be asymptotically negligible. For example, simulation reveals that $n = O\{N(\log N)^{-1}\}$ is appealing. Therefore, we put our imbalanced FACT test as

$$\mathsf{FACT}_{j\mathsf{N/n}} = \mathsf{FACT}_{j,0} \;\; \text{with} \;\; \widehat{\sigma}_{j\mathsf{N/n}} = \widehat{\sigma}_{j0}, \tag{2}$$

Here the subscript involves N/n in order to emphasize imbalancing.

Remark.

The test statistic here is the same to the previous basic test. However, in order for the previous test to be effective, we need stronger assumption on the expectation. But here we only need to adjust N and n for theoretical guarantee.

FACT test with conditioning

The second motivation let us replace the $\mathbb{E}[g(X_j)]$ with $\mathbb{E}[g(X_j\boldsymbol{X}_{[} - j]|)]$.

- Let $\widehat{Y}(\boldsymbol{X}_{-ij})$ be the random forests estimate of $\mathbb{E}[Y_i|\boldsymbol{X}_{-ij}]$.
- Let $\widehat{g}(\boldsymbol{X}_{-ij})$ be the random forests estimate of $\mathbb{E}[g(X_ij)|\boldsymbol{X}_{-ij}]$,
- Denote $d_{i2} = \left[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\right]\left[g(X_{ij}) - \widehat{g}(\boldsymbol{X}_{-ij})\right]$.
- Conditional variance estimate: $\widehat{\sigma}^2_{j|\boldsymbol{X}_{-j}} = n^{-1}\sum_{i=1}^{n}\left(d_{i2} - n^{-1}\sum_{i=1}^{n} d_{i2}\right)^2$.

Formally, we define the FACT test statistic with conditioning for each feature X_j as

$$\text{FACT}_{j|\boldsymbol{X}_{-j}} := n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\left[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\right]\left[g(X_{ij}) - \widehat{g}(\boldsymbol{X}_{-ij})\right]}{\widehat{\sigma}_{j|\boldsymbol{X}_{-j}}}, \tag{3}$$

FACT test with conditioning

Rejection criterion

- For each given test threshold $t > 0$, we reject the null hypothesis H_0 if $\left|\mathsf{FACT}_{j|\boldsymbol{X}_{-j}}\right| > t$.
- The asymptotic size is $2\Phi(-t)$.
- The asymptotic p-value is $2\Phi(-\left|\mathsf{FACT}_{j|\boldsymbol{X}_{-j}}\right|)$.

FACT test with conditioning

In that case, the bias can be bounded as

$$\begin{aligned}\mathsf{Bias}_2(N) &:= \mathbb{E}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\big[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\big]\big[g(X_{ij}) - \widehat{g}(\boldsymbol{X}_{-ij})\big]\Big\} \\ &= \sqrt{n}\mathbb{E}\big\{\big[\mathbb{E}(Y|\boldsymbol{X}_{-j}) - \widehat{Y}(\boldsymbol{X}_{-j})\big]\big[\mathbb{E}(g(X_j)|\boldsymbol{X}_{-j}) - \widehat{g}(\boldsymbol{X}_{-j})\big]\big\}\end{aligned}$$

With the consistency of random forests estimate, the RHS can be bounded as $\sqrt{n}B_1$, much faster than the basic one.

FACT test with ensemble

One of the flexibility of the FACT test is that g is user-specified. That means we can aggregate multiple g to improve the selection power.

- Let $g^{(l)}, l \in L = [m_1]$ be a sequence of transformations.
- Define the FACT test statistic with ensemble for each feature X_j as

$$\max_{l \in L} \left| \mathsf{FACT}_j^{(l)} \right|, \tag{4}$$

where each $\mathsf{FACT}_j^{(l)}$ for feature transformation $g^{(l)}(\cdot)$ with $l \in L$ is given as in (1) by

$$\mathsf{FACT}_j^{(l)} := n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\left[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\right]\left[g^{(l)}(X_{ij}) - n^{-1}\sum_{i=1}^{n} g^{(l)}(X_{ij})\right]}{\widehat{\sigma}_j^{(l)}},$$

FACT test with ensemble

Here

- $(\widehat{\sigma}_j^{(l)})^2 = n^{-1} \sum_{i=1}^{n} \left(d_{i3}^{(l)} - n^{-1} \sum_{i=1}^{n} d_{i3}^{(l)} \right)^2$
- $d_{i3}^{(l)} = [Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})][g^{(l)}(X_{ij}) - n^{-1} \sum_{i=1}^{n} g^{(l)}(X_{ij})]$.

FACT test with ensemble

Rejection criterion

- For given threshold $t > 0$, we reject null when $\max_{l \in L} \left| \mathsf{FACT}_j^{(l)} \right| > t$.
- When $t = -\Phi^{-1}(\alpha/(2|L|))$, then the asymptotic size is α.
- The asymptotic power can be conservatively estimated as $\min\left\{1, 2|L|\Phi(-\max_{l \in L} |\mathsf{FACT}_j^{(l)}|)\right\}$.

General FACT test

The general FACT test incorporates all the ideas presented above.

- From now on let $N = n$, $L = 2$, $g^{(1)}(x) = x$ and $g^{(2)}(x) = x^2$.
- Define $\widehat{Y}(\boldsymbol{X}_{-j})$ and $\widehat{g}^{(l)}(\boldsymbol{X}_{-j})$ as the random forests estimate using training data.
- Define

$$\mathsf{FACT}_j := \max_{l \in L,\, q \in Q} \left| \mathsf{FACT}_{j,q}^{(l)} \right|, \tag{5}$$

where

$$\mathsf{FACT}_{j,q}^{(l)} := \sum_{i \in \mathcal{N}_q} \frac{\left[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\right]\left[g^{(l)}(X_{ij}) - \widehat{g}^{(l)}(\boldsymbol{X}_{-ij})\right]}{|\mathcal{N}_q|^{1/2}\widehat{\sigma}_{j|\boldsymbol{X}_{-j}}^{(l)}}$$

General FACT test

And

- $Q = \{1, \ldots, k_n\}$ with k_n slowly diverge
- $\{\mathcal{N}_1, \ldots, \mathcal{N}_{k_n}\}$ is a random partition of set $\{1, \ldots, n\}$ with $||\mathcal{N}_k| - |\mathcal{N}_l|| \leq 1$.
- Let $d_{i4}^{(l)} = \big[Y_i - \widehat{Y}(\boldsymbol{X}_{-ij})\big]\big[g^{(l)}(X_{ij}) - \widehat{g}^{(l)}(\boldsymbol{X}_{-ij})\big]$.
- Variance estimate: $(\widehat{\sigma}_{j|\boldsymbol{X}_{-j}}^{(l)})^2 = n^{-1}\sum_{i=1}^n (d_{i4} - n^{-1}\sum_{i=1}^n d_{i4}^{(l)})^2$.

General FACT test

Rejection criterion

- For fixed threshold $t > 0$, we reject H_0 when $\mathsf{FACT}_j > t$.
- If we choose $t = \Phi^{-1}(\alpha/(4|Q|))$, then the test has asymptotic size α.
- The asymptotic p-value is conservatively estimated as $\min\{1, 4|Q|\Phi(-\mathsf{FACT}_j)\}$.

General FACT test

- The imbalancing technique lies in partition the inference sample to k_n groups. Therefore, for each. Also, as mentioned before, the choice $k_n = \log n$ could provide appealing performance. individual test statistic, the training sample size is way larger than the inference sample size.
- The conditional and ensemble technique are naturally deployed.

Out-of-bag estimation

Problem when the sample size is not large, doing splitting might lose power. Out-of-bag estimation may serve as alternative.
Let's say we have K tree estimates, and let $a_k \subset [n]$ be the index set of the samples used to fit the k-th tree.
For each $i \in [n]$, denote $A(i) \subset [K]$ be the collection of the trees k such that $i \notin a_k$. Then we can replace the corresponding random forest estimate in the FACT statistic with the average of the tree estimate in A_i.

Outline

Preliminaries

FACT inference method

- Vanilla version
- Bias issue and debiasing techniques
- Ensemble test and general version

Theoretical guarantees

- Technical conditions
- Null behavior
- Power analysis

Reference

Technical conditions

Let's begin with some conditions that simplifies our proof.

Condition 3.

Assume that $\mathbb{E}\big\{\big[\mathbb{E}(Y|\boldsymbol{X}_{-j}) - \widehat{Y}(\boldsymbol{X}_{-j})\big]\big[g(X_j) - \mathbb{E}(g(X_j))\big]\big|\mathcal{X}_0\big\} = 0$ a.s.

Condition 4.

Assume that the measurable transformation $g(\cdot)$ is bounded between 0 and 1 on its domain. In addition, assume that Var$(g(X_j)|\boldsymbol{X}_{-j}) \geq \varsigma_1$, Var$(Y|\boldsymbol{X}) \geq \varsigma_2$, and Var$(Y|\boldsymbol{X}_{-j}) \leq D$ almost surely, and that $\mathbb{E}Y^4 \leq D_2$ for some constants $\varsigma_1, \varsigma_2, D, D_2 > 0$.

The condition 3 provides condition for the basic FACT test to be effective. And the condition 4 assures that the denominator doesn't vanish.

Theoretical guarantees for basic p-value

Theorem 1.
For all large n, each consistency rate $0 < B_1 < 1$, and any $1 \le j \le p$ such that 1) conditions 1–4 hold, 2) null hypothesis is true for X_j, then for each $t > 0$,

$$\mathbb{P}\big(|\mathsf{FACT}_{j,0}| > t\big) \le 2\Phi(-t) + \frac{8c}{5\sqrt{\varsigma_2}\varsigma_1} + 2(\log n)(B_1^{1/4} + n^{-1/4}), \qquad (6)$$

where $\mathsf{FACT}_{j,0}$ is the basic FACT statistic defined in (1), $c = (\log n)(2B_1^{1/4} + n^{-1/4}\log n)(1 + |t|)$, ς_1 and ς_2 are given in condition 4.

Theoretical guarantees for basic p-value

Observations

- The theorem is non-asymptotic and is true for all n sufficiently large.
- The value of t is independent to n and p.
- If we choose $t = -\Phi^{-1}(\alpha/2)$, then it suffices to let

  $$B_1^{1/4} = o(1/\log n)$$

  to let the asymptotic size fall below α.
- Last requirement is true when $n = N$, because $B_1 = O(N^{-c})$ for some c.
- The non-asymptotic p-value is the upper bound on the RHS, which involves n, p and B_1. For simplicity we may only consider the asymptotic p-value.

Theoretical guarantees for imbalancing p-value

Unbiasedness condition 3 is too strong to be used. And the next theorem provides non-asymptotic guarantee under general cases.

Theorem 2.

For all large n, each consistency rate $0 < B_1 < 1$, each $t > 0$, and any $1 \leq j \leq p$ such that 1) *Conditions 1 and 4 hold and X_j is a null feature, it holds that*

$$\mathbb{P}\left(\left|\mathsf{FACT}_{j\mathsf{N/n}}\right| > t\right) \leq 2\Phi(-t) + \frac{8c}{5\sqrt{\varsigma_2\varsigma_1}} + (\log n)(B_1^{1/4} + n^{-1/4}) + (-\log B_1)^{-1},$$

where $\mathsf{FACT}_{j\mathsf{N/n}}$ is the imbalancing form of the FACT statistic and $c = (\log n)(2B_1^{1/4} + n^{-1/4}\log n)(1+t) + \sqrt{nB_1}(-\log B_1)$.

Theoretical guarantees for imbalancing

Observations

- The condition 3 is too stringent, especially for the dependent features. Last theorem however, doesn't require the condition 3.
- In order for the test to be non-trivial, it suffices to let $\sqrt{nB_1}(-\log B_1) = o(1)$, which requires $n = o(N)$, since B_1 is generally slower than $O(N^{-1})$.

Theoretical guarantees for conditioning

Condition 5.

Assume that $\mathbb{E}\{\mathbb{E}(g(X_j)|\boldsymbol{X}_{-j}) - \widehat{g}(\boldsymbol{X}_{-j})\}^2 \leq B_2$ for some small $B_2 > 0$.

Theorem 3.

For all large n, each $t > 0$, all consistency rates $0 < B_1, B_2 < 1$, and null feature j conditions 1 and 4–5 hold random forests estimates $\widehat{Y}(\boldsymbol{X}_{-j})$ and $\widehat{g}(\boldsymbol{X}_{-j})$ are constructed with $\mathcal{X}_0$, we have

$$\mathbb{P}\left(\left|\mathsf{FACT}_{j|\boldsymbol{X}_{-j}}\right| > t\right) \leq 2\Phi(-t) + \frac{8c}{5\sqrt{\varsigma_2\varsigma_1}} + (\log n)(n^{-1/4} + B_1^{1/4} + B_2^{1/4}) + (-\log(B_1B_2))$$

where $\mathsf{FACT}_{j|\boldsymbol{X}_{-j}}$ *is the conditioning form of the FACT statistic defined in* (3) *and* $c = tn^{-1/4}\log n + (2t+1)(2B_1^{1/4} + B_2^{1/4}) + \sqrt{nB_1B_2}(-\log(B_1B_2))$.

Theoretical guarantees for conditioning

In order for the conditioning test to be valid, it suffices to let $B_1 + B_2 = o(1)$ and $\sqrt{nB_1B_2}(-\log(B_1B_2)) = o(1)$. The latter condition is actually less restrictive than the one required before.

Theoretical gurantees for ensembling

Continue with the notations before, let $g^{(l)}, l \in L$ be the collection of all the possible transformations. We have following theorem

Theorem 4.

For all large n, when condition 1 holds and Conditions 3–4 are satisfied for each transformation $g^{(l)}(X_j)$ and X_j is a null feature, we have

$$\mathbb{P}\left(\cup_{l \in L}\left\{\left|\mathsf{FACT}_j^{(l)}\right| > t\right\}\right) \leq 2|L|\Phi(-t) + \frac{8|L|c}{5\sqrt{\varsigma_2}\varsigma_1} + (\log n)(B_1^{1/4} + n^{-1/4}), \quad (7)$$

where $c = (\log n)(2B_1^{1/4} + n^{-1/4}\log n)(1+t)$.

Theoretical guarantees for general test

Theorem 5.

For all large n, each Q with $1 \leq |Q| < n$, each $t > 0$, all consistency rates $0 < B_1, B_2 < 1$, and any $1 \leq j \leq p$ such that 1) *Condition 1 holds and Conditions 4–5 are satisfied for $g^{(l)}(X_j)$ and $\widehat{g}^{(l)}(\boldsymbol{X}_{-j})$ for each $l \in \{1, 2\}$ with $g^{(1)}(x) = x$ and $g^{(2)}(x) = x^2$, and X_j is a null feature, we have*

$$
\begin{aligned}
\mathbb{P}\left(\mathsf{FACT}_j > t\right) \leq 4|Q|\Phi(-t) + \frac{16|Q|c}{5\sqrt{\varsigma_2}\varsigma_1} + 2(-\log\left(B_1 B_2\right))^{-1} \\
+ |Q|(\log n)\left(\left(\frac{|Q|}{n - |Q|}\right)^{1/3} + n^{-1/4} + B_1^{1/4} + B_2^{1/4}\right),
\end{aligned}
\tag{8}
$$

where FACT_j *is the general FACT statistic defined in* (5) *and* $c = tn^{-1/4}\log n + (2t+1)(2B_1^{1/4} + B_2^{1/4}) + \sqrt{nB_1B_2}(-\log\left(B_1 B_2\right))$.

Power analysis: ensemble test

We begin with two population quantities.

$$\begin{aligned}\kappa_j^{(l)} &\coloneqq \mathbb{E}\big\{[Y - \mathbb{E}(Y|\boldsymbol{X}_{-j})][g^{(l)}(X_j) - \mathbb{E}(g^{(l)}(X_j))]\big\},\\ \kappa_{j|\boldsymbol{X}_{-j}}^{(l)} &\coloneqq \mathbb{E}\big\{[Y - \mathbb{E}(Y|\boldsymbol{X}_{-j})][g^{(l)}(X_j) - \mathbb{E}(g^{(l)}(X_j)|\boldsymbol{X}_{-j})]\big\}\end{aligned} \tag{9}$$

These quantities characterize the mean of the numerators in our test statistic.

Power analysis: ensemble test

Theorem 6.

Let $C > 0$ be some sufficiently large constant. For all $n \geq 1$, each $B_1 > 0$, each $t > 0$, and any $1 \leq j \leq p$ such that 1) *Condition 1 holds and Conditions 3–4 are satisfied for each measurable transformation $g^{(l)}(X_j)$,* 2) *the random forests estimate $\widehat{Y}(\boldsymbol{X_{-j}})$ is constructed based on the independent training sample $\mathcal{X}_0$, and* 3) $\sum_{l \in L} |\kappa_j^{(l)}| > 0$*, it holds that*

$$\mathbb{P}\left(\cap_{l \in L}\left\{\left|\mathsf{FACT}_j^{(l)}\right| \leq t\right\}\right) \leq \frac{(C+t)\left[\sqrt{B_1} + \sqrt{\mathsf{Var}(Y)}\right]}{\sqrt{n}\sum_{l \in L} |\kappa_j^{(l)}|}, \qquad (10)$$

where $\mathsf{FACT}_j^{(l)}$*'s are the ensemble form of the FACT statistics defined in* (4).

Power analysis: ensemble test

Observations

- In the last result, we can let $|L| = 1$ to get the power result for the basic FACT test.
- In order for the asymptotic power to approach one, it suffices to have

$$\sum_{l \in L} |\kappa_j^{(l)}| \gg n^{-1/2}.$$

- Later we will give some lower bound for some specific models.

Power analysis: general test

Theorem 7.

Let $C > 0$ be some sufficiently large constant. For all $n \geq 2$, each $|Q| < n$, each $B_1, B_2 > 0$, each $t > 0$, and null feature X_j such that Condition 1 holds and Conditions 4–5, and $|\kappa^{(1)}_{j|\boldsymbol{X}_{-j}}| + |\kappa^{(2)}_{j|\boldsymbol{X}_{-j}}| > 0$, it holds that

$$\mathbb{P}\left(\mathsf{FACT}_j \leq t\right) \leq \frac{\sqrt{|Q|}(C+t)\big(\mathsf{Var}(Y) + \sqrt{B_1} + \sqrt{B_2} + \sqrt{nB_1B_2}\big)}{\sqrt{n-|Q|}\sum_{l=1}^{2}\left|\kappa^{(l)}_{j|\boldsymbol{X}_{-j}}\right|},$$

where FACT_j is the general FACT test statistic defined in (5)*.*

Power analysis: additive model

We give some lower bound result for the quantities defined above, in this additive model.

Condition 6.

Assume that the nonparametric regression model is given by $Y = h(X_j) + H(\boldsymbol{X}_{-j}) + \varepsilon$, where $h(\cdot)$ and $H(\cdot)$ are some measurable functions and ε is the model error that is of zero mean and independent of the random feature vector $\boldsymbol{X}$. In addition, assume that the distribution of feature vector $\boldsymbol{X}$ has a density function.

Power analysis: additive model

Proposition 3.1.

Assume that Condition 6 holds, $g^{(1)}(x) = x$, $\mathbb{E}|H(\boldsymbol{X}_{-j})| < \infty$, $h(\cdot)$ is monotonic, and the derivative of function $h(\cdot)$ is integrable and bounded in absolute value. Then we have

$$|\kappa_j^{(1)}| \geq \left(\inf_{x \in [0,1]} |h'(x)| \right) \mathbb{E}\{\mathsf{Var}(X_j | \boldsymbol{X}_{-j})\}.$$

Power analysis: additive model

Proposition 3.2.

Assume that Condition 6 holds for some $1 \le j \le p$, $h(x) = a_0 + \sum_{l \in L} a_l x^l$ *for some* $a_l \in \mathbb{R}$ *with* $\sum_{l \in L} |a_l| > 0$, $\mathbb{E}|H(\boldsymbol{X}_{-j})| < \infty$, $\boldsymbol{X}$ *is uniformly distributed on* $[0,1]^p$, *and* $g^{(l)}(x) = x^l$ *for* $l \in L$. *Then, there exists some positive constant* $c_{|L|}$ *depending on* $|L|$ *such that*

$$\min\Big\{ \sum_{l \in L} |\kappa_j^{(l)}|, \sum_{l \in L} |\kappa_{j|\boldsymbol{X}_{-j}}^{(l)}| \Big\} \ge \frac{\sum_{l \in L} |a_l|}{|L|} c_{|L|} > 0.$$

Particularly, for $1 \le |L| \le 2$, *we have* $\min\big\{ \sum_{l \in L} |\kappa_j^{(l)}|, \sum_{l \in L} |\kappa_{j|\boldsymbol{X}_{-j}}^{(l)}| \big\} \ge 0.001 \times \sum_{l \in L} |a_l|$.

Outline

Preliminaries

FACT inference method

- Vanilla version
- Bias issue and debiasing techniques
- Ensemble test and general version

Theoretical guarantees

- Technical conditions
- Null behavior
- Power analysis

Reference

Reference