Significance tests for feature relevance of a black-box learner

Ben Dai

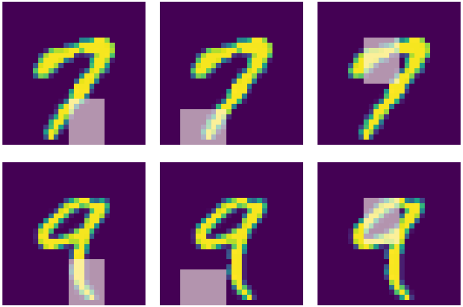
Department of Statistics
The Chinese University of Hong Kong

(Joint work with Xiaotong Shen and Wei Pan)

https://arxiv.org/abs/2103.04985

https://dnn-inference.readthedocs.io

Illustrative example



- **Question**: Can we provide a **valid** p**-value** for **any pre-specified region** (features) based on **a black-box model**, such as a convolutional neural network?

Motivation

- **Why significance tests?** Hypothesis testing, feature interpretation, XAI, make black-box models more reliable ...
- **Why region tests?** For image analysis, the impact of each pixel is negligible but a pattern of a collection of pixels (e.g. a region) may instead become salient.
- **Why black-box models?** Significant improvement in prediction performance, which enforce us to believe that a black-box model is a better option to model real data. For example, use a CNN to formulate image data.

Difficulty

- **Black-box models**. It is infeasible (or difficult) to "open the box", that is, we only access the input and output for a black-box model, and do not know its inner structure.
- **Feature-param correspondence**. The feature-parameter correspondence is unclear for black-box models, such as CNNs and RNNs.
- **High-dimensional hypothesized features**. The dimension of the hypothesized features could be extremely large.
- **Computationally expensive**. Refitting a deep learning model is computationally expensive.

Difficulty

- **Overparametrized models**. When the number of parameters increase, both training / testing errors decrease, and the training error could be zero.

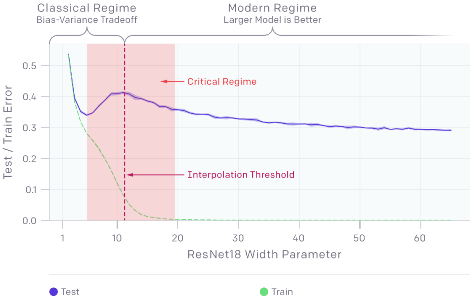


Figure 1: Source[1]

[1] https://openai.com/blog/deep-double-descent/

Issues for existing methods

- Likelihood Ratio Test (LRT)
  - **black-box models and overparam**: Taylor expansion is infeasible, and the training loss could be very small.
  - **feature-param relation** is unclear: LRT works for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ vs. $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$.

Issues for existing methods

- Likelihood Ratio Test (LRT)
  - **black-box models and overparam**: Taylor expansion is infeasible, and the training loss could be very small.
  - **feature-param relation** is unclear: LRT works for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ vs. $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$.
- Conditional randomization test (CRT; [Candès et al., 2018]) and holdout randomization test (HRT; [Tansey et al., 2018])
  - significance testing for a single feature.
  - require conditional Prob of hypothesized feature given the others. It is usually difficult to estimate for real complex datasets, or **high-dimensional hypothesized features**.

Issues for existing methods

- Likelihood Ratio Test (LRT)
  - **black-box models and overparam**: Taylor expansion is infeasible, and the training loss could be very small.
  - **feature-param relation** is unclear: LRT works for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ vs. $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$.
- Conditional randomization test (CRT; [Candès et al., 2018]) and holdout randomization test (HRT; [Tansey et al., 2018])
  - significance testing for a single feature.
  - require conditional Prob of hypothesized feature given the others. It is usually difficult to estimate for real complex datasets, or **high-dimensional hypothesized features**.
- Leave-one-covariate-out (LOCO; [Lei et al., 2018])
  - significance testing for a single feature.
  - finite-sample hypothesis testing.

Goodness

- Input and output: $\boldsymbol{X} \in \mathbb{R}^d$ and $\boldsymbol{Y} \in \mathbb{R}^K$;
  - **large-scale dataset** $(\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^N$
- Black-box model: $f : \mathbb{R}^d \to \mathbb{R}^K$;
  - **good performance, or small generalization error, or reasonable convergence rate**
- Flexible computing platform for a general loss function $l(f(\boldsymbol{X}), \boldsymbol{Y})$
  - **TensorFlow**, **Keras**, **Pytorch**

The proposed risk-based testing

- **Goal**: testing the relevance of a sub-feature $\boldsymbol{X}_{\mathcal{S}} = \{X_j : j \in \mathcal{S}\}$ to the outcome $\boldsymbol{Y}$ without specifying any form of the prediction function, where $\mathcal{S}$ is an index set of hypothesized features.
- **Our testing**: directly compare perf w/- or w/o hypothesized features

The proposed risk-based testing

- **Goal**: testing the relevance of a sub-feature $\boldsymbol{X}_{\mathcal{S}} = \{X_j : j \in \mathcal{S}\}$ to the outcome $\boldsymbol{Y}$ without specifying any form of the prediction function, where $\mathcal{S}$ is an index set of hypothesized features.
- **Our testing**: directly compare perf w/- or w/o hypothesized features
  - Masked data $(\boldsymbol{Z}, \boldsymbol{Y})$, with permutation of $\boldsymbol{Z}_{\mathcal{S}}$ or $\boldsymbol{Z}_{\mathcal{S}} = \mathbf{0}$, and $\boldsymbol{Z}_{\mathcal{S}^c} = \boldsymbol{X}_{\mathcal{S}^c}$.
  - Risk functions:

$$R(f) = \mathbb{E}\big(l(f(\boldsymbol{X}), \boldsymbol{Y})\big), \quad R_{\mathcal{S}}(g) = \mathbb{E}\big(l(g(\boldsymbol{Z}), \boldsymbol{Y})\big)$$

  - Population minimizer:

$$f^* = \operatorname*{argmin}_f R(f), \quad g^* = \operatorname*{argmin}_g R_{\mathcal{S}}(g)$$

$$H_0 : R(f^*) - R_{\mathcal{S}}(g^*) = 0, \quad \text{versus} \quad H_a : R(f^*) - R_{\mathcal{S}}(g^*) < 0. \quad (1)$$

The proposed risk-based testing

Relationships among the risk invariance hypothesis in (2), marginal independence, and conditional independence; the latter two are defined as:

$$\text{Marginal indep: } \boldsymbol{Y} \perp \boldsymbol{X}_{\mathcal{S}}, \quad \text{conditional indep : } \boldsymbol{Y} \perp \boldsymbol{X}_{\mathcal{S}} \mid \boldsymbol{X}_{\mathcal{S}^c}.$$

Lemma 1 (Relation to independence)

For any loss function, conditional independent implies risk invariance, or

$$\boldsymbol{Y} \perp \boldsymbol{X}_{\mathcal{S}} \mid \boldsymbol{X}_{\mathcal{S}^c} \implies R(f^*) - R_{\mathcal{S}}(g^*) = 0.$$

Moreover, if the cross-entropy loss $l(f(\boldsymbol{X}), Y) = -\mathbf{1}_Y^\intercal \log(f(\boldsymbol{X}))$ *is used in* (2), *then* H_0 *is equivalent to conditional independence almost surely under the marginal distribution of* $\boldsymbol{X}$.

The proposed risk-based testing

- (Constant loss): $l(f(\boldsymbol{X}), Y) = C$.
- (L_2-loss): $l(f(\boldsymbol{X}), Y) = \mathbb{E}(Y - f(\boldsymbol{X}))^2$.
- (Cross-entropy loss): $l(f(\boldsymbol{X}), Y) = \mathbf{1}_{\boldsymbol{Y}}^{\intercal} \log(f(\boldsymbol{X}))$.

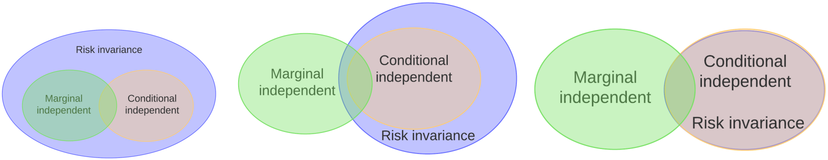
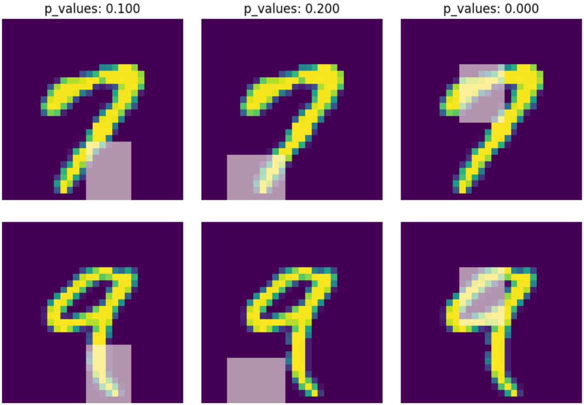


Figure 2: Three cases illustrate relationships among marginal independence, conditional independence, and risk invariance.

Our solution

- The proposed test is able to produce a valid p-value for (2).
- Python library `dnn-inference` (`https://dnn-inference.readthedocs.io`)



Splitting data

- Recall the proposed hypothesis:

$$H_0 : R(f^*) - R_{\mathcal{S}}(g^*) = 0, \quad \text{versus} \quad H_a : R(f^*) - R_{\mathcal{S}}(g^*) < 0. \quad (2)$$

- **Empirically** $\{$ estimate (f^*, g^*), evaluate $(R, R_{\mathcal{S}})$ $\}$

Splitting data

- Recall the proposed hypothesis:

$$H_0 : R(f^*) - R_{\mathcal{S}}(g^*) = 0, \quad \text{versus} \quad H_a : R(f^*) - R_{\mathcal{S}}(g^*) < 0. \quad (2)$$

- **Empirically** { estimate (f^*, g^*), evaluate $(R, R_{\mathcal{S}})$ }

Question: do we need to split data? **Yes!**

Splitting data

- Recall the proposed hypothesis:

$$H_0 : R(f^*) - R_{\mathcal{S}}(g^*) = 0, \quad \text{versus} \quad H_a : R(f^*) - R_{\mathcal{S}}(g^*) < 0. \quad (2)$$

- **Empirically** { estimate (f^*, g^*), evaluate $(R, R_{\mathcal{S}})$ }

Question: do we need to split data? **Yes!**

- *Deep neural networks easily fit shuffled pixels, random pixels.* See Figure 1 in [Zhang et al., 2016]: training loss converge to zero under random pixels, yet the testing loss is still sensible.
- Theoretically, it is not easy to find a limiting distribution based on a black-box model.

Splitting data

- Splitting data into **estimation** and **inference** sets

$$\text{Total set } (\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^N \to \text{Est set } (\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^n + \text{Inf set } (\boldsymbol{X}_{n+j}, \boldsymbol{Y}_{n+j})_{j=1}^m$$

$$(\boldsymbol{Z}_i, \boldsymbol{Y}_i)_{i=1}^n \qquad (\boldsymbol{Z}_{n+j}, \boldsymbol{Y}_{n+j})_{j=1}^m$$

- Obtain estimator $(\widehat{f}_n, \widehat{g}_n)$ based on estimation set, then plug into evaluation on an inference sample:

$$\widehat{R}(\widehat{f}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n)$$

- **Question**: Is it good estimation of $R(f^*) - R_{\mathcal{S}}(g^*)$? Asymptotic null distribution?

Decomposition

- Compare $\widehat{R}(\widehat{f}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n)$ with $R(f^*) - R_{\mathcal{S}}(g^*)$
- Consider the following decomposition

$$\begin{aligned}\widehat{R}(\widehat{f}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n) &= \widehat{R}(\widehat{f}_n) - R(\widehat{f}_n) + R_{\mathcal{S}}(\widehat{g}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n) \\ &+ R(\widehat{f}_n) - R(f^*) + R_{\mathcal{S}}(g^*) - R_{\mathcal{S}}(\widehat{g}_n) \\ &+ R(f^*) - R_{\mathcal{S}}(g^*) = T_1 + T_2 + T_3\end{aligned}$$

- T_1 is a conditional IID sum

$$\begin{aligned}T_1 = \widehat{R}(\widehat{f}_n) - R(\widehat{f}_n) + R_{\mathcal{S}}(\widehat{g}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n) \\ = \frac{1}{m}\sum_{j=1}^{m}\Big(\Delta_{n,j} - \mathbb{E}\big(\Delta_{n,j}\big|(\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^{n}\big)\Big),\end{aligned}$$

where $\Delta_{n,j} = l(\widehat{f}_n(\boldsymbol{X}_{n+j}), \boldsymbol{Y}_{n+j}) - l(\widehat{g}_n(\boldsymbol{Z}_{n+j}), \boldsymbol{Y}_{n+j})$

Decomposition

- T_2 converges to zero in probability for **good estimators** (peak performance for black-box models)

$$
\begin{aligned}
T_2 &= R(\widehat{f}_n) - R(f^*) + R_{\mathcal{S}}(g^*) - R_{\mathcal{S}}(\widehat{g}_n) \\
&\leq max\{R(\widehat{f}_n) - R(f^*), R_{\mathcal{S}}(\widehat{g}_n) - R_{\mathcal{S}}(g^*)\} \quad \underbrace{= O_P(n^{-\gamma})}_{\text{reasonable assumption}}
\end{aligned}
$$

In the literature, the convergence rate $\gamma > 0$ has been extensively investigated [Wasserman, 2006, Schmidt-Hieber et al., 2020].

- T_3 is related to H_0

$$
T_3 = R(f^*) - R(g^*) = 0, \quad \text{under } H_0
$$

Motivation

- **Main idea**:
  - **Normalize** T_1 by its standard derivation, which can be estimated by a sample standard derivation of evaluations on the inference set. Then, the normalized T_1 follows $N(0,1)$ asymptotically by CLT.
  - After normalization, T_2 is **convergence in probability** when $n \to \infty$, and $T_3 = 0$ under H_0.
- Consider the following test statistic:

$$\frac{\sqrt{m}}{\widehat{\sigma}_n}\big(\widehat{R}(\widehat{f}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n)\big) = \frac{\sum_{j=1}^{m} \Delta_{n,j}}{\sqrt{m}\widehat{\sigma}_n} = \frac{\sqrt{m}}{\widehat{\sigma}_n} T_1 + \frac{\sqrt{m}}{\widehat{\sigma}_n} T_2 + \frac{\sqrt{m}}{\widehat{\sigma}_n} T_3,$$

where $\widehat{\sigma}_n$ is a sample standard deviation of differenced evaluations on inference set, that is, $\{\Delta_{n,j}\}_{j=1}^{m}$.

Motivation

- **Main idea**:
  - **Normalize** T_1 by its standard derivation, which can be estimated by a sample standard derivation of evaluations on the inference set. Then, the normalized T_1 follows $N(0,1)$ asymptotically by CLT.
  - After normalization, T_2 is **convergence in probability** when $n \to \infty$, and $T_3 = 0$ under H_0.
- Consider the following test statistic:

$$\frac{\sqrt{m}}{\widehat{\sigma}_n}\big(\widehat{R}(\widehat{f}_n) - \widehat{R}_{\mathcal{S}}(\widehat{g}_n)\big) = \frac{\sum_{j=1}^{m} \Delta_{n,j}}{\sqrt{m}\widehat{\sigma}_n} = \frac{\sqrt{m}}{\widehat{\sigma}_n} T_1 + \frac{\sqrt{m}}{\widehat{\sigma}_n} T_2 + \frac{\sqrt{m}}{\widehat{\sigma}_n} T_3,$$

  where $\widehat{\sigma}_n$ is a sample standard deviation of differenced evaluations on inference set, that is, $\{\Delta_{n,j}\}_{j=1}^{m}$.
- Asymptotically Normally Distributed? **It may be WRONG!!**

Bias-sd-ratio issue

One unusual issue for the test statistic is varnishing standard deviation:

$$\text{Under } H_0, \text{ if } \widehat{f}_n \xrightarrow{p} f^*, \widehat{g}_n \xrightarrow{p} g^*, \text{ and } f^* = g^*, \text{ then } \widehat{\sigma}_n \xrightarrow{p} 0$$

Issues:

- **CLT may not hold for T_1.** CLT requires a standard derivation is fixed, or bounded away from zero.
- **Bias-sd-ratio.** Both bias and sd are convergence to zeros:

$$\frac{\sqrt{m}T_2}{\widehat{\sigma}_n} = \sqrt{m}\Big(\frac{R(\widehat{f}_n) - R(f^*) + R_{\mathcal{S}}(g^*) - R_{\mathcal{S}}(\widehat{g}_n)}{\widehat{\sigma}_n}\Big) = \sqrt{m}\Big(\frac{bias \xrightarrow{p} 0}{sd \xrightarrow{p} 0}\Big).$$

  - If T_2 and $\widehat{\sigma}_n$ are in the same order, $\sqrt{m}\widehat{\sigma}_n^{-1} T_2 = O_P(\sqrt{m})$, kills the null distribution.

Solution

- The issue is caused by vanishing standard deviation, we can address it by perturbation.
- **One-split test**. The proposed test statistic is given as:

$$\Lambda_n^{(1)} = \frac{\sum_{j=1}^m \Delta_{n,j}^{(1)}}{\sqrt{m}\widehat{\sigma}_n}, \quad \Delta_{n,j}^{(1)} = \Delta_{n,j} + \rho_n \varepsilon_j, \tag{3}$$

  where $\widehat{\sigma}_n$ is the sample standard derivation based on $\{\Delta_{n,j}^{(1)}\}_{j=1}^m$ conditional on $\widehat{f}_n$ and $\widehat{g}_n$, $\rho_n \to \rho$ is a level of perturbation.
- Note that $\widehat{\sigma}_n^{(1)} \to \sigma^{(1)} \geq \rho > 0$.

Decomposition

Reconsider the decomposition of $\Lambda_n^{(1)}$:

$$\Lambda_n^{(1)} = \underbrace{\frac{\sqrt{m}}{\widehat{\sigma}_n^{(1)}}\Big(\frac{1}{m}\sum_{j=1}^{m}\big(\Delta_{n,j}^{(1)} - \mathbb{E}(\Delta_{n,j}^{(1)}|\mathcal{E}_n)\big)\Big)}_{\to N(0,1) \text{ by conditional CLT of triangular array}}$$

$$+ \underbrace{\frac{\sqrt{m}}{\widehat{\sigma}_n^{(1)}}\Big(R(\widehat{f}_n) - R(f^*) - \big(R_{\mathcal{S}}(\widehat{g}_n) - R_{\mathcal{S}}(g^*)\big)\Big)}_{= O_p(m^{1/2}n^{-\gamma}) \text{ by prediction consistency}} + \underbrace{\frac{\sqrt{m}}{\widehat{\sigma}_n^{(1)}}\big(R(f^*) - R_{\mathcal{S}}(g^*)\big)}_{= 0 \text{ under } H_0}.$$

- If the splitting condition $m^{1/2}n^{-\gamma} = o_p(1)$ is satisfied, then $\Lambda_n^{(1)} \xrightarrow{d} N(0,1)$ under H_0

Asymptotic null distribution

- **Assumption A** (Prediction consistency). For some constant $\gamma > 0$, $(\widehat{f}_n, \widehat{g}_n)$ satisfies

$$\big(R(\widehat{f}_n) - R(f^*)\big) - \big(R_{\mathcal{S}}(\widehat{g}_n) - R_{\mathcal{S}}(g^*)\big) = O_p(n^{-\gamma}). \tag{4}$$

- **Assumptions B-C** are standard assumptions for CLT under triangle arrays [Cappé et al., 2006].

Theorem 2 (Asymptotic null distribution of $\Lambda_n^{(1)}$)

In addition to Assumptions A, B, and C, if $m = o(n^{2\gamma})$, then under H_0,

$$\Lambda_n^{(1)} \xrightarrow{d} N(0,1), \quad as \quad n \to \infty. \tag{5}$$

According to the asymptotic null distribution of $\Lambda_n^{(1)}$ in Theorem 2, we calculate the p-value $P^{(1)} = \Phi(\Lambda_n^{(1)})$.

Power analysis

Consider an alternative hypothesis $H_a : R(f^*) - R_{\mathcal{S}}(g^*) = -m^{-1/2}\delta < 0$ for $\delta > 0$. The power functions of the one-split test and its combined test can be written as

$$\pi_n(\delta) = \mathbb{P}(P^{(1)} \leq \alpha | H_a), \quad \bar{\pi}_n(\delta) = \mathbb{P}(\bar{P}^{(1)} \leq \alpha | H_a).$$

Theorem 3 (Local limiting power of the one-split test)

Suppose that the one-split test (3) *satisfies Assumptions A-C and* $m = o(n^{2\gamma})$*, then*

$$\lim_{n\to\infty} \inf \pi_n(\delta) = \Phi\Big(\frac{\delta}{\sigma^{(1)}} - z_\alpha\Big), \quad \textit{and} \quad \lim_{\delta\to\infty} \lim_{n\to\infty} \inf \pi_n(\delta) = 1, \qquad (6)$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ *is the z-multiplier of the standard normal distribution.*

Splitting condition

- **Question**: How to determine the estimation / inference ratio? $m = o(n^{2\gamma})$ for an unknown $\gamma > 0$.
- **Log-ratio sample splitting scheme.** Specifically, given a sample size $N \geq N_0$, the estimation and inference sizes n and m are obtained:

$$n = \lceil x_0 \rceil, \quad m = N - n,$$

$$\text{where } x_0 \text{ is a solution of} \quad \{x + \frac{N_0}{2\log(N_0/2)} \log(x) = N\}. \tag{7}$$

- Splitting ratio condition is automatically satisfied!

Lemma 4 (Log-ratio sample splitting scheme)

The estimation and inference sample sizes (n, m)*, determined by the log-ratio sample splitting formula* (7)*, satisfies* $m = o(n^{2\gamma})$ *for any* $\gamma > 0$ *in Assumption A.*

More comments

- **Power.** Heuristic data-adaptive sample splitting scheme.
- **Two-split test.** One-split test is valid for any perturbation $\rho > 0$, if you don't like a custom parameter, use **two-split test** (further splitting an inference sample into two equal subsamples yet the perturbation is not required).
- **CV.** Combining p-values over repeated random splitting.

Algorithm

Algorithm 1 One-split test for feature relevance to prediction

Input: Data: $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^N$; Set of hypothesized feats: $\mathcal{S}$; #splitting: U
Output: p-value for testing (2)
Determine the splitting ratio ξ and the perturbation level ρ (log-ratio or data-adaptive scheme)
for $u = 1, \cdots, U$ **do**
    Shuffle the data
    Split the data into estimation / inference sets, where $m = \hat{\zeta}N$ and $n = N - m$
    Compute $\Lambda_u^{(1)}$ from (3)
    Compute p-value $P_u^{(1)} = \Phi(\Lambda_u^{(1)})$
end for
Compute the combined p-value $\bar{P}^{(1)}$

- Just fit a DL model U-times, U can be as small as 1.

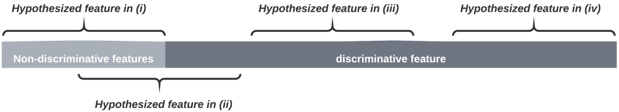
Numerical experiments

- **Comparison with existing black-box tests.**
  - Sample size $N = 1000$, dimension $p = 5$.
  - $\boldsymbol{X} = (X_1, \cdots, X_5)^\intercal$ follows a uniform distribution on $[-1, 1]$ with a pairwise correlation $\rho_{ij} = 0.5^{|i-j|}$.
  - $Y = 0.02(X_1 + X_2 + X_3) + 0.05\epsilon$

Test	Return	H_0	
One-split	p-value	risk-invariance $R(f^*) = R_{\mathcal{S}}(g^*)$,	0.003
Two-split	p-value	risk-invariance $R(f^*) = R_{\mathcal{S}}(g^*)$	0.018
HRT	p-values for all feats	conditional indep $\boldsymbol{X}_j \perp \boldsymbol{Y} \mid \boldsymbol{X}_{-j}$	(0.840, 0.045, 0.064, 0.900, 0.158)
LOCO	p-values for all feats	equal errors with/without feat j for a given dataset	(0.132, 0.791, 0.180, 0.435, 0.342)
PT	p-value	marginal indep $\boldsymbol{X}_{\mathcal{S}} \perp \boldsymbol{Y}$	0.010
HPT	p-value	marginal indep $\boldsymbol{X}_{\mathcal{S}} \perp \boldsymbol{Y}$	0.001

Numerical experiments

- Simulation for a neural network: $Y = f^*(\boldsymbol{X}) + \epsilon$.
  - $f^*(\boldsymbol{x})$ is a neural network. $\boldsymbol{X} \sim N(\boldsymbol{0}, B\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}_{ij} = r^{|i-j|}$, $r \in [0, 1)$
  - $f^*(\boldsymbol{x}) = g^*(\boldsymbol{z})$ only depends on a subset of features of $\boldsymbol{x}$, in which $\boldsymbol{z}_{\mathcal{S}_0} = \boldsymbol{0}$ and $\boldsymbol{z}_{\mathcal{S}_0^c} = \boldsymbol{x}_{\mathcal{S}_0^c}$ with $\mathcal{S}_0 = \{1, \cdots, |\mathcal{S}_0|\}$.
  - Given a hypothesized index set $\mathcal{S}$, our goal is to test if $\boldsymbol{X}_{\mathcal{S}}$ is relevant to predicting the outcome Y.



- (i) $\mathcal{S} \cup \mathcal{S}_0 = \mathcal{S}_0$ for Type I error. (ii)-(iv): $\mathcal{S} \cup \mathcal{S}_0 \neq \mathcal{S}_0$ for power.
- (ii) $\rightarrow$ (iv), the distance (or correlation) between $\mathcal{S}$ and $\mathcal{S}_0$ is increasing (or decreasing), thus the power is expected to go up.

Numerical experiments

Example 1. (*Impact of the sample size*) This example (Table 1) concerns the performance of the proposed tests in relation to the sample size N based on *data-adaptive* tuning methods, where N ranges from 2000 to 10000, $B = 0.4$, $r = 0.25$, $p = 100$, $\varpi = 128$, $\tau = 2$, $L = 3$, $|\mathcal{S}_0| = 5$.

Test	Sample size	Type I error	Power	Time (Second)
One-split	2000	0.043	(0.25, 0.79, 0.85)	15.2(0.1)
	6000	0.050	(0.61, 0.99, 1.00)	41.2(0.3)
	10000	0.049	(0.89, 1.00, 1.00)	66.0(0.4)
Two-split	2000	0.050	(0.11, 0.26, 0.31)	14.0(0.1)
	6000	0.035	(0.18, 0.51, 0.58)	37.0(0.2)
	10000	0.040	(0.19, 0.77, 0.75)	61.6(0.4)
Comb. one-split	2000	0.034	(0.26, 1.00, 0.95)	37.9(0.1)
	6000	0.046	(0.86, 1.00, 1.00)	68.3(0.3)
	10000	0.045	(1.00, 1.00, 1.00)	107.2(0.7)
Comb. two-split	2000	0.015	(0.09, 0.26, 0.29)	38.0(0.1)
	6000	0.030	(0.10, 0.70, 0.65)	76.3(0.5)
	10000	0.014	(0.13, 0.93, 0.92)	110.3(0.5)

Table 1: Empirical Type I errors and powers of the one-split and two-split tests, their combined tests in Example 1 at a nominal level $\alpha = 0.05$.

Numerical experiments

Example 2. (*Impact of the strength of features of interest*) This example (Table 2) concerns the performance of the proposed tests with respect to the magnitude of hypothesized features B, where $B = 0.2, 0.4, 0.6$, $N = 6000$, $p = 100$, $r = 0.25$, $\varpi = 128$, $\tau = 2$, $L = 3$, and $|\mathcal{S}_0| = 5$.

Test	B	Type I error	Power
One-split	0.2	0.057	(0.24, 0.68, 0.78)
	0.4	0.050	(0.61, 0.99, 1.00)
	0.6	0.057	(0.97, 1.00, 1.00)
Two-split	0.2	0.049	(0.06, 0.12, 0.14)
	0.4	0.035	(0.18, 0.51, 0.58)
	0.6	0.041	(0.37, 0.97, 0.98)
Comb. one-split	0.2	0.027	(0.27, 0.93, 0.93)
	0.4	0.046	(0.86, 1.00, 1.00)
	0.6	0.033	(1.00, 1.00, 1.00)
Comb. two-split	0.2	0.019	(0.00, 0.00, 0.03)
	0.4	0.030	(0.10, 0.70, 0.65)
	0.6	0.012	(0.45, 1.00, 1.00)

Table 2: Empirical Type I errors and powers of the one-split and two-split tests, and their combined tests in Example 2 at a nominal level $\alpha = 0.05$. The data-adaptive tuning scheme is applied.

Numerical experiments

- **Example 3.** (*Impact of the depth and width of a neural network*) This example concerns the performance of the proposed tests in terms of the width ϖ and depth L of a neural network, where $N = 6000$, $L = 2, 3, 4$, $\varpi = 32, 64, 128$, $B = 0.4$, $r = 0.25$, $p = 100$, $\tau = 2$, $L = 3$, and $|\mathcal{S}_0| = 5$.
- **Example 4.** (*Impact of the number of hypothesized features*) This example concerns the proposed tests with respect to the number of hypothesized features $|\mathcal{S}_0| = 3, 5, 10$.
- **Example 5.** (*Impact of feature correlations*) This example concerns the proposed tests in terms of the feature correlation $r = .2, .4, .6$.
- **Example 6.** (*Impact of different modes of combining p-values*) This example concerns the combined tests with different ways of combining p-values, including the Hommel, the Bonferroni, the first quantile, the median, the Cauchy, and the harmonic methods.

Numerical experiments

- **Role of perturbation**.
  - $\mathcal{S}_0 = \{1, 2, 3\}$, $\boldsymbol{X} \sim N(\mathbf{0}, B\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}_{ij} = r^{|i-j|}$, $r \in [0, 1)$
  - $\Sigma_{1j} = \Sigma_{j1} = .1$; $j = 1, \cdots, p$, and $\Sigma_{ij} = 0$, if $i, j \neq 1$ and $i \neq j$.
  - Only partial features are observed in a sample $(\boldsymbol{x}_i^{(N)}, y_i^{(N)})_{i=1}^N$, where $\boldsymbol{x}_i^{(N)} = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{id_N})^\intercal$ and $y_i^{(N)} = f^*(\tilde{\boldsymbol{x}}_i^{(N)}) + \epsilon_i$, $d_N \to d$ as $N \to \infty$, and $\tilde{\boldsymbol{x}}_i^{(N)} = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{id_N}, 0, \cdots, 0)^\intercal$ is a d-dimensional vector.

Test	$N = 2000$	$N = 6000$	$N = 10000$
One-split without perturbation	0.083	0.109	0.193
One-split with perturbation	0.057	0.053	0.061
Two-split	0.048	0.051	0.047

Table 3: Type I errors of the one-split tests with and without perturbation and the two-split test in Section 6.4 at a nominal level $\alpha = 0.05$.

Summary

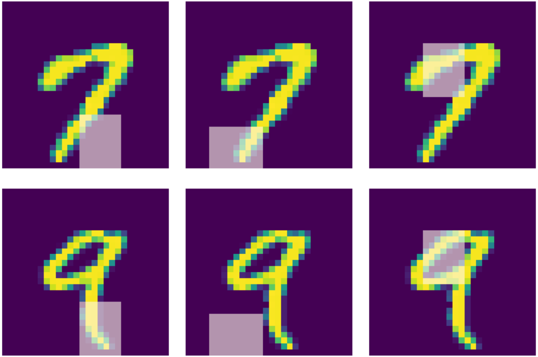
- Summarize the simulation results.

		Advantage	Evidence
Test	One-split	*More powerful*	Tables 3-5
	Two-split	*No need to perturb data*	Equation (14)
Combine	Comb.	*More powerful*	Tables 3-5
	Non-comb.	*Less computation time*	Table 3
Ratio	Data-adaptive	*More powerful*	Tables 3-5
	Log-ratio	*No need to tune the ratio, and less computation time*	Lemma 4, Table 3

Table 4: Advantage for different tests, combining, and tuning methods.

Real application

- The MNIST handwritten digits dataset [LeCun et al., 1998]. In particular, we extract $14,251$ images from the dataset with labels '7' and '9' to discriminate between these two digits.



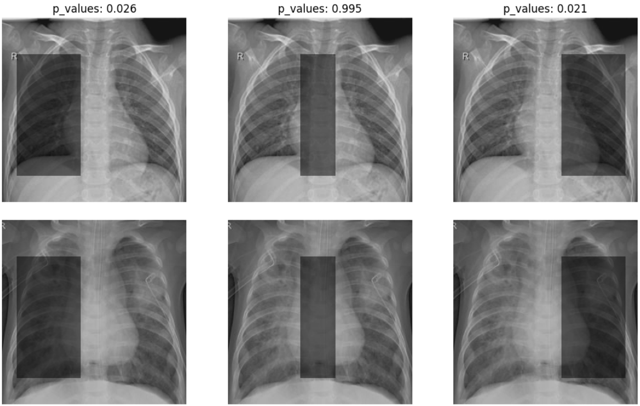
Real application

Test	p-values (case 1, case 2, case 3)	Time(Second)
One-split	(0.174, 0.329, 0.000)	4289
Two-split	(0.959, 0.569, 0.000)	4772
Comb. one-split	(0.385, 1.000, 0.000)	11404
Comb. two-split	(0.544, 0.192, 0.000)	13060

Table 5: P-values and runtimes of the one-split and two-split tests, their combined tests, and the permutation test in the MNIST benchmark example at a nominal level $\alpha = 0.05$.

Real application

- pneumonia diagnosis dataset [Kermany et al., 2018]. This dataset consists of 5,863 X-ray images, each labeled as "Pneumonia" or "Normal."



Real application

Test	p-values (case 1, case 2, case 3)	Time(Second)
One-split	(0.026, 0.995, 0.021)	15242
Two-split	(0.212, 0.561, 0.065)	14020
Comb. one-split	(0.041, 0.635, 0.075)	64416
Comb. two-split	(0.053, 0.754, 0.084)	64761

Table 6: P-values and runtimes of the one-split and two-split tests, and their combined tests in the chest X-ray dataset at a nominal level $\alpha = 0.05$.

Contribution

- A **novel risk-based hypothesis** is proposed in (2), as well as its relation to conditional independence tests.
- We derive the **one-split/two-split tests** based on the differenced empirical loss with and without hypothesized features. Theoretically, we show that the one-split and two-split tests, as well as their combined tests, can **control the Type I error** while being **consistent in terms of power**;
- The proposed tests only require **a limited number of refitting**, and we develop the Python library `dnn-inference` and examine the utility of the proposed tests on various simulated examples and two real datasets.

Thank you!

References I

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018).
Panning for gold: Model-x knockoffs for high dimensional controlled variable selection.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(3):551–577.

Cappé, O., Moulines, E., and Rydén, T. (2006).
Inference in hidden Markov models.
Springer Science & Business Media.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018).
Identifying medical diagnoses and treatable diseases by image-based deep learning.
Cell, 172(5):1122–1131.

References II

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998).
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2324.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018).
Distribution-free predictive inference for regression.
Journal of the American Statistical Association, 113(523):1094–1111.

Schmidt-Hieber, J. et al. (2020).
Nonparametric regression using deep neural networks with relu activation function.
Annals of Statistics, 48(4):1875–1897.

References III

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2018).
The holdout randomization test: Principled and easy black box feature selection.
arXiv preprint arXiv:1811.00645.

Wasserman, L. (2006).
All of nonparametric statistics.
Springer Science & Business Media.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016).
Understanding deep learning requires rethinking generalization.
arXiv preprint arXiv:1611.03530.