Testing Conditional Mean Independence Using Generative Neural Networks

Yi Zhang, Linjun Huang, Yun Yang & Xiaofeng Shao

UNIVERSITY OF MARYLAND · WashU · UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Conditional mean independence (CMI)

- For random vectors $X \in \mathbb{R}^{d_X}$, $Y \in \mathbb{R}^{d_Y}$ and $Z \in \mathbb{R}^{d_Z}$, we test the null hypothesis

$$H_0 : \mathbb{E}[Y|X = x, Z = z] = \mathbb{E}[Y|Z = z] \ a.e. \ (x, z) \in \mathbb{R}^{d_X+d_Z}$$

against

$$H_1 : \mathbb{P}\big(\mathbb{E}[Y|X, Z] \neq \mathbb{E}[Y|Z]\big) > 0$$

given iid samples $(X_i, Y_i, Z_i)_{i=1}^n$.

- Conditional mean independence (CMI) testing plays an important role in various areas of statistics and machine learning.
  1. In traditional statistical applications, such as nonparametric regression, CMI testing identifies subsets or functions of covariates that are useful to predict the response variable (omitted variable testing, significance testing).
  2. Variable importance measure is related to CMI [10].
  3. In machine learning, CMI testing has broad applications in areas like interpretable machine learning [8] and representation learning [1, 6].

Challenges and Motivations

1, Performance deterioration in high dimensional setting.

- This issue primarily arises from the estimation of the conditional mean functions $r(z) := \mathbb{E}[Y|Z = z]$ and $m(x, z) := \mathbb{E}[Y|X = x, Z = z]$.
- Early CMI tests, such as those in [5, 4], relied on kernel smoothing methods.
- Consequently, these CMI tests suffer from the curse of dimensionality: their performance declines significantly as the dimensions d_Z, $d_X + d_Z$ are moderate or large [11, Section 1].

2. Theoretical size guarantee.

- Most existing CMI tests rely on sample estimation of the population CMI measure $\Gamma := \mathbb{E}\big[(r(Z) - m(X, Z))^2 w(X, Z)\big]$ or its equivalent forms, where w is a positive weight function.
- Γ uniquely characterize CMI: $\Gamma = 0$ if and only if H_0 holds.
- A common plug-in estimator of Γ is given by

$$\widehat{\Gamma}(\widehat{r}, \widehat{m}) = n^{-1} \sum_{i=1}^{n} (\widehat{r}(Z_i) - \widehat{m}(X_i, Z_i))^2 w(X_i, Z_i),$$

where $\widehat{r}$ and $\widehat{m}$ are nonparametric estimators of the conditional mean functions.

- Two key issues:
  1. $\widehat{\Gamma}(\widehat{r}, \widehat{m})$ suffers from a degeneracy problem: under H_0, $\widehat{\Gamma}(\widehat{r}, \widehat{m})$ converges to zero at a rate faster than the $n^{-1/2}$ rate at which $\widehat{\Gamma}(r, m) - \Gamma$ converges to a non-degenerate limiting distribution under the alternative [5, Section 1].
  2. The nonparametric estimation errors for $r(z)$ and $m(x, z)$ typically decay slower than the $n^{-1/2}$ parametric rate, and the convergence rate of $\widehat{\Gamma}(\widehat{r}, \widehat{m})$ under H_0 depends heavily on how quickly these errors decay.

3, Weak power against local alternatives.

CMI tests in [5, 10, 3] fail to detect local alternatives with signal strength $\Delta_n := \sqrt{\mathbb{E}[(r(Z) - m(X, Z))^2]}$ of order $n^{-1/2}$.

1. Test in [5] takes the form $nh^{s/2}\widehat{\Gamma}$, where $h \to 0$ is a kernel smoothing bandwidth parameter, $s = d_Z$ or $d_X + d_Z$, and it cannot detect local alternatives converging to the null faster than $n^{-1/2}h^{-s/4}$.
2. Tests in [10, 3] use the population CMI measure $\Gamma_0 = \Gamma_1 - \Gamma_2$, where $\Gamma_1 = \mathbb{E}[(Y - r(Z))^2]$ and $\Gamma_2 = \mathbb{E}[(Y - m(X, Z))^2]$, which is equivalent to Γ. Since the quadratic terms Γ_1 and Γ_2 can only be estimated at the $n^{-1/2}$ rate, these tests can only detect local alternatives with Δ_n of order $n^{-1/4}$.
3. [2] employs an unequal sample splitting approach, with proportionally more data dedicated to conditional mean functions estimation, which may result in significant power loss in practice.

Our Approach

A new test that addresses all three challenges based on a novel population CMI measure.

1. The sample version of the population measure is in multiplicative form, which is key to mitigating the impact of estimation errors in nonparametric nuisance parameters (i.e., the conditional mean functions).
2. Our test requires estimating $r(z) = \mathbb{E}[Y|Z = z]$ and the conditional mean embedding (CME) of X given Z into a reproducing kernel Hilbert space (RKHS) on the space of X.
3. The CME is estimated using the Monte Carlo method with samples generated from a trained generative neural network (GNN).

Appealing Features of our method

1. Good empirical performance when d_X, d_Y, d_Z are large.
2. The test achieves asymptotic size control under H_0.
3. The test exhibits nontrivial power against local alternatives in an $n^{-1/2}$-neighborhood of H_0.

Background: CME

- Let $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\| \cdot \|_{\mathbb{H}}$ denote the associated inner product and the induced norm of a generic RKHS $\mathbb{H}$ with kernel $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

$$\mathbb{H} := \overline{\text{span}}\{f(x) = \sum_{i=1}^{n} a_i \mathcal{K}(x_i, x) : a_i \in \mathbb{R}, x_i \in \mathbb{R}^d, n = 1, 2, \dots\}$$

- For two generic random variables W, V taking values in the domain of $\mathbb{H}$. $\mathbb{E}[\mathcal{K}(W, \cdot)]$ is the (kernel) mean embedding of P_W into $\mathbb{H}$, which is the unique element in $\mathbb{H}$ such that $\mathbb{E}[f(W)] = \langle f, \mathbb{E}[\mathcal{K}(W, \cdot)]\rangle_{\mathbb{H}}$ for any $f \in \mathbb{H}$.

Background: Conditional Sampling

Noise-outsourcing Lemma

For any integer $m \geq 1$, there exist measurable function $\mathbb{G}_X$ such that for any $\eta \sim N(0, I_m)$ that is independent of (X, Z), we have $\mathbb{G}_X(\eta, Z) \,|\, Z \sim P_{X|Z}$.

- $\mathbb{G}_X$ can be estimated by GNN $\widehat{\mathbb{G}}_X : \mathbb{R}^m \times \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$.
- To estimate the CME $\mathbb{E}[\mathcal{K}_X(X, \cdot) \,|\, Z = z]$ for any $z \in \mathbb{R}^{d_Z}$, one can first generate M i.i.d. samples of $\{\eta_i\}_{i=1}^M$ from $N(0, I_m)$, and then estimate the CME by the sample average $\widehat{\mathbb{E}}[\mathcal{K}_X(X, \cdot) \,|\, Z = z] := M^{-1} \sum_{i=1}^M \mathcal{K}_X(\widehat{\mathbb{G}}_X(\eta_i, z), \cdot)$.

Background: GMMN

- The generative moment matching networks (GMMN) (we call it conditional generator) $\widehat{\mathbb{G}}_X$ for approximating $P_{X|Z}$ is obtained by minimizing the sample version of the squared Maximum Mean Discrepancy (MMD) between P_{XZ} and the induced joint distribution $P_{\widehat{X}Z}$ from the estimated $\widehat{X} = \widehat{\mathbb{G}}_X(\eta, Z)$ based on a generic set of training data $\{(X_i, Z_i)\}_{i=1}^{n_T}$ with training sample size n_T and Mn_T latent variables $\{\eta_i^m : i = 1, \dots, n_T, \ m = 1, \dots, M\}$:

$$\widehat{\mathbb{G}}_X = \mathop{\arg\min}_{\mathbb{G}_X \in \mathcal{G}_X} \frac{1}{n_T(n_T-1)} \sum_{\substack{k \neq \ell \\ k,\ell \in [n_T]}} \widehat{U}(X_k, X_\ell) \cdot \mathcal{K}_Z(Z_k, Z_\ell), \quad (1)$$

$$\text{with} \quad \widehat{U}(X_k, X_\ell) = \mathcal{K}_X(X_k, X_\ell) - \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_X\big(X_k, \mathbb{G}_X(\eta_\ell^m, Z_\ell)\big) - \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_X\big(X_\ell, \mathbb{G}_X(\eta_k^m, Z_k)\big) + \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_X\big(\mathbb{G}_X(\eta_k^m, Z_k), \mathbb{G}_X(\eta_\ell^m, Z_\ell)\big),$$

where $\mathcal{G}_X$ is an approximation family, such as (deep) neural networks, for the conditional generators.

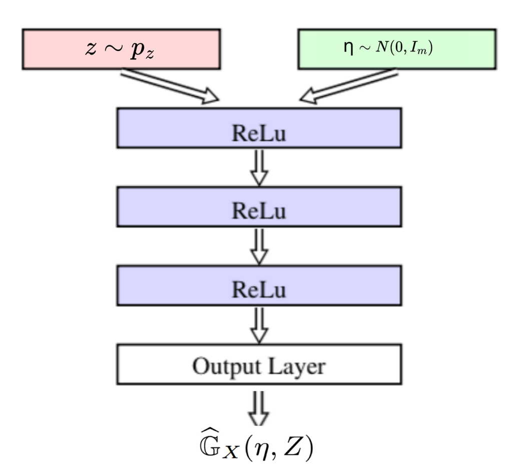


Figure 1. Example architecture of GMMN.

Population CMI Measure

Proposition 1

If $\mathbb{E}[\|Y\|_2^2] < \infty$, then the following properties are equivalent to each other:

a $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z]$ a.s.-$\mathbb{P}_{XZ}$.
b $\mathbb{E}\big[\big(f(X, Z) - \mathbb{E}[f(X, Z)|Z]\big)\, Y\big] = 0$ for any $f \in L_2(\mathbb{R}^{d_X+d_Z}, \mathbb{P}_{XZ})$.
c $\mathbb{E}\big[\big(f(X, Z) - \mathbb{E}[f(X, Z)|Z]\big)\, \big(Y - \mathbb{E}[Y|Z]\big)\big] = 0$ for any $f \in L_2(\mathbb{R}^{d_X+d_Z}, \mathbb{P}_{XZ})$.

- Let $\mathcal{K}_X : \mathbb{R}^{d_X} \times \mathbb{R}^{d_X}$ and $\mathcal{K}_Z : \mathbb{R}^{d_Z} \times \mathbb{R}^{d_Z}$ denote two symmetric positive-definite kernel functions that define two reproducing kernel Hilbert spaces (RKHS) $\mathbb{H}_X$ and $\mathbb{H}_Z$ over the spaces of X and Z, respectively.
- Let $\mathcal{K}_0 = \mathcal{K}_X \times \mathcal{K}_Z$ with $\mathbb{H}_0$ being the corresponding RKHS induced by $\mathcal{K}_0$.
- Define linear operator $\Sigma : \mathbb{R}^{d_Y} \to \mathbb{H}_0$,

$$\Sigma\, c = \mathbb{E}\Big\{\Big[\mathcal{K}_0\big((X, Z), \cdot\big) - \mathbb{E}\big[\mathcal{K}_0\big((X, Z), \cdot\big)|Z\big]\Big]\big[Y - \mathbb{E}[Y|Z]\big]^\top c\Big\}, \quad \text{for any } c \in \mathbb{R}^{d_Y}.$$

From the reproducing property, we see that for any $f \in \mathbb{H}_0$ and any $c \in \mathbb{R}^{d_Y}$,

$$\langle f, \Sigma\, c\rangle_{\mathbb{H}_0} = \mathbb{E}\Big\{\big[f(X, Z) - \mathbb{E}[f(X, Z)|Z]\big]\big[Y - \mathbb{E}[Y|Z]\big]^\top c\Big\}.$$

- Assume $\mathbb{H}_0$ is dense in $L_2(\mathbb{R}^{d_X+d_Z}, P_{XZ})$, which holds if $\mathcal{K}_X$ and $\mathcal{K}_Z$ are L_2- or c_0-universal kernels [9, Theorem 5], such as the Gaussian and Laplacian kernels.
- H_0 holds if and only if Σ is the zero operator (i.e., $\Sigma c = 0 \in \mathbb{H}_0$ for any $c \in \mathbb{R}^{d_Y}$).

Our proposed population CMI measure is defined as

$$\Gamma^* = \mathbb{E}\big[U(X, X')\, V(Y, Y')\, \mathcal{K}_Z(Z, Z')\big], \quad (2)$$

where $V(Y, Y') = \big[Y - g_Y(Z)\big]^\top \big[Y' - g_Y(Z')\big]$ and $U(X, X') = \mathcal{K}_X(X, X') - \langle g_X(Z), \mathcal{K}_X(X', \cdot)\rangle_{\mathbb{H}_X} - \langle g_X(Z'), \mathcal{K}_X(X, \cdot)\rangle_{\mathbb{H}_X} + \langle g_X(Z), g_X(Z')\rangle_{\mathbb{H}_X}$. Here, (X', Y', Z') is an independent copy of (X, Y, Z), $g_Y(\cdot) = \mathbb{E}[Y|Z = \cdot] \in \mathbb{R}^{d_Y}$ and $g_X(\cdot) = \mathbb{E}[\mathcal{K}_X(X, \cdot) \,|\, Z = \cdot] \in \mathbb{H}_X$.

- The squared Hilbert-Schmidt norm of Σ satisfies $\|\Sigma\|_{\text{HS}}^2 = \Gamma^*$.
- H_0 holds if and only if $\Gamma^* = 0$.

Sample Estimation

- Sample version of Γ^* takes the form of a U-statistic.
- $\langle g_X(Z_i), \mathcal{K}_X(X_j, \cdot)\rangle_{\mathbb{H}_X} = \mathbb{E}\big[\mathcal{K}_X(X_i, X_j)\big|Z_i, X_j\big]$ can be estimated by:

$$\frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_X(X_i^{(m)}, X_j),$$

where $\{X_i^{(m)}\}_{m=1}^M$ are sampled from the (estimated) conditional distribution $P_{X_i|Z_i}$.

- $\langle g_X(Z_i), g_X(Z_j)\rangle_{\mathbb{H}_X} = \mathbb{E}\big[\mathcal{K}_X(X_i, X_j)\big|Z_i, Z_j\big]$ can be estimated by:

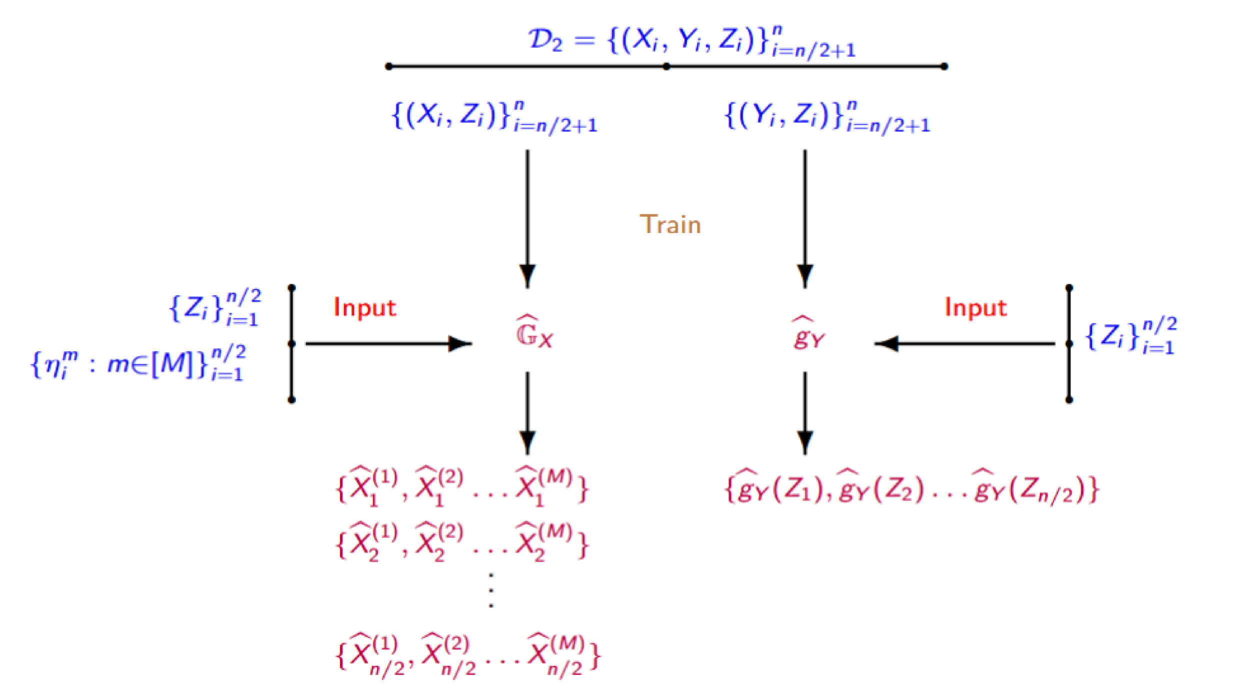
$$\frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_X(X_i^{(m)}, X_j^{(m)}),$$

- $g_Y : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_Y}$ is estimated by a DNN $\widehat{g}_Y$.

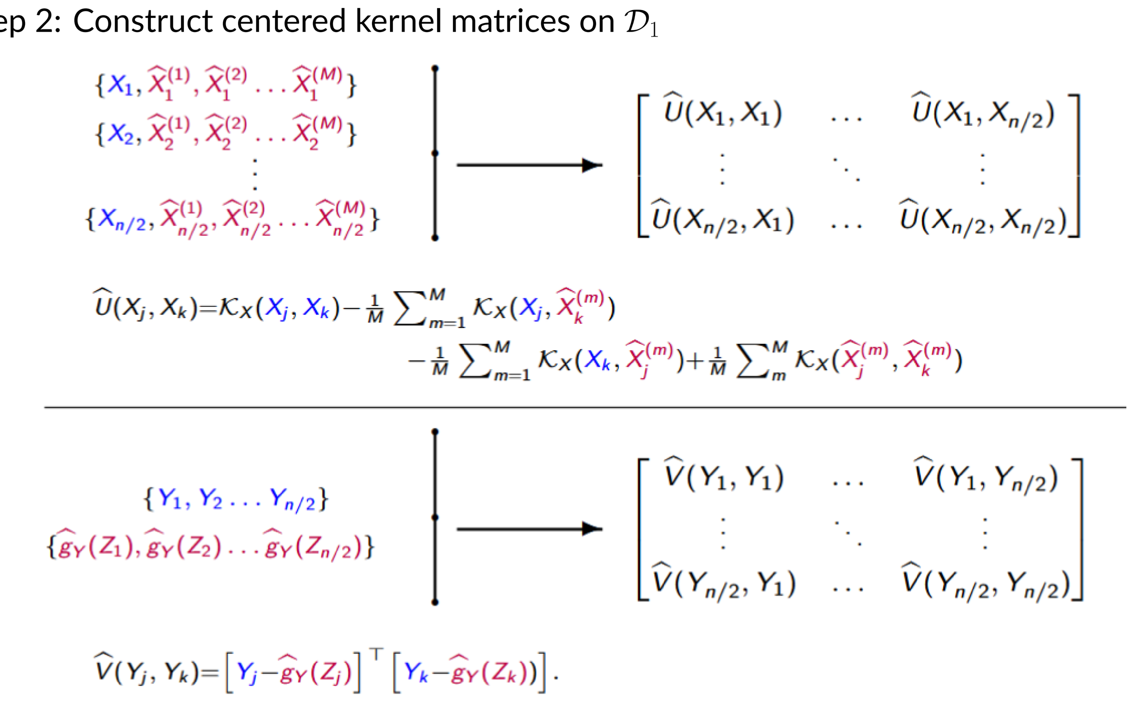
Implementation

Adopt a sample splitting and cross fitting framework to train GNNs. Divide the samples into two folds: $\mathcal{D}_1 = \{(X_i, Y_i, Z_i)\}_{i=1}^{n/2}$ and $\mathcal{D}_2 = \{(X_i, Y_i, Z_i)\}_{i=n/2+1}^{n}$.

Step 1: Networks training + synthetic data generation



Step 2: Construct centered kernel matrices on $\mathcal{D}_1$



Step 3: Calculate sample version of Γ^*

$$\widehat{T}_1 = \frac{1}{\frac{n}{2}(\frac{n}{2} - 1)} \sum_{j,k \in [n/2], j \neq k} \widehat{U}(X_j, X_k)\widehat{V}(Y_j, Y_k)\mathcal{K}_Z(Z_j, Z_k) \quad (3)$$

Step 4: Switch the role of $\mathcal{D}_1$ and $\mathcal{D}_2$ to claculate $\widehat{T}_2$, then our statistic is

$$\widehat{T}_n = (\widehat{T}_1 + \widehat{T}_2)/2 \quad (4)$$

A Wild Bootstrap Procedure for Test Calibration

- For each $b = 1, 2, \dots, B$, generate n i.i.d. random multipliers $\{e_{bi}\}_{i=1}^n$ from the standard normal distribution $N(0, 1)$.
- A bootstrap version of $\widehat{T}_n$ is then defined as

$$\widehat{T}_n^b = \frac{1}{2}\sum_{s=1}^{2}\Bigg\{\frac{1}{\frac{n}{2}(\frac{n}{2} - 1)} \sum_{\substack{j \neq k \\ X_j, X_k \in \mathcal{D}_s}} \widehat{U}(X_j, X_k)\widehat{V}(Y_j, Y_k)\mathcal{K}_Z(Z_j, Z_k)e_{bj}\, e_{bk}\Bigg\}. \quad (5)$$

- We then reject H_0 at level $\gamma \in (0, 1)$ if $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{\widehat{T}_n^b > \widehat{T}_n\}} < \gamma$.

Real Application

We examine whether covering specific regions of a facial image X affects the prediction of facial expression Y using the FER2013 dataset.

Figure 2. Original facial images in FER2013 (first column) and the covered images with HRs: TL, nose, right eye, mouth, left eye, eyes, face (Columns 2-8). From row 1 to 7, the expressions are 'angry', 'disgust', 'fear', 'happy', 'sad', 'surprise', 'neutral'.

- Following [3], we consider covering 7 hypothesized regions (HRs): top left corner (TL), nose, right eye, mouth, left eye, eyes, and face.
- We use 11,700 image-label pairs $\{(X_i, Y_i)\}_{i=1}^{11700}$.
  - X_i are 48×48 grayscale images.
  - $Y_i \in$ {'angry', 'disgust', 'fear', 'happy', 'sad', 'surprise', 'neutral'} represented by $\{e_i\}_{i=1}^7 \subset \mathbb{R}^7$: vectors with the ith component being one and the rest being zero.
  - Z_i is X_i with some HR covered in black.
- Test $H_0 : \mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z]$ for different HR.
- $\widehat{T}_n$ is calculated 10 times on different samples (size $n = 2000$) generated using stratified sampling.
- DSP$_M$ statistics [3] are evaluated under 0-1 loss and CE loss.
- Compare test accuracies from a VGG network [7] trained on (Y_i, Z_i) against the baseline accuracy from VGG net trained on (Y_i, X_i).

$\widehat{T}_n$ correctly identifies the nose and TL as non-discriminative regions, while rejecting H_0 for other HRs, consistent with their lower test accuracies. DSP$_M$ p-values vary by loss function, with CE loss exhibiting stronger detecting power but inflated type-I error for TL and nose.

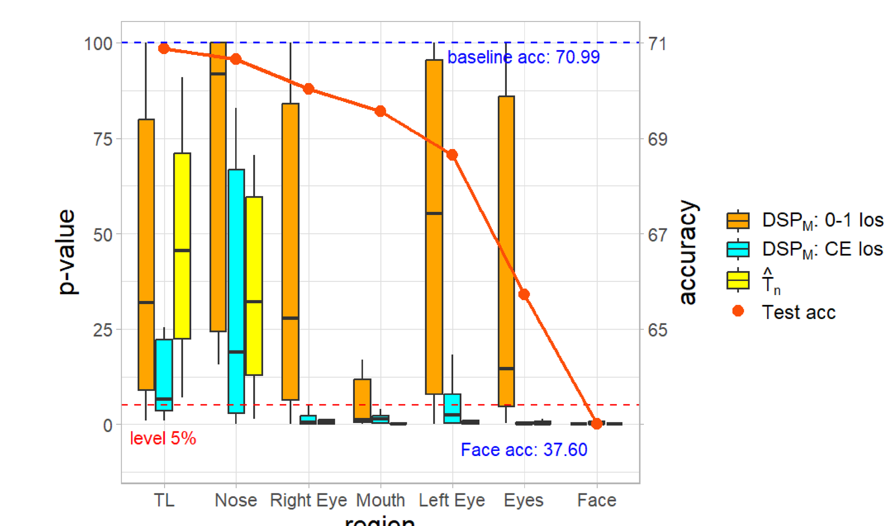


Figure 3. Box plot of the p-values (left y-axis) and the test accuracies (red line, right y-axis) for different HRs.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[2] Leheng Cai, Xu Guo, and Wei Zhong. Test and measure for partial mean dependence based on machine learning methods. *Journal of the American Statistical Association*, 0(ja):1–32, 2024.

[3] Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a black-box learner. *IEEE transactions on neural networks and learning systems*, 35(2):1898–1911, 2022.

[4] Miguel A Delgado and Wenceslao González Manteiga. Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics*, 29(5):1469–1507, 2001.

[5] Yanqin Fan and Qi Li. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, pages 865–890, 1996.

[6] Jian Huang, Yuling Jiao, Xu Liao, Jin Liu, and Zhou Yu. Deep dimension reduction for supervised representation learning. *IEEE Transactions on Information Theory*, 2024.

[7] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021.

[8] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[9] Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[10] Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

[11] Xingyu Zhou, Yuling Jiao, Jin Liu, and Jian Huang. A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1–12, 2022.

shaox@wustl.edu