An Asymptotic Test for Conditional Independence using Analytic Kernel Embeddings

M. Scetbon*

L. Meunier*

Y. Romano

ENSAE IP PARIS

FACEBOOK AI

Dauphine UNIVERSITÉ PARIS | PSL

TECHNION Israel Institute of Technology

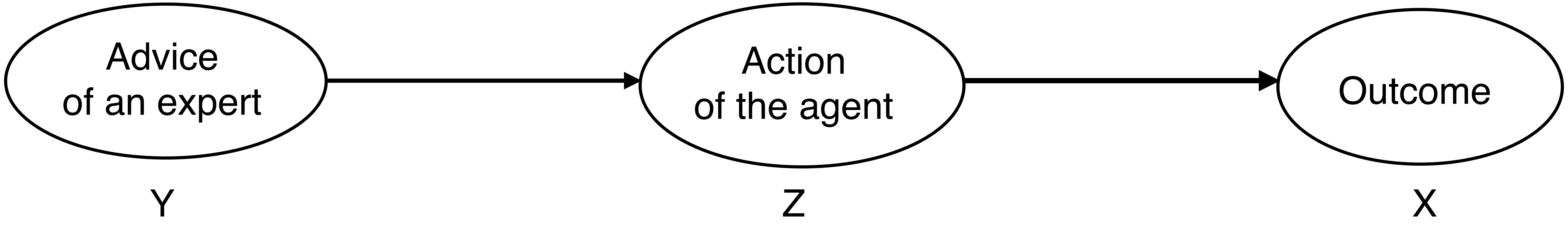
ICML International Conference On Machine Learning

Thirty-ninth International Conference on Machine Learning

* Authors contributed equally.

Conditional Independence Testing

A Simple Example:



This graph shows that the outcome does not depend on the advice given the action taken by the agent:

$$X \perp\!\!\!\perp Y \mid Z$$

Question: How to infer from data such relationships between random variables?

Test for Conditional Independence:

Goal: Given i.i.d samples $(X_i, Z_i, Y_i)_{i=1}^n \sim P_{XZY}$ where P_{XZY} is the law of (X, Z, Y) a random vector, we aim at testing the null Hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ against $H_1 : X \not\!\perp\!\!\!\perp Y \mid Z$.

→ We design a new kernel-based test statistic to test for conditional independence

ℓ^p Distance Between Mean Embeddings

Definition:

Let k be a definite positive, characteristic, continuous, bounded and ***analytic*** kernel on $\mathbb{R}^d$ and $p \geq 1$ an integer. Let also P, Q two probability distributions on $\mathbb{R}^d$ and denote respectively $\mu_{P,k}$ and $\mu_{Q,k}$ their mean embeddings. Then

$$d_{p,J}(P,Q) := \left[\frac{1}{J}\sum_{j=1}^{J} |\mu_{P,k}(\mathbf{t}_j) - \mu_{Q,k}(\mathbf{t}_j)|^p\right]^{\frac{1}{p}}$$

where $(\mathbf{t}_j)_{j=1}^J$ are sampled independently from any absolutely continuous Borel probability measure is random metric on the space of probability measures.

A First Characterization of the Conditional Independence:

- Let $d_x, d_y, d_z \geq 1$, $\mathcal{X} := \mathbb{R}^{d_x}$, $\mathcal{Y} := \mathbb{R}^{d_y}$, and $\mathcal{Z} := \mathbb{R}^{d_z}$. Let (X, Z, Y) be a random vector on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ with law P_{XZY}.
- Denote $\ddot{X} := (X, Z)$, $\ddot{\mathcal{X}} := \mathcal{X} \times \mathcal{Z}$ and let us define for all mesurable $(A, B) \in \mathcal{B}(\ddot{\mathcal{X}}) \times \mathcal{B}(\mathcal{Y})$:

$$P_{\ddot{X}\otimes Y|Z}(A \times B) := \mathbb{E}_Z\left[\mathbb{E}_{\ddot{X}}[\mathbf{1}_A | Z]\mathbb{E}_Y[\mathbf{1}_B | Z]\right].$$

Proposition: $d_{p,J}(P_{XZY}, P_{\ddot{X}\otimes Y|Z}) = 0$ if and only if $X \perp Y | Z$ a.s.

A first Oracle Statistic

- For all $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$, we have $\mu_{P_{\ddot{X} \otimes Y|Z}, k_{\ddot{\mathcal{X}}} \cdot k_{\mathcal{Y}}}(\mathbf{t}^{(1)}, t^{(2)}) = \mathbb{E}_Z \left[\mathbb{E}_{\ddot{X}} \left[k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{X}) \,|\, Z \right] \mathbb{E}_Y \left[k_{\mathcal{Y}}(t^{(2)}, Y) \,|\, Z \right] \right]$
- For all $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$, we have $\mu_{P_{XZY}, k_{\ddot{\mathcal{X}}} \cdot k_{\mathcal{Y}}}(\mathbf{t}^{(1)}, t^{(2)}) = \mathbb{E} \left[k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{X}) k_{\mathcal{Y}}(t^{(2)}, Y) \right]$
- For all $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$, we define the witness function:

$$\boxed{\Delta(\mathbf{t}^{(1)}, t^{(2)}) := \mu_{P_{\ddot{X} \otimes Y|Z}, k_{\ddot{\mathcal{X}}} \cdot k_{\mathcal{Y}}}(\mathbf{t}^{(1)}, t^{(2)}) - \mu_{P_{XZY}, k_{\ddot{\mathcal{X}}} \cdot k_{\mathcal{Y}}}(\mathbf{t}^{(1)}, t^{(2)})}$$

Reformulation of the Witness Function:

$$\Delta(\mathbf{t}^{(1)}, t^{(2)}) = \mathbb{E} \left[\left(k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{X}) - \mathbb{E}_{\ddot{X}} \left[k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{X}) \,|\, Z \right] \right) \left(k_{\mathcal{Y}}(t^{(2)}, Y) - \mathbb{E}_Y \left[k_{\mathcal{Y}}(t^{(2)}, Y) \,|\, Z \right] \right) \right]$$

A First Estimate of the Witness Function:

$$\Delta_n(\mathbf{t}^{(1)}, t^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} \left(k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{x}_i) - \mathbb{E}_{\ddot{X}} \left[k_{\ddot{\mathcal{X}}}(\mathbf{t}^{(1)}, \ddot{X}) \,|\, z_i \right] \right) \left(k_{\mathcal{Y}}(t^{(2)}, y_i) - \mathbb{E}_Y \left[k_{\mathcal{Y}}(t^{(2)}, Y) \,|\, z_i \right] \right)$$

Definition of Our Oracle Statistic

$$\mathrm{CI}_{n,p} := \sum_{j=1}^{J} \left| \Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)}) \right|^p$$

Asymptotic Distribution

Proposition:

- Under H_0, $\sqrt{n}\mathrm{CI}_{n,p} \to \|X\|_p^p$ where $X \sim \mathcal{N}(0_J, \Sigma)$, $\Sigma := \mathbb{E}(\mathbf{u}_1\mathbf{u}_1^T)$, $\mathbf{u}_1 := (u_1(1), \ldots, u_1(J))^T$, $u_i(j) := \left(k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - \mathbb{E}_{\ddot{X}}\left[k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, \ddot{X}) \,|\, Z = z_i\right]\right) \times \left(k_{\mathcal{Y}}(t_j^{(2)}, y_i) - \mathbb{E}_Y\left[k_{\mathcal{Y}}(t_j^{(2)}, Y) \,|\, Z = z_i\right]\right)$, and the convergence is in law.
- Under H_1, $\lim_{n\to\infty} P(n^{p/2}\mathrm{CI}_{n,p} \geq q) = 1$ for any $q \in \mathbb{R}$.

Consistency of the test

Problems:

- The oracle statistic involves unknown conditional means: $\mathbb{E}_{\ddot{X}}\left[k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, \ddot{X}) \,|\, Z = \cdot\,\right]$ and $\mathbb{E}_Y\left[k_{\mathcal{Y}}(t_j^{(2)}, Y) \,|\, Z = \cdot\,\right]$
- The asymptotic distributions involved an unknown covariance matrix Σ

Approximation of the Oracle Statistic

We estimate these conditional means using ***Regularized Least-squares Estimators***:

$$h_{j,r}^{(2)} := \min_{h \in H_{\mathcal{Z}}^{2,j}} \frac{1}{r} \sum_{i=1}^{r} \left(h(z_i) - k_{\mathcal{Y}}(t_j^{(2)}, y_i) \right)^2 + \lambda_{j,r}^{(2)} \|h\|_{H_{\mathcal{Z}}^{2,j}}^2$$

$$h_{j,r}^{(1)} := \min_{h \in H_{\mathcal{Z}}^{1,j}} \frac{1}{r} \sum_{i=1}^{r} \left(h(z_i) - k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, (x_i, z_i)) \right)^2 + \lambda_{j,r}^{(1)} \|h\|_{H_{\mathcal{Z}}^{1,j}}^2$$

Approximate Estimate of the Witness Function

$$\widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)}) := \frac{1}{n} \sum_{i=1}^{n} \left(k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - h_{j,r}^{(1)}(z_i) \right) \times \left(k_{\mathcal{Y}}(t_j^{(2)}, y_i) - h_{j,r}^{(2)}(z_i) \right)$$

Definition of our Approximate Statistic

$$\widetilde{\mathrm{CI}}_{n,r,p} := \sum_{j=1}^{J} \left| \widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)}) \right|^p$$

Proposition:

Under some *mild assumptions* on the family of distributions considered and for well chosen r_n, we obtain:

- Under H_0, $\sqrt{n}\widetilde{\mathrm{CI}}_{n,r_n,p} \to \|X\|_p^p$ where $X \sim \mathcal{N}(0_J, \Sigma)$
- Under H_1, $\lim_{n\to\infty} P(n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p} \geq q) = 1$ for any $q \in \mathbb{R}$.

It still involves the unknown covariance matrix

<u>Normalized Version of Our Test Statistic</u>

Denote $\widetilde{u}_{i,r}(j) := (k_{\ddot{\mathcal{X}}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - h_{j,r}^{(1)}(z_i))(k_{\mathcal{Y}}(t_j^{(2)}, y_i) - h_{j,r}^{(2)}(z_i))$, $\widetilde{\mathbf{S}}_{n,r} := \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{u}}_{i,r}$ and $\mathbf{\Sigma}_{n,r} := \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{u}}_{i,r}\widetilde{\mathbf{u}}_{i,r}^T$

$$\widetilde{\mathrm{NCI}}_{n,r,p} := \|(\Sigma_{n,r} + \delta_n \mathrm{Id}_J)^{-1/2}\widetilde{\mathbf{S}}_{n,r}\|_p^p.$$

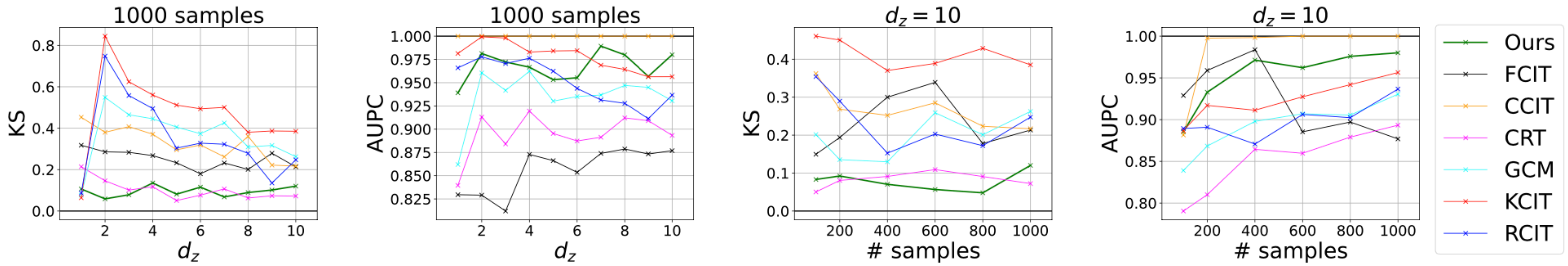
Proposition:

Under some *mild assumptions* on the family of distributions considered and for well chosen r_n, we obtain:

- Under H_0, $\sqrt{n}\widetilde{\mathrm{NCI}}_{n,r_n,p} \to \|X\|_p^p$ where $X \sim \mathcal{N}(0_J, \mathrm{Id}_J)$
- Under H_1, $\lim_{n\to\infty} P(n^{p/2}\widetilde{\mathrm{NCI}}_{n,r_n,p} \geq q) = 1$ for any $q \in \mathbb{R}$.

Now we have a simple null asymptotic distribution

Experimental results



Results: We show that our test is the only one able to demonstrate that our method consistently controls the type-I error and obtains a power similar to the best SoTA tests.

Thank you

Other results:

We show experimentally our theoretical findings where our approximate statistic is able to recover the asymptotic distribution.

We show the effect of the parameter r which allows in practice to deal with the tradeoff between the computational time and the control of the type-I error.

We also explore the effects of p and J and show that our method is robust to the choice of p, and the performances of the test do not necessarily increase as J increases.