Nyström M-Hilbert-Schmidt Independence Criterion

Florian Kalinke[1] and Zoltán Szabó[2]

[1]Karlsruhe Institute of Technology (KIT) [2]London School of Economics (LSE)

Overview

In a Nutshell

- Motivation:
  - HSIC (Hilbert-Schmidt independence criterion, a.k.a. distance covariance): popular dependency measure, various applications:
    - Independence testing [Gretton et al., 2008, Pfister et al., 2018, Albert et al., 2022], feature selection [Camps-Valls et al., 2010, Song et al., 2012, Wang et al., 2022] with applications in biomarker detection [Climente-González et al., 2019] and wind power prediction [Bouche et al., 2023], clustering [Song et al., 2007, Climente-González et al., 2019], and causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021].
  - Bottleneck: quadratic runtime.
  - Existing speedup: $M = 2$ components (= random variables), no guarantees.
- Contributions ($M \geq 2$):
  - Improved runtime: $\mathcal{O}\left(n^2\right)$ to $\mathcal{O}\left(n^{3/2}\right)$,
  - convergence rate: $\mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right)$; optimal in a minimax sense.
- Experiments: causal discovery, dependency testing of media annotations.

Dependency Intuition

- Given samples from a distribution $\mathbb{P}_{X_1 X_2}$,
- are X_1 and X_2 independent, that is, $\mathbb{P}_{X_1 X_2} \stackrel{?}{=} \mathbb{P}_{X_1} \otimes \mathbb{P}_{X_2}$.
- Think of correlation (e.g., height and weight, $[-1, 1]$) but for all kinds of dependence, also non-linear.

X_1	X_2
x_1^1: Ich hoffe, daß dort in Ihrem Sinne entschieden wird.	x_2^1: It will, I hope, be examined in a positive light.
x_1^2: Frau Präsidentin, können Sie mir sagen, warum sich dieses Parlament nicht an die Arbeitsschutzregelungen hält, die es selbst verabschiedet hat?	x_2^2: Madam President, can you tell me why this Parliament does not adhere to the health and safety legislation that it actually passes?
x_1^3: Weshalb wurde die Luftqualität in diesem Gebäude seit unserer Wahl nicht ein einziges Mal überprüft?	x_2^3: Why has no air quality test been done on this particular building since we were elected?
x_1^4: Weshalb ist der Arbeitsschutzausschuß seit 1998 nicht ein einziges Mal zusammengetreten?	x_2^4: Why has there been no Health and Safety Committee meeting since 1998?
x_1^5: Warum hat weder im Brüsseler noch im Straßburger Parlamentsgebäude eine Brandschutzübung stattgefunden?	x_2^5: Why has there been no fire drill, either in the Brussels Parliament buildings or the Strasbourg Parliament buildings?
x_1^6: Warum finden keine Brandschutzbelehrungen statt?	x_2^6: Why are there no fire instructions?

Motivation Kernel Methods

- Kernel methods are applicable to a large number of domains.
  - E.g., strings [Watkins, 1999, Lodhi et al., 2002] or more generally for sequences [Király and Oberhauser, 2019], sets [Haussler, 1999, Gärtner et al., 2002], rankings [Jiao and Vert, 2016], fuzzy domains [Guevara et al., 2017], and graphs [Borgwardt et al., 2020].
- Well-understood structure of the Hilbert space of functions (reproducing kernel Hilbert space; RKHS) associated to a kernel [Aronszajn, 1950, Schölkopf and Smola, 2002, Steinwart and Christmann, 2008].
  - Permits statistical analysis.
  - Well-suited for computations.
- Kernels allow representing probability measures as elements of RKHSs [Berlinet and Thomas-Agnan, 2004].
  - Mapping is injective if the RKHS is "rich enough" [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
  - Typically permits closed-form estimators.

Reproducing Kernel Hilbert Space (RKHS)

Definition (RKHS)

A Hilbert space $\mathcal{H}_k$ of functions $\mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space if there exists a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $x \in \mathcal{X}$ and $f \in \mathcal{H}_k$ it holds that

- $k(\cdot, x) \in \mathcal{H}_k$ ("generators"),
- $\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x)$ (reproducing property).

- For all $x, x' \in \mathcal{X}$, $k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k}$.
- We call $\phi_k(x) = k(\cdot, x)$ the (canonical) feature map and $\mathcal{H}_k$ the feature space; $\phi_k : \mathcal{X} \to \mathcal{H}_k$. Explicit form:

$$\mathcal{H}_k = \overline{\operatorname{span}\{\phi_k(x) \mid x \in \mathcal{X}\}}.$$

- Due to the reproducing property, one can express everything in terms of $k(x, y)$; actually computable.

RKHS and Kernel Examples

- RKHSs:
  - Euclidean space $\mathbb{R}^d$, $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{R}^d} = \mathbf{u}^\mathsf{T} \mathbf{v}$.
  - Square summable sequences:

$$\ell_2 = \left\{ u \in \mathbb{R}^{\mathbb{N}} \mid \sum_{j \in \mathbb{N}} u_j^2 < \infty \right\}.$$

  - Many other common spaces are RKHSs: Polynomials, splines, Sobolev spaces on $[0, 1]$.
- Some kernels on $\mathbb{R}^d$:
  - Linear:

$$k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^d}.$$

  - Polynomial:

$$k(\mathbf{x}, \mathbf{y}) = \left(\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^d} + c_0 \right)^{c_1}, \quad c_0 \geq 0, c_1 \in \mathbb{N}.$$

  - RBF / Gaussian:

$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^d}^2}, \quad \gamma > 0.$$

Kernel Mean Embedding Intuition

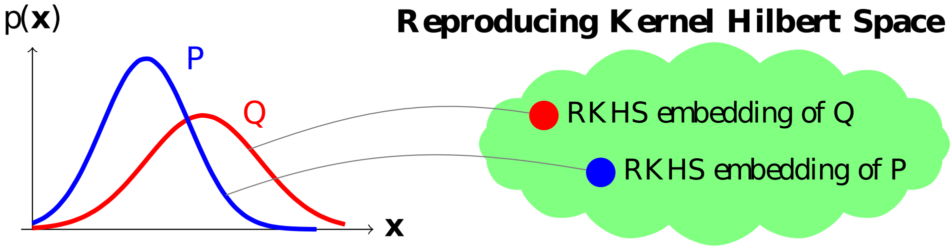


Figure: Embedding of marginal distributions: each distribution is mapped into a reproducing kernel Hilbert space via an expectation operation. Source: [Muandet et al., 2017].

Kernel mean embedding

- Extend the feature map ϕ_k to distributions, e.g., $\mathbb{P}$, and define

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(x,\cdot)}_{=\phi_k(x)} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

  - Integral is meant in Bochner's sense (properties similar to Lebesgue integral).
  - Boundedness of k, that is, $\sup_{x\in\mathcal{X}} k(x,x) < \infty$, is sufficient for $\mu_k(\mathbb{P})$ to exist.
- Mean reproducing property ($f \in \mathcal{H}_k$):

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{X\sim\mathbb{P}}\left[\langle f, \phi_k(X)\rangle_{\mathcal{H}_k}\right] = \langle f, \mathbb{E}_{X\sim\mathbb{P}}[\phi_k(X)]\rangle_{\mathcal{H}_k} = \langle f, \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k}.$$

  - For a Dirac measure centered at a particular $x_0 \in \mathcal{X}$ one recovers the reproducing property.
- Injectivity of the embedding: do we lose information?
  - Polynomial kernels lose information.
  - Mean embedding can be "rich enough" (= "characteristic"); like characteristic functions or MGFs.
    - E.g., Gaussian kernel.

Cross-covariance matrix $\to$ Cross-covariance operator ($M = 2$)

- Cross-covariance matrix:

$$C_{XY} = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[(X - \mathbb{E}_{X\sim\mathbb{P}_X}X)(Y - \mathbb{E}_{Y\sim\mathbb{P}_Y}Y)^{\mathsf{T}}\right],$$
$$\|C_{XY}\|_{\mathrm{F}} \stackrel{?}{=} 0 \text{ (“linearly independent”)}.$$

- Cross-covariance operator: consider feature maps of X and Y:

$$\begin{aligned} C_{XY} &= \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[(\phi_k(X) - \mathbb{E}_{X\sim\mathbb{P}_X}\phi_k(X)) \otimes (\phi_\ell(Y) - \mathbb{E}_{Y\sim\mathbb{P}_Y}\phi_\ell(Y))\right],\\ &= \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[(\phi_k(X) - \mu_k(\mathbb{P}_X)) \otimes (\phi_\ell(Y) - \mu_\ell(\mathbb{P}_Y))\right],\\ \|C_{XY}\|_{\mathrm{HS}} &=: \mathrm{HSIC}(\mathbb{P}_{XY}). \end{aligned}$$

Intuition HSIC $M \geq 2$

- Kullback-Leibler divergence (p is p.d.f. of $\mathbb{P}$, q is p.d.f. of $\mathbb{Q}$):

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x.$$

- Mutual information:

$$\mathrm{MI}(\mathbb{P}) = \mathrm{KL}\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

- Idea: quantify the "distance" of the joint distribution to the product of its marginal distributions.

Hilbert-Schmidt Independence Criterion

- Maximum mean discrepancy (MMD):

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \left\|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\right\|_{\mathcal{H}_k}.$$

- Previously $M = 2$; we need tuples. Let $x = (x_m)_{m=1}^M, y = (y_m)_{m=1}^M \in \times_{m=1}^M \mathcal{X}_m =: \mathcal{X}$, k_m-s be kernels on $\mathcal{X}_m$-s with feature maps ϕ_{k_m}-s and associated RKHSs $\mathcal{H}_{k_m}$. Define the product kernel

$$k(x,y) = \prod_{m=1}^M k_m(x_m, y_m).$$

- Hilbert-Schmidt independence criterion (HSIC):

$$\begin{aligned}\mathrm{HSIC}_k(\mathbb{P}) &= \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right) \\ &= \Bigg\|\underbrace{\mu_{\otimes_{m=1}^M k_m}(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)}_{\text{cross-covariance operator}}\Bigg\|_{\otimes_{m=1}^M \mathcal{H}_{k_m}}.\end{aligned}$$

- Alternative to mutual information.

HSIC Estimators

- Let $\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \in \mathcal{X}^n$ be an i.i.d. sample of M-tuples from $\mathbb{P}$ of size n.
- The closed-form quadratic time estimator

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) := \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n - \frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right)$$

  with Gram matrices $\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ can be computed in $O(n^2 M)$.
- Our proposed estimator is

$$\mathrm{HSIC}_{k,\mathrm{N}}^2\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}\right)\boldsymbol{\alpha}_k + \prod_{m\in[M]}\boldsymbol{\alpha}_{k_m}^\mathsf{T}\mathbf{K}_{k_m,n'n'}\boldsymbol{\alpha}_{k_m} - 2\boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}\boldsymbol{\alpha}_{k_m}\right),$$

  with Gram matrices $\mathbf{K}_{k_m} = \left[k_m\left(\tilde{x}_m^i, \tilde{x}_m^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}$, $\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{k_m}$-s $\in \mathbb{R}^{n'}$.
  - How to compute the estimator?

Classical Nyström Approach

- Idea: Reduce sample size.
- HSIC consists of different means and feature maps, we abstract away from the specifics by using $\mathbb{Q}, \ell$.
- Nyström points: $\tilde{\mathbb{Q}}_{n'} = \left\{\tilde{y}^1, \ldots, \tilde{y}^{n'}\right\}$ is a subsample of $\hat{\mathbb{Q}}_n = \left\{y^1, \ldots, y^n\right\} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.
- Typically:

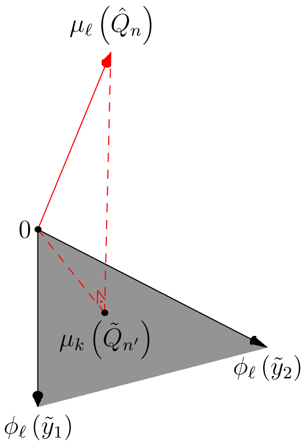
$$\mu_\ell(\mathbb{Q}) = \int_{\mathcal{Y}} \phi_\ell(y) \mathrm{d}\mathbb{Q}(y) \approx \frac{1}{n} \sum_{i \in [n]} \phi_\ell(y^i) = \mu_\ell(\hat{\mathbb{Q}}_n).$$

- Nyström approach:

$$\mu_\ell(\hat{\mathbb{Q}}_n) = \frac{1}{n} \sum_{i=1}^{n} \phi_\ell(y^i) \approx \sum_{i \in [n']} \alpha_i \phi_\ell(\tilde{y}^i) =: \mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) \in \mathcal{H}_\ell^{Nys},$$

where $\mathcal{H}_\ell^{Nys} := \operatorname{span}\left(\phi_\ell\left(\tilde{y}^i\right) \,:\, i \in [n']\right) \subset \mathcal{H}_\ell$.

Geometric Interpretation



- Compare to linear regression.
- Question: can we actually compute the projection?

Optimal Weights for Nyström Approximation

- The coefficients $\boldsymbol{\alpha}_\ell = (\alpha_\ell^1, \ldots, \alpha_\ell^{n'}) \in \mathbb{R}^{n'}$ are obtained by the minimum norm solution of

$$\min_{\boldsymbol{\alpha}_\ell \in \mathbb{R}^{n'}} \Bigg\| \underbrace{\mu_\ell\left(\hat{\mathbb{Q}}_n\right)}_{=\frac{1}{n}\sum_{i=1}^n \phi_\ell(y^i)} - \sum_{i\in[n']} \alpha_i \phi_\ell\left(\tilde{y}^i\right) \Bigg\|_{\mathcal{H}_\ell}^2 .$$

- Computable by (pseudo-)matrix inversion:

Lemma (Nyström mean embedding, [Laub, 2004, Chatalic et al., 2022])

For a kernel ℓ with corresponding feature map ϕ_ℓ, an i.i.d. sample $\hat{\mathbb{Q}}_n$ of distribution $\mathbb{Q}$, and a subsample $\tilde{\mathbb{Q}}_{n'}$ of $\hat{\mathbb{Q}}_n$, the Nyström estimate of $\mu_\ell(\mathbb{Q})$ is given by

$$\mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) = \sum_{i\in[n']} \alpha_\ell^i \phi_\ell\left(\tilde{y}^i\right), \qquad \boldsymbol{\alpha}_\ell = \frac{1}{n}\left(\mathbf{K}_{\ell,n'n'}\right)^{-}\mathbf{K}_{\ell,n'n}\mathbf{1}_n,$$

with Gram matrix $\mathbf{K}_{\ell,n'n'} = \left[\ell(\tilde{x}^i, \tilde{x}^j)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}$, *and* $\mathbf{K}_{\ell,n'n} = \left[\ell(\tilde{x}^i, x^j)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}$.

Contribution: Accelerating HSIC

- Recall:

$$\mathrm{HSIC}_k(\mathbb{P}) = \left\| \mu_{\otimes_{m=1}^M k_m}(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m) \right\|_{\otimes_{m=1}^M \mathcal{H}_{k_m}}.$$

  - $\rightarrow$ There are $M+1$ means in this expression.
- Proposed estimator: Compute each mean separately and combine, giving
  - $M+1$ weights:

$$\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) = \sum_{i\in[n']} \alpha^i_{k_m} \phi_{k_m}\left(\tilde{x}^i_m\right), \qquad \boldsymbol{\alpha}_{k_m} = \frac{1}{n}\left(\mathbf{K}_{k_m,n'n'}\right)^{-}\mathbf{K}_{k_m,n'n}\mathbf{1}_n,$$

$$\mu_k\left(\tilde{\mathbb{P}}_{n'}\right) = \sum_{i\in[n']} \alpha^i_k \otimes_{m=1}^M \phi_{k_m}\left(\tilde{x}^i_m\right), \qquad \boldsymbol{\alpha}_k = \frac{1}{n}\left(\mathbf{K}_{k,n'n'}\right)^{-}\left(\mathbf{K}_{k,n'n}\right)\mathbf{1}_n.$$

  - Runtime is $\mathcal{O}\left(Mn'^3 + Mn'n\right)$, saving if $n' = o\left(n^{2/3}\right)$.
    - Recall HSIC: $\mathcal{O}\left(Mn^2\right)$.

Contribution: Consistency

- For bounded kernels $(k_m)_{m=1}^M$, it holds that

$$\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,\mathrm{N}}\left(\hat{\mathbb{P}}_n\right)\right| = \mathcal{O}_P\left(n^{-1/2}\right),$$

  assuming that the effective dimension[3] either decays
  - polynomially ($< c\lambda^{-\gamma}$, $c > 0, \gamma \in (0,1]$) and $n' = \tilde{\mathcal{O}}\left(n^{1/(2-\gamma)}\right)$, or
  - exponentially ($< \log(1 + c/\gamma)/\beta$, $c, \beta > 0$) and $n' = \tilde{\mathcal{O}}\left(\sqrt{n}\right)$.
- Matches the bound that we obtain on the quadratic time estimator.

[3] $\mathcal{N}_X(\lambda) = \mathrm{trace}\left[\mu_{k\otimes k}(\mathbb{P})\left(\mu_{k\otimes k}(\mathbb{P}) + \lambda I\right)^{-1}\right]$.

Proof Sketch

- Known [Chatalic et al., 2022]: $\left\|\mu_k(\mathbb{P}) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\| = \mathcal{O}_P\left(n^{-1/2}\right)$.
- HSIC is expressed in terms of tensor products.
- Key is the following lemma:

Lemma (Error propagation on tensor products)

Let $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ bounded kernels ($\exists a_{k_m} \in (0,\infty)$ such that $\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \leq a_{k_m}$, $m \in [M]$), $k = \otimes_{m=1}^M k_m$, $\mathcal{H}_k$ the RKHS associated to k, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, $\mathbb{P}_m$ the m-th marginal of $\mathbb{P}$ ($m \in [M]$), $n' \leq n$, and $\tilde{\mathbb{P}}_{m,n'}$ the Nyström sample of the m-th marginal. Then

$$\left\|\otimes_{m=1}^M \mu_{k_m}\left(\mathbb{P}_m\right) - \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k} \leq \prod_{m\in[M]} \left(a_{k_m} + d_{k_m}\right) - \prod_{m\in[M]} a_{k_m},$$

where $d_{k_m} = \left\|\mu_{k_m}\left(\mathbb{P}_m\right) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}$.

Minimax Risk Idea

- We want to find an upper and a lower bound, that is,

$$L_n \leq R_n \leq U_n.$$

- $\rightarrow$ If both are close, we have succeeded.
- In our case (simplified): $R_n = \left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,\mathrm{N}}\left(\hat{\mathbb{P}}_n\right)\right|$, $U_n = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

Example (Minimax rate of convergence)

If $L_n = cn^{-\alpha}$ and $U_n = Cn^{-\alpha}$ for some positive constants c, C, and α, then the minimax rate of convergence is $n^{-\alpha}$.

Lower Bound (Unpublished)

Theorem (Lower bound for HSIC estimation)

Let $\mathcal{P}$ be a class of Borel probability measures over $\mathbb{R}^d$ containing the d-dimensional Gaussian distributions. Let $d = \sum_{m\in[M]} d_m$, $k_m(\mathbf{x}_m, \mathbf{x}'_m) = e^{-\frac{\gamma}{2}\|\mathbf{x}_m - \mathbf{x}'_m\|^2_{\mathbb{R}^{d_m}}}$ $(m \in [M])$ be Gaussian kernels on $\mathbb{R}^{d_m}$ with common bandwidth parameter $\gamma > 0$, $k = \otimes_{m=1}^{M} k_m$, and $\hat{F}_n$ denote any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ with n i.i.d. samples from $\mathbb{P} \in \mathcal{P}$. Then it holds that

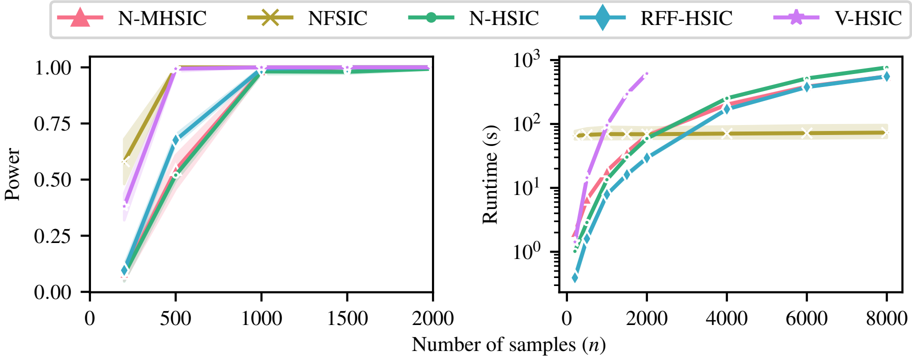
$$\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k\left(\mathbb{P}\right) - \hat{F}_n \right| \geq \frac{a}{\sqrt{n}} \right\} \geq \frac{1 - \sqrt{\frac{5}{8}}}{2},$$

for a constant $a = \frac{\gamma}{2(2\gamma+1)^{\frac{d}{4}+1}} > 0$ (depending on γ and d only).

- $\rightarrow$ with positive probability, the best estimator can not converge faster than $n^{-1/2}$: There exists a distribution $\mathbb{P} \in \mathcal{P}$ which is sufficiently difficult to estimate.
- Proof idea: construct adversarial pair of distributions that are close w.r.t. KL but sufficiently different when considering HSIC (framework: minimax theory); we consider Gaussians.

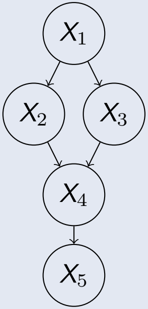
Experiments: Dependencies of Media Annotations ($M = 2$)

- Test for dependence of X and Y ($H_0 : \mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$, H_1 actually holds):
  - X: 90 acoustic features (timbre average (12), timbre covariance (78)).
  - Y: year of release.
  - $M = 2$ allows comparing to existing algorithms.



Experiments: Causality [Pearl, 2009, Schölkopf, 2022]

Example (A simple graph with its SCM)



- $\mathcal{G}$ induces the causal factorization

$$\mathbb{P}(X_1, \ldots, X_5) = \mathbb{P}(X_1)\,\mathbb{P}(X_2 \mid X_1)\,\mathbb{P}(X_3 \mid X_1)\,\mathbb{P}(X_4 \mid X_2, X_3)\,\mathbb{P}(X_5 \mid X_4),$$

by repeated application of

$$X_i = f_i(\mathrm{PA}_i, U_i),$$

and by using the **joint independence** of the U_i-s ($i = 1, \ldots, 5$).

Experiments: Additive and non-linear function class

- Consider an additive noise model

$$X_i = \sum_{k \in \mathrm{PA}_i} f_{i,k}(X_k) + U_i, \quad i = 1, \dots, M,$$

with U_i independent Gaussian, and $f_{i,k}$ non-linear.

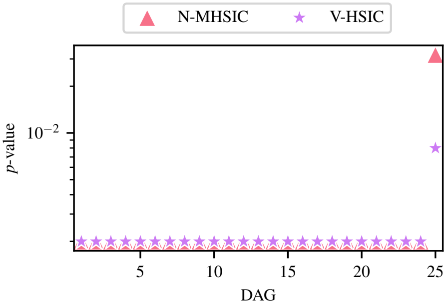
Algorithm (DAG verification method; [Pfister et al., 2018])

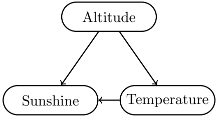
Given observations $\mathbf{x}_1, \dots, \mathbf{x}_n$, and a candidate DAG $\mathcal{G}$

- *Use generalized additive model regression to regress each node X_i on all its parents PA_i and denote the resulting vector of residuals by ϵ_i.*
- *Perform a M-variable **joint independence test** to test whether $(\epsilon_1, \dots, \epsilon_M)$ is jointly independent.*
- *If $(\epsilon_1, \dots, \epsilon_M)$ is jointly independent, the DAG $\mathcal{G}$ is not rejected.*

Experiments: Weather Causal Discovery ($M = 3$)

- 349 measurements of weather data in Germany [Mooij et al., 2016, Pfister et al., 2018].
- We want to infer the most plausible DAG with three nodes out of 25 possible DAGs ($3^3 - 2 = 25$, two graphs contain a cycle).





Summary

- Acceleration of dependency estimation with HSIC.
- Upper bound assuming appropriate effective dimension decay:

$$\left\|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,N}\left(\hat{\mathbb{P}}_n\right)\right\| = \mathcal{O}_P\left(n^{-1/2}\right).$$

- Matching lower bound.
  - Proposed algorithm is optimal in a minimax-sense (with the considered priors).
- Experiments on real-world data.
- Corresponding article: [Kalinke and Szabó, 2023], GitHub: https://github.com/FlopsKa/nystroem-mhsic/.

References I

Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879, 2022.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Riec. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020.

Dimitri Bouche, Rémi Flamary, Florence d'Alché Buc, Riwal Plougonven, Marianne Clausel, Jordi Badosa, and Philippe Drobinski. Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection. *Renewable Energy*, 2023.

Gustavo Camps-Valls, Joris M. Mooij, and Bernhard Schölkopf. Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591, 2010.

References II

Shubhadeep Chakraborty and Xianyang Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528):1638–1650, 2019.

Antoine Chatalic, Nicolas Schreuder, Alessandro Rudi, and Lorenzo Rosasco. Nyström kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, pages 3006–3024, 2022.

Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 2019.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496, 2008.

Thomas Gärtner, Peter Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592, 2008.

References III

Jorge Guevara, Roberto Hirata, and Stéphane Canu. Cross product kernels for fuzzy set similarity. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017.

David Haussler. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, 1999. (`http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf`).

Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. In *International Conference on Machine Learning (ICML)*, pages 2982–2990, 2016.

Florian Kalinke and Zoltán Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023.

Franz J. Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20:1–45, 2019.

Alan J Laub. *Matrix analysis for scientists and engineers*. SIAM, 2004.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

References IV

Joris Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10 (1-2):1–141, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 5–31, 2018.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022.

Bernhard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

References V

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Le Song, Alexander J. Smola, Arthur Gretton, and Karsten M. Borgwardt. A dependence maximization view of clustering. In *International Conference on Machine Learning (ICML)*, pages 815—822, 2007.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, pages 1517–1561, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Andi Wang, Juan Du, Xi Zhang, and Jianjun Shi. Ranking features to promote diversity: An approach based on sparse distance correlation. *Technometrics*, 64(3):384–395, 2022.

Chris Watkins. Dynamic alignment kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50, 1999.