Nyström M-Hilbert-Schmidt Independence Criterion*

Florian Kalinke[1] and Zoltán Szabó[2]

[1]Institute for Program and Data Structures, Karlsruhe Institute of Technology [2]Department of Statistics, London School of Economics

Quick Summary

- Faster estimation of Hilbert-Schmidt independence criterion (HSIC; $M = 2$: [2], $M \geq 2$: [5, 6, 4], validness: [7]).
- Guarantee: same convergence rate as the quadratic time estimator.
- Existing accelerations: $M = 2$, works efficiently in practice but without theoretical guarantees [8].
- Experiments on synthetic examples, dependency testing of media annotations, and causal discovery.

HSIC

- Given $X = (X_m)_{m=1}^M \sim \mathbb{P}$ on $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $\mathcal{X}_m$ is equipped with kernel k_m and feature map $\phi_{k_m} : \mathcal{X}_m \to \mathcal{H}_{k_m}$, HSIC takes the form

$$\mathrm{HSIC}_k(\mathbb{P}) = \left\| \mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right) \right\|_{\mathcal{H}_k}, \qquad k := \otimes_{m=1}^M k_m$$

with $\otimes_{m=1}^M \mathbb{P}_m$ the product of the marginal distributions $\mathbb{P}_m$, $m \in [M] := \{1, \ldots, M\}$, and $\mu_k(\mathbb{P}) = \mathbb{E}_{X \sim \mathbb{P}}[\phi_k(X)]$.

- Given an i.i.d. sample of M-tuples of size n

$$\hat{\mathbb{P}}_n := \left\{ \left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right) \right\} \subset \mathcal{X}^n,$$

from $\mathbb{P}$, the V-statistic based estimator takes the form

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

with Gram matrices

$$\mathbf{K}_{k_m,n,n} = \left[k_m\left(x_m^i, x_m^j\right) \right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \tag{1}$$

and can be computed in $\mathcal{O}\left(n^2\right)$ time.

Proposed Nyström-based estimator

- Let $\tilde{\mathbb{P}}_{n'} = \left\{ \left(\tilde{x}_1^1, \ldots, \tilde{x}_M^1\right), \ldots, \left(\tilde{x}_1^{n'}, \ldots, \tilde{x}_M^{n'}\right) \right\}$ be a subsample of $\hat{\mathbb{P}}_n$.
- Our proposed Nyström-based estimator is given by

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

$$\boldsymbol{\alpha}_{k_m} = \frac{1}{n} \left(\mathbf{K}_{k_m,n',n'} \right)^- \mathbf{K}_{k_m,n',n} \mathbf{1}_n,$$

$$\boldsymbol{\alpha}_k = \frac{1}{n} \left(\circ_{m\in[M]} \mathbf{K}_{k_m,n',n'} \right)^- \left(\circ_{m\in[M]} \mathbf{K}_{k_m,n',n} \right) \mathbf{1}_n,$$

where $\circ$ is the Hadamard product, $\mathbf{K}_{k_m,n',n'}$ is defined in (1), $\mathbf{K}_{k_m,n',n} = \left[k_m\left(\tilde{x}_m^i, x_m^j\right) \right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}$, and $(\cdot)^-$ denotes pseudo-inverse.

- Runtime complexity of $\mathcal{O}\left(Mn'^3 + Mn'n\right)$, saving if $n' = o\left(n^{2/3}\right)$.
- Code: `https://github.com/FlopsKa/nystroem-mhsic/`.

Main Result

- For bounded kernels $(k_m)_{m=1}^M$ and the effective dimension $\mathcal{N}_X(\lambda) = \operatorname{tr}\left[\mu_{k\otimes k}(\mathbb{P}) \left(\mu_{k\otimes k}(\mathbb{P}) + \lambda I \right)^{-1} \right]$, 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 either

- decays polynomially:

$$\max_{m\in[M]} \left(\mathcal{N}_X(\lambda), \mathcal{N}_{X_m}(\lambda) \right) \leq c\lambda^{-\gamma}, \qquad n' = n^{1/(2-\gamma)} \log(n/\delta),$$

for some $c > 0$ and $\gamma \in (0, 1]$ (computational savings if $\gamma < 1/2$), or

- decays exponentially:

$$\max_{m\in[M]} \left(\mathcal{N}_X(\lambda), \mathcal{N}_{X_m}(\lambda) \right) \leq \log(1 + c/\lambda)/\beta,$$

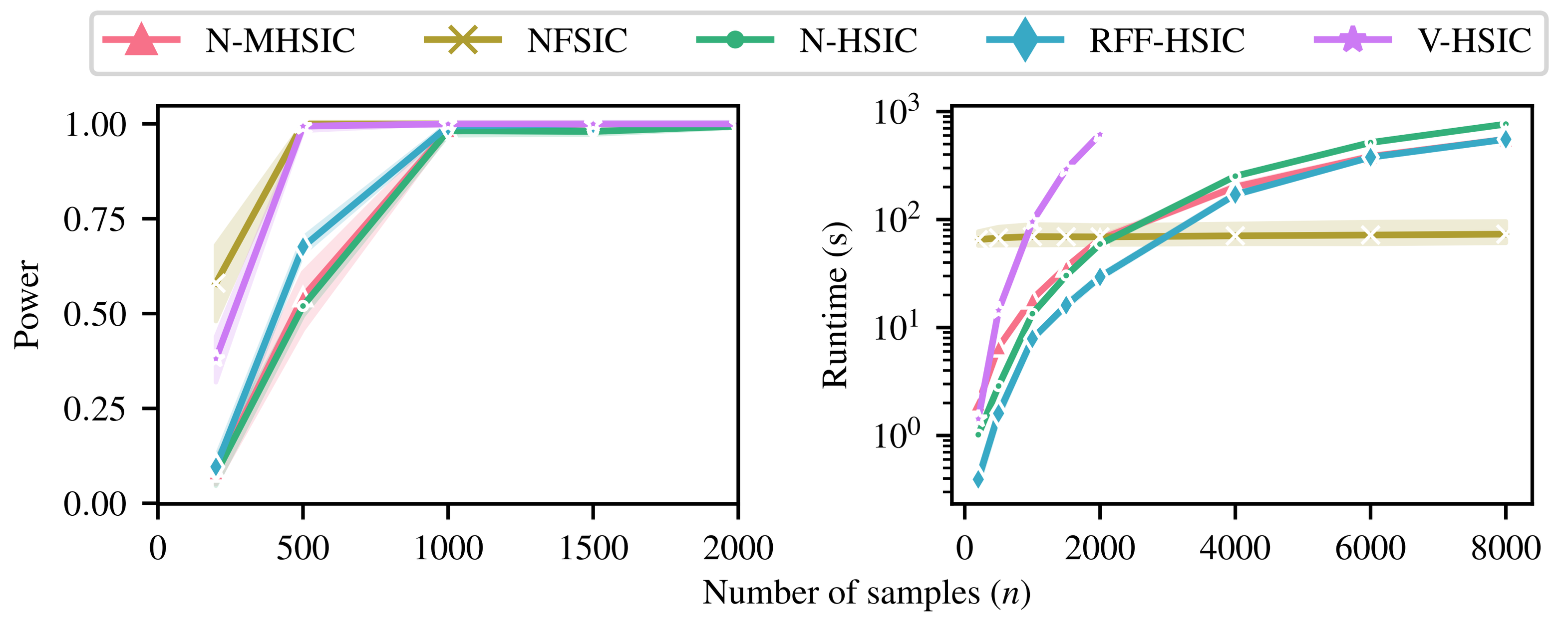
$$n' = \sqrt{n} \log\left(\sqrt{n} \max_{m\in[M]} \left(\frac{1}{\delta}, \frac{c}{6a_k^2}, \frac{c}{6a_{k_m}^2} \right) \right)$$

for some $c > 0$, $\beta > 0$, a_k, a_{k_m} bounds on the kernels k, k_m ($m \in [M]$).

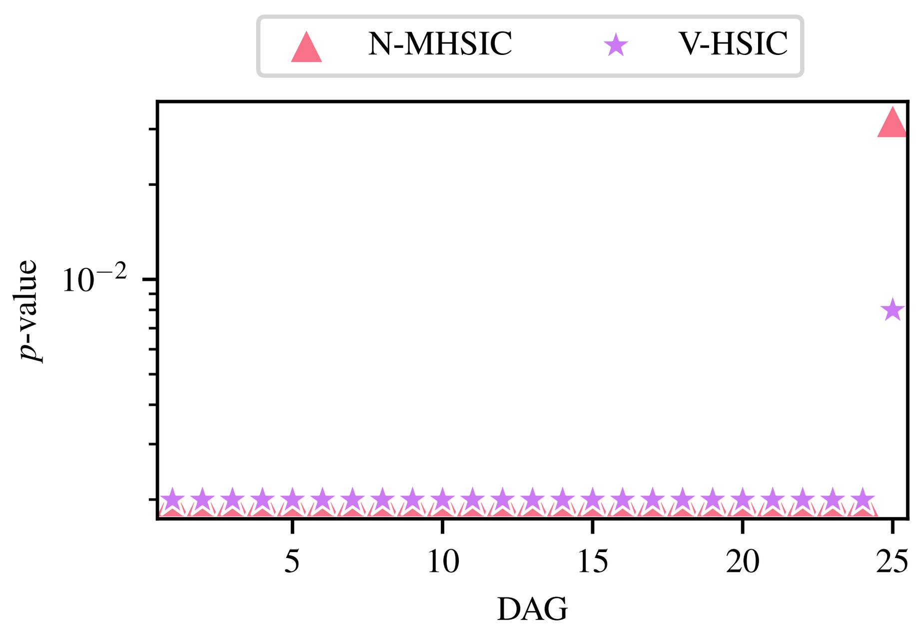
- The decay of the effective dimension can be linked to the decay of the eigenvalues of the covariance operator $\mu_{k\otimes k}(\mathbb{P})$ [1, Proposition 4, 5].

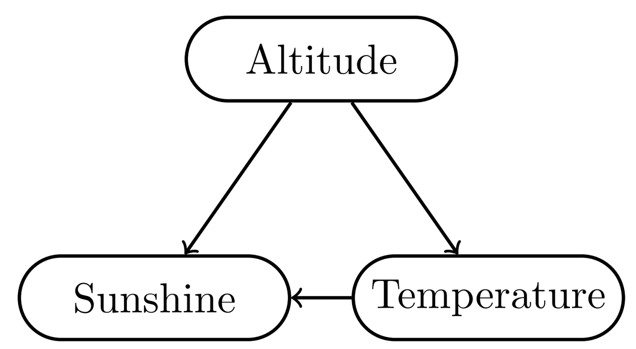
Example Applications

- Dependency estimation of media annotations ($M = 2$).



- Weather causal discovery [3] ($M = 3$).





References

[1] Andrea Della Vecchia, Jaouad Mourtada, Ernesto De Vito, and Lorenzo Rosasco. Regularized ERM on random subspaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4006–4014, 2021.

[2] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

[3] 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.

[4] 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.

[5] Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2009.

[6] Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132, 2013.

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

[8] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2018.

*UAI, Pittsburgh, PA, USA, 31st July–4th August 2023. Acknowledgements. This work was supported by the German Research Foundation (DFG) Research Training Group GRK 2153: Energy Status Data — Informatics Methods for its Collection, Analysis and Exploitation.