An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum,[1] Zoltán Szabó,[2] Arthur Gretton[1]

[1]Gatsby Unit, University College London [2]CMAP, École Polytechnique

Summary

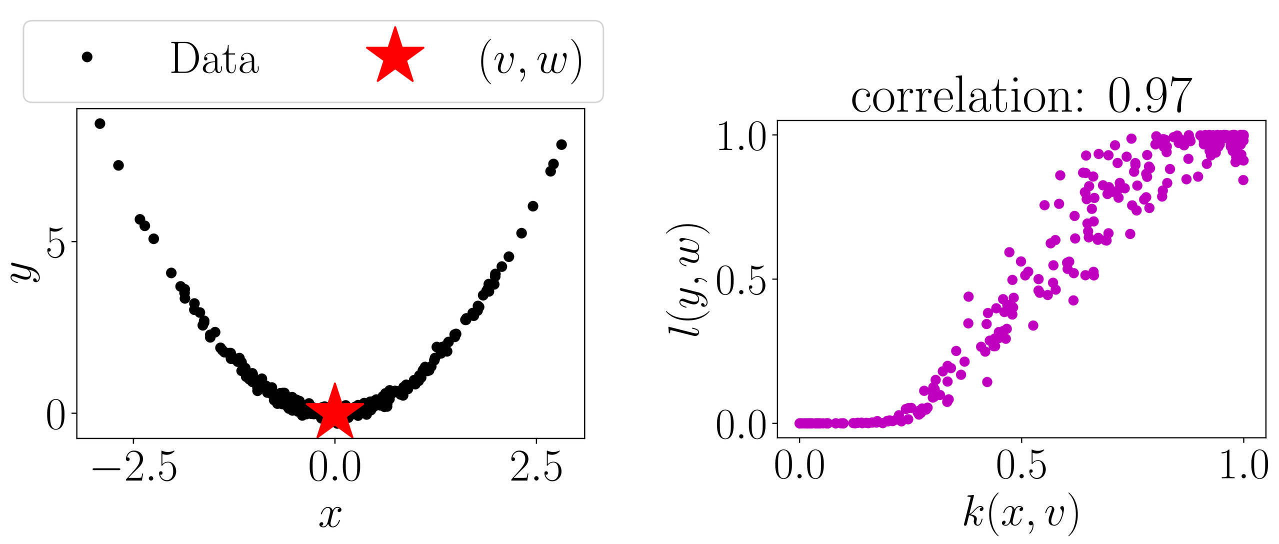
- **Observe**: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim P_{xy}$ (unknown distribution).
- **Goal**: Test $H_0 : P_{xy} = P_x P_y$ vs $H_1 : P_{xy} \neq P_x P_y$ quickly.
- **New multivariate independence test**:
  1. Nonparametric: arbitrary P_{xy}. $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$.
  2. Linear-time: $\mathcal{O}(n)$ runtime complexity.
  3. Adaptive: hyperparameters automatically tuned.

The Finite-Set Independence Criterion

$\mathrm{FSIC}(\mathbf{x}, \mathbf{y})$ = new efficient dependence measure.

1. Pick 2 positive definite kernels: k for $\mathbf{x}$, and l for $\mathbf{y}$.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick J **features** $\{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$
3. Compute $u_j := \mathrm{cov}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}[k(\mathbf{x}, \mathbf{v}_j), l(\mathbf{y}, \mathbf{w}_j)]$.

$$\mathrm{FSIC}^2(\mathbf{x}, \mathbf{y}) := \frac{1}{J}\mathbf{u}^\top \mathbf{u}, \quad \text{where } \mathbf{u} := (u_1, \ldots, u_J)^\top$$



Proposition. *Assume*

1. *Kernels k and l satisfy some smoothness conditions e.g. Gaussian kernels.*
2. *Features $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ are drawn from a distribution with a density e.g., normal distribution.*

Then, $\mathrm{FSIC}(\mathbf{x}, \mathbf{y}) = 0$ *iff* $P_{xy} = P_x P_y$*, for* **any** $J \geq 1$.

✗ But, under H_0, distribution of empirical $\widehat{\mathrm{FSIC}^2}$ is intractable. Hard to get test threshold. ☹

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}^2}(\mathbf{x}, \mathbf{y}) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with regularizer $\gamma_n \geq 0$, and $\hat{\boldsymbol{\Sigma}}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Proposition (NFSIC test is consistent)**.** *Assume $\gamma_n \to 0$, and same conditions on k and l as before. As $n \to \infty$, ...*

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$. ✓Easy to get test threshold.* ☺
2. *Under H_1, $\mathbb{P}$(reject H_0) $\to 1$. ✓Eventually reject if H_1 true.* ☺

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small J.

Test Power Lower Bound

- In practice, optimizing the features will improve performance.

Proposition. *The test power $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right)$ is at least*

$$L(\lambda_n) = 1 - 62e^{-\xi_1 \gamma_n^2 (\lambda_n - T_\alpha)^2/n} - 2e^{-\lfloor 0.5n \rfloor (\lambda_n - T_\alpha)^2/\left[\xi_2 n^2\right]} - 2e^{-\left[(\lambda_n - T_\alpha)\gamma_n(n-1)/3 - \xi_3 n - c_3 \gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2 (n-1)\right]},$$

where $\xi_1, \ldots, \xi_4, c_3 > 0$ are constants. For large n, $L(\lambda_n)$ is increasing in $\lambda_n := \mathrm{NFSIC}^2(\mathbf{x}, \mathbf{y}) = n\mathbf{u}^\top \boldsymbol{\Sigma}^{-1} \mathbf{u}$ (population NFSIC).

Proposal: Optimize features and kernel bandwidths by $\arg\max L(\lambda_n) = \arg\max \lambda_n$. Optimization is $\mathcal{O}(n)$ time.

- **Key**: Parameters chosen to maximize test power lower bound.
- Use a **separate** training set to estimate λ_n. Does not overfit.
- Splitting train/test sets keeps false rejection rate well-controlled.

We thank the Gatsby Charitable Foundation for the financial support.

Contact: wittawat@gatsby.ucl.ac.uk

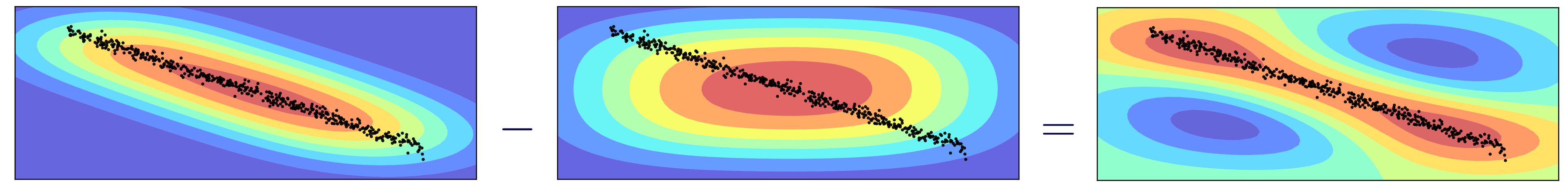
Code: github.com/wittawatj/fsic-test

Paper: http://proceedings.mlr.press/v70/jitkrittum17a.html

Witness Function View of FSIC

$$\begin{aligned} u(\mathbf{v}, \mathbf{w}) &= \mathrm{cov}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})] \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}[k(\mathbf{x}, \mathbf{v}) l(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathbf{x}\sim P_x}[k(\mathbf{x}, \mathbf{v})]\mathbb{E}_{\mathbf{y}\sim P_y}[l(\mathbf{y}, \mathbf{w})] \\ &:= \mu_{xy}(\mathbf{v}, \mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}) \end{aligned}$$

- $u(\mathbf{v}, \mathbf{w})$ is known as the **witness function**, capturing the diff. of P_{xy} and $P_x P_y$.

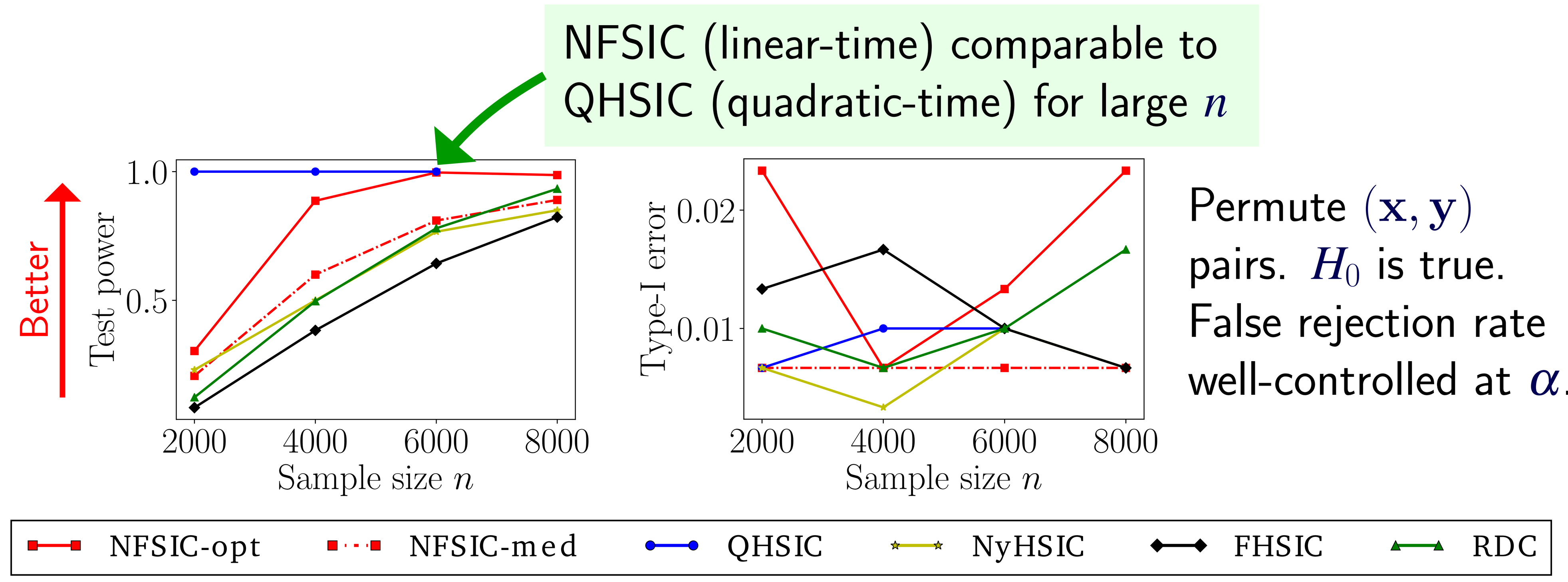


$\mu_{xy}(\mathbf{v}, \mathbf{w})$ − $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$ = Witness $u(\mathbf{v}, \mathbf{w})$

- $\mathrm{HSIC}(\mathbf{x}, \mathbf{y})$ = RKHS norm of the witness function. The norm costs $\mathcal{O}(n^2)$.
- $\mathrm{FSIC}(\mathbf{x}, \mathbf{y})$ = evaluates the witness at J locations (features). Costs only $\mathcal{O}(Jn)$.
- FSIC is good when P_{xy} and $P_x P_y$ differ locally. Pinpoint with the features.

Youtube Video ($\mathbf{x}$) vs. Text Caption ($\mathbf{y}$)

- $\mathbf{x} \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $\mathbf{y} \in \mathbb{R}^{1878}$: Bag of words. Term frequency. Significance level of the test $\alpha = 0.01$.



Permute $(\mathbf{x}, \mathbf{y})$ pairs. H_0 is true. False rejection rate well-controlled at α.

- NFSIC-opt = proposed test. Full optimization. $J = 10$.
- NFSIC-med = proposed test. Random $\{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J$. $J = 10$.
- QHSIC [Gretton et al., 2005] = quadratic-time state-of-the-art HSIC test.
- NyHSIC, FHSIC = HSIC tests with Nyström and random Fourier features. $\mathcal{O}(n)$.
- RDC [Lopez-Paz et al., 2013] = CCA with cosine basis. $\mathcal{O}(n \log n)$.