An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton

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

Summary

- Observe: \( \{ (x_i, y_i) \}_{i=1}^{n} \sim P_{xy} \) (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 \( P_{xy} \), \( x \in \mathbb{R}^d \) and \( y \in \mathbb{R}^d \).
  2. Linear-time: \( \mathcal{O}(n) \) runtime complexity.

Normalized FSIC (NFSIC)

\[
\widehat{NFSIC}^2(x, y) = \hat{\lambda}_{xy} := \mu_y^\top (\Sigma + \gamma I)^{-1} \mu_x
\]

with regularizer \( \gamma \geq 0 \), and \( \Sigma_{ij} \) = covariance of \( \hat{u}_i \) and \( \hat{u}_j \).

Proposition (NFSIC test is consistent). Assume \( \gamma_0 \to 0 \), and same conditions on \( k \) and \( l \) as before. As \( n \to \infty \),

1. Under \( H_0 \), \( \hat{\lambda}_{xy} \to \chi^2(J) \). Easy to get test threshold.
2. Under \( H_1 \), \( \mathbb{P}(\text{reject } H_0) \to 1 \). Eventually reject if \( H_1 \) true.

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

Test Power Lower Bound

- In practice, optimizing the features will improve performance.

Proposition. The test power \( P_{H_1} (\hat{\lambda}_{xy} \geq T_0) \) is at least

\[
L(\lambda_0) = 1 - \frac{P_{xy}}{2} - 2e^{-\left(\lambda_0 - T_0\right)/\gamma} - \frac{\left(\lambda_0 - T_0\right)^2}{\gamma^2}.
\]

where \( \xi \), \( \xi_0 \), \( \xi_1 \), \( \xi_2 \) are constants. For large \( n \), \( L(\hat{\lambda}_0) \) is increasing in \( \hat{\lambda}_0 := NFSIC^2(x, y) = \mu_y^\top \Sigma^{-1} \mu_x \) (population NFSIC).

Proposal: Optimize features and kernel bandwidths by \( \arg \max L(\hat{\lambda}_0) \). Optimization is \( \mathcal{O}(n) \) time.

- Key: Parameters chosen to maximize test power lower bound.
- Use a separate training set to estimate \( \hat{\lambda}_0 \). 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


Witness Function View of FSIC

\[
u(v, w) = \mu_x(v) \mu_y(w)
\]

\[
\mu_x(v) \mu_y(w)
\]

FSIC is good when \( P_{xy} \) and \( P_x P_y \) differ locally. Point out with the features.

YouTube Video (x) vs. Text Caption (y)

- \( x \in \mathbb{R}^{2000} \): Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- \( y \in \mathbb{R}^{1787} \): Bag of words. Term frequency. Significance level of the test \( \alpha = 0.01 \).
- NFSIC (linear-time) comparable to QHSIC (quadratic-time) for large \( n \).

Contact: Wittawat@gatsby.ucl.ac.uk

Code: github.com/wittawatj/fsic-test


We thank the Gatsby Charitable Foundation for the financial support.

Contact: Wittawat@gatsby.ucl.ac.uk

Code: github.com/wittawatj/fsic-test