The Finite-Set Independence Criterion

Wittawat Jitkrittum*, Zoltán Szabó† , Arthur Gretton*

*Gatsby Computational Neuroscience Unit, University College London.
†Center for Applied Mathematics (CMAP), École Polytechnique
wittawat@gatsby.ucl.ac.uk

1 Introduction

We consider the design of adaptive, nonparametric statistical tests of dependence: that is, tests of whether a joint distribution P_{xy} factorizes into the product of marginals $P_x P_y$ with the null hypothesis that $H_0 : X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are independent. While classical tests of dependence, such as Pearson's correlation and Kendall's τ, are able to detect monotonic relations between univariate variables, more modern tests can address complex interactions. Key to many recent tests is to examine covariance or correlation between data features. These interactions become significantly harder to detect, and the features are more difficult to design, when the data reside in high dimensions.

The approach we take is most closely related to HSIC [1] on a finite set of features. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are positive definite kernels, satisfying some smoothness conditions. The Finite Set Independence Criterion (FSIC) is defined as

$$\mathrm{FSIC}^2(\mathrm{X}, \mathrm{Y}) := \frac{1}{J} \sum_{i=1}^{J} \mathrm{cov}^2_{(\mathbf{x},\mathbf{y}) \sim P_{xy}} [k(\mathbf{x}, \mathbf{v}_i), l(\mathbf{y}, \mathbf{w}_i)],$$

which is an average of covariances of smooth functions defined on each of X and Y, parametrized by some *features* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J} \subset \mathcal{X} \times \mathcal{Y}$. With some mild conditions, we can show that $\mathrm{FSIC}^2(X, Y) = 0$ if and only if X and Y are independent. Also, a normalized version of the statistic (NFSIC) yields an asymptotic test threshold independent of P_{xy}.

Our test is consistent, despite a finite number (J) of features being used, via a generalization of arguments in [2]. As in recent work on two-sample testing by [3], our test is *adaptive* in the sense that we choose our features on a held-out validation set to optimize a lower bound on the test power; the result is a parsimonious and interpretable indication of how and where the null hypothesis is violated. The computational complexity of our tests is linear in the sample size.

2 Experiment

We consider a subset of the Million Song Data,[1] in which each song (X) out of 515,345 is represented by 90 features, of which 12 features are timbre average (over all segments) of the song, and 78 features are timbre covariance. The goal is to detect the dependency between each song and its year of release (Y). We use Gaussian kernels, set the significance level $\alpha := 0.01$, and repeat for 300 trials where the full sample is randomly subsampled to n points in each trial. We compare the proposed test with automatic parameter optimization (NFSIC-opt) to five multivariate nonparametric tests. The NFSIC test without optimization (NFSIC-med) acts as a baseline, allowing the effect of parameter optimization to be clearly seen. QHSIC is the quadratic-time HSIC test of [1]. Nyström HSIC (NyHSIC), and Finite-feature HSIC (FHSIC) are other variants of HSIC which run in linear time. Finally, the Randomized Dependence Coefficient (RDC) (an $\mathcal{O}(n \log n)$ test) proposed in [4] is also considered.

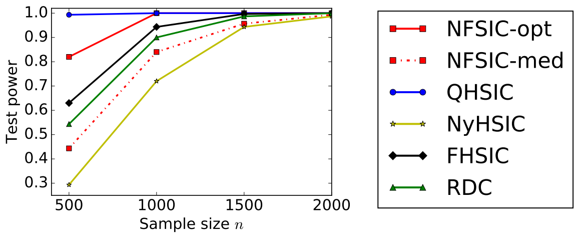

Figure 1: Probability of rejecting H_0 as n increases in the Million Song Data problem.

Figure 1 shows the test powers (rejection rate) as the sample size varies. We observe that NFSIC-opt has the highest test power among all the linear-time tests for all the sample sizes. Its test power is second to only QHSIC. The fact that there is a vast power gain from 0.4 (NFSIC-med) to 0.8 (NFSIC-opt) at $n = 500$ suggests that the optimization procedure can perform well even at a lower sample sizes.

In the full paper [5], we further demonstrate the performance of our tests on several other challenging problems, including detection of dependence between videos and captions, and artificial problems with interacting features. In these experiments, we outperform competing linear and $\mathcal{O}(n \log n)$ time tests. Also, when H_0 holds, the proposed test has correct false-positive rate.

References

[1] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *ALT*, pages 63–77. 2005.

[2] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *NIPS*, pages 1981–1989. 2015.

[3] W. Jitkrittum, Z. Szabó, K. Chwialkowski, and A. Gretton. Interpretable Distribution Features with Maximum Testing Power. In *NIPS*, pages 181–189. 2016.

[4] D. Lopez-Paz, P. Hennig, and B. Schölkopf. The Randomized Dependence Coefficient. In *NIPS*, pages 1–9. 2013.

[5] W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. *arXiv:1610.04782*, 2016.

[1] Million Song Data subset: https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD.