A low variance consistent test of relative dependency

Wacha Bounliphone, Arthur Gretton, Matthew Blaschko

A low variance consistent test of relative dependency

Wacha Bounliphone
Supélec
Gif-sur-Yvette, France
wacha.bounliphone@supelec.fr

Arthur Gretton
Gatsby Unit, UCL
United Kingdom
arthur.gretton@gmail.com

Matthew B. Blaschko
Équipe GALEN
Inria Saclay, France
matthew.blaschko@inria.fr

Abstract

We describe a novel non parametric statistical hypothesis test of relative dependence between a source variable and two candidate target variables. Such a test enables one to answer whether one source variable is significantly more dependent on the first target variable or the second. Dependence is measured via the Hilbert-Schmidt Independence Criterion (HSIC), resulting in a pair of empirical dependence measures (source-target 1, source-target 2). Modeling the covariance between these HSIC statistics leads to a provably more powerful test than the construction of independent HSIC statistics by sub sampling. The resulting test is consistent and unbiased, and (being based on U-statistics) has favorable convergence properties. The test can be computed in quadratic time, matching the computational complexity of standard empirical HSIC estimators. We demonstrate the effectiveness of the test on a real-world linguistics problem of identifying language groups from a multilingual corpus.

1 Introduction

Tests of dependence are an important tool in statistical analysis, and are widely applied in many data analysis contexts. Classical criteria include Spearman's rho and Kendall's tau, which can detect only linear dependencies. More recent research on dependence measurement has focused on non parametric measures of dependence, which apply even when the dependence is nonlinear, or the variables are multivariate or non-euclidean (for instance images, strings, and graphs). The statistics for such tests are diverse, and include kernel measures of covariance [6, 19] and correlation [3, 4], distance covariances (which are instances of kernel tests) [18, 15], rankings [8], and space partitioning approaches [7, 13, 10].

For many problems in data analysis, however, the question of whether dependence exists is secondary: there may be multiple dependencies, and the question becomes which dependence is the strongest. For instance, in neuroscience, multiple stimulus modalities may be present (e.g. visual and audio), and it may be of interest to determine which of the two has a stronger influence on brain activity [9]. In cross-language information retrieval and automated translation [12], it may be of interest to determine whether documents in a source language language are a significantly better match to those in one target language than to another target language, either as a measure of difficulty of the respective learning tasks, or of the quality of the language features used.

We present a statistical test which determines whether two target variables have a significant difference in their dependence on a third, source variable. The dependence between each of the target

variables and the source is computed using the Hilbert-Schmidt Independence Criterion [5, 6]. Care must be taken in analyzing the asymptotic behavior of the test statistics, since the two measures of dependence will themselves be correlated: they are both computed with respect to the same source. Thus, we derive the *joint* asymptotic distribution of both dependencies, and design our test so as to take this correlation into account. We prove this approach to have greater statistical power than constructing two uncorrelated statistics on the same data by subsampling, and testing on these. In experiments, we are able to successfully test which of two variables is most strongly related to a third, in synthetic examples and in a language group identification task.

To our knowledge, there do not exist competing non-parametric tests to determining which of two dependencies are strongest. One related area is that of multiple regression analysis (e.g. [1]). In this case a linear model is assumed, and it is determined whether individual inputs have a statistically significant effect on an output variable. The procedure does not address the question of whether the influence of one variable is higher than that of another to a statistically significant degree. The problem of variable selection has also been investigated in the case of nonlinear relations between the inputs and outputs [2, 17], however this again does not address which of two variables most strongly influences a third. A less closely related area is that of detecting three-variable interactions [14], where it is determined whether there exists any factorization of the joint distribution over three variables. This test does not address the issue of finding which connections are strongest, however.

2 Definitions and description of HSIC

We base our underlying notion of dependence on the Hilbert-Schmidt Independence Criterion [5]. With this choice, our problem is described as follows:

Problem 1 *We have a sample of size m from distribution P_x, P_y and P_z, where P_x, P_y and P_z are the respective marginal distributions on domains $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$. Let P_{xy} be a Borel probability measure defined on a domain $\mathcal{X} \times \mathcal{Y}$ and P_{xz} be a Borel probability measure defined on a domain $\mathcal{X} \times \mathcal{Z}$. We want to test the null hypothesis $\mathcal{H}_0 : HSIC(\mathcal{F}, \mathcal{H}, P_{xz})$ is greater or equal to $HSIC(\mathcal{F}, \mathcal{G}, P_{xy})$.*

We begin with a description of our kernel dependence criterion, the Hilbert-Schmidt Independence Criterion (HSIC) [5, 6], and then describe its asymptotic behaviour for dependent variables.

Definition 1 *Given separable RKHSs $\mathcal{F}$ and $\mathcal{G}$, and a joint measure P_{xy} over $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$, we define the Hilbert-Schmidt Independence Criterion (HSIC) as the square HS-norm of the associated cross-covariance operator $\|C_{xy}\|^2_{HS}$. HSIC can be expressed in terms of expectations of kernel functions,*

$$\begin{aligned} &HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) := \|C_{xy}\|^2_{HS} \\ &= \mathbb{E}_{xx'yy'}\left[k(x,x')l(y,y')\right] + \mathbb{E}_{xx'}\left[k(x,x')\right]\mathbb{E}_{yy'}\left[l(y,y')\right] \\ &- 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}[k(x,x')]\mathbb{E}_{y'}[l(y,y')]\right] \end{aligned} \tag{1}$$

Theorem 1 (Biased estimator of HSIC) *We denote by $\mathcal{S} = (X, Y)$ the set of observations $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ which are drawn iid from P_{xy}, the Borel probability measure defined on a domain $(\mathcal{X} \times \mathcal{Y})$. An estimator of HSIC, written $HSIC_0(\mathcal{F}, \mathcal{G}, \mathcal{S})$, is given by the following expression*

$$HSIC_0(\mathcal{F}, \mathcal{G}, \mathcal{S}) := (m-1)^{-2}\operatorname{Tr}(\mathbf{KHLH}). \tag{2}$$

where $\mathbf{K}$, $\mathbf{L} \in R^{m\times m}$ are kernel matrices containing $k_{ij} = k(x_i, x_j)$ and $l_{ij} = l(y_i, y_j)$ and $H = \boldsymbol{I} - m^{-1}\mathbb{1}\mathbb{1}^T \in \mathbb{R}^{m\times m}$ is a centering matrix.
The estimator $HSIC_0(\mathcal{F}, \mathcal{G}, \mathcal{S})$ has bias $O(m^{-1})$, that is, $HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) - \mathbb{E}_{\mathcal{S}}\left[HSIC_0(\mathcal{F}, \mathcal{G}, \mathcal{S})\right] = O(m^{-1})$

Theorem 2 (Unbiaised estimator of HSIC [17]) *The unbiaised estimator $HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S})$ is given by,*

$$HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}) := \frac{1}{m(m-3)}\left[\operatorname{Tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{\mathbb{1}'\tilde{\mathbf{K}}\mathbb{1}\mathbb{1}'\tilde{\mathbf{L}}\mathbb{1}}{(m-1)(m-2)} - \frac{2}{m-2}\mathbb{1}'\tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbb{1}\right] \tag{3}$$

where $\tilde{\mathbf{K}}$ *and* $\tilde{\mathbf{L}}$ *are the centered kernel matrices of* $\mathbf{K}$ *and* $\mathbf{L}$.

Theorem 3 (U-statistic of HSIC) *$HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S})$ can be written in terms of a U-statistic,*

$$HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}) = (m)_4^{-1} \sum_{(i,j,q,r) \in i_4^m} h_{ijqr} \tag{4}$$

where $(m)_n := \dfrac{m!}{(m-n)!}$*, the index set* i_r^m *denotes the set of all* $r-$*tuples drawn without replacement from the set* $\{1, \dots m\}$ *and the kernel h of the U-statistic is defined by*

$$h_{ijqr} = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st} + l_{uv} - 2l_{su}) \tag{5}$$

We will denote the finite sample biased estimator of HSIC as $HSIC_0(\mathcal{F}, \mathcal{G}, \mathcal{S}) := HSIC_0^{XY}$ and the finite sample unbiased estimator of HSIC as $HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}) := HSIC_1^{XY}$

Theorem 4 (Asymptotic distribution of $HSIC_1$ [17]) *If* $\mathbb{E}[h^2] < \infty$*, and data and labels are not independent, then as* $m \to \infty$*,* $HSIC_1^{XY}$ *converges in distribution to a Gaussian random variable with mean* $HSIC(\mathcal{F}, \mathcal{G}, P_{xy})$ *and estimated variance* $\sigma^2_{HSIC_1^{XY}}$

$$\sigma^2_{HSIC_1^{XY}} := \sigma^2_{XY} = \frac{1}{m}\left(R_{XY} - (HSIC_1^{XY})^2\right) \tag{6}$$

where $R_{XY} = \dfrac{1}{m}\sum_{i=1}^{m}\left((m-1)_3^{-1} \sum_{(j,q,r) \in i_3^m \setminus \{i\}} h_{ijqr}\right)^2$ *and the index set* $i_r^m \setminus \{i\}$ *denotes the set of all* $r-$*tuples drawn without replacement from the set* $\{1, \dots m\} \setminus \{i\}$.

3 A test of relative dependence

In this section we first construct a simple consistent test for Problem 1. This is done by computing two independent HSIC statistics and deriving a sound test threshold. We subsequently prove that this strategy is strictly less powerful than a test based on two dependent HSIC statistics. We therefore derive the joint asymptotic distribution of these dependent quantities, which is then used to construct a more powerful, consistent test.

3.1 A simple, consistent approach

From the result in Equation (6), we can construct a consistent relative test as follows : split the samples from P_x into two equal sized sets denoted by X' and X'', drop the second half of the samples from P_y and the first half of the samples from P_z. We will denote the remaining samples as Y' and Z''. We can now estimate the joint distribution of $[HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}), HSIC_1(\mathcal{F}, \mathcal{H}, \mathcal{S})]^T := [HSIC_1^{XY}, HSIC_1^{XZ}]^T$ as

$$\mathcal{N}\left(\begin{pmatrix} HSIC_1^{X'Y'} \\ HSIC_1^{X''Z''} \end{pmatrix}, \begin{pmatrix} \sigma^2_{X'Y'} & 0 \\ 0 & \sigma^2_{X''Z''} \end{pmatrix}\right) \tag{7}$$

which we will write as $\mathcal{N}(\mu', \Sigma')$. Given this joint distribution, we need to integrate the density of the distribution below the line defined by $HSIC_1^{X'Y'} = HSIC_1^{X''Z''}$. A simple way of achieving this is to rotate the distribution by $\frac{\pi}{4}$ rad counter-clockwise about the origin, and to integrate the resulting distribution projected onto the first axis. The resulting projection of the rotated distribution onto the primary axis is

$$\mathcal{N}\left([R\mu']_1, [R\Sigma' R^T]_{11}\right) = \mathcal{N}\left(\frac{\sqrt{2}}{2}(HSIC_1^{X'Y'} - HSIC_1^{X''Z''}), \frac{1}{2}(\sigma^2_{X'Y'} + \sigma^2_{X''Z''})\right) \tag{8}$$

where $R = \dfrac{\sqrt{2}}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ is the rotation matrix by $\frac{\pi}{4}$. From this empirically estimated distribution, it is relatively straightforward to construct a consistent test (cf. Equation (18)). The power of this test varies inversely with the variance of the distribution in Equation (8).

3.2 A lower variance approach

The above test is suboptimal in the sense that we have dropped half of the available samples in order to guarantee independence of the two empirical HSIC estimates. While discarding half the samples leads to a consistent test, we may achieve a lower-variance, unbiased, consistent test by finding an unbiased estimate of the off-diagonal covariance term using all available samples. In this section, we prove that such a test is always more powerful than the previous strategy, regardless of P_x, P_y and P_z.

We prove that the resulting test is asymptotically lower variance as follows : we first note that, given an unbiased and consistent estimate of σ_{XYXZ}, the empirical estimate of the joint distribution of $[HSIC_1^{XY} HSIC_1^{XZ}]^T$ has the form

$$\mathcal{N}\left(\begin{pmatrix} HSIC_1^{XY} \\ HSIC_1^{XZ} \end{pmatrix}, \begin{pmatrix} \sigma_{XY}^2 & \sigma_{XYXZ} \\ \sigma_{XYXZ} & \sigma_{XZ}^2 \end{pmatrix}\right) \tag{9}$$

which we denote as $\mathcal{N}(\mu, \Sigma)$. The resulting distribution from rotating by $\frac{\pi}{4}$ rad and projecting onto the primary axis is

$$\mathcal{N}\left([R\mu]_1, [R\Sigma R^T]_{11}\right) = \mathcal{N}\left(\frac{\sqrt{2}}{2}(HSIC_1^{XY} - HSIC_1^{XZ}), \frac{1}{2}(\sigma_{XY}^2 + \sigma_{XZ}^2 - 2\sigma_{XYXZ})\right) \tag{10}$$

Theorem 5 *A dependent test for Problem 1 is always more powerful than an independent test.*

Proof 1 *The two following lemmas are used for the proof of the Theorem 5.*

Lemma 1 *(Same Mean) As $HSIC_1(\cdot,\cdot)$ is an unbiased empirical estimator, we have that*

$$\begin{aligned}\mathbb{E}\left[\frac{\sqrt{2}}{2}(HSIC_1^{X'Y'} - HSIC_1^{X''Z'})\right] &= \mathbb{E}\left[\frac{\sqrt{2}}{2}(HSIC_1^{XY} - HSIC_1^{XZ})\right] \\ &= \frac{\sqrt{2}}{2}(HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz}))\end{aligned} \tag{11}$$

Lemma 2 *(Lower Variance) From the convergence of moments in the application of the central limit theorem [16], we have that $\sigma_{X'Y'}^2 \to 2\sigma_{XY}^2$ as the sample size $n \to \infty$. Using the property $\sigma_{X'Y'}^2 \to 2\sigma_{XY}^2$ we find that the variance of the estimate summarized in Equation* (10) *is $\frac{1}{2}(\sigma_{XY}^2 + \sigma_{XZ}^2 - 2\sigma_{XYXZ})$ while the variance of the estimate summarized in Equation* (8) *approaches $\frac{1}{2}(2\sigma_{XY}^2 + 2\sigma_{XZ}^2)$. We have that the variance of the dependent test is smaller than the variance of the indpendent test when*

$$\frac{1}{2}(\sigma_{XY}^2 + \sigma_{XZ}^2 - 2\sigma_{XYXZ}) < \frac{1}{2}(2\sigma_{XY}^2 + 2\sigma_{XZ}^2) \iff -2\sigma_{XYXZ} < \sigma_{XY}^2 + \sigma_{XZ}^2 \tag{12}$$

which is implied by the positive definiteness of Σ. □

To summarize, after we have observed a sufficient sample size $n > \tau$ for some distribution dependent τ, we always have that the statistical test using the distribution in Equation (10) is lower variance than the statistical test using the distribution in Equation (8), and the dependent test is therefore more powerful than the independent test.

We now derive the joint asymptotic distribution of the dependent HSIC estimates and show that it can be computed in quadratic time.

3.3 Joint asymptotic distribution of HSIC

To solve Problem 1, we denote by $\mathcal{S}_1 = (X, Y, Z)$ the joint sample from three sets of the observations (X, Y, Z) which are drawn *iid* with respective Borel probability measure P_x, P_y and P_z defined on domains $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$. The kernels k, l and m are associated uniquely with respective reproducing kernel Hilbert spaces $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$. Moreover, $\mathbf{K}$, $\mathbf{L}$ and

$\mathbf{M} \in R^{m\times m}$ are kernel matrices containing $k_{ij} = k(x_i, x_j)$, $l_{ij} = k(y_i, y_j)$ and $m_{ij} = m(z_i, z_j)$. Let $HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}_1)$ and $HSIC_1(\mathcal{F}, \mathcal{H}, \mathcal{S}_1)$ be respectively an unbiased estimator of $HSIC_1(\mathcal{F}, \mathcal{G}, P_{xy})$ and $HSIC_1(\mathcal{F}, \mathcal{H}, P_{xz})$ written as a sum of U-statistics with kernels h_{ijqr} and g_{ijqr} as described in (5),

$$h_{ijqr} = \frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st} + l_{uv} - 2l_{su})$$
$$g_{ijqr} = \frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(m_{st} + m_{uv} - 2m_{su}) \tag{13}$$

Theorem 6 (Joint asymptotic distribution of HSIC) *If $\mathbb{E}[h^2] < \infty$, then as $m \to \infty$, the joint asymptotic distribution of $[HSIC_1(\mathcal{F}, \mathcal{G}, \mathcal{S}_1), HSIC_1(\mathcal{F}, \mathcal{H}, \mathcal{S}_1)]^T := [HSIC_1^{XY}, HSIC_1^{XZ}]^T$ is a multivariate Gaussian,*

$$\mathcal{N}\left\{\begin{pmatrix} HSIC_1^{XY} \\ HSIC_1^{XZ} \end{pmatrix}, \begin{pmatrix} \sigma_{XY}^2 & \sigma_{XYXZ} \\ \sigma_{XYXZ} & \sigma_{XZ}^2 \end{pmatrix}\right\} \tag{14}$$

where σ_{XY}^2 and σ_{XZ}^2 are as in Theorem 4, and $\sigma_{XYXZ}^2 = \frac{16}{m}\left(R_{XYXZ} - HSIC_1^{XY} HSIC_1^{XZ}\right)$, where R_{XYXZ} is equal to

$$R_{XYXZ} = \frac{1}{m}\sum_{i=1}^{m}\left((m-1)_3^{-1}\sum_{(j,q,r)\in i_3^m\setminus\{i\}} h_{ijqr}g_{ijqr}\right). \tag{15}$$

Proof 2 *First, $HSIC_0^{XY}$ and $HSIC_0^{XZ}$ in Equation* (2) *can be writing as $V-statistics$, respectively,*

$$HSIC_{0,Vstat}^{XY} = \frac{1}{m^4}\sum_{i,j,q,r}^{m} h_{ijqr}, \qquad HSIC_{0,Vstat}^{XZ} = \frac{1}{m^4}\sum_{i,j,q,r}^{m} g_{ijqr}, \tag{16}$$

where h_{ijqr} and g_{ijqr} defined in (13) *do not change with permutation of their indices. Since the difference between the covariance of $HSIC_{0,Vstat}^{XY}$ and $HSIC_{0,Vstat}^{XZ}$ and the covariance of $HSIC_{1,Ustat}^{XY}$ and $HSIC_{1,Ustat}^{XZ}$ drops as $\frac{1}{m}$, then the joint asymptotic distribution of $[HSIC_1^{XY}, HSIC_1^{XZ}]^T$ converges asymptotically .* □

Unfortunately equation (15) is expensive to compute, because even computing the kernels h_{ijqr} and g_{ijqr} of the $U-$statistic itself is a non trivial task. Following [17], however, we first form a vector $\mathbf{h_{XY}}$ with its entry corresponding to $\sum_{(j,q,r)\in i_3^m\setminus\{i\}} h_{ijqr}$, and a vector $\mathbf{h_{XZ}}$ with its entry corresponding to $\sum_{(j,q,r)\in i_3^m\setminus\{i\}} g_{ijqr}$. Collecting terms in Equation (5) related to kernel matrices $\mathbf{K}$ and $\mathbf{L}$, $\mathbf{h_{XY}}$ and $\mathbf{h_{XZ}}$ can be written as

$$\begin{aligned}\mathbf{h_{XY}} = (m-2)^2\left(\tilde{K}\odot\tilde{L}\right)\mathbb{1} + (m-2)\left((Tr(\tilde{K}\tilde{L}))\mathbb{1} - \tilde{K}(\tilde{L}\mathbb{1}) - \tilde{L}(\tilde{K}\mathbb{1})\right)\\ - m(\tilde{K}\mathbb{1})\odot(\tilde{L}\mathbb{1}) + (\mathbb{1}'\tilde{L}\mathbb{1})\tilde{K}\mathbb{1} + (\mathbb{1}'\tilde{K}\mathbb{1})\tilde{L}\mathbb{1} - ((\mathbb{1}'\tilde{K})(\tilde{L}\mathbb{1}))\mathbb{1},\end{aligned} \tag{17}$$

where $\odot$ denotes element wise matrix multiplication. Then R_{XYXZ} in Equation (15) can be computed as $R_{XYXZ} = (4m)^{-1}(m-1)_3^{-2}h_{XY}^T h_{XZ}$. Using the order of operations implied by parentheses in Equation (17), the computation time of the cross covariance term is $O(m^2)$. Combining this with the unbiased estimator of HSIC in Equation (3) leads to a final computational complexity of $O(m^2)$.

4 Statistical test description

4.1 Dependence test for pairs of HSIC statistics

We can now describe a statistical test of dependence for two sets of observations, based on the joint asymptotic distribution of HSIC described in Theorem 6. Given a sample $\mathcal{S}_1$ as described in

section 3.3 earlier, the dependence statistical test $\mathcal{T}(\mathcal{S}_1) : \{(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^m\} \to \{0, 1\}$ is used to test the null hypothesis $\mathcal{H}_0 : HSIC_1^{XY}$ is less than or equal to $HSIC_1^{XZ}$ versus the alternative hypothesis $\mathcal{H}_1 : HSIC_1^{XY}$ is greater than or equal to $HSIC_1^{XZ}$ at a given significance level α. Following the empirical distribution from Equation (10), we determine the p-value to be

$$1 - \mathbf{\Phi}\left(\frac{(HSIC_1^{XY} - HSIC_1^{XZ})}{\sqrt{\sigma_{XY}^2 + \sigma_{XZ}^2 - 2\sigma_{XYXZ}}}\right) \tag{18}$$

where $\mathbf{\Phi}$ is the CDF of a standard normal distribution.

4.2 Generalizing to arbitrary numbers of HSIC statistics

The generalization of this dependence test to more than three domains follows from the derivation of the above test by applying successive rotations to a higher dimensional joint Gaussian distribution over multiple HSIC statistics. We assume a sample $\mathcal{S}$ of size m over d domains with kernels $k_1, \ldots, k_d$ associated uniquely with respective reproducing kernel Hilbert spaces, $\mathcal{F}_1, \ldots, \mathcal{F}_d$. We define a generalized statistical test, $\mathcal{T}_g(\mathcal{S}) \to \{0, 1\}$ to test the null hypothesis $\mathcal{H}_0 : \sum_{(x,y)\in\{1,\ldots,d\}} v_{(x,y)} HSIC_1(\mathcal{F}_x, \mathcal{F}_y, \mathcal{S}) \leq 0$ versus the alternative hypothesis $\mathcal{H}_1 : \sum_{(x,y)\in\{1,\ldots,d\}} v_{(x,y)} HSIC_1(\mathcal{F}_x, \mathcal{F}_y, \mathcal{S}) > 0$ where v is a vector of weights on each HSIC statistic. We may recover the test in the previous section by setting $v = [+1, -1]^T$.

The derivation of the test follows the general strategy used in the previous section: we construct a rotation matrix such that the test may project the joint Gaussian distribution onto the first axis, and read the p-value from a standard normal table. To construct the rotation matrix, we simply need to rotate v such that it is aligned with the first axis. Such a rotation can be computed by composing d 2-dimensional rotation matrices as in Algorithm 1.

Algorithm 1: Successive rotation for generalized high-dimensional relative tests of dependency (cf. Section 4.2)

Require: $v \in \mathbb{R}^d$
Ensure: $[Rv]_i = 0 \ \ \forall i \neq 1, R^T R = I$
  $R = I$
  for i = 2 **to** d **do**
    $\theta = -\tan^{-1} \frac{v_i}{[Rv]_1}$
    $R_i = I$
    $[R_i]_{11} = \cos(\theta)$
    $[R_i]_{1i} = -\sin(\theta)$
    $[R_i]_{i1} = \sin(\theta)$
    $[R_i]_{ii} = \cos(\theta)$
    $R = R_i R$
  end for

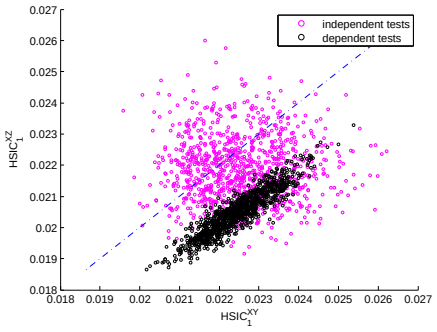


Figure 1: Empirical HSIC values for dependent and independent tests with 1000 samples. The dependent distribution converges faster than the independent distribution, resulting in a more powerful statistical test.

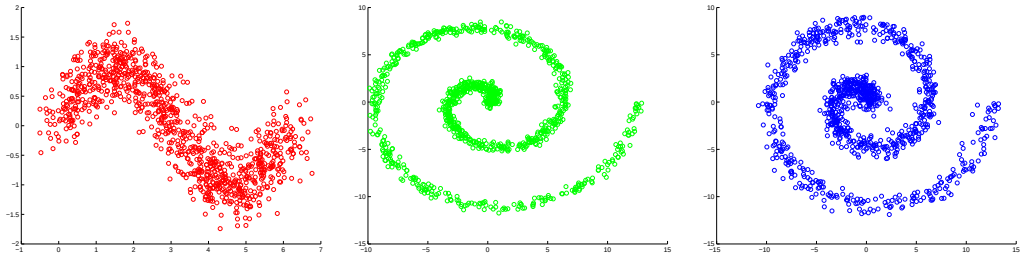
5 Experiments

We apply our estimates of statistical dependence to two problems. The first is a synthetic data experiment, in which we can directly control the relative degree of functional dependence between variates. The second experiment uses a multilingual corpus to determine the relative relation between European languages, demonstrating that the test is able to identify membership in major language groups (Romance languages vs. Germanic languages).

5.1 Synthetic experiment

We constructed 3 artificial distributions, as illustrated in figure 2. These data sets are constructed so that we can control the relative degree of functional dependence between the variates by varying the relative size of noise scaling parameters γ_2 and γ_3. In these experiments, we fixed $\gamma_1 = \gamma_2 = 0.3$,

while we varied γ_3. Figure 3 shows the empirical p-values for 100 repeated tests for 101 regularly spaced values of $\gamma_3 \in [0, 1]$ and 1000 samples per distribution. Figure 1 shows an empirical scatter plot of estimated HSIC values for the dependent and independent tests, showing a much tighter distribution of dependent tests. As predicted by the theory (cf. Theorem 5) the dependent test is substantially more powerful than an independent test.



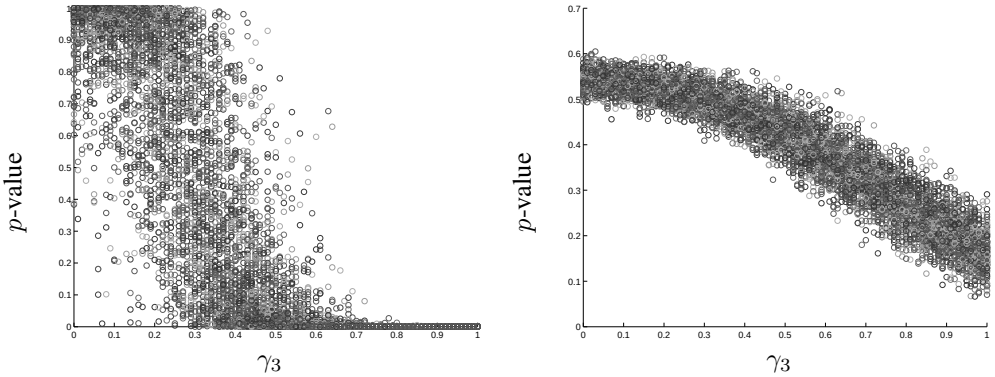
(a) (b) (c)

Figure 2: Artificial data sets. Let $t \sim \mathcal{U}[(0, 2\pi)]$:

(a) : $x_1 \sim t + \gamma_1 \mathcal{N}(0, 1)$ $\quad y_1 \sim \sin(t) + \gamma_1 \mathcal{N}(0, 1)$

(b) : $x_2 \sim t \cos(t) + \gamma_2 \mathcal{N}(0, 1)$ $\quad y_2 \sim t \sin(t) + \gamma_2 \mathcal{N}(0, 1)$

(c) : $x_3 \sim t \cos(t) + \gamma_3 \mathcal{N}(0, 1)$ $\quad y_3 \sim t \sin(t) + \gamma_3 \mathcal{N}(0, 1)$



(a) dependent test (b) independent test

Figure 3: γ_3 vs. empirical p-values for the synthetic experiments described in Section 5.1 and Figure 2. The dependent test (a) described in Section 4.1 is substantially more powerful than a simple consistent independent test (b) constructed by subsampling the data.

5.2 Multilingual data

In this section, we demonstrate dependence testing on different languages. We use real world data sets taken from the European Parliament corpus [11], which includes translations in 21 European languages. We choose 1000 random documents written in : French (fr), German (de), Dutch (nl), Spanish (sp), Portuguese (pt), Italian (it), Swedish (sv), and Danish (da). These languages can be broadly categorized into either the Germanic or Romance families of European languages. Our goal was to test if the statistical dependence between two languages in the same group is greater than the statistical dependence between languages in different groups. For preprocessing, we removed stop-words (`http://www.nltk.org`) and performed stemming (`http://snowball.tartarus.org`). We applied the bag-of-words model (BoW model) as a feature representation and used a χ^2 exponential kernel with the bandwidth set per language as the median pairwise χ^2 distance between documents. A selection of three-language tests is given in Table 1. All p-values strongly support that the dependent tests find the different language groups with very high significance. In contrast, the independent test yielded p-values on the order of 0.10, which would not reject the null hypothesis at standard significance levels.

In our next tests, we evaluate our more general framework for testing relative dependencies with more than two HSIC statistics. We chose four languages, and tested whether the dependence be-

tween two languages in the same group is higher than an average of dependencies between groups. The results of these tests are give in Table 2. As before, our generalized test is able to distinguish language groups with high significance.

Table 1: Relative dependency test between two pairs of HSIC statistics for the multilingual corpus data.

Source	Targets	p-value
nl	de fr	0.0001^{***}
it	pt sw	$< 0.0001^{***}$

Table 2: Relative dependency test of four pairs of HSIC statistics for the multilingual corpus data.

Source	Targets	p-value
es	it pt sw	0.0013^{***}
es	it pt da	0.0198^{***}

6 Conclusions

We have described a novel statistical test that determines whether a source random variable is more strongly dependent on one target random variable or another. This test, built on the Hilbert-Schmidt Independence Criterion, is low variance, consistent, and unbiased. We have shown that our test is strictly more powerful than a test that does not exploit the covariance between the HSIC from the source to each of the targets. We have empirically demonstrated the test performance on synthetic data, where the degree of dependence could be controlled; and on the challenging problem of identifying language groups from a multilingual corpus. The computation and memory requirements of the test are quadratic in the sample size, matching the performance of HSIC and related tests for dependence between two random variables. We have generalized the test framework to more than two HSIC statistics, and have given an algorithm to construct a consistent, low-variance, unbiased test when comparing multiple HSIC dependencies.

References

[1] M. Srivastava A. Sen. *Regression Analysis – Theory, Methods, and Applications*. Springer-Verlag, 2011.

[2] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *JMLR*, 13:795–828, 2012.

[3] J. Dauxois and G. M. Nkiet. Nonlinear canonical analysis and independence tests. *Ann. Statist.*, 26(4):1254–1278, 1998.

[4] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. pages 489–496. MIT Press, 2008.

[5] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pages 63–77, 2005.

[6] A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *NIPS*, pages 585–592, 2008.

[7] A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[8] R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

[9] K. Kording J. Trommershauser and M. S. Landy. *Sensory Cue Integration*. Oxford University Press, 2011.

[10] J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *PNAS*, 2014.

[11] P. Koehn. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, 2005.

[12] C. Peters, M. Braschler, and P. Clough. *Multilingual Information Retrieval: From Research to Practice*. Springer, 2012.

[13] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large datasets. *Science*, 334(6062), 2011.

[14] D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *NIPS*, 2013.

[15] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2702, 2013.

[16] R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 1981.

[17] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[18] G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Stat.*, 35(6):2769–2794, 2007.

[19] K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.