A low variance consistent test of relative dependency

Wacha Bounliphone WACHA.BOUNLIPHONE@CENTRALESUPELEC.FR
CentraleSupélec & Inria, Grande Voie des Vignes, 92295 Châtenay-Malabry, France

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, University College London, United Kingdom

Arthur Tenenhaus ARTHUR.TENENHAUS@CENTRALESUPELEC.FR
CentraleSupélec, 3 rue Joliot-Curie, 91192 Gif-Sur-Yvette, France

Matthew B. Blaschko MATTHEW.BLASCHKO@INRIA.FR
Inria & CentraleSupélec, Grande Voie des Vignes, 92295 Châtenay-Malabry, France

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 us to determine whether one source variable is significantly more dependent on a first target variable or a 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). We test whether the first dependence measure is significantly larger than the second. Modeling the covariance between these HSIC statistics leads to a provably more powerful test than the construction of independent HSIC statistics by subsampling. 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. The effectiveness of the test is demonstrated on several real-world problems: we identify language groups from a multilingual corpus, and we prove that tumor location is more dependent on gene expression than chromosomal imbalances. Source code is available for download at `https://github.com/wbounliphone/reldep`.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

Tests of dependence are important tools in statistical analysis, and are widely applied in many data analysis contexts. Classical criteria include Spearman's ρ and Kendall's τ, which can detect non-linear monotonic 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 (Gretton et al., 2008; Zhang et al., 2011) and correlation (Dauxois & Nkiet, 1998; Fukumizu et al., 2008), distance covariances (which are instances of kernel tests) (Székely et al., 2007; Sejdinovic et al., 2013b), kernel regression tests (Cortes et al., 2009; Gunn & Kandola, 2002), rankings (Heller et al., 2013), and space partitioning approaches (Gretton & Gyorfi, 2010; Reshef et al., 2011; Kinney & Atwal, 2014). Specialization of such methods to univariate linear dependence can yield similar tests to classical approaches such as Darlington (1968); Bring (1996).

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 stimuli may be present (e.g. visual and audio), and it is of interest to determine which of the two has a stronger influence on brain activity (Trommershauser et al., 2011). In automated translation (Peters et al., 2012), it is of interest to determine whether documents in a source 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 as a basic tool for comparative linguistics.

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 (Gretton et al., 2005; 2008).[1] 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. The derivation of our test utilizes classical results of U-statistics (Hoeffding, 1963; Serfling, 1981; Arcones & Gine, 1993). In particular, we make use of results by Hoeffding (1963) and Serfling (1981) to determine the asymptotic joint distributions of the statistics (see Theorem 4). Consequently, we derive the *lowest* variance unbiased estimator of the test statistic.

We prove our 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, in a language group identification task, and in a task for identifying the relative strength of factors for Glioma type in a pediatric patient population.

To our knowledge, there do not exist competing non-parametric tests to determine which of two dependencies is strongest. One related area is that of multiple regression analysis (e.g. (Sen & Srivastava, 2011)). 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 (Cortes et al., 2009; 2012; Song et al., 2012), 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 (Sejdinovic et al., 2013a), where it is determined whether there exists any factorization of the joint distribution over three variables. This test again does not address the issue of finding which connections are strongest, however.

[1] Dependency can also be tested with the correlation operator. However, Fukumizu et al., (2007) show that unlike the covariance operator, the asymptotic distribution of the norm of the correlation operator is unknown, so the construction of a computationally efficient test of relative dependence remains an open problem.

2. Definitions and description of HSIC

We base our underlying notion of dependence on the Hilbert-Schmidt Independence Criterion (Gretton et al., 2005; 2008; Song et al., 2012). All results in this section except for Problem 1 can be found in these previous works.

Definition 1. *(Gretton et al., 2005, Definition 1,Lemma 1: Hilbert-Schmidt Independence Criterion)*

Let P_{xy} be a Borel probability measure over over $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$ with Γ and Λ the respective Borel sets on $\mathcal{X}$ and $\mathcal{Y}$, and P_x and P_y the marginal distributions on domains $\mathcal{X}$ and $\mathcal{Y}$. Given separable RKHSs $\mathcal{F}$ and $\mathcal{G}$, the Hilbert-Schmidt Independence Criterion (HSIC) is defined as the squared HS-norm of the associated cross-covariance operator C_{xy}. When the kernels k, l are associated uniquely withs respective RKHSs $\mathcal{F}$ and $\mathcal{G}$ and bounded, HSIC can be expressed in terms of expectations of kernel functions

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

HSIC determines independence: HSIC = 0 iff $P_{xy} = P_x P_y$ when kernels k and l are characteristic on their respective marginal domains (Gretton, 2015).

With this choice, the problem we would like to solve is described as follows:

Problem 1. *Given separable RKHSs $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ with $HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) > 0$ and $HSIC(\mathcal{F}, \mathcal{H}, P_{xz}) > 0$, we test the null hypothesis $\mathcal{H}_0$: $HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) \leq HSIC(\mathcal{F}, \mathcal{H}, P_{xz})$ versus the alternative hypothesis $\mathcal{H}_1$: $HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) > HSIC(\mathcal{F}, \mathcal{H}, P_{xz})$ at a given significance level α.*

We now describe the asymptotic behavior of the HSIC for dependent variables.

Theorem 1. *(Song et al., 2012, Theorem 2: Unbiased estimator for $HSIC(\mathcal{F}, \mathcal{G}, P_{xy})$) We denote by $\mathcal{S}$ the set of observations $\{(x_1, y_1), ..., (x_m, y_m)\}$ of size m drawn i.i.d. from P_{xy}. The unbiased estimator $HSIC_m(\mathcal{F}, \mathcal{G}, \mathcal{S})$ is given by*

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

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ are related to $\mathbf{K}$ and $\mathbf{L}$ by $\tilde{\mathbf{K}}_{ij} = (1 - \delta_{ij})\tilde{\mathbf{K}}_{ij}$ and $\tilde{\mathbf{L}}_{ij} = (1 - \delta_{ij})\tilde{\mathbf{L}}_{ij}$.

Theorem 2. *(Song et al., 2012, Theorem 3: U-statistic of HSIC) This finite sample unbiased estimator of $HSIC_m^{XY}$*

can be written as a U-statistic,

$$HSIC_m^{XY} = (m)_4^{-1} \sum_{(i,j,q,r)\in i_4^m} h_{ijqr} \tag{3}$$

where $(m)_4 := \frac{m!}{(m-4)!}$*, the index set* i_4^m *denotes the set of all* $4-$*tuples drawn without replacement from the set* $\{1,\dots m\}$*, and the kernel h of the U-statistic is defined as*

$$h_{ijqr} = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st} + l_{uv} - 2l_{su}) \tag{4}$$

where the kernels k *and* l *are associated uniquely with respective reproducing kernel Hilbert spaces* $\mathcal{F}$ *and* $\mathcal{G}$*.*

Theorem 3. *(Gretton et al., 2008, Theorem 1: Asymptotic distribution of* $HSIC_m$*) If* $\mathbf{E}[h^2] < \infty$*, and source and targets are not independent, then, under* $\mathcal{H}_1$*, as* $m \to \infty$*,*

$$\sqrt{m}(HSIC_m^{XY} - HSIC(\mathcal{F},\mathcal{G},P_{xy})) \xrightarrow{d} \mathcal{N}(0,\sigma_{XY}^2) \tag{5}$$

where $\sigma_{XY}^2 = 16\left(\mathbb{E}_i\left(\mathbf{E}_{j,q,r}h_{ijqr}\right)^2 - HSIC(\mathcal{F},\mathcal{G},P_{xy}))\right)$ *with* $\mathbf{E}_{j,q,r} := \mathbf{E}_{S_j,S_q,S_r}$*. Its empirical estimate is* $\hat{\sigma}_{XY} = 16\left(R_{XY} - (HSIC_m^{XY})^2\right)$ *where* $R_{XY} = \frac{1}{m}\sum_{i=1}^m \left((m-1)_3^{-1} \sum_{(j,q,r)\in i_3^m\setminus\{i\}} h_{ijqr}\right)^2$ *and the index set* $i_3^m\setminus\{i\}$ *denotes the set of all* $3-$*tuples drawn without replacement from the set* $\{1,\dots m\}\setminus\{i\}$*.*

3. A test of relative dependence

In this section we calculate two dependent HSIC statistics and derive the joint asymptotic distribution of these dependent quantities, which is used to construct a consistent test for Problem 1. We next construct a simpler consistent test, by computing two independent HSIC statistics on sample subsets. While the simpler strategy is superficially attractive and less effort to implement, we prove the dependent strategy is strictly more powerful.

3.1. Joint asymptotic distribution of HSIC and test

In the present section, we compute each HSIC estimate on the full dataset, and explicitly obtain the correlations between the resulting empirical dependence measurements $HSIC_m^{XY}$ and $HSIC_m^{XZ}$. We denote by $\mathcal{S}_1 = (X,Y,Z)$ the joint sample of observations which are drawn *i.i.d.* with respective Borel probability measure P_{xyz} defined on the domain $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$. The kernels k, l and d are associated uniquely with respective reproducing kernel Hilbert spaces $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$. Moreover, $\mathbf{K}$, $\mathbf{L}$ and $\mathbf{D} \in R^{m\times m}$ are kernel matrices containing $k_{ij} = k(x_i,x_j)$, $l_{ij} = l(y_i,y_j)$ and $d_{ij} = d(z_i,z_j)$. Let $HSIC_m^{XY}$ and $HSIC_m^{XZ}$ be respectively the unbiased estimators of $HSIC(\mathcal{F},\mathcal{G},P_{xy})$ and $HSIC(\mathcal{F},\mathcal{H},P_{xz})$, written as a sum of U-statistics with respective kernels h_{ijqr} and g_{ijqr} as described in (4),

$$h_{ijqr} = \frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st}+l_{uv}-2l_{su}),$$
$$g_{ijqr} = \frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(d_{st}+d_{uv}-2d_{su}). \tag{6}$$

Theorem 4. (Joint asymptotic distribution of HSIC) *If* $\mathbb{E}[h^2] < \infty$ *and* $\mathbb{E}[g^2] < \infty$*, then*

$$\sqrt{m}\left(\begin{pmatrix} HSIC_m^{XY} \\ HSIC_m^{XZ}\end{pmatrix} - \begin{pmatrix} HSIC(\mathcal{F},\mathcal{G},P_{xy}) \\ HSIC(\mathcal{F},\mathcal{H},P_{xz})\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{XY}^2 & \sigma_{XYXZ} \\ \sigma_{XYXZ} & \sigma_{XZ}^2\end{pmatrix}\right), \tag{7}$$

where σ_{XY}^2 *and* σ_{XZ}^2 *are as in Theorem 3. The empirical estimate of* σ_{XYXZ} *is* $\hat{\sigma}_{XYXZ} = \frac{16}{m}\left(R_{XYXZ} - HSIC_m^{XY}HSIC_m^{XZ}\right)$*, where*

$$R_{XYXZ} = \frac{1}{m}\sum_{i=1}^m \left((m-1)_3^{-2} \sum_{(j,q,r)\in i_3^m\setminus\{i\}} h_{ijqr}g_{ijqr}\right). \tag{8}$$

Proof. Eq. (8) is constructed with the definition of variance of a U-statistic as given by Serfling, Ch. 5 (1981), where one variable is fixed. Eq. (7) follows from the application of Hoeffding, Theorem 7.1 (1963), which gives the joint asymptotic distribution of U-statistics. □

Based on the joint asymptotic distribution of HSIC described in Theorem 4, we can now describe a statistical test to solve Problem 1: given a sample $\mathcal{S}_1$ as described in Section 3.1, $\mathcal{T}(\mathcal{S}_1) : \{(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})^m\} \to \{0,1\}$ is used to test the null hypothesis $\mathcal{H}_0 : HSIC(\mathcal{F},\mathcal{G},P_{xy}) \le HSIC(\mathcal{F},\mathcal{H},P_{xz})$ versus the alternative hypothesis $\mathcal{H}_1 : HSIC(\mathcal{F},\mathcal{G},P_{xy}) > HSIC(\mathcal{F},\mathcal{H},P_{xz})$ at a given significance level α. This is achieved by projecting the distribution to 1D using the statistic $HSIC_m^{XY} - HSIC_m^{XZ}$, and determining where the statistic falls relative to a conservative estimate of the the $1-\alpha$ quantile of the null. We now derive this conservative estimate. A simple way of achieving this is to rotate the distribution by $\frac{\pi}{4}$ counter-clockwise about the origin, and to integrate the resulting distribution projected onto the first axis (cf. Fig. 3). Denote the asymptotically normal distribution of $\sqrt{m}[HSIC_m^{XY}HSIC_m^{XZ}]^T$ as $\mathcal{N}(\mu,\Sigma)$. The distribution resulting from rotation and projection is

$$\mathcal{N}\left([Q\mu]_1, [Q\Sigma Q^T]_{11}\right), \tag{9}$$

where $Q = \frac{\sqrt{2}}{2}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$ is the rotation matrix by $\frac{\pi}{4}$ and

$$[Q\mu]_1 = \frac{\sqrt{2}}{2}\left(HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz})\right), \tag{10}$$

$$[Q\Sigma Q^T]_{11} = \frac{1}{2}(\sigma^2_{XY} + \sigma^2_{XZ} - 2\sigma_{XYXZ}). \tag{11}$$

Following the empirical distribution from Eq. (9), a test with statistic $HSIC^{XY}_m - HSIC^{XZ}_m$ has p-value

$$p \leq 1 - \mathbf{\Phi}\left(\frac{(HSIC^{XY}_m - HSIC^{XZ}_m)}{\sqrt{\sigma^2_{XY} + \sigma^2_{XZ} - 2\sigma_{XYXZ}}}\right), \tag{12}$$

where $\mathbf{\Phi}$ is the CDF of a standard normal distribution, and we have made the most conservative possible assumption that $HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz}) = 0$ under the null (the null also allows for the difference in population dependence measures to be negative).

To implement the test in practice, the variances of $\sigma^2_{XY}, \sigma^2_{XZ}$ and σ^2_{XYXZ} may be replaced by their empirical estimates. The test will still be consistent for a large enough sample size, since the estimates will be sufficiently well converged to ensure the test is calibrated. Eq. (8) is expensive to compute naïvely, because even computing the kernels h_{ijqr} and g_{ijqr} of the U-statistic itself is a non trivial task. Following (Song et al., 2012, Section 2.5), we first form a vector $\mathbf{h_{XY}}$ with entries corresponding to $\sum_{(j,q,r)\in i^m_3\setminus\{i\}} h_{ijqr}$, and a vector $\mathbf{h_{XZ}}$ with entries corresponding to $\sum_{(j,q,r)\in i^m_3\setminus\{i\}} g_{ijqr}$. Collecting terms in Eq. (4) related to kernel matrices $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$, $\mathbf{h_{XY}}$ can be written as

$$\begin{aligned}\mathbf{h_{XY}} = {} & (m-2)^2\left(\tilde{\mathbf{K}}\odot\tilde{\mathbf{L}}\right)\mathbb{1} - m(\tilde{\mathbf{K}}\mathbb{1})\odot(\tilde{\mathbf{L}}\mathbb{1}) \\ & + (m-2)\left((\mathrm{Tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}))\mathbb{1} - \tilde{\mathbf{K}}(\tilde{\mathbf{L}}\mathbb{1}) - \tilde{\mathbf{L}}(\tilde{\mathbf{K}}\mathbb{1})\right) \\ & + (\mathbb{1}^T\tilde{\mathbf{L}}\mathbb{1})\tilde{\mathbf{K}}\mathbb{1} + (\mathbb{1}^T\tilde{\mathbf{K}}\mathbb{1})\tilde{\mathbf{L}}\mathbb{1} - ((\mathbb{1}^T\tilde{\mathbf{K}})(\tilde{\mathbf{L}}\mathbb{1}))\mathbb{1}\end{aligned} \tag{13}$$

where $\odot$ denotes the Hadamard product. Then R_{XYXZ} in Eq. (8) can be computed as $R_{XYXZ} = (4m)^{-1}(m-1)^{-2}_3\mathbf{h_{XY}}^T\mathbf{h_{XZ}}$. Using the order of operations implied by the parentheses in Eq. (13), the computational cost of the cross covariance term is $\mathcal{O}(m^2)$. Combining this with the unbiased estimator of HSIC in Eq. (2) leads to a final computational complexity of $\mathcal{O}(m^2)$.

In addition to the asymptotic consistency result, we provide a finite sample bound on the deviation between the difference of two population HSIC statistics and the difference of two empirical HSIC estimates.

Theorem 5 (Generalization bound on the difference of empirical HSIC statistics)**.** *Assume that k, l, and d are bounded almost everywhere by 1, and are non-negative. Then for $m > 1$ and all $\delta > 0$ with probability at least $1-\delta$, for all p_{xyz}, the generalization bound on the difference of empirical HSIC statistics is*

$$\begin{aligned}&\left|\left\{HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz})\right\}\right. \\ &\qquad\left. - \left\{HSIC^{XY}_m - HSIC^{XZ}_m\right\}\right| \\ &\leq 2\left\{\sqrt{\frac{log(6/\delta)}{\alpha^2 m}} + \frac{C}{m}\right\}\end{aligned} \tag{14}$$

where $\alpha > 0.24$ and C are constants.

Proof. In Gretton et al., (2005) a finite sample bound is given for a single HSIC statistic. Eq. (14) is proved by using a union bound. □

Corollary 1. *$HSIC^{XY}_m - HSIC^{XZ}_m$ converges to the population statistic at rate $\mathcal{O}(\sqrt{m})$.*

3.2. A simple consistent test via uncorrelated HSICs

From the result in Eq. (5), a simple, consistent test of relative dependence can be constructed as follows: split the samples from P_x into two equal sized sets denoted by X' and X'', and drop the second half of the sample pairs with Y and the first half of the sample pairs with Z. We will denote the remaining samples as Y' and Z''. We can now estimate the joint distribution of $\sqrt{m}[HSIC^{X'Y'}_{m/2}, HSIC^{X''Z''}_{m/2}]^T$ as

$$\mathcal{N}\left(\begin{pmatrix}HSIC(\mathcal{F},\mathcal{G},P_{xy})\\ HSIC(\mathcal{F},\mathcal{H},P_{xz})\end{pmatrix}, \begin{pmatrix}\sigma^2_{X'Y'} & 0\\ 0 & \sigma^2_{X''Z''}\end{pmatrix}\right), \tag{15}$$

which we will write as $\mathcal{N}(\mu',\Sigma')$. Given this joint distribution, we need to determine the distribution over the half space defined by $HSIC(\mathcal{F},\mathcal{G},P_{xy}) < HSIC(\mathcal{F},\mathcal{H},P_{xz})$. As in the previous section, we achieve this by rotating the distribution by $\frac{\pi}{4}$ counter-clockwise about the origin, and integrating the resulting distribution projected onto the first axis (cf. Fig. 3). The resulting projection of the rotated distribution onto the primary axis is

$$\mathcal{N}\left([Q\mu']_1, [Q\Sigma'Q^T]_{11}\right) \tag{16}$$

where

$$[Q\mu']_1 = \frac{\sqrt{2}}{2}\left(HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz})\right), \tag{17}$$

$$[Q\Sigma'Q^T]_{11} = \frac{1}{2}(\sigma^2_{X'Y'} + \sigma^2_{X''Z''}). \tag{18}$$

From this empirically estimated distribution, it is straightforward to construct a consistent test (cf. Eq. (12)). The power of this test varies inversely with the variance of the distribution in Eq. (16).

3.3. The dependent test is more powerful

While discarding half the samples leads to a consistent test, we might expect some loss of power over the approach in Section 3.1, due to the increase in variance with lower sample size. In this section, we prove the Section 3.1 test is more powerful than that of Section 3.2, regardless of P_{xy} and P_{xz}.

We call the simple and consistent approach in Section 3.2, the *independent approach*, and the lower variance approach in Section 3.1, the *dependent approach*. The following theorem compares these approaches.

Theorem 6. *The asymptotic relative efficiency (ARE) of the independent approach relative to the dependent approach is always greater than 1.*

Remark 1. *The asymptotic relative efficiency (ARE) is defined in e.g. Serfling (1981, Chap.5, Section 1.15.4). If m_A and m_B are the sample sizes at which tests "perform equivalently" (i.e. have equal power), then the ratio $\frac{m_A}{m_B}$ represents the relative efficiency. When m_A and m_B tend to $+\infty$ and the ratio $\frac{m_A}{m_B} \to L$ (at equivalent performance), then the value L represents the asymptotic relative efficiency of procedure B relative to procedure A. This example is relevant to our case since we are comparing two test statistics with different asymptotically Normal distributions.*

The following lemma is used for the proof of Theorem 6.

Lemma 1. *(Lower Variance) The variance of the dependent test statistic is smaller than the variance of the independent test statistic.*

Proof. From the convergence of moments in the application of the central limit theorem (von Bahr, 1965), we have that $\sigma^2_{X'Y'} = 2\sigma^2_{XY}$. Then the variance summary in Eq. (11) is $\frac{1}{2}(\sigma^2_{XY} + \sigma^2_{XZ} - 2\sigma_{XYXZ})$ and the variance summary in Equation (18) is $\frac{1}{2}(2\sigma^2_{XY} + 2\sigma^2_{XZ})$ where in both cases the statistic is scaled by $\sqrt{m}$. We have that the variance of the independent test statistic is smaller than the variance of the dependent test statistic when

$$\frac{1}{2}(\sigma^2_{XY} + \sigma^2_{XZ} - 2\sigma_{XYXZ}) < \frac{1}{2}(2\sigma^2_{XY} + 2\sigma^2_{XZ})$$
$$\iff -2\sigma_{XYXZ} < \sigma^2_{XY} + \sigma^2_{XZ} \tag{19}$$

which is implied by the positive definiteness of Σ. □

Proof of Theorem 6. The Type II error probability of the independent test at level α is

$$\Phi\left[\Phi^{-1}(1-\alpha) - \frac{m^{-1/2}\big(HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz})\big)}{\sqrt{\sigma^2_{X'Y'} + \sigma^2_{X''Z''}}}\right], \tag{20}$$

where we again make the most conservative possible assumption that $HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz}) = 0$ under the null. The Type II error probability of the dependent test at level α is

$$\Phi\left[\Phi^{-1}(1-\alpha) - \frac{m^{-1/2}\big(HSIC(\mathcal{F},\mathcal{G},P_{xy}) - HSIC(\mathcal{F},\mathcal{H},P_{xz})\big)}{\sqrt{\sigma^2_{XY} + \sigma^2_{XZ} - 2\sigma_{XYXZ}}}\right] \tag{21}$$

where Φ is the CDF of the standard normal distribution. The numerator in Eq. (20) is the same as the numerator in Eq. (21), and the denominator in Eq. (21) is smaller due to Lemma 1. The lower variance dependent test therefore has higher ARE, i.e., for a sufficient sample size $m > \tau$ for some distribution dependent $\tau \in \mathbb{N}_+$, the dependent test will be more powerful than the independent test. □

4. Generalizing to more than two HSIC statistics

The generalization of the dependence test to more than three random variables follows from the earlier derivation 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 n domains with kernels $k_1, \ldots, k_n$ associated uniquely with respective reproducing kernel Hilbert spaces $\mathcal{F}_1, \ldots, \mathcal{F}_n$. We define a generalized statistical test, $\mathcal{T}_g(\mathcal{S}) \to \{0,1\}$ to test the null hypothesis $\mathcal{H}_0 : \sum_{(x,y)\in\{1,\ldots,n\}^2} v_{(x,y)} HSIC(\mathcal{F}_x,\mathcal{F}_y,P_{xy}) \leq 0$ versus the alternative hypothesis $\mathcal{H}_m : \sum_{(x,y)\in\{1,\ldots,n\}^2} v_{(x,y)} HSIC(\mathcal{F}_x,\mathcal{F}_y,P_{xy}) > 0$, where v is a vector of weights on each HSIC statistic. We may recover the test in the previous section by setting $v_{(1,2)} = +1$ $v_{(1,3)} = -1$ and $v_{(i,j)} = 0$ for all $(i,j) \in \{1,2,3\}^2 \setminus \{(1,2),(1,3)\}$.

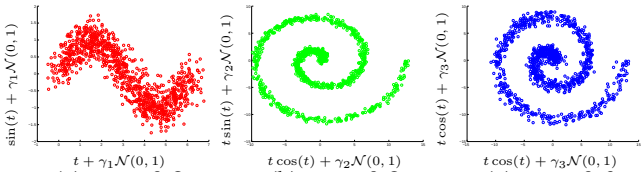
The derivation of the test follows the general strategy used in the previous section: we construct a rotation matrix so as to project the joint Gaussian distribution onto the first axis, and read the p-value from a standard normal table. To construct the rotation matrix, we simply need to rotate v such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

5. Experiments

We apply our estimates of statistical dependence to three challenging 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 relations between European languages. The last exper-

Algorithm 1 Successive rotation for generalized high-dimensional relative tests of dependency (cf. Section 4)

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



(a) $\gamma_1 = 0.3$ (b) $\gamma_2 = 0.3$ (c) $\gamma_3 = 0.6$

Figure 1. Illustration of a synthetic dataset sampled from the distribution in Eq. (22).

iment is a 3-block dataset which combines gene expression, comparative genomic hybridization, and a qualitative phenotype measured on a sample of Glioma patients.

5.1. Synthetic experiment

We constructed 3 distributions as defined in Eq. (22) and illustrated in Figure 1.

$$\text{Let } t \sim \mathcal{U}[(0, 2\pi)], \tag{22}$$
$$(a)\ x_1 \sim t + \gamma_1\mathcal{N}(0,1) \quad y_1 \sim \sin(t) + \gamma_1\mathcal{N}(0,1)$$
$$(b)\ x_2 \sim t\cos(t) + \gamma_2\mathcal{N}(0,1) \quad y_2 \sim t\sin(t) + \gamma_2\mathcal{N}(0,1)$$
$$(c)\ x_3 \sim t\cos(t) + \gamma_3\mathcal{N}(0,1) \quad y_3 \sim t\sin(t) + \gamma_3\mathcal{N}(0,1)$$

These distributions are specified so that we can control the relative degree of functional dependence between the variates by varying the relative size of noise scaling parameters γ_1, γ_2 and γ_3. The question is then whether the dependence between (a) and (b) is larger than the dependence between (a) and (c). In these experiments, we fixed $\gamma_1 = \gamma_2 = 0.3$, while we varied γ_3, and used a Gaussian kernel with bandwidth σ selected as the median pairwise distance between data points. This kernel is sufficient to obtain good performance, although others choices exist (Gretton et al., 2012).

Figure 2 shows the power of the dependent and the independent tests as we vary γ_3. It is clear from these results that the dependent test is far more powerful than the independent test over the great majority of γ_3 values considered. Figure 3 demonstrates that this superior test power arises due to the tighter and more concentrated distribution of the dependent statistic.

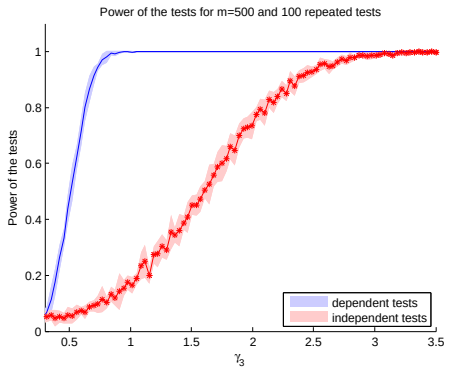


Figure 2. Power of the dependent and independent test as a function of γ_3 on the synthetic data described in Section 5.1. For values of $\gamma_3 > 0.3$ the distribution in Fig. 1(a) is closer to 1(b) than to 1(c). The problem becomes difficult as $\gamma_3 \to 0.3$. As predicted by theory, the dependent test is significantly more powerful over almost all values of γ_3 by a substantial margin.

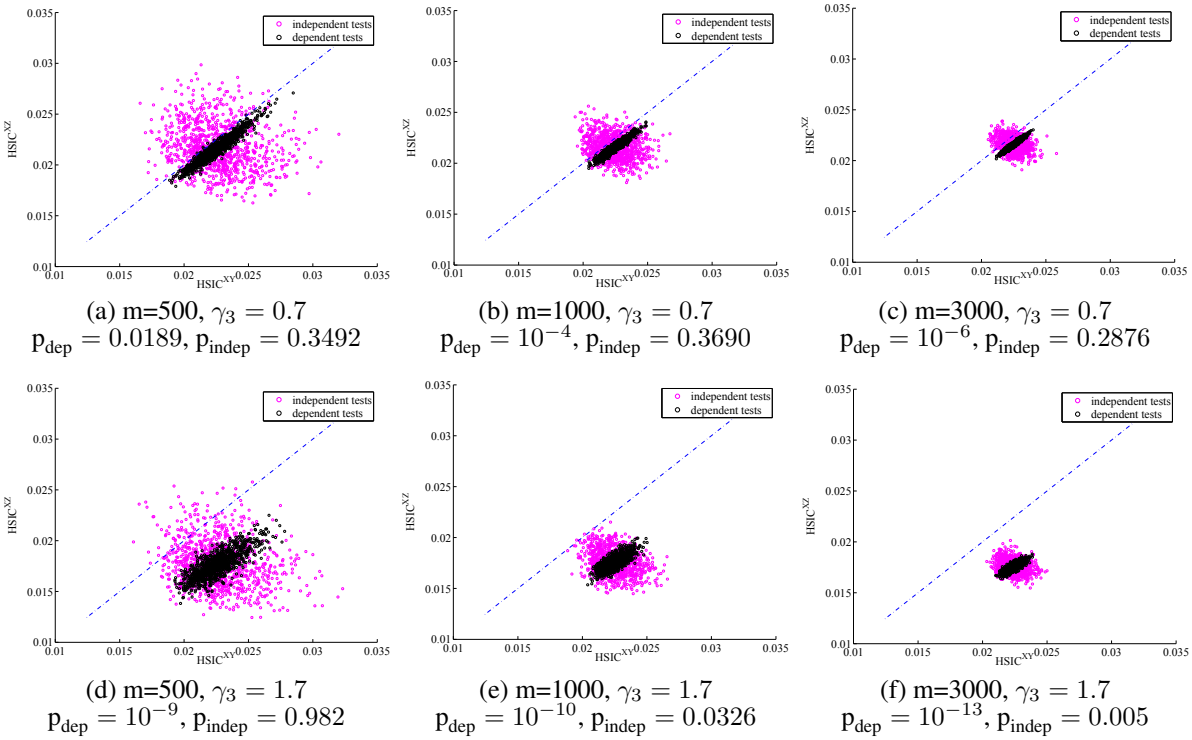
5.2. Multilingual data

In this section, we demonstrate dependence testing to predict the relative similarity of different languages. We use a real world dataset taken from the parallel European Parliament corpus (Koehn, 2005). We choose 3000 random documents in common written in: Finnish (fi), Italian (it), French (fr), Spanish (es), Portuguese (pt), English (en), Dutch (nl), German (de), Danish (da) and Swedish (sv). These languages can be broadly categorized into either the Romance, Germanic or Uralic groups (Gray & Atkinson, 2003). In this dataset, we considered each language as a random variable and each document as an observation.

Our first goal is 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 pre-processing, we removed stop-words (`http://www.nltk.org`) and performed stemming (`http://snowball.tartarus.org`). We applied the TF-IDF model as a feature representation and used a Gaussian kernel with the bandwidth σ set per language as the median pairwise distance between documents.

In Table 1, a selection of tests between language groups (Germanic, Romance, and Uralic) is given: all p-values strongly support that our relative dependence test finds the different language groups with very high significance.

Further, if we focus on the Romance family, our test enables one to answer more fine-grained questions about the relative similarity of languages within the same group. As before, we determine the ground truth similarities from the topology of the tree of European languages determined by the linguistics community (Gray & Atkinson, 2003; Bouckaert et al., 2012) as illustrated in Fig. 4 for the Romance



(a) m=500, $\gamma_3 = 0.7$
$p_{dep} = 0.0189$, $p_{indep} = 0.3492$

(b) m=1000, $\gamma_3 = 0.7$
$p_{dep} = 10^{-4}$, $p_{indep} = 0.3690$

(c) m=3000, $\gamma_3 = 0.7$
$p_{dep} = 10^{-6}$, $p_{indep} = 0.2876$

(d) m=500, $\gamma_3 = 1.7$
$p_{dep} = 10^{-9}$, $p_{indep} = 0.982$

(e) m=1000, $\gamma_3 = 1.7$
$p_{dep} = 10^{-10}$, $p_{indep} = 0.0326$

(f) m=3000, $\gamma_3 = 1.7$
$p_{dep} = 10^{-13}$, $p_{indep} = 0.005$

Figure 3. For the synthetic experiments described in Section 5.1, we plot empirical HSIC values for dependent and independent tests for 100 repeated draws with different sample sizes. Empirical p-values for each test show that the dependent distribution converges faster than the independent distribution even at low sample size, resulting in a more powerful statistical test.

Source	Target 1	Target 2	p-value
es	pt	fi	$\mathbf{0.0066}$
fr	it	da	$\mathbf{0.0418}$
it	es	fi	$\mathbf{0.0169}$
pt	es	da	$\mathbf{0.0173}$
de	nl	fi	$\mathbf{< 10^{-4}}$
nl	en	es	$\mathbf{< 10^{-4}}$
da	sv	fr	$\mathbf{< 10^{-6}}$
sv	en	it	$\mathbf{< 10^{-4}}$
en	de	es	$\mathbf{< 10^{-4}}$

Table 1. A selection of relative dependency tests between two pairs of HSIC statistics for the multilingual corpus data. Low p-values indicate a source is closer to target 1 than to target 2. In all cases, the test correctly identifies that languages within the same group are more strongly related than those in different groups.

group. We have run the test on all triplets from the corpus for which the topology of the tree specifies a correct ordering of the dependencies. In a fraction of a second (excluding kernel computation), we are able to recover certain features of the subtree of relationships between languages present in the Romance language group (Table 2). The test always indicates the correct relative similarity of languages when nearby languages (pt,es) are compared with those further away (ft,it), however errors are made when comparing triplets of languages for which the nearest common ancestor is more than one link removed.

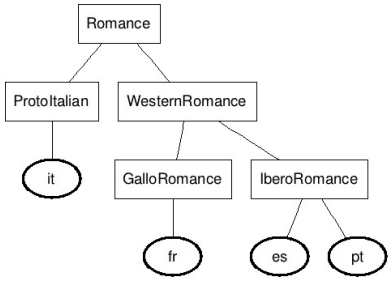


Figure 4. Partial tree of Romance languages adapted from (Gray & Atkinson, 2003).

Source	Target 1	Target 2	p-value
fr	es	it	$\mathbf{0.0157}$
fr	pt	it	0.1882
es	fr	it	0.2147
es	pt	it	$\mathbf{< 10^{-4}}$
es	pt	fr	$\mathbf{< 10^{-4}}$
pt	fr	it	0.7649
pt	es	it	$\mathbf{0.0011}$
pt	es	fr	$\mathbf{< 10^{-8}}$

Table 2. Relative dependency tests between Romance languages. The tests are ordered such that a low p-value corresponds with a confirmation of the topology of the tree of Romance languages determined by the linguistics community (Gray & Atkinson, 2003).

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 average dependence between languages in the same group is higher than the dependence between groups. The results of these tests are in Table 3. As before, our test is able to distinguish language groups with high significance.

Source	Targets	p-value
da	de sv fi	$< \mathbf{10^{-9}}$
da	sv en fr	$< \mathbf{10^{-9}}$
de	sv en it	$< \mathbf{10^{-5}}$
fr	it es sv	$< \mathbf{10^{-5}}$
es	fr pt nl	$\mathbf{0.0175}$

Table 3. Relative dependency test between four pairs of HSIC statistics for the multilingual corpus data. These tests show the ability of the relative dependence test to generalize to arbitrary numbers of HSIC statistics by constructing a rotation matrix using Algorithm 1. In all cases $v = [1\ 1\ -2]$.

5.3. Pediatric glioma data

Brain tumors are the most common solid tumors in children and have the highest mortality rate of all pediatric cancers. Despite advances in multimodality therapy, children with pediatric high-grade gliomas (pHGG) invariably have an overall survival of around 20% at 5 years. Depending on their location (e.g. brainstem, central nuclei, or supratentorial), pHGG present different characteristics in terms of radiological appearance, histology, and prognosis. The hypothesis is that pHGG have different genetic origins and oncogenic pathways depending on their location. Thus, the biological processes involved in the development of the tumor may be different from one location to another.

In order to evaluate such hypotheses, pre-treatment frozen tumor samples were obtained from 53 children with newly diagnosed pHGG from Necker Enfants Malades (Paris, France) from Puget et al, (2012). The 53 tumors are divided into 3 locations: supratentorial (HEMI), central nuclei (MIDL), and brain stem (DIPG). The final dataset is organized in 3 blocks of variables defined for the 53 tumors: X is a block of indicator variables describing the location category, the second data matrix Y provides the expression of 15 702 genes (GE). The third data matrix Z contains the imbalances of 1229 segments (CGH) of chromosomes.

For X, we use a linear kernel, which is characteristic for indicator variables, and for Y and Z, the kernel was chosen to be the Gaussian kernel with σ selected as the median of pairwise distances. The p-value of our relative dependency test is $< 10^{-5}$. This shows that the tumor location in the brain is more dependent on gene expression than on chromosomal imbalances. By contrast with Section 5.1, the independent test was also able to find the same order-

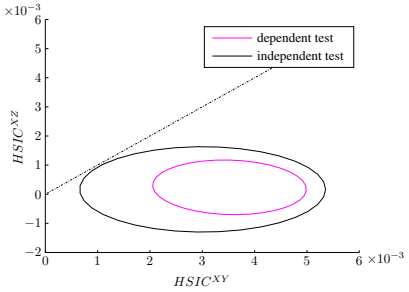


Figure 5. 2σ iso-curves of the Gaussian distributions estimated from the pediatric Glioma data. As before, the dependent test has a much lower variance than the independent test. The tests support the stronger dependence on the tumor location to gene expression than chromosomal imbalances.

ing of dependence, but with a p-value that is three orders of magnitude larger ($p = 0.005$). Figure 5 shows iso-curves of the Gaussian distributions estimated in the independent and dependent tests. The empirical relative dependency is consistent with findings in the medical literature, and provides additional statistical support for the importance of tumor location in Glioma (Gilbertson & Gutmann, 2007; Palm et al., 2009; Puget et al., 2012).

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 HSIC statistics, and empirically achieves p-values several orders of magnitude smaller. We have empirically demonstrated the test performance on synthetic data, where the degree of dependence could be controlled; on the challenging problem of identifying language groups from a multilingual corpus; and for finding the most important determinant of Glioma type. 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. The test is therefore scalable to the wide range of problem instances where non-parametric dependency tests are currently applied. 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 in this setting.

Acknowledgements

We thank Ioannis Antonoglou for helpful discussions. The first author is supported by a fellowship from Centrale-

Supélec. This work is partially funded by the European Commission through ERC Grant 259112 and FP7-MCCIG334380.

References

Arcones, M. A. and Gine, E. Limit theorems for U-processes. *The Annals of Probability*, pp. 1494–1542, 1993.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012.

Bring, J. A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1): 57–62, 1996.

Cortes, C., Mohri, M., and Rostamizadeh, A. Learning non-linear combinations of kernels. In *Neural Information Processing Systems*, 2009.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

Darlington, Richard B. Multiple regression in psychological research and practice. *Psychological bulletin*, 69(3): 161, 1968.

Dauxois, J. and Nkiet, G. M. Nonlinear canonical analysis and independence tests. *Annals of Statistics*, 26(4): 1254–1278, 1998.

Fukumizu, K., Bach, F. R., and Gretton, A. Statistical consistency of kernel canonical correlation analysis. *The Journal of Machine Learning Research*, 8:361–383, 2007.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, pp. 489–496. MIT Press, 2008.

Gilbertson, R. J. and Gutmann, D. H. Tumorigenesis in the brain: location, location, location. *Cancer research*, 67 (12):5579–5582, 2007.

Gray, R. D. and Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

Gretton, A. A simpler condition for consistency of a kernel independence test. arXiv:1501.06103, 2015.

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

Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pp. 63–77, 2005.

Gretton, A., Fukumizu, K., Teo, C.-H., Song, L., Schölkopf, B., and Smola, A. J. A kernel statistical test of independence. In *Neural Information Processing Systems*, pp. 585–592, 2008.

Gretton, A., Sejdinovic, D., Strathmann, H.and Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pp. 1205–1213, 2012.

Gunn, S. R. and Kandola, J. S. Structural modelling with sparse kernels. *Machine Learning*, 48(1):137–163, 2002.

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

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 2014.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pp. 79–86, 2005.

Palm, T., Figarella-Branger, D., Chapon, F., Lacroix, C., Gray, F., Scaravilli, F., Ellison, D. W., Salmon, I., Vikkula, M., and Godfraind, C. Expression profiling of ependymomas unravels localization and tumor grade-specific tumorigenesis. *Cancer*, 115(17):3955–3968, 2009.

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

Puget, S., Philippe, C., Bax, D., Job, B., Varlet, P., Junier, M. P., Andreiuolo, F., Carvalho, D., Reis, R., and Guerrini-Rousseau, L. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. *PloS one*, 7(2):e30313, 2012.

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

Sejdinovic, D., Gretton, A., and Bergsma, W. A kernel test for three-variable interactions. In *Neural Information Processing Systems*, 2013a.

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

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

Serfling, R. J. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley, 1981.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

Székely, G., Rizzo, M., and Bakirov, N. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

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

von Bahr, Bengt. On the convergence of moments in the central limit theorem. *The Annals of Mathematical Statistics*, 36(3):808–818, 06 1965.

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