Statistical Dependence Test with Hilbert-Schmidt Independence Criterion

Xinle Tian¹, *, Jiapeng He², Yueming Shi³

¹Columbian College of Arts and Sciences, The George Washington University, Washington, DC 20052, USA
²School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China
³The School of Arts and Sciences, Rutgers University–New Brunswick, New Brunswick, NJ 08901, USA

*Corresponding Author’s E-mail: tianxl1997@gwu.edu

Abstract. This paper studies an approach of nonlinear independence test between random variables. The method is Hilbert–Schmidt independence criterion (HSIC), one of the popular kernel methods for testing statistical dependence between random variables. There are several advantages of using HSIC method. It is easier compared with other kernel methods like the canonical correlation, because it requires no extra regularization terms for an appropriate finite sample behavior. Also, the learning rates of HSIC method is faster because there exists a well-defined population quantity in the estimate. The paper studies the detailed concepts and criterion associated with HSIC, its p-value and sensitivity maps. Results in the experiment about the wine data from UCI data set show the good performance of the approach.

1. Introduction

Estimating dependencies among various random variables is well-studied in statistics. Some traditional methods are widely used, such as Pearson’s correlation[1] and Spearman’s rank correlation[2]. These two approaches are instances of non-parametric statistics that most researchers are familiar with and are widely available in many statistical software packages[3]. However, the former can only be used in linear associations between random variables, while the latter shares the same drawback with the former: it can only be used among pairs of scalar variables. With these two methods, the multidimensional cases of dependence test are limited by repeating the testing process for all the pairwise variables and combining the results to get the dependence matrix.

To solve the problems in nonlinear dependence measures, researchers find various advantages with kernel dependence methods[4]. Several advantages are revealed such as stable properties with higher dimensional variables, small number of data samples setting required, the criteria relations between variables in higher-order and it is simple to execute the empirical estimates[5]. Under pairwise independence with multivariate components situations, some approaches are introduced. Szekely et al.[6], Szekely and Rizzo[7] proposed a test on the basis of distance covariance with fixed probabilities and infinite sample size. Also, Szekely and Rizzo[8] proposed a t-test on an account of a modified distance covariance method. Furthermore, Gretton et al.[9] introduced with a method corresponding to Hilbert–Schmidt independence criterion (HSIC). The HSIC is one of the kernel methods to evaluate the statistical dependence between various random variables[10] and the paper will be developed about this topic.
In this paper, some details of the HSIC methods are explored. In section 2, the paper introduces linear dependence, the HSIC estimate, its decision threshold, an appropriate kernel selection method, the concept of sensitivity maps and the setup background of experiment using the data of wines\[11\]. In section 3, the paper discusses the process and results of the experiment. Finally, the work makes a conclusion in section 4 with some commentaries and further potential research advice.

2. Experimental Section
This section presents an experiment with a criterion for the HSIC method. There are two main parts in the section. One is to define the estimator for the HSIC, its decision threshold and its corresponding sensitivity maps. The other part is to introduce the experiment setup for the wine data\[11\].

2.1. Linear Dependence
To make it clear, here some notations are defined. Let us define \( \mathcal{X} \subseteq \mathbb{R}^d \) and \( \mathcal{Y} \subseteq \mathbb{R}^d \) as two spaces that gather the sample pairs \((x, y)\) from the joint distribution \( P_{xy} \). Then, the covariance matrix is defined as

\[
C_{xy} = E_{xy}(xy^T) - E_x(x)E_y(y^T)
\]

where \( E_{xy} \) is the expectation of \( xy \) with respect to \( P_{xy} \), \( E_x \) is the expectation of the marginal distribution \( P_x \). Also, all the variables and quantities are assumed to be existed so that the following expressions can be defined. The covariance matrix contains the first-order dependences between variables \( x \) and \( y \). A statistic that is well studied to summary this matrix is called the Hilbert–Schmidt norm\[9\]. The squared norm is equal to the summation of its squared singular values of eigenvalues:

\[
\|C_{xy}\|_{HS}^2 = \sum_j \gamma_j^2.
\]

If there does not exist a first-order dependence between \( x \) and \( y \), then the value of that quantity is zero.

2.2. Kernel Dependence
Some nonlinear extension of the covariance notation was introduced in\[9\]. A mapping \( \phi: \mathcal{X} \to \mathcal{F} \) is defined and \( k_{xx}(x, x') = \langle \phi(x), \phi(x') \rangle \) is a positive kernel function that shows the inner product between features. The feature space \( \mathcal{F} \) shows a structure of a reproducing kernel Hilbert space (RKHS). Then another map \( \psi: \mathcal{Y} \to \mathcal{G} \) with the corresponding positive definite kernel function \( k_{yy}(y, y') = \langle \psi(y), \psi(y') \rangle \) can be defined, which shares a similar process with the mapping before. Then, there exists a cross-covariance operator between the two maps, which shares the similarities with the covariance matrix in (1). Then the cross-covariance operator can be a linear operator \( C_{xy}: \mathcal{G} \to \mathcal{F} \) such that

\[
C_{xy} = E_{xy}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)]
\]

where \( \otimes \) is a tensor product between two vector spaces, \( \mu_x = E_x[\phi(x)] \) and \( \mu_y = E_y[\psi(y)] \). Interest readers can find some detailed research proposed by Baker\[12\] and Fukumizu et al.\[13\]. Then a squared norm of the cross-covariance operator \( \|C_{xy}\|_{HS}^2 \) is named as the HSIC and can be written as

\[
HSIC(F, G, P_{xy}) = \|C_{xy}\|_{HS}^2 = E_{xx'}E_{yy'}[k_x(x, x')k_y(y, y')] + E_{xx'}[k_x(x, x')E_{yy'}[k_y(y, y')]] - 2E_{xy}[E_x[k_x(x, x')]E_y[k_y(y, y')]]
\]

In the equation, \( E_{xx'}E_{yy'} \) is the expectation of both \((x, y) \sim P_{xy} \) and \((x', y') \sim P_{xy} \) where \((x', y') \) is another pair of variables drawn independently under the same condition. More details are studied by Gretton et al.\[9\].

Given sample datasets, \( X \in \mathbb{R}^{m \times d_x} \) and \( Y \in \mathbb{R}^{m \times d_y} \) with \( m \) pairs data sets \( Z = \{(x_1, y_1), ..., (x_m, y_m)\} \) drawn from \( P_{xy} \). Then the empirical estimator of HSIC is
\[ HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) = \frac{1}{m^2} \text{Tr}(K_xHK_yH) \]  \hspace{1cm} (4)

where \( \text{Tr} \) is the trace of the matrix, \( K_x \) and \( K_y \) represent the corresponding kernel matrices for \( x \) and \( y \) with \( (k_x)_{i,j} = k_x(x_i, x_j) \) and \( (k_y)_{i,j} = k_y(y_i, y_j) \) and \( H = I - \frac{1}{n} \mathbb{1} \mathbb{1}^T \) centers the \( x \) and \( y \) features in \( \mathcal{F} \) and \( \mathcal{G} \). Here, we define the size of the empirical HSIC with size \( m \times m \).

### 2.3. Decision Threshold

The next important issue is to set the decision threshold \( \theta \) for the kernel-based measures of independence test. Song et al.\cite{14} proposed some statistical independence tests based on empirical HSIC estimators (3). The test focuses on a test between a null hypothesis \( H_0: P_{xy} = P_xP_y \) and an alternative hypothesis \( H_1: P_{xy} \neq P_xP_y \). This hypothesis decision can be completed through the comparison between a test statistic and a threshold. Within all the definitions of the threshold for the HSIC estimate, one of the approaches is through approximation and define the null distribution as a two-parameter gamma distribution so that the estimator becomes:

\[ HSIC \sim \frac{x^{\alpha-1}e^{-x/b}}{mb^{\alpha}f(a)} \]  \hspace{1cm} (5)

where \( \alpha = E[HSIC]^2/V[HSIC] \) and \( b = V[HSIC]/E[HSIC] \). Then, the threshold \( \theta \) can be calculated by an inverse cumulative density function with \( 1 - \alpha \), where \( \alpha \) is a defined significance level. Usually \( \alpha \) is taken to be 0.05 so a 95% confidence interval about the true value can be obtained. The two random variables are then considered dependent if \( HSIC \geq \theta \), in the case the null hypothesis is rejected and be independent otherwise. An alternative method is through the computation of the HSIC's p-value directly from the corresponding HSIC estimate and cumulative density function using a permutation test to verify the dependency.

### 2.4. Kernel Selection

The method contains two kernel functions: the \( k_x \) and \( k_y \). About the data, common kernels are usually used, such as radial basis functions or polynomial functions. Here the radial basis function expressed as \( k(x_i, x_j) = \exp(-||x_i - x_j||^2/2\sigma^2), \sigma \in \mathbb{R}^+ \). In the dependence estimation, if \( \mathcal{F} \) and \( \mathcal{G} \) are RKHSs with its corresponding universal kernels \( k_x \) and \( k_y \), which is studied by Steinwart\cite{15}. Then, \( HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) = 0 \) if \( x \) and \( y \) are independent. Therefore, radial basis functions or Laplacian kernels are universal and polynomial kernels may be only preferred in some cases.

Another approach to select the kernel functions to maximize the power of the test, i.e. to maximize the probability of rejecting the null hypothesis \( H_0: P_{xy} = P_xP_y \) when \( x \) and \( y \) are actually dependent. One method is to use the test statistic \( \frac{1}{m^2} \text{Tr}(K_xHK_yH) \) as a surrogate for power. However, the problem in this case is that using the surrogate to do the test is invalid because it also depends on the data. So, an alternative method is to pick half of the data \( \langle (x_i, y_i) \rangle_{i=1}^{m/2} \) and use them as the kernel function to do the test on the rest half of the data \( \langle (x_i, y_i) \rangle_{i=m/2}^m \).

### 2.5. Sensitivity Maps

The learned relations for any kernel method are related to the kernel feature mapping. The study of this part can help researchers set a visualized interpretation of dependence estimation. The definition about sensitivity maps was proposed by Kjems et al.\cite{16} and Camps-Valls\cite{5}. The concept of sensitivity maps is related to leverage and influential points. A function \( f(Z): \mathbb{R}^d \rightarrow \mathbb{R} \) is defined with parametrization of \( Z = [z_1, ..., z_d] \). Then the sensitivity map of the \( j^{th} \) feature \( z_j \) is equal to the squared derivative of \( f(Z) \) with respect to its arguments, which is showed below:

\[ s_j = \int_Z \left( \frac{\partial f(Z)}{\partial z_j} \right)^2 p(Z) \text{d}Z \]  \hspace{1cm} (6)
where \( p(Z) \) is a probability density function of input \( Z \in \mathbb{Z} \). Therefore, the goal of its sensitivity map is to calculate the variation of \( f(Z) \) in the \( j \)th direction of the space of its input. The derivative parts are squared so as to avoid the cancellation of possible terms because of the signs. Because of this, some other transformations, such as taking absolute values can also be used. The empirical sensitivity map is calculated by using a summation over \( m \) samples instead of using the integral sign in (5)

\[
s_j \approx \frac{1}{m} \sum_{i=1}^{m} \left( \frac{\partial f(Z)}{\partial z_j} \right)_{z_j}^2
\]

which can be represented by putting in a vector to get the definition of a sensitivity vector \( s := [s_1, ..., s_d] \).

### 2.6. HSIC Sensitivity Maps

The previous section is the general definition of sensitivity maps. However, the defined sensitivity map is constrained in some areas. First, it cannot be applied directly to some general functions, which are dependent on matrices. Second, it estimates the feature relevant \( s_j \), so it ignores the individual samples. The former definition limits the HSIC since it is based on matrices of data. In the following contents, the HSIC is derived with respect to its input samples and features to fix the disadvantages.

Let \( f := HSIC \) and \( \frac{\partial f(Z)}{\partial z_j} = \frac{\partial f(Z)}{\partial z_{ij}} \). So, the pairwise HSIC is composed by \( Z \sim X, Y | \sigma \) with a matrix parametrized to \( z \sim x, y | \sigma \). HSIC is derived with respect to its input data matrix entries \( X_{ij} \) and \( Y_{ij} \) in order to get the sensitivity maps of HSIC. Meanwhile, the first-order derivatives and chain rule are used on the results for \( X_{ij} \)

\[
S^x_{ij} = \frac{\partial HSIC}{\partial X_{ij}} = -\frac{2}{\sigma^2 m^2} Tr \left( HK_x H \left( K_x \circ M \right) \right)
\]

where the entries in \( M \) is \( M_{ik} = X_{ij} - X_{kj} \) \((1 \leq i, k \leq n)\), and \( \circ \) is the Hadamard product between matrices. With the similar process, the corresponding expression for \( Y_{ij} \) can be obtained

\[
S^y_{ij} = \frac{\partial HSIC}{\partial Y_{ij}} = -\frac{2}{\sigma^2 m^2} Tr \left( HK_y H \left( K_y \circ N \right) \right)
\]

with entries \( N_{ik} = Y_{ij} - Y_{kj} \) in \( N \).

Noted that though each sensitivity map can be independently used, the result is in a vector form. So, the result of the sensitivity map should be treated jointly. Therefore, the general sensitivity map for all the features and samples is \( S := [S^x, S^y] \in \mathbb{R}^{m \times (d_x + d_y)} \). The empirical sensitivity maps with \( S \) for both features and samples can be computed by integration in the corresponding domains, where each corresponding empirical estimates are \( s_i \approx \frac{1}{m} \sum_{j=1}^{d_x} n^2_j \) and \( s_j \approx \frac{1}{m} \sum_{i=1}^{d_y} S^2_{ij} \). This shows the information about the directions that have the most influence on the estimate and we can get a quantitively geometric evaluation on the test result.

### 2.7. Proposed Criterion

This is a study of an alternative criterion for dependence based on HSIC’s sensitivity in two directions: forward and backward:

\[
\overline{C}_f := (S^x_b + S^y_f) - (S^x_f + S^y_b)
\]

where \( f \) represents the forward and \( b \) is backward direction. The superscripts represent the sensitivities of corresponding observations \( x \) and \( y \), and corresponding residuals, \( rb \) and \( rf \).

### 2.8. Experimental Setup

HSIC method is used to analyze the dependence of 13 attributes on a set of red wine data[11]. The data of red wine is divided into three groups A, B, C in total and the data of each group is standardized.
measuring dependence, eight kernel functions are used for the dependence test. Two kernel functions $k$ and $l$ are selected to calculate for the two selected attributes $i$ and $j$, and finally get the result of the calculation formula defined in equation (3)

$$R(i,j) = \text{Tr}(\tilde{K}\tilde{L})$$  \hspace{1cm} (11)

where $\tilde{K} = HKH$, $K$ is the kernel matrix of function $k$, and $H = I - \frac{1}{n} \mathbb{I}\mathbb{I}^T$ with $n$ is the size of the data set. Since the 13 attributes are in the same space, any two of the attributes or kernel functions can be selected arbitrarily, and the correspondence between the two attributes and the kernel function would be changed with the selection. After rearranging the correspondence between the two attributes to obtain a new set of data that is theoretically independent. Through fitting and distribution test, the distribution and the result corresponds, the two attributes’ $p$-value is

$$P = F(R(i,j))$$  \hspace{1cm} (12)

where $F$ is the distribution function corresponding to the corresponding distribution of results. Through the combination of different kernel functions $k$ and $l$, we finally obtained the independence test results of 13 attributes.

3. Result and Discussion

3.1. Experiment Results and Problems

The paper measured the dependence of 13 attributes for each group. In what follows, some notations are defined to represent the 13 attributes respectively:

- A==Alcohol;
- B==Malic acid;
- C==Ash;
- D==Alkalinity of ash;
- E==Magnesium;
- F==Total phenols;
- G==Flavonoids;
- H==Nonflavonoid phenols;
- I==Proanthocyanins;
- J==Color intensity;
- K==Hue;
- L==OD280/OD315 of diluted wines;
- M==Proline;

Table 1. Shows the P values between the 13 attributes in group A:

|    | A       | B       | C       | D       | E       | F       | G       | H       | I       | J       | K       | L       | M       |
|----|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| A  | 0       | 0.003324| 0.001144| 0.051751| 0.169366| 3.84E-09| 4.65E-08| 0.572028| 0.00697 | 6.03E-05| 0.016863| 0.914677| 6.09E-07|
| B  | 0.00133 | 0.001144| 0.00362  | 0.37E-02| 0.001296| 0.024253| 0.082662| 0.00112 | 2.08E-05| 0.640654| 0.032473| 0.012499| 0.061914|
| C  | 3.1E-05 | 1.9E-05  | 0       | 4.72E-05| 0.000251| 0.097884| 0.061131| 0.082662| 0.012499| 0.012499| 0.061914| 0.012499| 0.061914|
| D  | 0.034421| 0.000661| 3.37E-13| 0       | 0.000029| 0.099568| 0.796912| 0.280606| 0.669711| 0.33046 | 0.231382| 0.057133| 0.015274|
| E  | 0.084062| 3.2E-06  | 6.49E-06| 0       | 0       | 0.044E-06| 0.016587| 0.639403| 0.108567| 0.731678| 0.000779| 0.668516|
| F  | 2.18E-07| 0.000565| 8.53E-05| 0.143379| 0.004E-06| 0.56E-15| 0.061587| 0.639403| 0.108567| 0.731678| 0.000779| 0.668516|
| G  | 1.84E-07| 0.044141| 0.065322| 0.891243| 0.010389| 2.1E-15  | 0.053898| 2.1E-09 | 3.96E-13| 0.270235| 0.587541| 0.958E-05|
| H  | 0.728938| 0.068799| 0.034141| 0.173724| 0.693676| 0.650025| 0.436805| 0.264761| 0.069592| 0.001188| 0.000181| 0.291303|
| I  | 0.023191| 0.001155| 0.712541| 0.693101| 0.097539| 0.307E-07| 0.768E-07| 0.012691| 6.59E-10 | 0.105485| 0.236984| 0.023984|
| J  | 3.09E-06| 0.000978| 0.022234| 0.577014| 3.73E-05| 3.76E-14 | 2.4E-14  | 0.012068| 1.58E-07 | 0.02086 | 0.796894| 3.1E-11 |
| K  | 0.016518| 0.493897| 0.067042| 0.76307 | 0.725877| 0.574255| 0.359722| 0.002681| 0.071777| 0.014326| 0.014228| 1.35E-06|
| L  | 0.932461| 4.7E-05 | 0.127836| 0.050309| 0.001594| 0.173546| 0.721067| 0.000417| 0.237312| 0.799824| 0.006407| 0.594194|
| M  | 1.26E-05| 0.072306| 0.002794| 0.01261 | 0.757381| 4.61E-06| 7.06E-06 | 0.343346| 0.018096| 6.51E-12| 1.4E-05 | 0.591846|

Table 1 shows the P values between the 13 attributes in group A.
Table 2. Shows the P values between the 13 attributes in group B:

| A       | B       | C       | D       | E       | F       | G       | H       | I       | J       | K       | L       | M       |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| A       | 0.000235 | 0.09E-07 | 2.23E-12 | 0.000795 | 0.099558 | 0.230941 | 0.389341 | 0.666741 | 0.00361 | 0.254131 | 5.21E-07 | 5.31172 |
| B       | 0.000285 | 0.58E-08 | 5.33E-06 | 4.61E-09 | 0.01631 | 0.0357-06 | 0.099341 | 0.666741 | 0.00361 | 0.254131 | 5.21E-07 | 5.31172 |
| C       | 1.52E-09 | 0.001119 | 0.002795 | 0.027995 | 0.002795 | 0.027995 | 0.002795 | 0.002795 | 0.002795 | 0.002795 | 0.002795 | 0.002795 |
| D       | 6.76E-14 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 |
| E       | 0.000371 | 2.48E-07 | 7.23E-07 | 3.8E-07  | 0.249E-06 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 |

Table 2 shows the P values between the 13 attributes in group B.

Table 3. Shows the P values between the 13 attributes in group C:

| A       | B       | C       | D       | E       | F       | G       | H       | I       | J       | K       | L       | M       |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| A       | 0.000235 | 0.09E-07 | 2.23E-12 | 0.000795 | 0.099558 | 0.230941 | 0.389341 | 0.666741 | 0.00361 | 0.254131 | 5.21E-07 | 5.31172 |
| B       | 0.000285 | 0.58E-08 | 5.33E-06 | 4.61E-09 | 0.01631 | 0.0357-06 | 0.099341 | 0.666741 | 0.00361 | 0.254131 | 5.21E-07 | 5.31172 |
| C       | 1.52E-09 | 0.001119 | 0.002795 | 0.027995 | 0.002795 | 0.027995 | 0.002795 | 0.002795 | 0.002795 | 0.002795 | 0.002795 | 0.002795 |
| D       | 6.76E-14 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 | 0.000119 |
| E       | 0.000371 | 2.48E-07 | 7.23E-07 | 3.8E-07  | 0.249E-06 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 | 0.002615 |

In theory, different kernel functions are selected in the HSIC method in order to measure the attributes in different spaces and the selection of kernel functions should not change the test results. Theoretically, the values of the symmetrical positions in the above table should be the same, which should be reasonable since they indicate the independence of the same pair of attributes. However, the tables shown above are not symmetrical. In order to study whether the combination of kernel functions will increase or decrease the difference between p-value in the symmetrical position, we divide the combination of kernel functions into three groups:

1) Function combinations are two kernel functions of the same kind. For example, if we choose two Gaussian kernel functions:

\[ k(x, x') = \exp \left( -\frac{1}{2\gamma^2} \|x - x'\|^2 \right), \gamma > 0 \]

the parameters of the two Gaussian kernel functions are not the same. The value of p in the symmetrical position will not be much different.

2) Function combinations are two different kinds of kernel functions, but both kernel functions are bounded. For example, if we choose the Gaussian kernel function and rational quadratic function:

\[ k(x, x') = (1 + \frac{\|x - x'\|^2}{\alpha \gamma^2})^{-\alpha}, \alpha, \gamma > 0 \]

There is still no significant difference of the p-value in the symmetrical position.

3) At least one unbounded function was selected in the combination of kernel function, there is no significant difference of the p-value in the symmetrical position. Data normalization is very important for unbounded functions, so we experienced on the data without normalization. However, there is still no significant difference of the p-value in the symmetrical position.
3.2. Analysis and Discussion of Problems

Since the calculation of the p-value heavily depends on the random sorting, different random ordering will indeed result in different distributions corresponding to the results. In fact, although the p-values of the symmetrical positions in the table are not the same to each other, they are in the same order of magnitude. At the same time, it is found in this experiment that the choices of kernel functions will not significantly affect the results, so the experimental process should be simplified, such as simplifying the amount of calculation. Although data normalization is very important for an experiment, it is found in the process that removing data normalization in the HSIC method would not have much impact. It can also be seen from the experiment that some p-values are near 0.05 or 0.01, which are the commonly used significance indexes when we make hypothesis tests. Therefore, in HSIC applications, some two attributes may get different results if the same experiment is repeated. It is because the selected significance index is too close to the two attributes’ p-value.

4. Conclusion

This report has presented details about one of the kernel methods for measuring nonlinear dependence between various variables. The result studied the HSIC independence criterion based on regression and dependence estimation and mentioned about the sensitivity maps of the estimates. The method was tested in the actual wine data problems and the performance of the method was reviewed with the interpretations. The empirical estimate was relatively easy to calculate and was useful for data inference. Due to the limitation of time and resources, the sensitivity maps part could have some further application. Therefore, some potential research about this would be subjects in some further studies.

References

[1] K. Pearson. Notes on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London, 1895, 58, 240-242.
[2] A. Szczepańska. Research design and statistical analysis. International Statistical Review, 2011, 79(3), 491–492.
[3] O. L. O. Astivia and B. D. Zumbo. Population models and simulation methods: The case of the Spearman rank correlation. British Journal of Mathematical and Statistical Psychology, 2017, 70, 347-367.
[4] A. Gretton, R. Herbrich, and A. Hyvarinen. Kernel methods for measuring independence. Journal of Machine Learning Research, 2005a, 6, 2075–2129.
[5] A. Pérez-Suay and G. Camps-Valls. Sensitivity maps of the Hilbert–Schmidt independence criterion. Applied Soft Computing Journal, 2018.
[6] G. J. Szekely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. The annals of statistics, 2007, 35(6), 2769–2794.
[7] G. J. Szekely and M. L. Rizzo. Brownian distance covariance. The annals of applied statistics, 2009, 3(4), 1236–1265.
[8] G. J. Szekely and M. L. Rizzo. The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 2013, 117, 193–213.
[9] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with hilbert-schmidt norms, 2005, 16, 63–78.
[10] G. Camps-Valls, J. Mooij, and B. Scholkopf. Remote sensing feature selection by kernel dependence measures, 2010, 7 (3), 587–591.
[11] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. University of California. www http://www.ics.uci.edu/~mlearn/MLRepository, 1998.
[12] C. Baker. Joint measures and cross-covariance operators, 1973, 186, 273–289.
[13] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, 2004, 5, 73–99.
[14] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. Journal of Machine Learning Research, 2012, 13, 1393–1434.
[15] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. 2001, 2, 67–93.

[16] U. Kjems, L. K. Hansen, J. Anderson, S. Frutiger, S. Muley, J. Sidtis, D. Rottenberg, and S. C. Strother. The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves. NeuroImage, 2002, 15 (4), 772–786.