LETTER

Least-Squares Independence Test

Masashi SUGIYAMA[†,††a)], ***Member and*** **Taiji SUZUKI**[†††b)], ***Nonmember***

SUMMARY Identifying the statistical independence of random variables is one of the important tasks in statistical data analysis. In this paper, we propose a novel non-parametric independence test based on a least-squares density ratio estimator. Our method, called *least-squares independence test* (LSIT), is distribution-free, and thus it is more flexible than parametric approaches. Furthermore, it is equipped with a model selection procedure based on cross-validation. This is a significant advantage over existing non-parametric approaches which often require manual parameter tuning. The usefulness of the proposed method is shown through numerical experiments.

key words: *independence test, density ratio estimation, unconstrained least-squares importance fitting, squared-loss mutual information*

1. Introduction

Identifying the statistical independence of random variables is one of the fundamental tasks in statistical data analysis. Independence tests can be used for various purposes such as feature selection [9] and causal inference [10].

A traditional independence measure is the *Pearson correlation coefficient*, which can be used for detecting linear dependency. Thus, it is useful for Gaussian data, although the Gaussian assumption is rarely fulfilled in practice. Recently, kernel-based independence measures have been studied in order to overcome the weakness of the Pearson correlation coefficient. The *Hilbert-Schmidt independence criterion* (HSIC) [4] utilizes *cross-covariance operators* on *universal* reproducing kernel Hilbert spaces (RKHSs) [7], which is an infinite-dimensional generalization of covariance matrices. HSIC allows efficient detection of non-linear dependency by making use of the reproducing property of RKHSs [1]. However, HSIC has a critical weakness that its performance depends on the choice of RKHSs and there is no theoretically justified way to determine the RKHS properly. In practice, using the Gaussian RKHS with width set to the median distance between samples is a popular heuristic [4].

Another popular independence criterion would be *mutual information* [2]. In this paper, we consider a squared-loss variant of mutual information and use its estimator *least-squares mutual information* (LSMI) [9] for independence test. LSMI is also distribution-free as HSIC, but it is equipped with a natural model selection procedure based on cross-validation, which is an advantage over HSIC. Through experiments, we show the usefulness of the LSMI-based independence test called *least-squares independence test* (LSIT).

Manuscript received January 6, 2011.

[†]The author is with Tokyo Institute of Technology, Tokyo, 152–8552 Japan.

[††]The author is with PRESTO, Japan Science and Technology Agency, Tokyo, 152–8552 Japan.

[†††]The author is with The University of Tokyo, Tokyo, 113–8656 Japan.

a) E-mail: sugi@cs.titech.ac.jp

b) E-mail: s-taiji@stat.t.u-tokyo.ac.jp

DOI: 10.1587/transinf.E94.D.1333

2. Least-Squares Independence Test

In this section, we propose a novel non-parametric independence test.

2.1 Formulation

Let $\boldsymbol{x}$ ($\in \mathcal{X} \subset \mathbb{R}^{d_{\mathrm{x}}}$) be an input feature and $\boldsymbol{y}$ ($\in \mathcal{Y} \subset \mathbb{R}^{d_{\mathrm{y}}}$) be an output feature, which follow a joint probability distribution with density $p(\boldsymbol{x}, \boldsymbol{y})$. Suppose we are given a set of i.i.d. paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. Our goal is to test whether $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent or not, based on the samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$.

For independence test, we employ the *squared-loss mutual information* (SMI) defined as follows:

$$\mathrm{SMI} := \frac{1}{2} \iint p(\boldsymbol{x})p(\boldsymbol{y}) \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \qquad (1)$$

where $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. SMI is the *Pearson divergence* [6] from the joint density $p(\boldsymbol{x}, \boldsymbol{y})$ to the product of marginals $p(\boldsymbol{x})p(\boldsymbol{y})$, and SMI is zero if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. Hence, SMI can be used for detecting the statistical independence of random variables.

SMI includes unknown probability densities $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$, and thus it cannot be directly computed. A naive approach is to estimate the densities $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$, and plug the estimated densities in Eq. (1). However, since density estimation is known to be a hard task and division by estimated densities can magnify the estimation error, we consider estimating the following *density ratio* function directly:

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}. \qquad (2)$$

Once an estimator of the density ratio $\widehat{r}(\boldsymbol{x}, \boldsymbol{y})$ is obtained, SMI can be approximated using samples as follows:

$$\widehat{\mathrm{SMI}} := \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i,\boldsymbol{y}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n}\widehat{r}(\boldsymbol{x}_i,\boldsymbol{y}_j)^2 - \frac{1}{2}. \tag{3}$$

2.2 Least-Squares Mutual Information

Here we explain the *least-squares mutual information* (LSMI) method [9], which directly learns $r(\boldsymbol{x},\boldsymbol{y})$ from data samples without going through density estimation of $p(\boldsymbol{x},\boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$.

Let us approximate the density ratio (2) using the following model:

$$\widehat{r}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\ell=1}^{b}\alpha_\ell\psi_\ell(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\alpha}^\top\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{y}),$$

where $\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{y}) : \mathbb{R}^{d_x}\times\mathbb{R}^{d_y} \to \mathbb{R}^b$ is a non-negative basis function vector, $\boldsymbol{\alpha}$ ($\in \mathbb{R}^b$) is a parameter vector, and $^\top$ denotes the transpose.

We determine the parameter $\boldsymbol{\alpha}$ so that the following squared-error J_0 is minimized:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &:= \frac{1}{2}\iint\left(\widehat{r}(\boldsymbol{x},\boldsymbol{y}) - r(\boldsymbol{x},\boldsymbol{y})\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \frac{1}{2}\iint\widehat{r}(\boldsymbol{x},\boldsymbol{y})^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &\quad - \iint\widehat{r}(\boldsymbol{x},\boldsymbol{y})p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} + \mathrm{Const}. \end{aligned}$$

Let us denote the first two terms by J. Since J contains the expectations over unknown densities $p(\boldsymbol{x},\boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$, we approximate the expectations by empirical averages. By including an ℓ_2-regularizer, the LSMI optimization problem is formulated as follows.

$$\widehat{\boldsymbol{\alpha}} := \underset{\boldsymbol{\alpha}\in\mathbb{R}^b}{\mathrm{argmin}}\left[\frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\widehat{\boldsymbol{h}} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\right], \tag{4}$$

where λ (≥ 0) is the regularization parameter that controls the strength of regularization, and

$$\widehat{\boldsymbol{H}} := \frac{1}{n^2}\sum_{i,j=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i,\boldsymbol{y}_j)\boldsymbol{\psi}(\boldsymbol{x}_i,\boldsymbol{y}_j)^\top, \quad \widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i,\boldsymbol{y}_i).$$

The solution $\widehat{\boldsymbol{\alpha}}$ can be analytically obtained as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}}, \tag{5}$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix. Finally, the density ratio estimator $\widehat{r}(\boldsymbol{x})$ is given by

$$\widehat{r}(\boldsymbol{x}) := \widehat{\boldsymbol{\alpha}}^\top\boldsymbol{\psi}(\boldsymbol{x}).$$

Once a density ratio estimator $\widehat{r}(\boldsymbol{x},\boldsymbol{y})$ is obtained, SMI can be approximated by Eq. (3).

Thanks to the analytic-form expression, LSMI is computationally very efficient. Furthermore, the above least-squares density ratio estimator was shown to possess the optimal non-parametric convergence rate and optimal numerical stability [5], [8].

2.3 Basis Function Choice and Model Selection

Given that both $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ are normalized so that every element of $\boldsymbol{x}$ and $\boldsymbol{y}$ has unit variance, we use the following basis functions:

$$\psi_\ell(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_\ell\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{y}_\ell\|^2}{2\sigma^2}\right) & \text{for regression,} \\ \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_\ell\|^2}{2\sigma^2}\right)\delta(\boldsymbol{y}=\boldsymbol{y}_\ell) & \text{for classification,} \end{cases}$$

where $\delta(c) = 1$ if the condition c is true; otherwise $\delta(c) = 0$. We may also include a constant basis function $\phi_0(\boldsymbol{x},\boldsymbol{y}) = 1$ to the above kernel basis functions.

The practical performance of LSMI depends on the choice of the kernel width σ and the regularization parameter λ. Model selection of LSMI is possible based on *cross-validation* with respect to the criterion J. More specifically, the sample set $\mathcal{Z} = \{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$ is divided into M disjoint sets $\{\mathcal{Z}_m\}_{m=1}^M$. Then an LSMI solution $\widehat{r}_m(\boldsymbol{x})$ is obtained using $\mathcal{Z}\backslash\mathcal{Z}_m$ (i.e., all samples without $\mathcal{Z}_m$), and its J-score for the hold-out samples $\mathcal{Z}_m$ is computed as

$$\widehat{J}_m^{\mathrm{CV}} := \frac{1}{2|\mathcal{Z}_m|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{Z}_m}\widehat{r}_m(\boldsymbol{x},\boldsymbol{y})^2 - \frac{1}{|\mathcal{Z}_m|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{Z}_m}\widehat{r}_m(\boldsymbol{x},\boldsymbol{y}),$$

where $|\mathcal{Z}|$ denotes the number of elements in the set $\mathcal{Z}$. This procedure is repeated for $m = 1,\ldots,M$, and the average score $\widehat{J}^{\mathrm{CV}} := \frac{1}{M}\sum_{m=1}^M\widehat{J}_m^{\mathrm{CV}}$ is computed. Finally, the model (the kernel width σ and the regularization parameter λ in the current setup) that minimizes $\widehat{J}^{\mathrm{CV}}$ is chosen as the most suitable one.

2.4 Permutation Test

Our independence test procedure is based on the *permutation test* [3].

We first run LSMI using the original datasets $\mathcal{Z} = \{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, and obtain an SMI estimate $\widehat{\mathrm{SMI}}(\mathcal{Z})$. Next, we randomly permute $\{\boldsymbol{y}_i\}_{i=1}^n$ and form a shuffled dataset $\widetilde{\mathcal{Z}} = \{(\boldsymbol{x}_i,\boldsymbol{y}_{\tau(i)})\}_{i=1}^n$, where $\tau(\cdot)$ is a randomly chosen permutation function. Then we run LSMI again using the randomly shuffled dataset $\widetilde{\mathcal{Z}}$, and obtain an SMI estimate $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$. Note that the random permutation eliminates the dependency between $\boldsymbol{x}$ and $\boldsymbol{y}$ (if exists), so $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$ would take a value close to zero.

This random permutation procedure is repeated many times, and the distribution of $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$ under the null-hypothesis (i.e., $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of $\widehat{\mathrm{SMI}}(\mathcal{Z})$ in the distribution of $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$. We refer to this procedure as the *least-squares independence test* (LSIT).

A MATLAB® implementation of LSIT is available

from http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSIT/.

3. Experiments

In this section, we report experimental results.

3.1 Numerical Illustration

First, we illustrate how the proposed LSIT method behaves using the following toy datasets with one-dimensional input x and one-dimensional output y:

(A) Regression: For $x \sim U(-20, 20)$ where $U(a, b)$ denotes the uniform distribution on (a, b),

$$y \sim \begin{cases} U(-1, 1) & \text{(Independent)}, \\ N(\sin(20x/\pi), 1) & \text{(Dependent)}, \end{cases}$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

(B) Classification: For $y \sim B(0.5)$ where $B(p)$ denotes the binomial distribution on $\{-1, +1\}$ with probability of having $+1$ being p,

$$x \sim \begin{cases} 0.5N(-1, 1) + 0.5N(1, 1) & \text{(Independent)}, \\ N(y, 1) & \text{(Dependent)}. \end{cases}$$

Examples of realized samples are plotted in Fig. 1 for $n = 100$, where $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are normalized to have unit variance. Figure 2 depicts the distributions of $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$ and the value of $\widehat{\mathrm{SMI}}(\mathcal{Z})$. The graphs show that reasonable p-values were obtained for all the four cases. Figure 3 depicts the p-values and the frequency of accepting the null hypothesis (i.e., x and y are independent) as functions of the sample size n. The graphs show that LSIT works reasonably well.

3.2 Performance Comparison

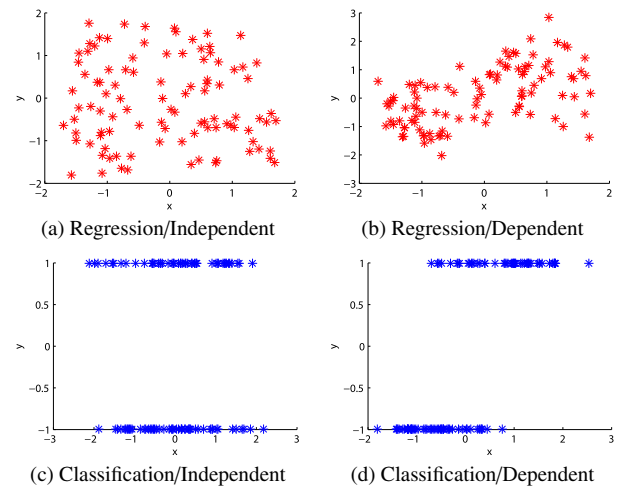
Here we compare the performance of LSMI and HSIC under the common permutation-test framework. HSIC is the state-of-the-art measure of statistical independence utilizing Gaussian kernels [4]. The performance of HSIC depends on the choice of the Gaussian width, and to the best of our knowledge, there is no theoretically justified method to determine the kernel width. Here we use a standard heuristic of setting the Gaussian width to the median distance between samples, which was also adopted in the original paper [4].

We generate data samples by

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix},$$

where

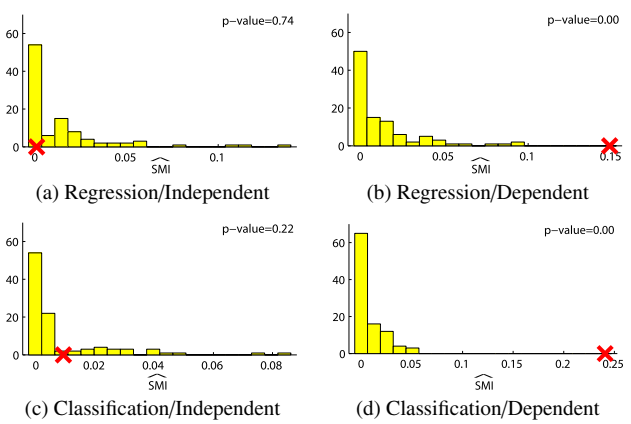
$$x' \sim 0.5N(-1, 1) + 0.5N(1, 1),$$
$$y' \sim 0.5N(-2, 1) + 0.5N(2, 1).$$



(a) Regression/Independent (b) Regression/Dependent

(c) Classification/Independent (d) Classification/Dependent

Fig. 1 Toy datasets.



(a) Regression/Independent (b) Regression/Dependent

(c) Classification/Independent (d) Classification/Dependent

Fig. 2 Distributions of $\widehat{\mathrm{SMI}}(\widetilde{\mathcal{Z}})$ (randomly shuffled samples) for the toy datasets. '×' denotes the value of $\widehat{\mathrm{SMI}}(\mathcal{Z})$ (original samples).

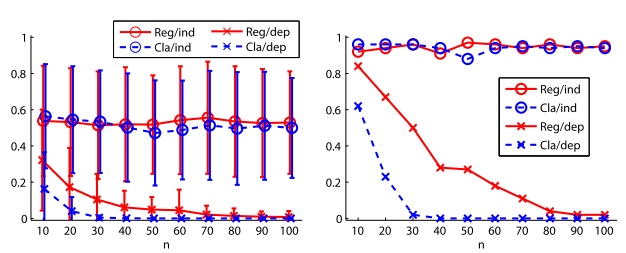


Fig. 3 Experimental results for the toy datasets. Left: Mean and standard deviation of p-values over 100 runs. Right: The frequency of accepting the null hypothesis over 100 runs under the significance level 0.05.

Thus, (x, y) are rotation of (x', y') by angle θ. We conduct experiments for $\theta = 0$ (i.e., x and y are independent) and $\theta = \pi/8$ (i.e., x and y are dependent). Data samples $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are normalized to have unit variance (see Fig. 4).

The results plotted in Fig. 5 show that the proposed LSIT has comparable type-I error (rejecting correct null-hypotheses) to HSIC, with far smaller type-II error (accepting incorrect null-hypotheses).

Next, we vary the Gaussian width of HSIC and evaluate how the performance is changed. In Fig. 5, 'HSIC(c)'

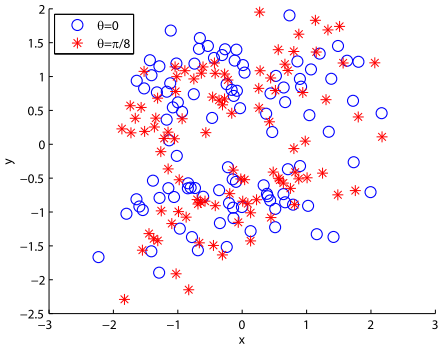


Fig. 4 Rotation datasets. x and y are independent if $\theta = 0$, and they are dependent when $\theta = \pi/8$.

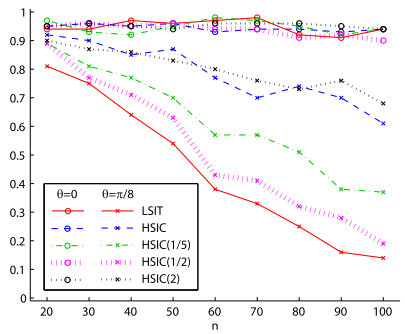


Fig. 5 Experimental results for the rotation datasets. Frequency of accepting the null hypothesis over 100 runs under the significance level 0.05 is depicted.

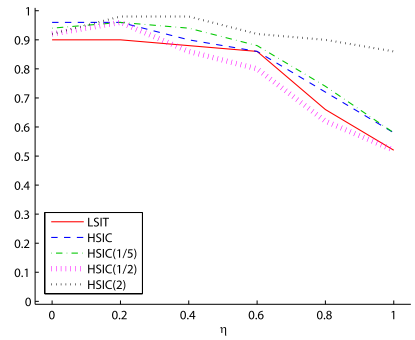


Fig. 6 Experimental results for the MNIST datasets. Frequency of accepting the null hypothesis over 50 runs under the significance level 0.05 is depicted.

denotes HSIC with Gaussian width multiplied by c. The results show that, although the type-I error does not really change with respect to c, the type-II error is heavily affected by the choice of the Gaussian kernel width. For this dataset, the median heuristic does not work well, and $c = 1/2$ works the best. However, the optimally-tuned HSIC is still outperformed by the proposed LSIT, which is automatically tuned based on cross-validation and does not involve manual parameter tuning.

Finally, we use the *MNIST handwritten digit dataset* for further performance evaluation. Each digit image (representing an integer in $\{0, 1, 2, \ldots, 9\}$) consists of 784 ($= 28 \times 28$) pixels, each of which takes an integer value between 0 to 255 representing its intensity level in gray-scale. Here, we label the data as 'small' for digits '0', '1', '2', '3', and '4', and 'large' for digits '5', '6', '7', '8', and '9'. We randomly choose 250 samples and randomly shuffle the label of $250(1 - \eta)$ samples. Thus, increasing η from 0 to 1 corresponds to increasing the dependence between digit patterns and labels.

The results are plotted in Fig. 6, showing that the proposed LSIT has slightly larger type-I error (i.e., lower acceptance rate when $\eta = 0$) than HSIC, but the type-II error of LSIT is slightly smaller than HSIC (i.e., lower acceptance rate when $\eta > 0$). For this dataset, the optimally-tuned HSIC ($c = 1/2$) slightly outperforms automatically-tuned LSIT.

4. Discussions and Conclusions

We proposed a novel non-parametric method of independence test based on an estimator of a squared-loss variant of mutual information called *least-squares mutual information* [9]. The proposed method, which we called the *least-squares independence test* (LSIT), can overcome the limitation of the state-of-the-art method, the *Hilbert-Schmidt independence criterion* (HSIC) [4], which is not equipped with a model selection procedure of the Gaussian kernel width. Through experiments, we confirmed that the proposed LSIT compares favorably with the HSIC-based independence test.

Acknowledgment

MS was supported by SCAT, AOARD, and the JST PRESTO program. TS was supported by MEXT Grant-in-Aid for Young Scientists (B) 22700289.

References

[1] N. Aronszajn, "Theory of reproducing kernels," Trans. American Mathematical Society, vol.68, pp.337–404, 1950.
[2] T.M. Cover and J.A. Thomas, Elements of Information Theory, 2nd ed., John Wiley & Sons, Inc., 2006.
[3] B. Efron and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, 1993.
[4] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A. Smola, "A kernel statistical test of independence," in Advances in Neural Information Processing Systems 20, pp.585–592, 2008.
[5] T. Kanamori, T. Suzuki, and M. Sugiyama, "Condition number analysis of kernel-based density ratio estimation," Tech. Rep., arXiv, 2009.
[6] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philosophical Magazine, vol.50, pp.157–175, 1900.
[7] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," J. Machine Learning Research, vol.2, pp.67–93, 2001.
[8] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," International Conference on Artificial Intelligence and Statistics, pp.804–811, 2010.
[9] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes," BMC Bioinformatics, vol.10, no.1, p.S52, 2009.
[10] M. Yamada and M. Sugiyama, "Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise," AAAI Conference on Artificial Intelligence, pp.643–648, 2010.