IEICE Transactions on Information and Systems,
vol.E95-D, no.10, pp.2564–2567, 2012.

On Kernel Parameter Selection in Hilbert-Schmidt Independence Criterion

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Makoto Yamada
Tokyo Institute of Technology, Japan.
yamada@sg.cs.titech.ac.jp

Abstract

The *Hilbert-Schmidt independence criterion* (HSIC) is a kernel-based statistical independence measure that can be computed very efficiently. However, it requires us to determine the kernel parameters heuristically because no objective model selection method is available. *Least-squares mutual information* (LSMI) is another statistical independence measure that is based on direct density-ratio estimation. Although LSMI is computationally more expensive than HSIC, LSMI is equipped with cross-validation, and thus the kernel parameter can be determined objectively. In this paper, we show that HSIC can actually be regarded as an approximation to LSMI, which allows us to utilize cross-validation of LSMI for determining kernel parameters in HSIC. Consequently, both computational efficiency and cross-validation can be achieved.

Keywords

Hilbert-Schmidt independence criterion, least-squares mutual information, cross-validation, Gaussian kernel

1 Introduction

Measuring statistical independence between random variables is an important challenge in machine learning, because it can be used for various purposes such as feature selection [17, 24], feature extraction [22, 25], clustering [16, 9, 21], statistical independence test [7, 19], independent component analysis [15, 23], object matching [13, 27], and causal inference [11, 26].

Among various statistical independence measures, the *Hilbert-Schmidt independence criterion* (HSIC) [6] is a powerful and computationally efficient method. The basic idea

of HSIC is to evaluate all possible non-linear correlations in universal reproducing kernel Hilbert spaces [18], which can be performed efficiently via the *kernel trick* [14]. However, HSIC requires us to choose kernel parameters manually because no objective model selection criterion is available. In practice, using Gaussian kernels with widths set to the median distances between samples is a popular heuristic [6, 7], although such a heuristic does not always work well.

Least-squares mutual information (LSMI) [24] is another statistical independence measure, which is an estimator of a squared-loss variant of mutual information. The basic idea of LSMI is to approximate the ratio of a joint density over the product of marginal densities directly in a single-shot process, allowing us to avoid density estimation systematically [20]. LSMI was shown to possess a superior non-parametric convergence property [22] and optimal numerical stability [8]. Furthermore, LSMI is equipped with cross-validation that can be used for objectively determining kernel parameters.

In this paper, we show that HSIC can actually be regarded as an approximation to LSMI. This interpretation allows us to employ cross-validation of LSMI to determine kernel parameters in HSIC, by which both computational efficiency and objective model selection can be achieved. Through numerical experiments, we show the usefulness of the proposed approach.

2 Measuring Statistical Independence between Random Variables

Suppose that we are given a set of paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ on $\mathcal{X} \times \mathcal{Y}$, which are independently drawn from a joint probability distribution with density $p(\boldsymbol{x}, \boldsymbol{y})$. Our goal is to evaluate statistical independence between $\boldsymbol{x}$ and $\boldsymbol{y}$.

2.1 Hilbert-Schmidt Independence Criterion (HSIC)

Here, we review a kernel-based statistical independence measure called the *Hilbert-Schmidt independence criterion* (HSIC) [6].

Let $\mathcal{F}$ be a *reproducing kernel Hilbert space* (RKHS) [2] with reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$, and $\mathcal{G}$ be another RKHS with reproducing kernel $L(\boldsymbol{y}, \boldsymbol{y}')$. Let us denote the inner products in $\mathcal{F}$ and $\mathcal{G}$ by $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$, respectively, and marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$ by $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$, respectively.

Let C be a *cross-covariance operator* from $\mathcal{G}$ to $\mathcal{F}$, which is defined such that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, Cg \rangle_{\mathcal{F}} = \iint \left(\left[f(\boldsymbol{x}) - \int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right] \left[g(\boldsymbol{y}) - \int g(\boldsymbol{y}) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right] \right) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}.$$

By the reproducing properties,

$$f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x}) \rangle_{\mathcal{F}} \quad \text{and} \quad g(\boldsymbol{y}) = \langle g, L(\cdot, \boldsymbol{y}) \rangle_{\mathcal{G}},$$

the cross-covariance operator C can be more explicitly expressed as

$$C := \iint \left(\left[K(\cdot, \boldsymbol{x}) - \int K(\cdot, \boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right] \left[L(\cdot, \boldsymbol{y}) - \int L(\cdot, \boldsymbol{y}) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right] \right) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$

where '$\otimes$' denotes the *tensor product*.

The cross-covariance operator is a generalization of the *cross-covariance matrix* between random vectors. When $\mathcal{F}$ and $\mathcal{G}$ are *universal RKHSs* [18] defined on compact domains $\mathcal{X}$ and $\mathcal{Y}$, respectively, the largest singular value of C is zero if and only $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. Gaussian RKHSs are examples of the universal RKHS.

HSIC is defined as the squared *Hilbert-Schmidt norm* (the sum of the squared singular values) of the cross-covariance operator C:

$$\begin{aligned} \mathrm{HSIC} := & \iiiint K(\boldsymbol{x}, \boldsymbol{x}') L(\boldsymbol{y}, \boldsymbol{y}') p(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}', \boldsymbol{y}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{y}' \\ & + \iint K(\boldsymbol{x}, \boldsymbol{x}') p(\boldsymbol{x}) p(\boldsymbol{x}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{x}' \iint L(\boldsymbol{y}, \boldsymbol{y}') p(\boldsymbol{y}) p(\boldsymbol{y}') \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{y}' \\ & - 2 \iiint K(\boldsymbol{x}, \boldsymbol{x}') p(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}' \int L(\boldsymbol{y}, \boldsymbol{y}') p(\boldsymbol{y}') \mathrm{d}\boldsymbol{y}' p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}. \end{aligned}$$

Its empirical estimator is given as

$$\begin{aligned} \widehat{\mathrm{HSIC}} := & \frac{1}{n^2} \sum_{i,i'=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) L(\boldsymbol{y}_i, \boldsymbol{y}_{i'}) + \frac{1}{n^4} \sum_{i,i',j,j'=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) L(\boldsymbol{y}_j, \boldsymbol{y}_{j'}) \\ & - \frac{2}{n^3} \sum_{i,j,k=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_k) L(\boldsymbol{y}_j, \boldsymbol{y}_k) \\ = & \frac{1}{n^2} \mathrm{tr}(\boldsymbol{K} \boldsymbol{\Gamma} \boldsymbol{L} \boldsymbol{\Gamma}), \end{aligned}$$

where $\boldsymbol{K}_{i,i'} = K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, $\boldsymbol{L}_{j,j'} = L(\boldsymbol{y}_j, \boldsymbol{y}_{j'})$, $\boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ is the "centering" matrix in RKHSs, $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix, and $\mathbf{1}_n$ denotes the n-dimensional vector with all ones.

$\widehat{\mathrm{HSIC}}$ depends on the choice of the universal RKHSs $\mathcal{F}$ and $\mathcal{G}$. In the original HSIC papers [6, 7], the Gaussian RKHSs with widths set to the median distances between samples were used. However, there is no theoretical justification for this choice.

2.2 Least-Squares Mutual Information (LSMI)

Next, we review another statistical independence measure called *least-squares mutual information* (LSMI) [24].

LSMI is an estimator of *squared-loss mutual information* (SMI) defined as

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

SMI is non-negative and zero if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. Hence, SMI can be used for detecting statistical independence between random variables[1].

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 separately estimate the densities $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$, and plug the estimated densities in Eq.(1). However, density estimation is known to be a hard task and division by estimated densities can magnify the estimation error. To cope with this problem, LSMI systematically avoids density estimation by directly estimating the following *density ratio* function:

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

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

$$r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \theta_i K(\boldsymbol{x}, \boldsymbol{x}_i) L(\boldsymbol{y}, \boldsymbol{y}_i).$$

The parameter $\boldsymbol{\theta}$ are determined so that the following squared-error J is minimized:

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \iint \left(r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) - r(\boldsymbol{x}, \boldsymbol{y})\right)^2 p(\boldsymbol{x})p(\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \iint r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})^2 p(\boldsymbol{x})p(\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - 2 \iint r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} + \mathrm{Const.} \end{aligned}$$

Since J contains the expectations over unknown densities $p(\boldsymbol{x})p(\boldsymbol{y})$ and $p(\boldsymbol{x}, \boldsymbol{y})$, the expectations are approximated by empirical averages. By including an ℓ_2-regularizer and ignoring the irrelevant constant, the LSMI optimization problem is given as follows:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^n} \left[\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right],$$

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

$$\begin{aligned} \widehat{\boldsymbol{H}}_{i',j'} &:= \frac{1}{n^2} \sum_{i,j=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) K(\boldsymbol{x}_i, \boldsymbol{x}_{j'}) L(\boldsymbol{y}_j, \boldsymbol{y}_{i'}) L(\boldsymbol{y}_j, \boldsymbol{y}_{j'}), \\ \widehat{\boldsymbol{h}}_j &:= \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_j) L(\boldsymbol{y}_i, \boldsymbol{y}_j). \end{aligned}$$

[1]Note that SMI is the *Pearson divergence* [12] from the joint density $p(\boldsymbol{x}, \boldsymbol{y})$ to the product of marginals $p(\boldsymbol{x})p(\boldsymbol{y})$, whereas ordinary mutual information [3], defined by

$$\mathrm{MI} := \iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

is the *Kullback-Leibler divergence* [10] from $p(\boldsymbol{x}, \boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$. The Pearson divergence and the Kullback-Leibler divergence both belong to the class of *Ali-Silvey-Csiszár divergences* (also known as f-divergences, see [1, 4]), which share similar properties.

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

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n)^{-1} \widehat{\boldsymbol{h}}, \tag{3}$$

with which the density ratio estimator $\widehat{r}(\boldsymbol{x}, \boldsymbol{y})$ is obtained as

$$\widehat{r}(\boldsymbol{x}, \boldsymbol{y}) := r_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \widehat{\theta}_i K(\boldsymbol{x}, \boldsymbol{x}_i) L(\boldsymbol{y}, \boldsymbol{y}_i).$$

Finally, SMI can be approximated as

$$\widehat{\mathrm{SMI}} := \frac{1}{n} \sum_{i,j=1}^{n} \widehat{\theta}_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) L(\boldsymbol{y}_i, \boldsymbol{y}_j) - 1, \tag{4}$$

which is based on the following expression of SMI:

$$\mathrm{SMI} = \iint r(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} - 1.$$

Practical performance of LSMI depends on the choice of kernel parameters in $K(\boldsymbol{x}, \boldsymbol{x}')$ and $L(\boldsymbol{y}, \boldsymbol{y}')$ 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 subsets $\{\mathcal{Z}_m\}_{m=1}^{M}$. 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

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

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 parameters and the regularization parameter λ in the current setup) that minimizes $\widehat{J}^{\mathrm{CV}}$ is chosen as the most suitable one.

3 HSIC as an Approximation to LSMI

For a centralized kernel $\widetilde{K}(\boldsymbol{x}, \boldsymbol{x}')$ where $\widetilde{\boldsymbol{K}} := \boldsymbol{\Gamma} \boldsymbol{K} \boldsymbol{\Gamma}$, $\widehat{\mathrm{HSIC}}$ can be expressed as

$$\widehat{\mathrm{HSIC}} = \frac{1}{n^2} \sum_{i,j=1}^{n} \widetilde{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) L(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

which is equivalent to $\widehat{\mathrm{SMI}}$ with centralized kernel $\widetilde{K}(\boldsymbol{x}, \boldsymbol{x}')$ and parameters $\{\widehat{\theta}_i\}_{i=1}^{n}$ approximated by $1/n$, up to an irrelevant constant -1 (see Eq.(4)). This implies that HSIC can actually be regarded as an approximation to LSMI.

An advantage of HSIC over LSMI is that HSIC is computationally more efficient than LSMI, because LSMI involves matrix inversion (see Eq.(3)), whereas HSIC only computes the sum of kernel values. On the other hand, a disadvantage of HSIC is that kernel parameters are determined heuristically, whereas kernel parameter selection in LSMI can be performed objectively via cross-validation. Thus, the view that HSIC is an approximation to LSMI allows us to use cross-validation also for HSIC. More specifically, we replace $\widehat{\mathrm{SMI}}$ given by Eq.(4) with

$$\widetilde{\mathrm{SMI}} = \widehat{\mathrm{HSIC}} - 1,$$

and perform cross-validation as described in Section 2.2. Consequently, advantages of HSIC (i.e., computational efficiency) and LSMI (i.e., objective model selection) can both be gained.

4 Numerical Examples

In this section, we consider statistical independence testing by the *permutation test* [5], and experimentally compare the performances of LSMI with the Gaussian width chosen by cross-validation (denoted as $\mathrm{LSMI}_{\mathrm{CV}}$), HSIC with the Gaussian widths set to the median distances between samples (denoted as $\mathrm{HSIC}_{\mathrm{med}}$), and HSIC with the Gaussian widths chosen by LSMI cross-validation (denoted as $\mathrm{HSIC}_{\mathrm{CV}}$). We use 5-fold cross-validation (i.e., $M = 5$).

We generate data samples $\{(x_i, y_i)\}_{i=1}^n$ by

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

i.e., (x, y) are a rotation of (x', y') by angle α. We generate (x', y') as

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

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. We conduct experiments for $\alpha = 0$ (i.e., x and y are independent) and $\alpha = \pi/8$ (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.

Figure 1 shows the rejection rates of $\mathrm{LSMI}_{\mathrm{CV}}$, $\mathrm{HSIC}_{\mathrm{med}}$, and $\mathrm{HSIC}_{\mathrm{CV}}$. When x and y are independent, all three methods successfully accept the correct null-hypothesis with roughly the designated significance level (i.e., rejection rate 5%). On the other hand, when x and y are dependent, $\mathrm{HSIC}_{\mathrm{med}}$ rejects the incorrect null-hypothesis less frequently, and $\mathrm{HSIC}_{\mathrm{CV}}$ performs much better than $\mathrm{HSIC}_{\mathrm{med}}$; its performance is close to state-of-the-art $\mathrm{LSMI}_{\mathrm{CV}}$, with about 80% reduction in computation time (see Table 1).

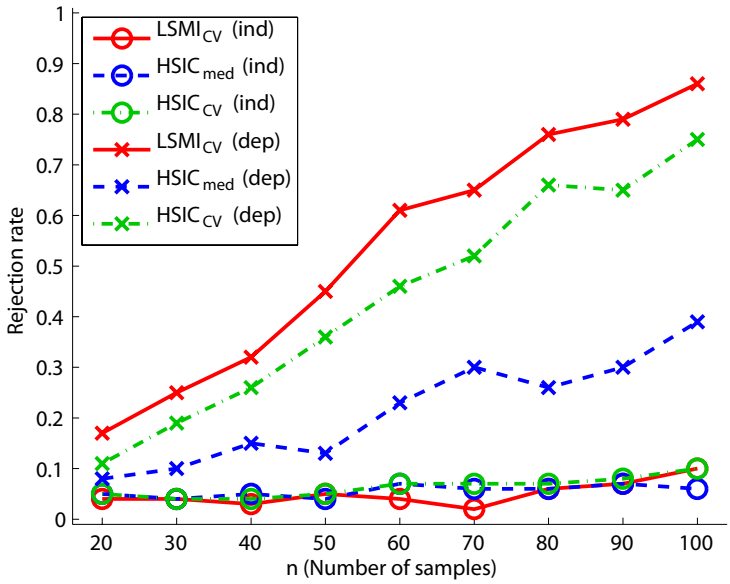

Figure 1: Results of independence test. Frequency of rejecting the null hypothesis (i.e., independent) over 100 runs under the significance level 5% is depicted.

Table 1: Normalized CPU computation time.

Method	LSMI$_{CV}$	HSIC$_{med}$	HSIC$_{CV}$
Time	1	0.022	0.193

5 Conclusions

In this paper, we showed that HSIC can be regarded as an approximation to LSMI, allowing us to employ LSMI cross-validation for kernel parameter choice in HSIC. Consequently, advantages of HSIC (i.e., computational efficiency) and LSMI (i.e., objective model selection) can both be gained. Experiments illustrated the validity of our approach.

MS was supported by AOARD and the JST PRESTO program.

References

[1] S.M. Ali and S.D. Silvey, "A general class of coefficients of divergence of one distribution from another," Journal of the Royal Statistical Society, Series B, vol.28, no.1, pp.131–142, 1966.

[2] N. Aronszajn, "Theory of reproducing kernels," Transactions of the American Mathematical Society, vol.68, pp.337–404, 1950.

[3] T.M. Cover and J.A. Thomas, Elements of Information Theory, 2nd ed., John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006.

[4] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," Studia Scientiarum Mathematicarum Hungarica, vol.2, pp.229–318, 1967.

[5] B. Efron and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, New York, NY, USA, 1993.

[6] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," Algorithmic Learning Theory, pp.63–77, 2005.

[7] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A. Smola, "A kernel statistical test of independence," Advances in Neural Information Processing Systems 20, pp.585–592, 2008.

[8] T. Kanamori, T. Suzuki, and M. Sugiyama, "Condition number analysis of kernel-based density ratio estimation," Tech. Rep. 0912.2800, arXiv, 2009.

[9] M. Kimura and M. Sugiyama, "Dependence-maximization clustering with least-squares mutual information," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol.15, no.7, pp.800–805, 2011.

[10] S. Kullback and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol.22, pp.79–86, 1951.

[11] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf, "Regression by dependence minimization and its application to causal inference in additive noise models," International Conference on Machine Learning pp.745–752, 2009.

[12] 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 Series 5, vol.50, no.302, pp.157–175, 1900.

[13] N. Quadrianto, A.J. Smola, L. Song, and T. Tuytelaars, "Kernelized sorting," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, pp.1809–1821, 2010.

[14] B. Schölkopf and A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, USA, 2002.

[15] H. Shen, S. Jegelka, and A. Gretton, "Fast kernel-based independent component analysis," IEEE Transactions on Signal Processing, vol.57, no.9, pp.3498–3511, 2009.

[16] L. Song, A. Smola, A. Gretton, and K. Borgwardt, "A dependence maximization view of clustering," International Conference on Machine Learning, pp.815–822, 2007.

[17] L. Song, A. Smola, A. Gretton, K.M. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation," International Conference on Machine Learning, pp.823–830, 2007.

[18] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," Journal of Machine Learning Research, vol.2, pp.67–93, 2001.

[19] M. Sugiyama and T. Suzuki, "Least-squares independence test," IEICE Transactions on Information and Systems, vol.E94-D, no.6, pp.1333–1336, 2011.

[20] M. Sugiyama, T. Suzuki, and T. Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, Cambridge, UK, 2012.

[21] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya, "On information-maximization clustering: Tuning parameter selection and analytic solution," International Conference on Machine Learning, pp.65–72, 2011.

[22] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," Neural Computation, to appear.

[23] T. Suzuki and M. Sugiyama, "Least-squares independent component analysis," Neural Computation, vol.23, no.1, pp.284–301, 2011.

[24] 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.

[25] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama, "Computationally efficient sufficient dimension reduction via squared-loss mutual information," Asian Conference on Machine Learning, pp.247–262, 2011.

[26] 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.

[27] M. Yamada and M. Sugiyama, "Cross-domain object matching with model selection," International Conference on Artificial Intelligence and Statistics, pp.807–815, 2011.