IEICE Transactions on Information and Systems,
vol.E96-D, no.10, pp.2282–2285, 2012.

Direct Approximation of Quadratic Mutual Information and Its Application to Dependence-Maximization Clustering

Janya Sainui
Tokyo Institute of Technology, Japan.
janya@sg.cs.titech.ac.jp

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

Abstract

Mutual information (MI) is a standard measure of statistical dependence of random variables. However, due to the log function and the ratio of probability densities included in MI, it is sensitive to outliers. On the other hand, the L^2-distance variant of MI called *quadratic MI* (QMI) tends to be robust against outliers because QMI is just the integral of the squared difference between the joint density and the product of marginals. In this paper, we propose a kernel least-squares QMI estimator called *least-squares QMI* (LSQMI) that directly estimates the density difference without estimating each density. A notable advantage of LSQMI is that its solution can be analytically and efficiently computed just by solving a system of linear equations. We then apply LSQMI to dependence-maximization clustering, and demonstrate its usefulness experimentally.

Keywords

Quadratic mutual information, Least-square density-difference estimation, Information theoretic clustering.

1 Introduction

Mutual information (MI) [4] between random variable $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\mathrm{MI} := \iint p(\boldsymbol{x},\boldsymbol{y}) \log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \tag{1}$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ denotes the joint probability density of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ denote the marginal probability densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. MI is non-negative and equal to zero if and only if $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$. Thus, MI can be used as a measure of statistical dependence of $\boldsymbol{x}$ and $\boldsymbol{y}$.

A naive approach to approximating MI from i.i.d. paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ is to separately approximate $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$, and then plug the estimated densities into Eq.(1). However, density estimation is known to be a hard task and taking the ratio of estimated densities can significantly magnify the estimation error [6]. To cope with this problem, an MI approximator that directly estimates the density ratio $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$ without density estimation, called *maximum-likelihood MI* (MLMI), was proposed [10]. Although MLMI was proved to achieve the optimal non-parametric convergence rate, it tends to be sensitive to outliers due to the log function and the ratio of probability densities [7]. Also, MLMI is computationally expensive due to the log function.

To overcome the excessive sensitivity of MI to outliers and high computational costs of MLMI, *squared-loss MI* (SMI) [9] has been proposed:

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

Whereas the ordinary MI is the Kullback-Leibler divergence [2] from $p(\boldsymbol{x}, \boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$, SMI is the Pearson divergence [3]. Because SMI does not include the log function, it is more robust against outliers [7] and its density-ratio approximator, called *least-squares MI* (LSMI), can be computed efficiently in an analytic form [9]. However, the density ratio function $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$ can diverge to infinity even for a simple case where two one-dimensional variables x and y follow a correlated Gaussian distribution. In such a case, LSMI (and also MLMI) loses its consistency. This implies that LSMI is still sensitive to outliers.

To cope with this problem, we consider an L^2-distance variant of MI called *quadratic MI* [11]:

$$\mathrm{QMI} := \iint \Big(p(\boldsymbol{x}, \boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y})\Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{2}$$

Because QMI includes neither the log function nor the density ratio, it is expected to be robust against outliers.

Although QMI can also be approximated from i.i.d. paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ via density estimation similarly to MI, this is unreliable due to the hardness of density estimation. In this paper, we give a method to directly approximate the density difference $p(\boldsymbol{x}, \boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y})$ without density estimation. This estimator can be regarded as an application of the least-squares density-difference estimator [8] to QMI, and possesses various excellent properties: Its solution can be computed analytically in a computationally efficient way, all tuning parameters can be systematically optimized via cross-validation, and it achieves the optimal non-parametric convergence rate.

We apply our QMI approximator, called *least-squares QMI* (LSQMI), to dependence-maximization clustering [5], which determines cluster labels so that an information measure between feature vectors and cluster labels is maximized. Through experiments, we demonstrate that LSQMI-based clustering tends to be more robust against outliers than the state-of-the-art LSMI-based clustering [1].

2 QMI Approximation Based on Density-Difference Estimation

In this section, we propose a new method to approximate QMI.

Density-Difference Estimation: Suppose that we are given a set of paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ on some domain, which are independently drawn from a joint probability distribution with density $p(\boldsymbol{x}, \boldsymbol{y})$. Following [8], we directly approximate the following density-difference function without density estimation of $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$:

$$f(\boldsymbol{x}, \boldsymbol{y}) := p(\boldsymbol{x}, \boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y}),$$

where $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ denote the marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

We approximate the density difference $f(\boldsymbol{x}, \boldsymbol{y})$ using the following linear-in-parameter model:

$$g(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\ell=1}^{b} \theta_\ell \phi_\ell(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)^\top$ is a parameter vector, $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) = (\phi_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, \phi_b(\boldsymbol{x}, \boldsymbol{y}))^\top$ is a basis function vector, b denotes the number of parameters, and $^\top$ denotes the transpose. We will explain how the basis functions are designed in practice later.

We fit the model g to the true density-difference function f by least-squares:

$$\min_{\boldsymbol{\theta}} \iint \Big(g(\boldsymbol{x}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y})\Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{3}$$

An empirical and regularized version of the above optimization problem is given as

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

where $\lambda \geq 0$ is the regularization parameter and $\boldsymbol{H}$ and $\widehat{\boldsymbol{h}}$ are defined as

$$\boldsymbol{H} := \iint \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})^\top \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

$$\widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \frac{1}{n^2}\sum_{i,i'=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_{i'}).$$

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

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

where $\boldsymbol{I}$ denotes the identity matrix. Finally, our density-difference estimator $\widehat{f}(\boldsymbol{x},\boldsymbol{y})$ is given by

$$\widehat{f}(\boldsymbol{x},\boldsymbol{y}) = \widehat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}).$$

Model Selection by CV: The above density-difference estimator depends on the choice of the regularization parameter λ and some parameters included in the basis function $\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})$. These tuning parameters (which we call a *model*) can be systematically optimized based on cross-validation (CV) with respect to the objective function (3) as follows: First, the sample set $\mathcal{Z} = \{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$ is divided into disjoint subsets $\{\mathcal{Z}_m\}_{m=1}^M$ of (approximately) the same size. Then a QMI estimator $\widehat{f}_m$ is obtained from $\mathcal{Z}\backslash\mathcal{Z}_m$ (i.e., all samples without $\mathcal{Z}_m$), and its objective value is evaluated using the hold-out samples $\mathcal{Z}_m$ as

$$\begin{aligned}\widehat{J}_m^{\mathrm{CV}} := & \iint \widehat{f}_m(\boldsymbol{x},\boldsymbol{y})^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - \frac{2}{|Z_m|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in Z_m} \widehat{f}_m(\boldsymbol{x},\boldsymbol{y}) \\ & + \frac{2}{|Z_m|^2}\sum_{\boldsymbol{x},\boldsymbol{y}\in Z_m} \widehat{f}_m(\boldsymbol{x},\boldsymbol{y}),\end{aligned}$$

where $|\mathcal{Z}_m|$ denotes the number of elements in the set $\mathcal{Z}_m$, $\sum_{(\boldsymbol{x},\boldsymbol{y})\in Z_m}$ indicates the summation over every paired sample $(\boldsymbol{x},\boldsymbol{y})$ in $\mathcal{Z}_m$ (i.e., summation over $|\mathcal{Z}_m|$ elements), and $\sum_{\boldsymbol{x},\boldsymbol{y}\in Z_m}$ indicates the summation over every unpaired sample $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{Z}_m$ (i.e., summation over $|\mathcal{Z}_m|^2$ combinations). This procedure is repeated for $m = 1,\ldots,M$, and the model that minimizes the average of the above hold-out error over all m is chosen as the best one.

QMI Approximation: QMI (2) can be expressed using the density difference $f(\boldsymbol{x},\boldsymbol{y})$ as

$$\mathrm{QMI} = \iint f(\boldsymbol{x},\boldsymbol{y})^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

Either replacing $f(\boldsymbol{x},\boldsymbol{y})^2$ with $\widehat{f}(\boldsymbol{x},\boldsymbol{y})^2$ or replacing $f(\boldsymbol{x},\boldsymbol{y})^2$ with $\widehat{f}(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{x},\boldsymbol{y})$ and using empirical approximation, we can approximate QMI as $\widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}$ or $\widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}}$. However, as discussed in [8], we use their linear combination as our QMI estimator, which has a smaller bias than the above naive estimators:

$$\widehat{\mathrm{QMI}} := 2\widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}.$$

We call this the *least-squares QMI* (LSQMI) estimator.

3 Dependence-Maximization Clustering with LSQMI

In this section, we apply LSQMI to dependence-maximization clustering and give its computationally efficient implementation.

Algorithm 1 LSQMIC

Input: Feature vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ and the number of clusters, c.
Output: Cluster assignments $\{y_i \mid y_i \in \{1, \ldots, c\}\}_{i=1}^n$.

```
Randomly permute $\{\boldsymbol{x}_i\}_{i=1}^n$;
Randomly initialize cluster assignments $\{y_i\}_{i=1}^n$;
repeat
    for $i' = 1, \ldots, n$ do
        for $y = 1, \ldots, c$ do
            Compute $\widehat{\mathrm{QMI}}_y$ from $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $y_{i'} = y$;
        end for
        $y_{i'} \leftarrow \mathrm{argmax}_{y=1,\ldots,c}\, \widehat{\mathrm{QMI}}_y$;
    end for
until $\{y_i\}_{i=1}^n$ do not change.
```

Formulation and Algorithm: Given feature vectors $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, the goal of dependence-maximization clustering is to find associated cluster labels $\{y_i | y_i \in \{1, \ldots, c\}\}_{i=1}^n$ that maximizes a certain information measure [5, 1], where the number of clusters, c, is assumed known below. Here we use LSQMI as our information measure. A greedy algorithm of LSQMI-based clustering is described in Algorithm 1, which we refer to as *LSQMI-based clustering* (LSQMIC).

Computationally Efficient Implementation: In LSQMIC, we use

$$g(\boldsymbol{x}, y) := \sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell) L(y, y_\ell) \tag{4}$$

as our density-difference model. Here, $K(\boldsymbol{x}, \boldsymbol{x}')$ and $L(y, y')$ denote the Gaussian kernel and the delta kernel, respectively:

$$K(\boldsymbol{x}, \boldsymbol{x}') := \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right),$$
$$L(y, y') := \begin{cases} 1 & (y = y'), \\ 0 & (y \neq y'), \end{cases}$$

where σ^2 denotes the Gaussian width which can be optimized by CV. When n is large, we may use only a subset of samples as kernel centers to reduce the number of kernel bases.

Thanks to the sparseness brought by the delta kernel for y, $\boldsymbol{H}$ becomes block-diagonal for model (4), if samples are sorted according to the cluster labels. This nice structure actually allows us to compute the density-difference estimator in a cluster-wise manner, which significantly contributes to reducing the computation time.

More specifically, let $\{\boldsymbol{x}_i^{(y)}\}_{i=1}^{n^{(y)}}$ be samples in cluster y, where $n^{(y)}$ denotes the number of samples in cluster y: $\sum_{y=1}^{c} n^{(y)} = n$. Let

$$g^{(y)}(\boldsymbol{x}) := \sum_{\ell=1}^{n^{(y)}} \theta_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)})$$

be the density-difference model for cluster y. Then the density-difference solution $\widehat{\boldsymbol{\theta}}^{(y)}$ for cluster y is given by

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{(y)} &:= \mathop{\mathrm{argmin}}_{\boldsymbol{\theta} \in \mathbb{R}^{n^{(y)}}} \left[\boldsymbol{\theta}^\top \boldsymbol{H}^{(y)} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}}^{(y)} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right] \\ &= (\boldsymbol{H}^{(y)} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}^{(y)},\end{aligned}$$

where

$$\begin{aligned}H_{\ell,\ell'}^{(y)} &:= \int K(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)}) K(\boldsymbol{x}, \boldsymbol{x}_{\ell'}^{(y)}) \mathrm{d}\boldsymbol{x} \\ &= (\pi\sigma^2)^{d/2} \exp\left(-\frac{\|\boldsymbol{x}_\ell^{(y)} - \boldsymbol{x}_{\ell'}^{(y)}\|^2}{4\sigma^2}\right), \\ \widehat{h}_\ell^{(y)} &:= \frac{1}{n} \sum_{i=1}^{n^{(y)}} K(\boldsymbol{x}_i^{(y)}, \boldsymbol{x}_\ell^{(y)}) - \frac{n^{(y)}}{n^2} \sum_{i=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_\ell^{(y)}).\end{aligned}$$

Finally, the density-difference estimator $\widehat{f}(\boldsymbol{x}, y)$ is given by

$$\widehat{f}(\boldsymbol{x}, y) = \sum_{\ell=1}^{n^{(y)}} \widehat{\theta}_\ell^{(y)} K(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)}),$$

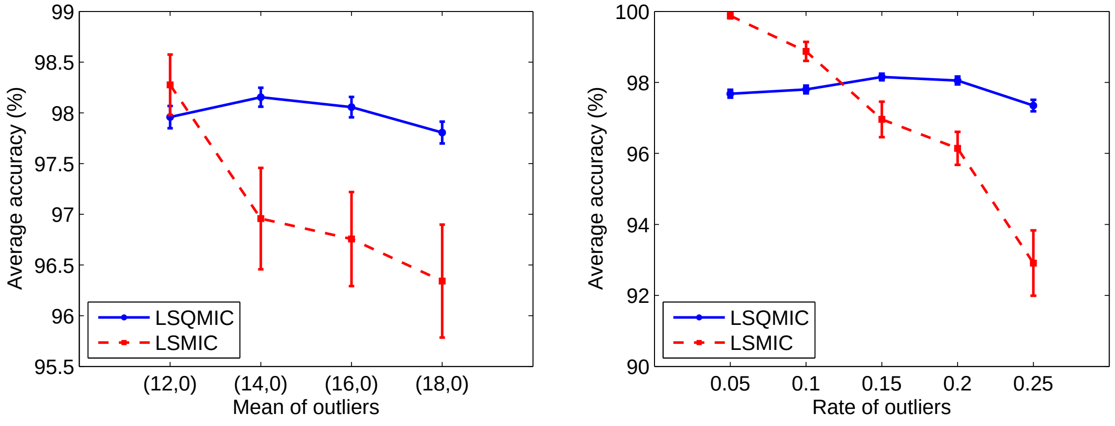
and the LSQMI estimator is given by

$$\widehat{\mathrm{QMI}} = \sum_{y=1}^{c} \left(2\widehat{\boldsymbol{\theta}}^{(y)\top} \widehat{\boldsymbol{h}}^{(y)} - \widehat{\boldsymbol{\theta}}^{(y)\top} \boldsymbol{H}^{(y)} \widehat{\boldsymbol{\theta}}^{(y)}\right).$$

4 Experiments

In this section, we experimentally compare the proposed LSQMIC with its LSMI counterpart called *LSMI-based clustering* (LSMIC) [1], which is a state-of-the-art dependence maximization clustering method. Because both methods are greedy algorithms, we run them 9 times with random initialization and choose the best solutions that maximize each information measure. Before feeding the data into clustering algorithms, we normalize the data so that element-wise variance is one.

Our focus in the following experiments is robustness against outliers.



(a) Outlier rate $\eta = 0.15$ (b) Outlier mean $\mu = 14$

Figure 1: Averages and standard errors of clustering accuracy on toy data over 100 runs.

Toy Datasets: Let the number of clusters be $c = 2$, the sample dimensionality be $d = 2$, and the sample size be $n = 500$. Inlier samples in each cluster are drawn from the Gaussian distributions with covariance matrix identity and mean $(-5, 0)^\top$ and $(5, 0)^\top$, respectively. With probability η, outlier samples are drawn from the Gaussian distribution with mean $(\mu, 0)^\top$ and covariance matrix $0.5\boldsymbol{I}$. We run the experiments 100 times with different random seeds and compare the average clustering accuracy for inlier samples.

Figure 1(a) depicts the results for outlier rate $\eta = 0.15$ and outlier mean $\mu = 12$, 14, 16, and 18, and Figure 1(b) depicts the results for outlier mean $\mu = 14$ and outlier rate $\eta = 0.05$, 0.1, 0.15, 0.2, and 0.25. These results show high robustness of the proposed LSQMIC: The accuracy of LSMIC tends to be decreased in both cases, while that of LSQMIC is almost unchanged.

Benchmark Datasets: Next, we employ 8 real-world datasets taken from the *UCI Repository*[1]. The experimental results are summarized in Table 1(a), showing that LSQMIC and LSMIC are comparable. Table 1(b) shows the results when 10% outliers are added to the same datasets (outliers are drawn from the Gaussian distribution with mean $\mathbf{21}$ and covariance matrix $0.1\boldsymbol{I}$, where $\mathbf{1}$ denotes the vector with all ones). This shows that LSQMIC tends to outperform LSMIC.

5 Conclusion

In this paper, we applied the general method of least-squares density-difference [8] to the difference between the joint density and the product of marginals, and derived an estimator of quadratic mutual information (QMI). We then applied the QMI estimator named LSQMI to dependence-maximization clustering and demonstrated its high robustness against outliers.

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Table 1: Averages (and standard errors in the brackets) of clustering accuracy on UCI datasets over 100 runs. The best and comparable methods by the t-test at the significance level 1% are described in boldface.

(a) Without outliers

Dataset	c	d	n	LSQMIC	LSMIC
Seismic	3	50	100	**55.4 (0.36)**	**56.1 (0.48)**
Sonar	3	60	100	55.4 (0.34)	**57.1 (0.47)**
Pima	2	8	100	**65.9 (0.45)**	**63.8 (0.54)**
Balance	2	4	100	**52.4 (0.81)**	**52.5 (0.63)**
Seeds	3	7	100	**90.2 (0.29)**	88.7 (0.46)
Liver-disorders	2	6	100	**54.5 (0.35)**	**54.5 (0.35)**
Shuttle	7	9	100	**56.3 (0.74)**	53.3 (0.75)
Vehicle	4	18	100	40.4 (0.39)	**42.0 (0.40)**

(b) With 10% outliers

Dataset	c	d	n	LSQMIC	LSMIC
Seismic	3	50	110	**55.5 (0.37)**	50.8 (0.74)
Sonar	3	60	110	**55.4 (0.30)**	**56.1 (0.41)**
Pima	2	8	110	**67.5 (0.42)**	**65.8 (0.63)**
Balance	2	4	110	**53.2 (0.93)**	**54.6 (0.90)**
Seeds	3	7	110	**89.4 (0.32)**	86.7 (0.67)
Liver-disorders	2	6	110	**54.0 (0.30)**	**54.6 (0.33)**
Shuttle	7	9	110	**58.6 (0.67)**	**58.3 (0.86)**
Vehicle	4	18	110	**40.2 (0.39)**	**40.9 (0.37)**

Acknowledgments

JS was supported by the Ph.D. scholarship from Prince of Songkla University, and MS is supported by MEXT KAKENHI 25700022 and AOARD.

References

[1] 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.

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

[3] 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.

[4] C. Shannon, “A mathematical theory of communication,” Bell Systems Technical Journal, vol.27, pp.379–423, 1948.

[5] L. Song, A. Smola, A. Gretton, and K. Borgwardt, “A dependence maximization view of clustering,” Proceedings of the 24th Annual International Conference on Machine Learning, pp.815–822, 2007.

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

[7] M. Sugiyama, T. Suzuki, and T. Kanamori, “Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation,” Annals of the Institute of Statistical Mathematics, vol.64, no.5, pp.1009–1044, 2012.

[8] M. Sugiyama, T. Suzuki, T. Kanamori, M.C. du Plessis, S. Liu, and I. Takeuchi, “Density-difference estimation,” Neural Computation, to appaer.

[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 (12 pages), 2009.

[10] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori, “Approximating mutual information by maximum likelihood density ratio estimation,” JMLR Workshop and Conference Proceedings, vol.4, pp.5–20, 2008.

[11] K. Torkkola, “Feature extraction by non-parametric mutual information maximization,” Journal of Machine Learning Research, vol.3, pp.1415–1438, 2003.