Sufficient Dimension Reduction via Squared-loss Mutual Information Estimation

Taiji Suzuki
The University of Tokyo
s-taiji@stat.t.u-tokyo.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

Abstract

The goal of sufficient dimension reduction in supervised learning is to find the low-dimensional subspace of input features that is 'sufficient' for predicting output values. In this paper, we propose a novel sufficient dimension reduction method using a squared-loss variant of mutual information as a dependency measure. We utilize an *analytic* approximator of squared-loss mutual information based on density ratio estimation, which is shown to possess suitable convergence properties. We then develop a natural gradient algorithm for sufficient subspace search. Numerical experiments show that the proposed method compares favorably with existing dimension reduction approaches.

1 Introduction

The purpose of *dimension reduction* in supervised learning is to construct a map from input features to their low-dimensional representation which has 'sufficient' information for predicting output values. Supervised dimension reduction methods can be divided broadly into two types—*wrappers* and *filters* (Guyon & Elisseeff, 2003). The wrapper approach performs dimension reduction specifically for a particular predictor (such as support vector classification or Gaussian process regression), while the filter approach is independent of the choice of successive predictors.

If one wants to enhance the prediction accuracy, the wrapper approach would be a suitable choice since predictors' characteristics could be taken into account in the dimension reduction phase. On the other hand, if one wants to interpret dimension-reduced features (e.g., in bioinformatics, computational chemistry, or brain analysis), the filter approach would be more appropriate since the extracted features are independent of the choice of successive predictors and therefore reliable in terms of interpretability. In this paper, we focus on the filter approach.

A standard formulation of filter-type dimension reduction is *sufficient dimension reduction* (SDR), which is aimed at finding a low-rank projection matrix such that, given the relevant subspace of input features, the rest becomes conditionally independent of output values (Cook, 1998; Chiaromonte & Cook, 2002; Fukumizu et al., 2009). A traditional dependency measure between random variables would be the *Pearson correlation coefficient* (PCC). PCC can be used for detecting linear dependency, so it is useful for Gaussian data. However, the Gaussian assumption may be rarely fulfilled in practice.

Recently, kernel-based dimension reduction has been studied in order to overcome the weakness of PCC. The *Hilbert-Schmidt independence criterion* (HSIC) (Gretton et al., 2005) utilizes *cross-covariance operators* on *universal* reproducing kernel Hilbert spaces (RKHSs) (Steinwart, 2001). Cross-covariance operators are an infinite-dimensional generalization of covariance matrices. HSIC allows one to efficiently detect non-linear dependency thanks to the *kernel trick* (Schölkopf & Smola, 2002). Its usefulness in feature selection scenarios has been shown in Song et al. (2007). However, HSIC has several weaknesses both theoretically and practically. Theoretically, HSIC evaluates independence between random variables, not *conditional* independence. Thus HSIC does not perform SDR in a strict sense. From the practical point of view, HSIC evaluates the covariance between random variables, not the correlation. This means that the change of input feature scaling affects the dimension reduction solution, which is not preferable in practice.

Kernel dimension reduction (KDR) (Fukumizu et al., 2004) can overcome these weaknesses. KDR evaluates

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

Table 1: Summary of existing and proposed dependency measures.

Methods	Non-linear dependency	Model selection	Distribution	Density estimation	Feature extraction
PCC	Not detectable	**Not necessary**	Gaussian	**Not involved**	**Possible**
HSIC	**Detectable**	Not available	**Free**	**Not involved**	**Possible**
KDR	**Detectable**	Not available	**Free**	**Not involved**	**Possible**
HIST	**Detectable**	**Available**	**Free**	Involved	Not available
KDE	**Detectable**	**Available**	**Free**	Involved	**Possible**
NN	**Detectable**	Not available	**Free**	**Not involved**	Not available
EDGE	**Detectable**	**Not necessary**	Near Gaussian	**Not involved**	**Possible**
MLMI	**Detectable**	**Available**	**Free**	**Not involved**	Not available
LSMI	**Detectable**	**Available**	**Free**	**Not involved**	**Possible**

conditional covariance using the kernel trick. Therefore, KDR directly performs SDR; furthermore, its theoretical properties such as consistency have been studied thoroughly (Fukumizu et al., 2009). However, KDR still has a weakness in practice—the performance of KDR (and HSIC) depends on the choice of kernel parameters (e.g., the Gaussian width) and the regularization parameter. So far, there seems no model selection method for KDR and HSIC (as discussed in Fukumizu et al. (2009))[1].

Another popular criterion for SDR is *mutual information* (MI) (Cover & Thomas, 1991). MI could be directly employed in the context of SDR since maximizing MI between output and projected input leads to conditional independence between output and input given the projected input. A great deal of effort has been made to estimate MI accurately, e.g., based on an adaptive histogram (HIST) (Darbellay & Vajda, 1999), kernel density estimation (KDE) (Torkkola, 2003), the nearest neighbor distance (NN) (Kraskov et al., 2004), the Edgeworth expansion (EDGE) (Hulle, 2005), and maximum likelihood MI estimation (MLMI) (Suzuki et al., 2008). Among them, MLMI has been shown to possess various advantages as summarized in Table 1.

So we want to employ the MLMI method for dimension reduction. However, this may not be possible since the MLMI estimator is not explicit (i.e., the MLMI estimator is implicitly defined as the solution of an optimization problem and is computed numerically)—in the dimension reduction scenarios, the projection matrix needs to be optimized over an MI approximator. To cope with this problem, we adopt a squared-loss variant of MI called the *squared-loss MI* (SMI) as our independence measure, and apply an SMI approximator called *least-squares MI* (LSMI) (Suzuki et al., 2009). LSMI inherits the good properties of MLMI, and moreover it provides an *analytic* MI estimator (see Table 1 again).

Based on LSMI, we develop a dimension reduction algorithm called *Least-squares dimension reduction* (LDR). LDR optimizes the projection matrix using a *natural gradient* algorithm (Amari, 1998) on the *Stiefel manifold*. Through numerical experiments, we show the usefulness of the LDR method.

2 Dimension Reduction via SMI Estimation

In this section, we first formulate the problem of *sufficient dimension reduction* (SDR) (Cook, 1998; Chiaromonte & Cook, 2002; Fukumizu et al., 2009) and show how *squared-loss mutual information* (SMI) could be employed in the context of SDR. Then we introduce a method of approximating SMI without going through density estimation and develop a dimension reduction method. Finally, several theoretical issues such as convergence properties of the proposed SMI estimator are investigated.

2.1 Sufficient Dimension Reduction

Let $\mathcal{D}_{\mathrm{X}}$ ($\subset \mathbb{R}^m$) be the domain of input features and $\mathcal{D}_{\mathrm{Y}}$ be the domain of output data. In the following, $\mathcal{D}_{\mathrm{Y}}$ could be multi-dimensional and either continuous (i.e., regression) or categorical (i.e., classification); structured outputs can also be handled in our framework as shown later.

The purpose of dimension reduction is to find a good low-dimensional representation of $\boldsymbol{x}$ which 'describes' output $\boldsymbol{y}$. Here we focus on linear dimension reduction, i.e.,

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x},$$

where $\boldsymbol{W}$ is a projection matrix onto a d-dimensional

[1] In principle, it is possible to choose the Gaussian width and the regularization parameter by cross validation over a successive predictor. However, this is not preferable due to the following two reasons. The first is significant increase of the computational cost. When cross validation is used, the tuning parameters in KDR (or HSIC) and hyper-parameters in the target predictor (such as the kernel parameters and the regularization parameter) should be optimized at the same time. This results in a deeply nested cross validation procedure and therefore this could be computationally very expensive. Another reason is that features extracted based on cross validation are no longer independent of predictors. Thus a merit of the filter approach (i.e., the obtained features are 'reliable') is lost.

subspace. That is, $\boldsymbol{W}$ is a member of the *Stiefel manifold* $\mathbb{S}_d^m(\mathbb{R})$:

$$\mathbb{S}_d^m(\mathbb{R}) := \{\boldsymbol{W} \in \mathbb{R}^{d\times m} \mid \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_d\},$$

where $\boldsymbol{I}_d$ is the d-dimensional identity matrix. We assume that d is known when developing a theory and an algorithm; practically d may be chosen by cross validation.

SDR is the problem of finding a projection matrix $\boldsymbol{W}$ such that

$$\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} \mid \boldsymbol{z}. \tag{1}$$

This means that, given the projected feature $\boldsymbol{z}$, the (remaining) feature $\boldsymbol{x}$ is conditionally independent of output $\boldsymbol{y}$ and therefore can be discarded without sacrificing the prediction ability.

Suppose that we are given n independent and identically distributed (i.i.d.) paired samples

$$D^n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid \boldsymbol{x}_i \in \mathcal{D}_{\mathrm{X}},\ \boldsymbol{y}_i \in \mathcal{D}_{\mathrm{Y}},\ i = 1, \ldots, n\}$$

drawn from a joint distribution with density $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$. Our goal is to estimate the subspace (or to find the projection matrix on it) such that Eq.(1) is fulfilled. We write $\boldsymbol{z}_i = \boldsymbol{W}\boldsymbol{x}_i$.

A direct approach to SDR would be to determine $\boldsymbol{W}$ so that Eq.(1) is fulfilled. To this end, we adopt SMI as our criterion to be maximized with respect to $\boldsymbol{W}$:

$$I_s(Y, Z) := \tfrac{1}{2}\int \left(\tfrac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})} - 1\right)^2 p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}, \tag{2}$$

where $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$ denotes the joint density of $\boldsymbol{y}$ and $\boldsymbol{z}$, and $p_{\mathrm{y}}(\boldsymbol{y})$ and $p_{\mathrm{z}}(\boldsymbol{z})$ denote the marginal densities of $\boldsymbol{y}$ and $\boldsymbol{z}$, respectively. $I_s(Y, Z)$ allows us to evaluate independence between $\boldsymbol{y}$ and $\boldsymbol{z}$ since $I_s(Y, Z)$ vanishes if and only if

$$p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z}) = p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z}).$$

Note that Eq.(2) corresponds to the *f-divergence* (Ali & Silvey, 1966; Csiszár, 1967) from $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$ to $p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})$ with the squared loss, while ordinary MI corresponds to the f-divergence with the log loss (i.e., the *Kullback-Leibler divergence*). Thus SMI could be regarded as a natural alternative to ordinary MI.

The rationale behind SMI in the context of SDR relies on the following lemma:

Lemma 1 *Let $p_{\mathrm{xy|z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$, $p_{\mathrm{x|z}}(\boldsymbol{x}|\boldsymbol{z})$, and $p_{\mathrm{y|z}}(\boldsymbol{y}|\boldsymbol{z})$ be conditional densities. Then we have*

$$I_s(X, Y) - I_s(Z, Y)$$
$$= \frac{1}{2}\int\left(1 - \frac{p_{\mathrm{xy|z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})}{p_{\mathrm{x|z}}(\boldsymbol{x}|\boldsymbol{z})p_{\mathrm{y|z}}(\boldsymbol{y}|\boldsymbol{z})}\right)^2 \frac{p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})^2 p_{\mathrm{x}}(\boldsymbol{x})}{p_{\mathrm{z}}(\boldsymbol{z})^2 p_{\mathrm{y}}(\boldsymbol{y})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$
$$\geq 0.$$

A proof of this lemma is given in Appendix. Lemma 1 implies that $I_s(X, Y) \geq I_s(Z, Y)$ and the equality holds if and only if

$$p_{\mathrm{xy|z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p_{\mathrm{x|z}}(\boldsymbol{x}|\boldsymbol{z})p_{\mathrm{y|z}}(\boldsymbol{y}|\boldsymbol{z}),$$

which is equivalent to Eq.(1). Thus, Eq.(1) can be achieved by maximizing $I_s(Z, Y)$ with respect to $\boldsymbol{W}$; then the 'sufficient' subspace can be identified.

Now we want to find the projection matrix $\boldsymbol{W}$ that maximizes $I_s(Z, Y)$. However, SMI is inaccessible in practice since densities $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$, $p_{\mathrm{y}}(\boldsymbol{y})$, and $p_{\mathrm{z}}(\boldsymbol{z})$ are unknown. Thus SMI needs to be estimated from data samples. Our key constraint when estimating SMI is that we want to avoid density estimation since this would be harder than dimension reduction itself. To accomplish this requirement, we employ an estimator of SMI proposed recently in Suzuki et al. (2009). The estimator utilizes *density ratio* estimation instead of density estimation itself based on the density ratio estimator proposed by Kanamori et al. (2009). Below, we explain the details.

2.2 SMI Approximation via Density Ratio Estimation

For the moment, we consider a fixed projection matrix $\boldsymbol{W}$ and let $\mathcal{D}_{\mathrm{Z}} = \boldsymbol{W}\mathcal{D}_{\mathrm{X}}$. Using *convex duality* (Boyd & Vandenberghe, 2004), we can express SMI as

$$I_s(Y, Z) = -\inf_g J(g) - \tfrac{1}{2},$$
$$J(g) = \frac{1}{2}\int g(\boldsymbol{y}, \boldsymbol{z})^2 p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z} - \int g(\boldsymbol{y}, \boldsymbol{z})p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z},$$

where $\inf_g$ is taken over all measurable functions. Thus computing I_s is reduced to finding the minimizer g^* of $J(g)$—it was shown that g^* is given as follows (Nguyen et al., 2008)[2]:

$$g^*(\boldsymbol{y}, \boldsymbol{z}) := \tfrac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}. \tag{3}$$

Thus, estimating $I_s(Y, Z)$ amounts to estimating the density ratio (3).

However, directly minimizing $J(g)$ is not possible due to the following two reasons. The first reason is that finding the minimizer over all measurable functions is not tractable in practice since the search space is too vast. To overcome this problem, we restrict the search space to some linear subspace $\mathcal{G}$:

$$\mathcal{G} := \{\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z}) \mid \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top \in \mathbb{R}^b\}, \tag{4}$$

[2] A more general result—the solution is given by Eq.(3) for any f-divergence—was obtained in Nguyen et al. (2008).

where $\boldsymbol{\alpha}$ is a parameter to be learned from samples, $^\top$ denotes the transpose of a matrix or a vector, and $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{z})$ is a basis function such that, for $\mathbf{0}_b$ being the b-dimensional vector with all zeros,

$$\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{z}) := (\varphi_1(\boldsymbol{y},\boldsymbol{z}),\ldots,\varphi_b(\boldsymbol{y},\boldsymbol{z}))^\top \geq \mathbf{0}_b \text{ for all } \boldsymbol{y},\boldsymbol{z}.$$

Note that $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{z})$ could be dependent on the samples $\{(\boldsymbol{y}_i,\boldsymbol{z}_i)\}_{i=1}^n$, i.e., *kernel* models are also allowed. Later, we explain how the basis functions $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{z})$ are designed in practice.

The second reason why directly minimizing $J(g)$ is not possible is that the true probability densities $p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})$, $p_{\mathrm{y}}(\boldsymbol{y})$, and $p_{\mathrm{z}}(\boldsymbol{z})$ contained in the density ratio (3) are unavailable. To cope with this problem, we approximate them by their empirical distributions—then we have

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{R}\boldsymbol{\alpha}\right], \quad (5)$$

where we included $\lambda\boldsymbol{\alpha}^\top\boldsymbol{R}\boldsymbol{\alpha}$ ($\lambda>0$) for regularization purposes; $\boldsymbol{R}$ is some positive definite matrix,

$$\widehat{\boldsymbol{H}} := \tfrac{1}{n^2}\textstyle\sum_{i,i'=1}^n \boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{z}_{i'})\boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{z}_{i'})^\top,$$
$$\text{and} \quad \widehat{\boldsymbol{h}} := \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{z}_i).$$

Differentiating the objective function (5) with respect to $\boldsymbol{\alpha}$ and equating it to zero, we obtain

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{R})^{-1}\widehat{\boldsymbol{h}}.$$

Thus, the solution can be computed *analytically* by solving a system of linear equations. Then we can analytically approximate SMI as follows, which is called *least-squares mutual information* (LSMI) (Suzuki et al., 2009):

$$\widehat{I}_s(Y,Z) := \widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\alpha}} - \tfrac{1}{2}\widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}} - \tfrac{1}{2}. \quad (6)$$

2.3 Least-squares Dimension Reduction

Next, we show how the LSMI approximator could be employed in dimension reduction scenarios. Since $\boldsymbol{W}$ is a projection matrix, dimension reduction involves an optimization problem over the Stiefel manifold $\mathbb{S}_d^m(\mathbb{R})$.

Here we employ a gradient ascent algorithm to find the maximizer of the LSMI approximator with respect to $\boldsymbol{W}$. After a few lines of calculations, we can show that the gradient is given by

$$\begin{aligned}\tfrac{\partial\widehat{I}_s}{\partial W_{\ell,\ell'}} =& \tfrac{\partial\widehat{\boldsymbol{h}}^\top}{\partial W_{\ell,\ell'}}(2\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\alpha}}^\top\tfrac{\partial\widehat{\boldsymbol{H}}}{\partial W_{\ell,\ell'}}(\tfrac{3}{2}\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}})\\ &+\lambda\widehat{\boldsymbol{\alpha}}^\top\tfrac{\partial\boldsymbol{R}}{\partial W_{\ell,\ell'}}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\alpha}}),\end{aligned}$$

where $\widehat{\boldsymbol{\beta}} := (\widehat{\boldsymbol{H}}+\lambda\boldsymbol{R})^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}}$.

In the Euclidean space, the ordinary gradient $\frac{\partial\widehat{I}_s}{\partial\boldsymbol{W}}$ gives the steepest direction. However, on a manifold, the *natural gradient* (Amari, 1998) gives the steepest direction. The natural gradient $\nabla\widehat{I}_s(\boldsymbol{W})$ at $\boldsymbol{W}$ is the projection of the ordinary gradient $\frac{\partial\widehat{I}_s}{\partial\boldsymbol{W}}$ to the tangent space of $\mathbb{S}_d^m(\mathbb{R})$ at $\boldsymbol{W}$. If the tangent space is equipped with the canonical metric $\langle\boldsymbol{G}_1,\boldsymbol{G}_2\rangle = \frac{1}{2}\mathrm{tr}(\boldsymbol{G}_1^\top\boldsymbol{G}_2)$, the natural gradient is given by

$$\nabla\widehat{I}_s(\boldsymbol{W}) = \tfrac{1}{2}\Big(\tfrac{\partial\widehat{I}_s}{\partial\boldsymbol{W}} - \boldsymbol{W}\tfrac{\partial\widehat{I}_s}{\partial\boldsymbol{W}}^\top\boldsymbol{W}\Big).$$

Then the *geodesic* from $\boldsymbol{W}$ to the direction of the natural gradient $\nabla\widehat{I}_s(\boldsymbol{W})$ over $\mathbb{S}_d^m(\mathbb{R})$ can be expressed using t ($\in\mathbb{R}$) as

$$\boldsymbol{W}_t := \boldsymbol{W}\exp\Big(t(\boldsymbol{W}^\top\tfrac{\partial\widehat{I}_s}{\partial\boldsymbol{W}} - \tfrac{\partial\widehat{I}_s}{\partial\boldsymbol{W}}^\top\boldsymbol{W})\Big),$$

where 'exp' for a matrix denotes the *matrix exponential*. Thus line search along the geodesic in the natural gradient direction is equivalent to finding the maximizer from $\{\boldsymbol{W}_t \mid t\geq 0\}$. More details of the geometric structure of the Stiefel manifold can be found in Nishimori and Akaho (2005).

For choosing the step size of each gradient update, we may use *Armijo's rule* (Patriksson, 1999). We call the proposed dimension reduction algorithm *Least-squares Dimension Reduction* (LDR). The entire algorithm is summarized in Figure 1.

2.4 Convergence Analysis

Here, we analyze convergence properties of LSMI for parametric and non-parametric setups.

Let us begin with the case where the function class $\mathcal{G}$ is a parametric model:

$$\mathcal{G} = \{g_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{z}) \mid \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^b\}.$$

Suppose that the true density ratio g^* is contained in the model $\mathcal{G}$, i.e., there exists $\boldsymbol{\theta}^*$ ($\in\Theta$) such that $g^* = g_{\boldsymbol{\theta}^*}$. Then we have the following theorem (its proof is omitted due to lack of space).

Theorem 1 *We have*

$$\widehat{I}_s(Y,Z) - I_s(Y,Z) = \mathcal{O}_p(1/\sqrt{n}),$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability. Furthermore, we have

$$\mathrm{E}_{D^n}[\widehat{I}_s(Y,Z) - I_s(Y,Z)] = \tfrac{1}{2n}\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{B}) + o(1/n),$$

where E_{D^n} denotes the expectation over data samples D^n. $\boldsymbol{A}$ and $\boldsymbol{B}$ are $b\times b$ matrices defined as

$$A_{\ell,\ell'} := \mathrm{E}_{p_{\mathrm{x}}p_{\mathrm{z}}}[\partial_\ell g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z})\partial_{\ell'}g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z})],$$

1. Initialize projection matrix $\boldsymbol{W}$.
2. Optimize Gaussian width σ and regularization parameter λ by CV (explained later).
3. Update $\boldsymbol{W}$ by $\boldsymbol{W} \leftarrow \boldsymbol{W}_\varepsilon$, where step-size ε may be chosen using Armijo's rule.
4. Repeat 2. and 3. until $\boldsymbol{W}$ converges.

Figure 1: The LDR algorithm.

$$
\begin{aligned}
B_{\ell,\ell'} :=& \mathrm{E}_{p_{\mathrm{yz}}}[(\partial_\ell g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z}) - \mathrm{E}_{p_{\mathrm{z}'|\mathrm{y}}}[\partial_\ell g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z}')] \\
&- \mathrm{E}_{p_{\mathrm{y}'|\mathrm{z}}}[\partial_\ell g_{\boldsymbol{\theta}^*}(\boldsymbol{y}',\boldsymbol{z})] + \mathrm{E}_{p_{\mathrm{y}'\mathrm{z}'}}[\partial_\ell g_{\boldsymbol{\theta}^*}(\boldsymbol{y}',\boldsymbol{z}')]) \\
&\times (\partial_{\ell'} g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z}) - \mathrm{E}_{p_{\mathrm{z}'|\mathrm{y}}}[\partial_{\ell'} g_{\boldsymbol{\theta}^*}(\boldsymbol{y},\boldsymbol{z}')] \\
&- \mathrm{E}_{p_{\mathrm{y}'|\mathrm{z}}}[\partial_{\ell'} g_{\boldsymbol{\theta}^*}(\boldsymbol{y}',\boldsymbol{z})] + \mathrm{E}_{p_{\mathrm{y}'\mathrm{z}'}}[\partial_{\ell'} g_{\boldsymbol{\theta}^*}(\boldsymbol{y}',\boldsymbol{z}')])],
\end{aligned}
$$

where $\boldsymbol{y}'$ and $\boldsymbol{z}'$ are copies of $\boldsymbol{y}$ and $\boldsymbol{z}$. The partial derivative ∂_ℓ is taken with respect to the ℓ-th element θ_ℓ of the parameter $\boldsymbol{\theta}$.

This theorem means that LSMI retains optimality in terms of the order of convergence in n since $\mathcal{O}_p(n^{-\frac{1}{2}})$ is the optimal convergence rate in the parametric setup.

Next, we consider non-parametric cases. Let $\mathcal{G}$ be a general set of functions on $\mathcal{D}_{\mathrm{Y}} \times \mathcal{D}_{\mathrm{Z}}$. For a function g ($\in \mathcal{G}$), let us consider a non-negative regularization functional $R(g)$ such that

$$
\sup_{\boldsymbol{y},\boldsymbol{z}}[g(\boldsymbol{y},\boldsymbol{z})] \leq R(g). \tag{7}
$$

If $\mathcal{G}$ is an RKHS with kernel $k(\cdot,\cdot)$ and there exists C such that $\sup_{\boldsymbol{y},\boldsymbol{z}} k((\boldsymbol{y},\boldsymbol{z}),(\boldsymbol{y},\boldsymbol{z})) \leq C$, $R(g) := \sqrt{C}\|g\|_{\mathcal{G}}$ satisfies Eq.(7):

$$
\begin{aligned}
g(\boldsymbol{y},\boldsymbol{z}) &= \langle k((\boldsymbol{y},\boldsymbol{z}),\cdot), g(\cdot)\rangle \\
&\leq \sqrt{k((\boldsymbol{y},\boldsymbol{z}),(\boldsymbol{y},\boldsymbol{z}))}\|g\|_{\mathcal{G}} \leq \sqrt{C}\|g\|_{\mathcal{G}},
\end{aligned}
$$

where we used the reproducing property of the kernel (Aronszajn, 1950) and Schwartz's inequality. Note that the Gaussian RKHS satisfies this with $C = 1$.

Let us consider a non-parametric version of the problem (5):

$$
\widehat{w} := \operatorname*{argmin}_{g\in\mathcal{G}}\Big[\frac{1}{2n^2}\sum_{i,j=1}^{n} g(\boldsymbol{y}_i,\boldsymbol{z}_j)^2 - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{y}_i,\boldsymbol{z}_i) + \frac{\lambda_n}{2}R(g)^2\Big].
$$

Note that if $\mathcal{G}$ is an RKHS, the above optimization problem is reduced to a form of the finite dimensional optimization problem (5). We assume that the true density ratio function $g^*(\boldsymbol{y},\boldsymbol{z})$ is contained in model $\mathcal{G}$ and is bounded from above:

$$
g^*(\boldsymbol{y},\boldsymbol{z}) < M_0 \quad \text{for all} \quad (\boldsymbol{y},\boldsymbol{z}) \in \mathcal{D}_{\mathrm{Y}} \times \mathcal{D}_{\mathrm{Z}}.
$$

We also assume that there exists γ ($0 < \gamma < 2$) such that

$$
\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p_{\mathrm{y}}p_{\mathrm{z}})) = O((M/\epsilon)^\gamma),
$$

$$
\mathcal{G}_M := \{g \in \mathcal{G} \mid R(g) \leq M\},
$$

where $\mathcal{H}_{[]}$ is the *bracketing entropy* of $\mathcal{G}_M$ with respect to the $L_2(p_{\mathrm{y}}p_{\mathrm{z}})$-norm (van der Vaart & Wellner, 1996). This quantity represents a complexity of function class $\mathcal{G}$—the larger γ is, the more complex the function class $\mathcal{G}$ is. The Gaussian RKHS satisfies this condition for arbitrarily small γ (Steinwart & Scovel, 2007). Then we have the following theorem (its proof is omitted due to lack of space; we used Theorem 5.11 in van de Geer (2000)).

Theorem 2 *Under the above setting, if $\lambda_n \to 0$ and $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$, then we have*

$$
\widehat{I}_s(Y,Z) - I_s(Y,Z) = \mathcal{O}_p(\max(\lambda_n, n^{-1/2})). \tag{8}
$$

The conditions $\lambda_n \to 0$ and $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$ roughly means that the regularization parameter λ_n should be sufficiently small but not too small. This theorem shows that the convergence rate of non-parametric LSMI is slightly slower than the parametric counterpart ($\mathcal{O}_p(n^{-1/2})$), but the non-parametric method would require a milder model assumption for eliminating the modeling error. According to Nguyen et al. (2008) where a log-loss version of the above theorem has been proven in the context of KL-divergence estimation, the above convergence rate achieves the optimal minimax rate under some setup. Thus the convergence property of non-parametric LSMI would also be optimal in the same sense.

2.5 Model Selection by Cross Validation

As shown above, LSMI has preferable convergence properties. Nevertheless, its practical performance depends on the choice of basis functions and the regularization parameter. In order to determine basis functions $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{z})$ and the regularization parameter λ, cross validation (CV) is available for the LSMI estimator: First, the samples $\{(\boldsymbol{y}_i,\boldsymbol{z}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{S}_k\}_{k=1}^K$ of (approximately) the same size. Then an estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}$ is obtained using $\{\mathcal{S}_j\}_{j\neq k}$ (i.e., without $\mathcal{S}_k$) and the approximation error for the hold-out samples $\mathcal{S}_k$ is computed; this procedure is repeated for $k = 1,\ldots,K$ and its mean $\widehat{J}^{(K\text{-CV})}$ is outputted:

$$
\widehat{J}^{(K\text{-CV})} := \tfrac{1}{K}\textstyle\sum_{k=1}^K \left(\frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}^\top \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k} - \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}\right). \tag{9}
$$

For model selection, we compute $\widehat{J}^{(K\text{-CV})}$ for all model candidates (the basis function $\boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z})$ and the regularization parameter λ) and choose the best model that minimizes $\widehat{J}^{(K\text{-CV})}$. We can show that $\widehat{J}^{(K\text{-CV})}$ is an almost unbiased estimator of the objective function J, where 'almost'-ness comes from the fact that the sample size is reduced in the CV procedure due to data splitting (Schölkopf & Smola, 2002).

For the parametric setup, we can derive an asymptotic unbiased estimator of J (a.k.a. an *information criterion* (Akaike, 1974)) based on Theorem 1, which could be employed for model selection. However, we omit the detail due to lack of space.

2.6 Design of Basis Functions

The above CV procedure would be useful when good candidates of basis functions are prepared. Here we propose to use the *product kernel* of the following form as basis functions:

$$\varphi_\ell(\boldsymbol{y}, \boldsymbol{z}) = \phi^{\mathrm{y}}_\ell(\boldsymbol{y})\phi^{\mathrm{z}}_\ell(\boldsymbol{z})$$

since the number of kernel evaluation when computing $\widehat{H}_{\ell,\ell'}$ is reduced from n^2 to $2n$:

$$\widehat{H}_{\ell,\ell'} = \tfrac{1}{n^2}\big(\textstyle\sum_{i=1}^n \phi^{\mathrm{y}}_\ell(\boldsymbol{y}_i)\phi^{\mathrm{y}}_{\ell'}(\boldsymbol{y}_i)\big)\big(\sum_{j=1}^n \phi^{\mathrm{z}}_\ell(\boldsymbol{z}_j)\phi^{\mathrm{z}}_{\ell'}(\boldsymbol{z}_j)\big).$$

In the regression scenarios where $\boldsymbol{y}$ is continuous, we use the Gaussian kernel as the 'base' kernels: $\phi^{\mathrm{y}}_\ell(\boldsymbol{y}) := \exp\big(-\|\boldsymbol{y} - \boldsymbol{u}_\ell\|^2/(2\sigma^2)\big)$ and $\phi^{\mathrm{z}}_\ell(\boldsymbol{z}) := \exp\big(-\|\boldsymbol{z} - \boldsymbol{v}_\ell\|^2/(2\sigma^2)\big)$, where $\{(\boldsymbol{u}_\ell, \boldsymbol{v}_\ell)\}_{\ell=1}^b$ are Gaussian centers randomly chosen from $\{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^n$—more precisely, we set $\boldsymbol{u}_\ell := \boldsymbol{y}_{c(\ell)}$ and $\boldsymbol{v}_\ell := \boldsymbol{z}_{c(\ell)}$, where $\{c(\ell)\}_{\ell=1}^b$ are randomly chosen from $\{1, \ldots, n\}$ without replacement.

The rationale behind this basis function choice is as follows: The density ratio tends to take large values if $p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})$ is small and $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$ is large (and vise versa). When a non-negative function is approximated by a Gaussian kernel model, many kernels may be needed in the region where the output of the target function is large; on the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero. Following this heuristic, we decided to allocate many kernels in the regions where $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$ is large; this can be achieved by setting the Gaussian centers at[3] $\{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^n$.

In the classification scenarios where $\boldsymbol{y}$ is categorical, we use the *delta kernel* for $\boldsymbol{y}$: $\phi^{\mathrm{y}}_\ell(\boldsymbol{y}) := \delta(\boldsymbol{y} = \boldsymbol{u}_\ell)$, where $\delta(\boldsymbol{y} = \boldsymbol{u}_\ell)$ is 1 if $\boldsymbol{y} = \boldsymbol{u}_\ell$ and 0 otherwise. More generally, when $\boldsymbol{y}$ is structured (e.g., strings, trees, and graphs), we may employ kernels for structured data (Gärtner, 2003) as $\phi^{\mathrm{y}}_\ell(\boldsymbol{y})$.

[3] Alternatively, we may locate n^2 Gaussian kernels at $\{(\boldsymbol{y}_i, \boldsymbol{z}_j)\}_{i,j=1}^n$. However, in our preliminary experiments, this did not further improve the performance, but significantly increased the computational cost.

3 Numerical Experiments

In this section, we experimentally investigate the performance of the proposed and existing dimension reduction methods using artificial and real datasets. In the proposed method, we use the Gaussian kernel as basis functions and employ the regularized kernel Gram matrix as the regularization matrix $\boldsymbol{R}$: $\boldsymbol{R} = \widetilde{\boldsymbol{K}} + \epsilon \boldsymbol{I}_b$, where $\widetilde{\boldsymbol{K}}$ is the kernel Gram matrix for the chosen centers: $\widetilde{K}_{\ell,\ell'} := \phi^{\mathrm{y}}_\ell(\boldsymbol{u}_{\ell'})\phi^{\mathrm{z}}_\ell(\boldsymbol{v}_{\ell'})$. $\epsilon \boldsymbol{I}_b$ is added to $\widetilde{\boldsymbol{K}}$ for avoiding non-degeneracy; we set $\epsilon = 0.01$. We fix the number of basis functions at $b = \min(100, n)$, and choose the Gaussian width σ and the regularization parameter λ based on 5-fold CV with grid search. we restart the natural gradient search 10 times with random initial points and choose the one having the minimum CV score (9).

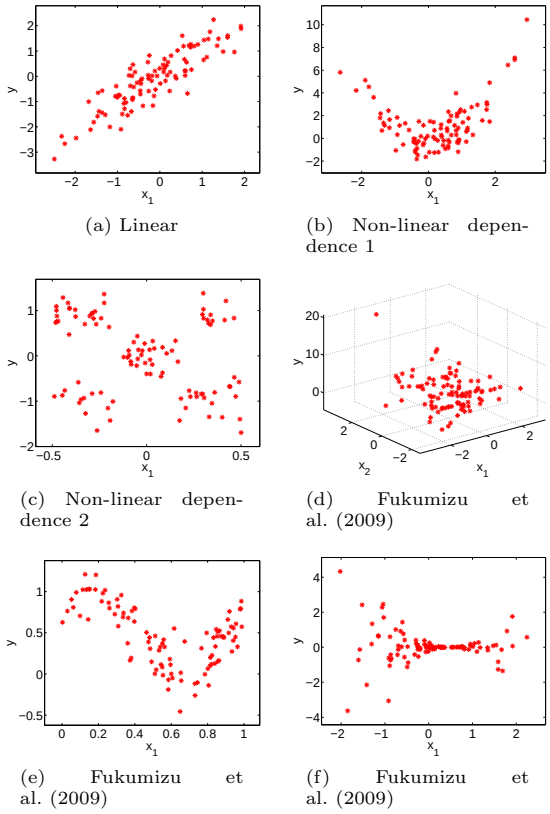
3.1 Dimension Reduction for Artificial Datasets

We use 6 artificial datasets (3 datasets designed by us and 3 datasets borrowed from Fukumizu et al. (2009); see Figure 2), and compare LDR with kernel dimension reduction (KDR) (Fukumizu et al., 2009), the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005), sliced inverse regression (SIR) (Li, 1991), and sliced average variance estimation (SAVE) (Cook, 2000). In KDR and HSIC, the Gaussian width is set to the median sample distance, following the suggestions in the original papers (Gretton et al., 2005; Fukumizu et al., 2009). We evaluate the performance of each method by

$$\tfrac{1}{\sqrt{2d}}\|\widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}} - \boldsymbol{W}^{*\top}\boldsymbol{W}^*\|_{\mathrm{Frobenius}},$$

where $\|\cdot\|_{\mathrm{Frobenius}}$ denotes the Frobenius norm, $\widehat{\boldsymbol{W}}$ is an estimated projection matrix, and $\boldsymbol{W}^*$ is the optimal projection matrix. Note that the above error measure takes its value in $[0, 1]$.

The performance of each method is summarized in Table 2, which depicts the mean and standard deviation of the above Frobenius-norm error over 50 trials when the number of samples is $n = 100$. LDR overall shows good performance; in particular, it performs the best for datasets (b), (c), and (e). KDR also tends to work reasonably well, but it sometimes performs poorly; this seems to be caused by the inappropriate choice of the Gaussian kernel width, implying that the heuristic of using the median sample distance as the kernel width is



(a) Linear (b) Non-linear dependence 1

(c) Non-linear dependence 2 (d) Fukumizu et al. (2009)

(e) Fukumizu et al. (2009) (f) Fukumizu et al. (2009)

Figure 2: Artificial datasets.

not always appropriate. On the other hand, LDR with CV performs stably well for various types of datasets.

3.2 Classification for Benchmark Datasets

Finally, we evaluate the classification performance after dimension reduction for several benchmark datasets. We use 'image', 'waveform', 'pima-indians-diabetes', and 'letter recognition' in the UCI repository. We randomly choose 200 samples from the dataset and apply LDR, KDR, and HSIC to obtain projections onto low-dimension subspaces with $d = \lceil m/4 \rceil$, $\lceil m/2 \rceil$, and $\lceil 3m/4 \rceil$. Then we train the support vector machine on the projected 200 training samples.

The generalization error is computed for the samples not used for training. Table 3 summarizes the mean and standard deviation of the classification error over 20 iterations. This shows that the proposed method overall compares favorably with the other methods.

4 Conclusions

In this paper, we proposed a new dimension reduction method utilizing a squared-loss variant of mutual information (SMI). The proposed method inherits several preferable properties of the SMI estimator, e.g., density estimation is not involved, it is distribution-free, and model selection by cross validation is available. The effectiveness of the proposed method over existing methods was shown through experiments.

Table 2: Mean and standard deviation of Frobenius-norm error for toy datasets. The best method in terms of the mean error and comparable ones based on the one-sided t-test at the significance level 1% are indicated by boldface.

	m	d	LDR	KDR	HSIC	SIR	SAVE
(a)	5	1	**.13(.04)**	.13(.05)	.17(.07)	**.11(.04)**	.19(.10)
(b)	5	1	**.15(.06)**	.25(.21)	.44(.36)	.83(.20)	.24(.08)
(c)	5	1	**.10(.05)**	.44(.32)	.68(.32)	.85(.20)	.31(.11)
(d)	4	2	**.20(.14)**	**.16(.06)**	**.18(.08)**	.29(.16)	.40(.17)
(e)	4	1	**.09(.06)**	.13(.06)	.16(.07)	.20(.10)	.21(.14)
(f)	10	1	**.35(.12)**	**.40(.12)**	.49(.17)	.64(.22)	.73(.20)

Table 3: Mean and standard deviation of misclassification rates for benchmark datasets.

	d	LDR	KDR	HSIC
image	5	**.078(.017)**	.115(.034)	.154(.040)
	9	**.091(.015)**	**.100(.031)**	.107(.025)
	14	**.090(.019)**	**.086(.017)**	**.088(.019)**
waveform	6	**.133(.015)**	**.132(.013)**	.157(.017)
	11	**.123(.012)**	.138(.013)	.162(.013)
	16	**.117(.009)**	.137(.010)	.159(.016)
pima	2	**.249(.022)**	**.246(.018)**	**.252(.021)**
	4	**.251(.019)**	**.255(.020)**	**.261(.027)**
	6	**.249(.021)**	**.246(.021)**	**.253(.021)**
letter (a, b & c)	4	**.029(.007)**	**.025(.009)**	.034(.010)
	8	.026(.008)	**.018(.009)**	**.019(.007)**
	12	**.015(.007)**	**.015(.007)**	**.015(.007)**

Acknowledgements

T.S. was supported in part by Global COE Program "The research and training center for new development in mathematics", MEXT, Japan. M.S. was supported by AOARD, SCAT, and the JST PRESTO program.

Proof of Lemma 1

Let $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{z}_\perp)$. By $d\boldsymbol{x} = d\boldsymbol{z}d\boldsymbol{z}_\perp$, we have

$$
\begin{aligned}
&I_s(X,Y) - I_s(Z,Y) \\
=&\frac{1}{2}\int \left(\frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})}\right)^2 p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \\
&- \frac{1}{2}\int \left(\frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{z}}(\boldsymbol{z})p_{\mathrm{y}}(\boldsymbol{y})}\right)^2 p_{\mathrm{z}}(\boldsymbol{z})p_{\mathrm{y}}(\boldsymbol{y})d\boldsymbol{z}d\boldsymbol{y} \\
=&\frac{1}{2}\int \left(\frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})} - \frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{z}}(\boldsymbol{z})p_{\mathrm{y}}(\boldsymbol{y})}\right)^2 p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{y},
\end{aligned}
$$

because $\int \frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})} p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = \int \left(\frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}\right)^2 p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})d\boldsymbol{y}d\boldsymbol{z}$. Noticing that $\frac{p_{\mathrm{xy}}}{p_{\mathrm{x}}p_{\mathrm{y}}} = \frac{p_{\mathrm{xy|z}}}{p_{\mathrm{x|z}}p_{\mathrm{y|z}}}\frac{p_{\mathrm{yz}}}{p_{\mathrm{y}}p_{\mathrm{z}}}$ where we used the relation $p_{\mathrm{x}}(\boldsymbol{x}) = p_{\mathrm{x|z}}(\boldsymbol{x}|\boldsymbol{z})p_{\mathrm{z}}(\boldsymbol{z})$, Then we have

$$
I_s(X,Y) - I_s(Z,Y) = \frac{1}{2}\int \left(1 - \frac{p_{\mathrm{xy|z}}}{p_{\mathrm{x|z}}p_{\mathrm{y|z}}}\right)^2 \frac{{p_{\mathrm{yz}}}^2 p_{\mathrm{x}}}{{p_{\mathrm{z}}}^2 p_{\mathrm{y}}} d\boldsymbol{x}d\boldsymbol{y}.
$$

■

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723.

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, *28*, 131–142.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251–276.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Chiaromonte, F., & Cook, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, *54*, 768–795.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, *29*, 2109–2121.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. N. Y.: John Wiley & Sons, Inc.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 229–318.

Darbellay, G. A., & Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, *45*, 1315–1321.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, *5*, 73–99.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, *37*, 1871–1905.

Gärtner, T. (2003). A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, *5*, 49–58.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory* (pp. 63–77). Berlin: Springer-Verlag.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hulle, M. M. V. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation*, *17*, 1903–1910.

Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*, 066138.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of American Statistical Association*, *86*, 316–342.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2008). Estimating divergence functions and the likelihood ratio by penalized convex risk minimization. *Advances in Neural Information Processing Systems 20* (pp. 1089–1096). Cambridge, MA: MIT Press.

Nishimori, Y., & Akaho, S. (2005). Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, *67*, 106–135.

Patriksson, M. (1999). *Nonlinear programming and variational inequality problems*. Dredrecht: Kluwer Academic.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., & Bedo, J. (2007). Supervised feature selection via dependence estimation. *Proceedings of the 24th International Conference on Machine learning* (pp. 823–830). New York, NY, USA: ACM.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, *35*, 575–607.

Suzuki, T., Sugiyama, M., Kanamori, T., & Sese, J. (2009). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, *10*, S52.

Suzuki, T., Sugiyama, M., Sese, J., & Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *JMLR Workshop and Conference Proceedings*, *4*, 5–20.

Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, *3*, 1415–1438.

van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer, New York.