Neural Computation. vol.25, no.3, pp.725–758, 2013.

Sufficient Dimension Reduction via Squared-loss Mutual Information Estimation

Taiji Suzuki
Department of Mathematical Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
t-suzuki@mist.i.u-tokyo.ac.jp

Masashi Sugiyama
Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
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 contains the whole information about the output values that the input features possess. In this paper, we propose a novel sufficient dimension reduction method using a squared-loss variant of mutual information as a dependency measure. We apply a density-ratio estimator for approximating squared-loss mutual information that is formulated as a minimum contrast estimator on parametric or non-parametric models. Since cross-validation is available for choosing an appropriate model, our method does not require any pre-specified structure on the underlying distributions. We elucidate the asymptotic bias of our estimator on parametric models and the asymptotic convergence rate on non-parametric models. The convergence analysis utilizes the uniform tail-bound of a U-process, and the convergence rate is characterized by the bracketing entropy of the model. We then develop a natural gradient algorithm on the Grassmann manifold for sufficient subspace search. The analytic formula of our estimator allows us to compute the gradient efficiently. Numerical experiments show that the proposed method compares favorably with existing dimension reduction approaches on artificial and benchmark datasets.

Keywords: dimension reduction, density ratio, convergence rate, mutual information.

1 Introduction

The purpose of *dimension reduction* in supervised learning is to find the low-dimensional subspace of input features which has 'sufficient' information for predicting output values. *Sufficient dimension reduction* (SDR) initiated by Li (1991) 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, 1998b; Chiaromonte & Cook, 2002). Such a low-dimensional subspace contains all the information of the output that the covariate contains. Finding such a subspace not only allows us to use the dimension-reduced features for estimating input-output relations, but also gives insights about which features are important.

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 by making use of the reproducing property of RKHSs (Aronszajn, 1950). 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 *conditional covariance* using the kernel trick, and thus KDR directly performs SDR. Through experiments, KDR was demonstrated to outperform other dimension reduction schemes such as *canonical correlation analysis* (Hotelling, 1936; Breiman & Friedman, 1985), *partial least-squares* (Wold, 1966; Goutis & Fearn, 1996; Durand & Sabatier, 1997; Reiss & Ogden, 2007), *sliced inverse regression* (Li, 1991; Bura & Cook, 2001; Cook & Ni, 2005; Zhu et al., 2006), and the *principal Hessian direction* (Li, 1992; Cook, 1998a; Li et al., 2000). Theoretical properties of KDR such as consistency have been studied thoroughly (Fukumizu et al., 2009). However, KDR still has a weakness in practice—the performance of KDR (and also 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].

[1] In principle, it is possible to choose the Gaussian width and the regularization parameter by cross-validation (CV) over a successive predictor. However, this is not preferable due to the following two

Table 1: Summary of existing and proposed dependency measures.

Methods	Non-linear dependence	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**

Another possible criterion for SDR is *mutual information* (MI) (Cover & Thomas, 2006). MI could be directly employed for SDR since maximizing MI between output and projected input leads to conditional independence between output and input given the projected input. So far, 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 practical advantages.

As summarized in Table 1, MLMI affords model selection by cross-variation, while there is no systematic method to choose tuning parameters for HSIC, KDR, and NN. MLMI does not require specific structures on the underlying distributions, while EDGE requires that the distributions are near Gaussian. MLMI does not involve density estimation of the underlying distributions so that it shows a good performance in practice.

Based on the above comparison, we want to employ the MLMI method for dimension reduction. However, this is not straightforward 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 (or feature extraction) 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 use an estimator of SMI called *least-squares MI* (LSMI) (Suzuki et al., 2009) for dimension reduction. LSMI inherits good properties from MLMI, and moreover it provides an *analytic* SMI estimator that gives an analytic formula for its derivative (see Table 1 again).

The goal of this paper is to develop a dimension reduction algorithm based on LSMI.

reasons. The first is significant increase of the computational cost. When CV is used, the tuning parameters in KDR (or HSIC) and hyper-parameters in the target predictor (such as basis parameters and the regularization parameter) should be optimized at the same time. This results in a deeply nested CV procedure, and therefore this could be computationally very expensive. Another reason is that features extracted based on CV are no longer independent of predictors, which is not preferable from the viewpoint of interpretability.

Our first contribution in this paper is to theoretically elucidate the rate of convergence of the LSMI estimator in parametric and non-parametric settings. Then we develop a practical dimension reduction algorithm based on LSMI, which we call *Least-Squares Dimension Reduction* (LSDR). LSDR optimizes the projection matrix using a *natural gradient* algorithm (Amari, 1998) on the *Grassmann manifold*. Finally, through numerical experiments, we show the usefulness of the LSDR method.

2 Dimension Reduction via SMI Estimation

In this section, we first formulate the problem of *sufficient dimension reduction* (SDR) (Cook, 1998b; Chiaromonte & Cook, 2002), and show how *squared-loss mutual information* (SMI) can be employed in the context of SDR. Then we introduce a method of approximating SMI without going through density estimation, and we elucidate convergence properties of the SMI estimator. Finally, we develop a dimension reduction method based on the SMI estimator.

2.1 Sufficient Dimension Reduction

Let $\mathcal{D}_{\mathrm{X}}$ $(\subset \mathbb{R}^d)$ be the domain of input feature $\boldsymbol{x}$, and $\mathcal{D}_{\mathrm{Y}}$ be the domain of output data[2] $\boldsymbol{y}$. We suppose that $\mathcal{D}_{\mathrm{Y}}$ is equipped with a σ-algebra $\mathcal{B}_Y$ and there is a base measure denoted by $\mathrm{d}\boldsymbol{y}$. As for $\mathcal{D}_{\mathrm{X}}$, we consider the standard σ-algebra $\mathcal{B}_X$ of the Lebesgue measurable sets and denote by $\mathrm{d}\boldsymbol{x}$ as the Lebesgue measure. We assume there is a joint density $p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})$ defined on the product space $(\mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}, \mathcal{B}_X \times \mathcal{B}_Y)$ with respect to $\mathrm{d}\boldsymbol{x} \times \mathrm{d}\boldsymbol{y}$.

To search a subspace of input space containing sufficient information about the output, we utilize the *Grassmann manifold* $\mathrm{Gr}_m^d(\mathbb{R})$ that is the set of all m-dimensional subspaces in $\mathbb{R}^d$. The Grassmann manifold $\mathrm{Gr}_m^d(\mathbb{R})$ is obtained by identifying those matrices in $d \times m$ orthonormal matrices whose columns span the same subspace:

$$\mathrm{Gr}_m^d(\mathbb{R}) := \{\boldsymbol{W} \in \mathbb{R}^{m\times d} \mid \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_m\}/\sim, \tag{1}$$

where $^\top$ denotes the transpose, $\boldsymbol{I}_m$ is the m-dimensional identity matrix, and $\sim$ is the equivalence relation such that $\boldsymbol{W} \sim \boldsymbol{W}'$ if the raws of both $\boldsymbol{W}$ and $\boldsymbol{W}'$ span the same space.

Let $\boldsymbol{W}^*$ be any projection matrix corresponding to a member of the Grassmann manifold $\mathrm{Gr}_m^d(\mathbb{R})$. Let $\boldsymbol{z}^*$ $(\in \mathbb{R}^m)$ be the orthogonal projection of input $\boldsymbol{x}$ given by $\boldsymbol{W}^*$:

$$\boldsymbol{z}^* = \boldsymbol{W}^*\boldsymbol{x}.$$

Suppose that $\boldsymbol{z}^*$ satisfies

$$\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} \mid \boldsymbol{z}^*. \tag{2}$$

[2] $\mathcal{D}_{\mathrm{Y}}$ can be multi-dimensional and either continuous (i.e., regression) or categorical (i.e., classification); structured outputs such as strings, trees, and graphs can also be handled in our framework, as explained later.

That is, given the projected feature $\boldsymbol{z}^*$, the (remaining) feature $\boldsymbol{x}$ is conditionally independent of output $\boldsymbol{y}$ and thus can be discarded without sacrificing the predictability of $\boldsymbol{y}$. Note that the conditionally independence is invariant against the choice of the representative $\boldsymbol{W}^*$.

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

$$D^n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid \boldsymbol{x}_i \in \mathcal{D}_{\mathrm{X}},\ \boldsymbol{y}_i \in \mathcal{D}_{\mathrm{Y}},\ i = 1, \ldots, n\}, \tag{3}$$

drawn from a joint distribution with density $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$. The goal of SDR is, from data D^n, to find a projection matrix whose range agrees with that of $\boldsymbol{W}^*$. For a projection matrix $\boldsymbol{W}$, we write

$$\boldsymbol{z}_i = \boldsymbol{W}\boldsymbol{x}_i.$$

We assume that m is known throughout this paper.

2.2 Squared-Loss Mutual Information

A direct approach to SDR would be to determine $\boldsymbol{W}$ so that Eq.(2) is fulfilled. Let us denote by $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$ for some projection matrix $\boldsymbol{W}$. To this end, we adopt SMI as our criterion to be maximized with respect to $\boldsymbol{W}$:

$$\mathrm{SMI}(Y, Z) := \frac{1}{2} \int \left(\frac{p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y}) p_{\mathrm{z}}(\boldsymbol{z})} - 1 \right)^2 p_{\mathrm{y}}(\boldsymbol{y}) p_{\mathrm{z}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z}, \tag{4}$$

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. $\mathrm{SMI}(Y, Z)$ allows us to evaluate independence between $\boldsymbol{y}$ and $\boldsymbol{z}$ since $\mathrm{SMI}(Y, Z)$ 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.(4) 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 (KL) divergence (Kullback & Leibler, 1951):

$$\mathrm{MI}(Y, Z) := \int \log \left(\frac{p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y}) p_{\mathrm{z}}(\boldsymbol{z})} \right) p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z}.$$

Thus SMI could be regarded as a natural alternative to ordinary MI.

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

Lemma 1. *Decompose $\boldsymbol{x}$ into $\boldsymbol{z}$ and the component orthogonal to $\boldsymbol{z}$ as $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{z}_\perp)$, i.e., $\boldsymbol{z}$ is a member of the image of $\boldsymbol{W}$ and $\boldsymbol{z}_\perp$ is a member of the subspace perpendicular to* *the image of* $\boldsymbol{W}$*. Let* $p_{\mathrm{z}_\perp \mathrm{y}|\mathrm{z}}(\boldsymbol{z}_\perp, \boldsymbol{y}|\boldsymbol{z})$*,* $p_{\mathrm{z}_\perp|\mathrm{z}}(\boldsymbol{z}_\perp|\boldsymbol{z})$*, and* $p_{\mathrm{y}|\mathrm{z}}(\boldsymbol{y}|\boldsymbol{z})$ *be conditional densities. Then we have*

$$\begin{aligned}&\mathrm{SMI}(X,Y) - \mathrm{SMI}(Z,Y)\\ &= \frac{1}{2}\int \left(1 - \frac{p_{\mathrm{z}_\perp \mathrm{y}|\mathrm{z}}(\boldsymbol{z}_\perp, \boldsymbol{y}|\boldsymbol{z})}{p_{\mathrm{z}_\perp|\mathrm{z}}(\boldsymbol{z}_\perp|\boldsymbol{z})p_{\mathrm{y}|\mathrm{z}}(\boldsymbol{y}|\boldsymbol{z})}\right)^2 \frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})^2 p_{\mathrm{x}}(\boldsymbol{x})}{p_{\mathrm{z}}(\boldsymbol{z})^2 p_{\mathrm{y}}(\boldsymbol{y})}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{z}_\perp\mathrm{d}\boldsymbol{y}\\ &\geq 0.\end{aligned}$$

A proof of this lemma is given in Appendix A. Lemma 1 implies

$$\mathrm{SMI}(X,Y) \geq \mathrm{SMI}(Z,Y),$$

and the equality holds if and only if

$$p_{\mathrm{z}_\perp \mathrm{y}|\mathrm{z}}(\boldsymbol{z}_\perp, \boldsymbol{y}|\boldsymbol{z}) = p_{\mathrm{z}_\perp|\mathrm{z}}(\boldsymbol{z}_\perp|\boldsymbol{z})p_{\mathrm{y}|\mathrm{z}}(\boldsymbol{y}|\boldsymbol{z}),$$

which is equivalent to Eq.(2). Thus, Eq.(2) can be achieved by maximizing $\mathrm{SMI}(Z,Y)$ with respect to $\boldsymbol{W}$; then the 'sufficient' subspace can be identified.

Now we want to find the projection matrix $\boldsymbol{W}$ that maximizes $\mathrm{SMI}(Z,Y)$. 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. Below, we introduce an SMI estimator.

2.3 SMI Approximation via Density-Ratio Estimation

Here, we consider a fixed projection matrix $\boldsymbol{W}$, and discuss the problem of approximating SMI from samples. The *convex duality* (Boyd & Vandenberghe, 2004, p.91) gives the variational representation (Keziou, 2003; Nguyen et al., 2010) of SMI as

$$\mathrm{SMI}(Y,Z) = -\inf_g J(g) - \frac{1}{2},$$

where $\inf_g$ is taken over all measurable functions on $(\mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}, \mathcal{B}_X \times \mathcal{B}_Y)$, and

$$J(g) = \frac{1}{2}\int g(\boldsymbol{y},\boldsymbol{z})^2 p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z} - \int g(\boldsymbol{y},\boldsymbol{z})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}. \tag{5}$$

This can be checked as follows: For $f(u) = \frac{1}{2}(u^2 - 1)$, we have

$$\begin{aligned}\mathrm{SMI}(Y,Z) &= \int f\left(\frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}\right)p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\ &= \sup_g \int p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})g(\boldsymbol{y},\boldsymbol{z}) - f^*\left(g(\boldsymbol{y},\boldsymbol{z})\right)p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\ &= \sup_g \int p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})g(\boldsymbol{y},\boldsymbol{z}) - \frac{1}{2}g^2(\boldsymbol{y},\boldsymbol{z})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z} - \frac{1}{2}\\ &= -\inf_g J(g) - \frac{1}{2},\end{aligned}$$

where f^* is the convex conjugate of f that satisfies $f(u) = \sup_{v\in\mathbb{R}}\{uv - f^*(v)\}$ (Boyd & Vandenberghe, 2004). Thus computing SMI is reduced to finding the minimizer g^* of $J(g)$. We can show that g^* is given by

$$g^*(\boldsymbol{y},\boldsymbol{z}) := \frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}. \tag{6}$$

Thus, estimating $\mathrm{SMI}(Y,Z)$ is reduced to estimating the above density ratio[3]. We do not choose a strategy to plug in density estimators of p_{yz}, p_{y}, and p_{z} into the formula (6). This is because, in a region with small $p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})$, small estimation error of $p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})$ is strongly amplified. To avoid the unstable behavior around the tail, we directly model the density ratio g^* itself and impose regularization to control instability of density-ratio estimators when needed.

Below, we consider parametric and non-parametric methods for estimating SMI.

2.3.1 Parametric Convergence Analysis

Let us consider the case where the function class $\mathcal{G}$ from which the function g is searched 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 $g^* = g_{\boldsymbol{\theta}^*}$. Approximating the probability densities $p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})$, $p_{\mathrm{y}}(\boldsymbol{y})$, and $p_{\mathrm{z}}(\boldsymbol{z})$ in Eq.(5) by their empirical distributions, we obtain the following optimization problem.

$$\widehat{\boldsymbol{\theta}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta} \left[\frac{1}{2n^2}\sum_{i,j=1}^{n} g_{\boldsymbol{\theta}}(\boldsymbol{y}_i,\boldsymbol{z}_j)^2 - \frac{1}{n}\sum_{i=1}^{n} g_{\boldsymbol{\theta}}(\boldsymbol{y}_i,\boldsymbol{z}_i)\right]. \tag{7}$$

Then an SMI approximator $\widehat{\mathrm{SMI}}(Y,Z)$ can be constructed as

$$\widehat{\mathrm{SMI}}(Y,Z) := \frac{1}{n}\sum_{i=1}^{n} g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_i,\boldsymbol{z}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n} g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_i,\boldsymbol{z}_j)^2 - \frac{1}{2}. \tag{8}$$

Suppose the standard regularity conditions for the consistency $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| \xrightarrow{p} 0$ is satisfied (see, for example, Section 3.2.1 of van der Vaart & Wellner, 1996). Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $b\times b$ matrices defined as

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

[3] This result can be generalized to a general f-divergence with small modification (Keziou, 2003; Nguyen et al., 2010).

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

Theorem 1. *Suppose that the matrix $\boldsymbol{A}$ is positive definite, then the SMI estimator* (8) *satisfies*

$$\widehat{\mathrm{SMI}}(Y,Z) - \mathrm{SMI}(Y,Z) = \mathcal{O}_p(n^{-1/2}), \tag{9}$$

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

$$\mathrm{E}_{D^n}[\widehat{\mathrm{SMI}}(Y,Z) - \mathrm{SMI}(Y,Z)] = \frac{1}{2n}\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{B}) + o(n^{-1}), \tag{10}$$

where E_{D^n} denotes the expectation over data samples D^n (see Eq.(3)*).*

A proof of Theorem 1 can be found in Appendix B. This theorem means that the above SMI estimator retains the optimality in terms of the order of convergence in n, since $\mathcal{O}_p(n^{-1/2})$ is the optimal convergence rate in the parametric setup (van der Vaart, 2000).

2.3.2 Non-Parametric Convergence Analysis

Next, we consider non-parametric cases. Let the function class $\mathcal{G}$ be a general set of functions on $\mathcal{D}_\mathrm{Y} \times \mathcal{D}_\mathrm{Z}$, where $\mathcal{D}_\mathrm{Z} = \boldsymbol{W}\mathcal{D}_\mathrm{X}$. Let us consider a non-parametric version of the empirical problem (cf. Eq.(7)):

$$\widehat{g} := \mathop{\mathrm{argmin}}_{g\in\mathcal{G}} \left[\frac{1}{2n^2}\sum_{i,j=1}^{n} g(\boldsymbol{y}_i,\boldsymbol{z}_j)^2 - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{y}_i,\boldsymbol{z}_i) + \frac{\lambda_n}{2}R(g)^2\right], \tag{11}$$

where $R(g)$ is a non-negative regularization functional such that

$$\sup_{\boldsymbol{y},\boldsymbol{z}}[g(\boldsymbol{y},\boldsymbol{z})] \leq R(g). \tag{12}$$

Then a non-parametric version of SMI approximator $\widehat{\mathrm{SMI}}(Y,Z)$ is given as

$$\widehat{\mathrm{SMI}}(Y,Z) := \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{y}_i,\boldsymbol{z}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n}\widehat{g}(\boldsymbol{y}_i,\boldsymbol{z}_j)^2 - \frac{1}{2}.$$

A useful example is to use a *reproducing kernel Hilbert space* (RKHS) (Aronszajn, 1950) as $\mathcal{G}$ and the RKHS norm as $R(g)$. Suppose $\mathcal{G}$ is an RKHS associated with bounded kernel $k(\cdot,\cdot)$:

$$\sup_{\boldsymbol{y},\boldsymbol{z}}[k((\boldsymbol{y},\boldsymbol{z}),(\boldsymbol{y},\boldsymbol{z}))] \leq C. \tag{13}$$

Let $\|\cdot\|_\mathcal{G}$ denote the norm in the RKHS $\mathcal{G}$. Then $R(g) = \sqrt{C}\|g\|_\mathcal{G}$ satisfies Eq.(12):

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

where we used the reproducing property of the kernel and Schwartz's inequality. Note that the Gaussian kernel satisfies this with $C = 1$. It is known that Gaussian kernel RKHS spans a dense subset in the set of continuous functions. Another example of RKHSs is Sobolev space. The canonical norm for this space is the integral of the squared derivatives of functions. Thus the regularization term $R(g) = \|g\|_{\mathcal{G}}$ imposes the solution to be smooth. The RKHS technique in Sobolev space has been well exploited in the context of spline models (Wahba, 1990). We intend that the regularization term $R(g)$ is a generalization of the RKHS norm. Roughly speaking, $R(g)$ is like a "norm" of the function space $\mathcal{G}$.

We assume that the true density-ratio function $g^*(\boldsymbol{y}, \boldsymbol{z})$ is contained in the model $\mathcal{G}$ and is bounded from above:

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

Let $\mathcal{G}_M$ be a *ball* of $\mathcal{G}$ with radius $M > 0$:

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

To derive the convergence rate of our estimator, we utilize the *bracketing entropy* that is a complexity measure of a function class (van der Vaart & Wellner, 1996, p. 83).

Definition 1. *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l(\boldsymbol{y}, \boldsymbol{z}) \leq f(\boldsymbol{y}, \boldsymbol{z}) \leq u(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{y}$ and $\boldsymbol{z}$. An ϵ-bracket is a bracket $[l, u]$ with $\|l - u\|_{L_2(p_{\mathrm{y}}p_{\mathrm{z}})} < \epsilon$. The bracketing entropy $\mathcal{H}_{[]}(\mathcal{F}, \epsilon, L_2(p_{\mathrm{y}}p_{\mathrm{z}}))$ is the logarithm of the minimum number of ϵ-brackets needed to cover a function set $\mathcal{F}$.*

We assume that there exists γ $(0 < \gamma < 2)$ such that, for all $M > 0$,

$$\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p_{\mathrm{y}}p_{\mathrm{z}})) = O\left(\left(\frac{M}{\epsilon}\right)^{\gamma}\right). \tag{14}$$

This quantity represents a complexity of function class $\mathcal{G}$—the larger γ is, the more complex the function class $\mathcal{G}$ is because, for larger γ, more brackets are needed to cover the function class. Gaussian RKHS satisfies this condition for arbitrarily small γ (Steinwart & Scovel, 2007). Note that when $R(g)$ is the RKHS norm, the condition (14) holds for all $M > 0$ if that holds for $M = 1$.

Then we have the following theorem.

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

$$\widehat{\mathrm{SMI}}(Y, Z) - \mathrm{SMI}(Y, Z) = \mathcal{O}_p(\max(\lambda_n, n^{-1/2})). \tag{15}$$

A proof of Theorem 2 can be found in Appendix C. The conditions $\lambda_n \to 0$ and $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$ roughly mean that the regularization parameter λ_n should be sufficiently small but not too small. This theorem shows that the convergence rate of the non-parametric version is also $\mathcal{O}_p(n^{-1/2})$ if we take λ_n as $n^{-2/(2+\gamma)+\epsilon} \lesssim \lambda_n \lesssim n^{-1/2}$ for sufficiently small ϵ. However, the non-parametric method requires a milder model assumption.

According to Nguyen et al. (2010) 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 the above non-parametric method would also be optimal in the same sense.

As stated above, Gaussian RKHS satisfies the bracketing number condition (14) for arbitrary $\gamma > 0$. Thus SMI with the Gaussian kernel achieves the convergence rate (15) with arbitrary $\gamma > 0$. However Gaussian RKHS is not sufficiently rich to estimate a function in Sobolev spaces with the optimal rate. To estimate a function in a Sobolev space with the optimal rate, we need to adjust the Gaussian width appropriately depending on the sample size and the regularity of the Sobolev space. To analyze the convergence rate for varying Gaussian widths, another technique than that used in this paper is required. What we have shown in the theorem works only for a fixed Gaussian width. To cope with a situation of varying Gaussian width, the techniques recently developed by Eberts and Steinwart (2011) are useful.

2.3.3 Practical Implementation

Here we describe a practical version of the above SMI approximation method, called *least-squares mutual information* (LSMI) (Suzuki et al., 2009).

Let us restrict the search space of function g to some linear subspace $\mathcal{G}$:

$$\mathcal{G} := \{\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z}) \mid \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top \in \mathbb{R}^b\}, \tag{16}$$

where $\boldsymbol{\alpha}$ is a parameter to be learned from samples. $\boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z})$ is a basis function such that

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

where $\boldsymbol{0}_b$ is the b-dimensional vector with all zeros, and the inequality for vectors is applied in the element-wise manner. $\boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z})$ may be dependent on the samples $\{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^n$, i.e., *kernel* models are also allowed. Later in Section 2.3.5, we explain how the basis functions $\boldsymbol{\varphi}(\boldsymbol{y}, \boldsymbol{z})$ are designed.

Let us approximate the probability densities $p_{\mathrm{yz}}(\boldsymbol{y}, \boldsymbol{z})$, $p_{\mathrm{y}}(\boldsymbol{y})$, and $p_{\mathrm{z}}(\boldsymbol{z})$ in Eq.(5) by their empirical distributions. Then we have

$$\widehat{\boldsymbol{\alpha}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha} \in \mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{R} \boldsymbol{\alpha}\right], \tag{17}$$

where we included $\lambda \boldsymbol{\alpha}^\top \boldsymbol{R} \boldsymbol{\alpha}$ $(\lambda > 0)$ for regularization purposes, and

$$\widehat{\boldsymbol{H}} := \frac{1}{n^2} \sum_{i,j=1}^n \boldsymbol{\varphi}(\boldsymbol{y}_i, \boldsymbol{z}_j) \boldsymbol{\varphi}(\boldsymbol{y}_i, \boldsymbol{z}_j)^\top, \quad \widehat{\boldsymbol{h}} := \frac{1}{n} \sum_{i=1}^n \boldsymbol{\varphi}(\boldsymbol{y}_i, \boldsymbol{z}_i). \tag{18}$$

Note that when $\mathcal{G}$ is an RKHS corresponding to a kernel k, $g(\boldsymbol{y}, \boldsymbol{z}) = \sum_{\ell=1}^n \alpha_\ell k((\boldsymbol{y}_\ell, \boldsymbol{z}_\ell), (\boldsymbol{y}, \boldsymbol{z}))$ is a member of the RKHS and the RKHS norm of g satisfies

$\|g\|_{\mathcal{G}}^2 = \boldsymbol{\alpha}^\top \boldsymbol{R}\boldsymbol{\alpha}$ where $\boldsymbol{R}$ is the Gram matrix, i.e., $R_{\ell,\ell'} = k((\boldsymbol{y}_\ell, \boldsymbol{z}_\ell), (\boldsymbol{y}_{\ell'}, \boldsymbol{z}_{\ell'}))$. Thus the regularization term $R(g) = \sqrt{C}\|g\|_{\mathcal{G}} = \sqrt{C}\sqrt{\boldsymbol{\alpha}^\top \boldsymbol{R}\boldsymbol{\alpha}}$ satisfies the condition (12) if the kernel function is bounded as Eq.(13).

Differentiating the objective function (17) with respect to $\boldsymbol{\alpha}$ and equating it to zero, we obtain

$$(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{R})\boldsymbol{\alpha} = \widehat{\boldsymbol{h}}. \tag{19}$$

Thus, the solution can be obtained just by solving the above system of linear equations. The solution $\widehat{\boldsymbol{\alpha}}$ is given *analytically* as

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

Then we can analytically approximate SMI as follows:

$$\widehat{\mathrm{SMI}}(Y, Z) := \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}} - \frac{1}{2}. \tag{20}$$

2.3.4 Model Selection by Cross-Validation

As shown in Sections 2.3.1 and 2.3.2, our SMI estimator was shown to possess 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 $\mathcal{S} = \{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^n$ 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 $\mathcal{S}\backslash\mathcal{S}_k$ (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 $\widehat{J}_{\mathrm{CV}}$ is outputted:

$$\widehat{J}_{\mathrm{CV}} := \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}^\top \widehat{\boldsymbol{H}}_{\mathcal{S}_k}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k} - \widehat{\boldsymbol{h}}_{\mathcal{S}_k}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}\right), \tag{21}$$

where $\widehat{\boldsymbol{H}}_{\mathcal{S}_k}$ and $\widehat{\boldsymbol{h}}_{\mathcal{S}_k}$ denote $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ computed on the hold-out samples $\mathcal{S}_k$ in (18). For model selection, we compute $\widehat{J}_{\mathrm{CV}}$ 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 $\widehat{J}_{\mathrm{CV}}$. We can show that $\widehat{J}_{\mathrm{CV}}$ is an almost unbiased estimator of the objective function J, where the '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 may 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 do not pursue this direction in this paper.

2.3.5 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 $2n$:

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

In the regression scenarios where $\boldsymbol{y}$ is continuous, we use the Gaussian kernel as the 'base' kernels:

$$\phi^{\mathrm{y}}_\ell(\boldsymbol{y}) := \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{u}_\ell\|^2}{2\sigma^2}\right), \quad \phi^{\mathrm{z}}_\ell(\boldsymbol{z}) := \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{v}_\ell\|^2}{2\sigma^2}\right),$$

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 (6) 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. 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[4] $\{(\boldsymbol{y}_i,\boldsymbol{z}_i)\}_{i=1}^{n}$.

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. Note that, in this case, the matrix $\widehat{\boldsymbol{H}}$ becomes block-diagonal, given that the samples are sorted according to the class labels. Then the linear equation (19) can be solved efficiently.

More generally, when $\boldsymbol{y}$ is structured (e.g., strings, trees, and graphs), we may employ kernels for structured data as $\phi^{\mathrm{y}}_\ell(\boldsymbol{y})$ (Lodhi et al., 2002; Collins & Duffy, 2002; Kashima & Koyanagi, 2002; Kondor & Lafferty, 2002; Kashima et al., 2003; Gärtner et al., 2003; Gärtner, 2003).

2.4 Least-Squares Dimension Reduction

Finally, we show how the SMI approximator is employed for dimension reduction. To find a sufficient subspace, the dimension reduction problem is casted as an optimization problem over the Grassmann manifold $\mathrm{Gr}^{d}_{m}(\mathbb{R})$ (see Eq.(1)).

[4] 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.

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

$$\frac{\partial \widehat{\mathrm{SMI}}}{\partial W_{\ell,\ell'}} = \frac{\partial \widehat{\boldsymbol{h}}^\top}{\partial W_{\ell,\ell'}}(2\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\alpha}}^\top \frac{\partial \widehat{\boldsymbol{H}}}{\partial W_{\ell,\ell'}}(\frac{3}{2}\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\beta}}) + \lambda \widehat{\boldsymbol{\alpha}}^\top \frac{\partial \boldsymbol{R}}{\partial W_{\ell,\ell'}}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\beta}} := (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{R})^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}}$ and we used the formula $\frac{\partial \boldsymbol{X}^{-1}}{\partial t} = -\boldsymbol{X}^{-1}\frac{\partial \boldsymbol{X}}{\partial t}\boldsymbol{X}^{-1}$ for a positive symmetric matrix $\boldsymbol{X}$ each element of which is a function of t.

In the Euclidean space, the ordinary gradient $\frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}$ gives the steepest direction. However, on a manifold, the *natural gradient* (Amari, 1998) gives the steepest direction. The natural gradient $\nabla \widehat{\mathrm{SMI}}(\boldsymbol{W})$ at $\boldsymbol{W}$ is the projection of the ordinary gradient $\frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}$ to the tangent space of $\mathrm{Gr}_d^m(\mathbb{R})$ 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 as follows (Edelman et al., 1998):

$$\nabla \widehat{\mathrm{SMI}}(\boldsymbol{W}) = \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}} - \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}\boldsymbol{W}^\top \boldsymbol{W} = \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}\boldsymbol{W}_\perp^\top \boldsymbol{W}_\perp, \tag{22}$$

where $\boldsymbol{W}_\perp$ is any $(d-m) \times d$ matrix such that $[W^\top \ W_\perp^\top]$ is orthogonal. Then the *geodesic* from $\boldsymbol{W}$ to the direction of the natural gradient $\nabla \widehat{\mathrm{SMI}}(\boldsymbol{W})$ over $\mathrm{Gr}_d^m(\mathbb{R})$ can be expressed using t $(\in \mathbb{R})$ as

$$\boldsymbol{W}_t := \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{O}_{d-m} \end{bmatrix} \exp\left(t \begin{bmatrix} \boldsymbol{O}_m & \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}\boldsymbol{W}_\perp^\top \\ -\boldsymbol{W}_\perp \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}}^\top & \boldsymbol{O}_{d-m} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{W}_\perp \end{bmatrix},$$

where 'exp' for a matrix denotes the *matrix exponential*, and $\boldsymbol{O}_b$ is the $b \times b$ zero matrix (note that the derivative $\partial_t \boldsymbol{W}_t|_{t=0}$ coincides with the natural gradient (22), see Edelman et al. (1998) for detailed derivation of the geodesic). Thus line search along the geodesic in the natural gradient direction is equivalent to finding the maximizer from $\{\boldsymbol{W}_t \mid t \geq 0\}$.

For choosing the step size of each gradient update, we may use some approximate line search method such as *Armijo's rule* (Patriksson, 1999, p.50) or *backtracking line search* (Boyd & Vandenberghe, 2004, p.464). In our setting, Armijo's rule finds the step size as the maximum t_k that satisfies

$$\begin{aligned} \widehat{\mathrm{SMI}}(\boldsymbol{W}_{t_k}) - \widehat{\mathrm{SMI}}(\boldsymbol{W}) &\geq t_k \mu \,\mathrm{tr}\left[\left.\frac{\partial \boldsymbol{W}_t}{\partial t}\right|_{t=0}^\top \frac{\partial \widehat{\mathrm{SMI}}(\boldsymbol{W})}{\partial \boldsymbol{W}} \right] \\ &= t_k \mu \,\mathrm{tr}\left[\boldsymbol{W}_\perp^\top \boldsymbol{W}_\perp \frac{\partial \widehat{\mathrm{SMI}}(\boldsymbol{W})}{\partial \boldsymbol{W}}^\top \frac{\partial \widehat{\mathrm{SMI}}(\boldsymbol{W})}{\partial \boldsymbol{W}} \right], \end{aligned}$$

where $t_0 = 1$, $t_k = \alpha t_{k-1}$ $(k = 1, 2, \ldots)$, $\alpha \in (0, 1)$, and $\mu \in (0, 1)$ are given parameters.

We call the proposed dimension reduction algorithm *Least-Squares Dimension Reduction* (LSDR). The entire algorithm is summarized in Figure 1. In practice, we performed CV once in several steps because executing CV at every step is computationally expensive.

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

Figure 1: The LSDR algorithm.

.

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 to $b = \min(100, n)$, 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 (21).

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):

(a) Linear dependence: $d = 5$, $m = 1$. y has a linear dependence on $\boldsymbol{x}$ as

$$y = x^{(1)} + \epsilon,$$

where $x^{(k)}$ denotes the k-th element of $\boldsymbol{x}$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_5)$ and $\epsilon \sim \mathcal{N}(y; 0, 0.25)$. Here $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal density with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The optimal projection is given by

$$\boldsymbol{W}^* = [1\ 0\ 0\ 0\ 0]\,. \tag{23}$$

(b) Non-linear dependence 1: $d = 5$, $m = 1$. y has a quadratic dependence on $\boldsymbol{x}$ as

$$y = (x^{(1)})^2 + \epsilon,$$

where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_5)$ and $\epsilon \sim \mathcal{N}(y; 0, 1)$. The optimal projection is given by Eq.(23).

(c) Non-linear dependence 2: $d = 5$, $m = 1$. y has a lattice-structured dependence on $\boldsymbol{x}$ as

$$\boldsymbol{x} \sim \mathcal{U}(\boldsymbol{x}; [-0.5, 0.5]^5),$$
$$y|\boldsymbol{x} \sim \begin{cases} \mathcal{N}(y; 0, 0.25) & \text{if } x^{(1)} \leq |\frac{1}{6}|, \\ \frac{1}{2}\mathcal{N}(y; 1, 0.25) + \frac{1}{2}\mathcal{N}(y; -1, 0.25) & \text{otherwise,} \end{cases}$$

where $\mathcal{U}(\boldsymbol{x}; \mathcal{S})$ denotes the uniform density on a set $\mathcal{S}$. The optimal projection is given by Eq.(23).

(d) Fukumizu et al. (2009): $d = 4$, $m = 2$. y has a non-linear dependence on $\boldsymbol{x}$ as

$$y = \frac{x^{(1)}}{0.5 + (x^{(2)} + 1.5)^2} + (1 + x^{(2)})^2 + 0.4\epsilon,$$

where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_4)$ and $\epsilon \sim \mathcal{N}(1; 0, 1)$. The optimal projection is $\boldsymbol{W}^* = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$.

(e) Fukumizu et al. (2009): $d = 4$, $m = 1$. y has a non-linear dependence on $\boldsymbol{x}$ as

$$y = \sin^2(\pi x^{(1)} + 1) + 0.4\epsilon,$$

where $\boldsymbol{x} \sim \mathcal{U}(\boldsymbol{x}; [0, 1]^4 \setminus \{\boldsymbol{x} \in \mathbb{R}^4 \mid x^{(i)} \leq 0.7 \ (i = 1, \ldots, 4)\})$ and $\epsilon \sim \mathcal{N}(1; 0, 1)$. The optimal projection is $\boldsymbol{W}^* = [1 \ 0 \ 0 \ 0]$.

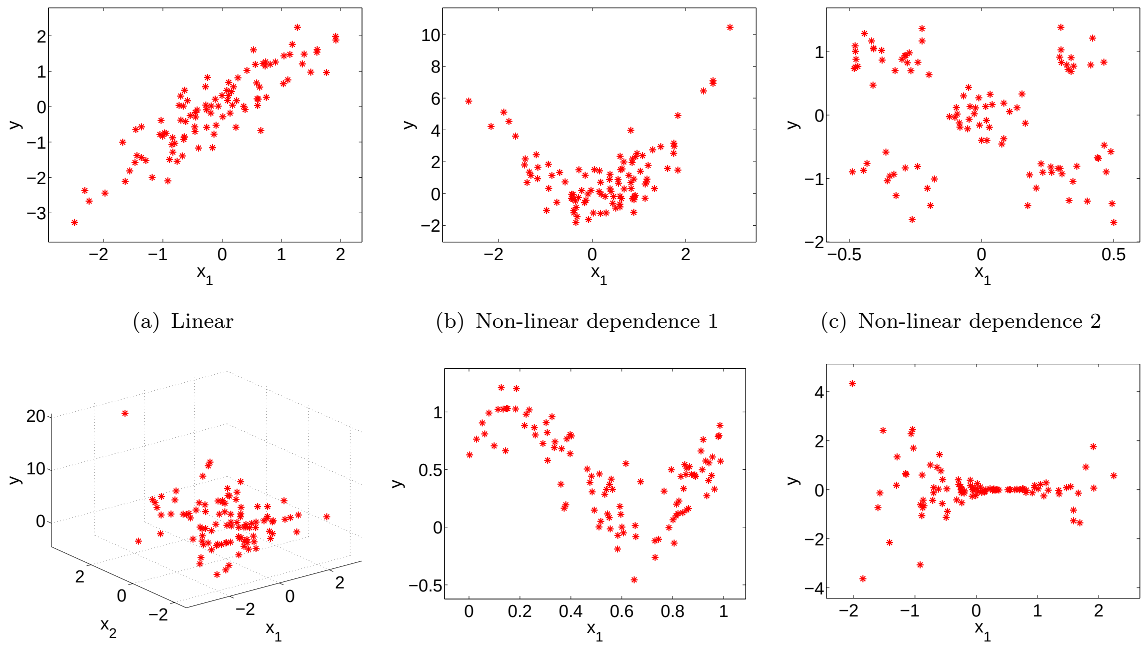
(f) Fukumizu et al. (2009): $d = 10$, $m = 1$. y has a non-linear dependence on $\boldsymbol{x}$ as

$$y = \frac{1}{2}(x^{(1)} - 1)^2\epsilon,$$

where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_{10})$ and $\epsilon \sim \mathcal{N}(1; 0, 1)$. The optimal projection is $\boldsymbol{W}^* = [1 \ 0 \ \cdots \ 0]$.

Let us compare the proposed LSDR with the following methods.

- Kernel dimension reduction (KDR) (Fukumizu et al., 2009),
- The Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005),
- Sliced inverse regression (SIR) (Li, 1991),
- Sliced average variance estimation (SAVE) (Cook, 2000),
- Principal Hessian direction (pHd) (Li, 1992).
- Minimum Average (conditional) Variance Estimation (dMAVE) (Xia, 2007).



(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.

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 used the dimension reduction package `dr` included in R for SIR, SAVE, and pHd. The parameters for these methods such as the number of slices were set to be the default values. To execute dMAVE, we used the publicly available code[5]. The principal directions estimated by SIR, SAVE, pHd, and dMAVE do not necessarily form an orthogonal system, i.e., if we let $\boldsymbol{F}$ be the matrix each row of which corresponds to each principal direction, then $\boldsymbol{F}$ is not necessarily a projection matrix. To recover a projection matrix $\boldsymbol{W}$, we performed singular decomposition of $\boldsymbol{F}$ as $\boldsymbol{F} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}$, and set $\boldsymbol{W} = \boldsymbol{U}$.

We evaluate the performance of each method by

$$\left\| \widehat{\boldsymbol{W}}^{\top}\widehat{\boldsymbol{W}} - \boldsymbol{W}^{*\top}\boldsymbol{W}^{*} \right\|_{\mathrm{Frobenius}}, \tag{24}$$

where $\|\cdot\|_{\mathrm{Frobenius}}$ denotes the Frobenius norm, $\widehat{\boldsymbol{W}}$ is an estimated projection matrix, and $\boldsymbol{W}^{*}$ is an optimal projection matrix.

The performance of each method is summarized in Table 2, which depicts the mean and standard deviation of the Frobenius-norm error (24) over 50 trials when the number of samples is $n = 100$. LSDR overall shows good performance; in particular, it performs the best for datasets (b) and (c). KDR also tends to work reasonably well, but it sometimes performs poorly; this seems to be caused by an inappropriate choice of the Gaussian kernel

[5] http://www.stat.nus.edu.sg/˜staxyc/dMAVE.m

Table 2: Mean (and standard deviation in the bracket) of the Frobenius-norm error (24) 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.

Data	d	m	LSDR	KDR	HSIC	SIR	SAVE	pHd	dMAVE
(a)	5	1	**.13(.04)**	.13(.05)	.17(.07)	**.11(.05)**	.37(.27)	.89(.12)	**.13(.04)**
(b)	5	1	**.15(.06)**	.25(.21)	.44(.36)	.83(.19)	.31(.11)	.24(.07)	.20(.09)
(c)	5	1	**.10(.05)**	.44(.32)	.68(.32)	.89(.14)	.48(.20)	.86(.12)	.25(.28)
(d)	4	2	.20(.14)	.16(.06)	.18(.08)	.30(.15)	.44(.18)	.50(.18)	**.10(.05)**
(e)	4	1	**.09(.06)**	.13(.06)	.16(.07)	.21(.10)	.34(.19)	.36(.14)	**.08(.04)**
(f)	10	1	.35(.12)	.40(.12)	.49(.17)	.68(.22)	.91(.13)	.83(.12)	**.26(.06)**

width, implying that the heuristic of using the median sample distance as the kernel width is not always appropriate. On the other hand, LSDR with CV performs stably well for various types of datasets. dMAVE also works well, and is competitive to LSDR for these artificial 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[6]. We randomly choose 200 samples from the dataset, and apply LSDR, KDR, HSIC, and dMAVE to obtain projections onto low-dimension subspaces with $m = \lceil d/4 \rceil, \lceil d/2 \rceil, \lceil 3d/4 \rceil$, where $\lceil c \rceil$ denotes the smallest integer not smaller than c. Then we train the support vector machine (Schölkopf & Smola, 2002) on the 200 projected training samples.

The misclassification rate is computed for 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). Our contributions were parametric and non-parametric analyses of the rate of convergence of the SMI approximator, and the proposal of a dimension reduction algorithm based on the SMI approximator. The proposed method is advantageous in several respects, 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.

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

Table 3: Mean (and standard deviation in the bracket) of misclassification rates for benchmark 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.

Data set	d	m	LSDR	KDR	HSIC	dMAVE
image	18	5	**.083(.019)**	.125(.038)	.158(.044)	.501(.134)
	18	9	**.088(.022)**	.106(.026)	.115(.035)	.468(.130)
	18	14	**.093(.018)**	**.091(.019)**	**.095(.023)**	.468(.130)
waveform	21	6	**.130(.014)**	**.127(.008)**	.160(.016)	.183(.016)
	21	11	**.119(.013)**	.135(.010)	.163(.016)	.184(.015)
	21	16	**.116(.007)**	.131(.008)	.159(.014)	.182(.021)
pima	8	2	**.249(.022)**	**.247(.024)**	**.252(.020)**	**.257(.017)**
	8	4	**.260(.016)**	**.250(.021)**	**.252(.017)**	**.265(.027)**
	8	6	**.244(.020)**	**.243(.019)**	**.251(.021)**	**.252(.019)**
letter (a,b, & c)	16	4	.031(.009)	.028(.012)	.035(.014)	**.018(.008)**
	16	8	.026(.008)	**.017(.007)**	**.020(.006)**	**.016(.006)**
	16	12	**.016(.006)**	**.014(.006)**	**.017(.008)**	**.013(.008)**

Acknowledgements

T.S. was supported in part by MEXT Kakenhi 22700289 and the Aihara Project, the FIRST program from JSPS, initiated by CSTP. M.S. was supported by AOARD, SCAT, and the JST PRESTO program.

A Proof of Lemma 1

Let $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{z}_\perp)$. By the relations $\mathrm{d}\boldsymbol{x} = \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{z}_\perp$ and $p_{\mathrm{x}}(\boldsymbol{x}) = p_{\mathrm{z}_\perp|\mathrm{z}}(\boldsymbol{z}_\perp|\boldsymbol{z})p_{\mathrm{z}}(\boldsymbol{z})$, we have

$$\begin{aligned}
&\int \frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})} p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\
&= \int \frac{p_{\mathrm{z}_\perp|\mathrm{zy}}(\boldsymbol{z}_\perp|\boldsymbol{z},\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{z}_\perp\mathrm{d}\boldsymbol{y} \\
&= \int \frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})^2}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{y} = \int \left(\frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})}\right)^2 p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Thus we obtain

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

Noticing that

$$\begin{aligned}\frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})} &= \frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})} = \frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{z_\perp|z}}(\boldsymbol{z}_\perp|\boldsymbol{z})p_{\mathrm{z}}(\boldsymbol{z})p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{y|z}}(\boldsymbol{y}|\boldsymbol{z})p_{\mathrm{z}}(\boldsymbol{z})}\\ &= \frac{p_{\mathrm{z_\perp y|z}}(\boldsymbol{z}_\perp,\boldsymbol{y}|\boldsymbol{z})}{p_{\mathrm{z_\perp|z}}(\boldsymbol{z}_\perp|\boldsymbol{z})p_{\mathrm{y|z}}(\boldsymbol{y}|\boldsymbol{z})}\frac{p_{\mathrm{yz}}(\boldsymbol{y},\boldsymbol{z})}{p_{\mathrm{y}}(\boldsymbol{y})p_{\mathrm{z}}(\boldsymbol{z})},\end{aligned}$$

we have

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

which concludes the proof of Lemma 1. ■

B Proof of Theorem 1

For notational simplicity, we define linear operators $Q, Q_n, \widetilde{Q}_n, P, P_n$ as

$$\begin{aligned}&Qf := \mathrm{E}_{p_{\mathrm{y}}p_{\mathrm{z}}}f, \quad Q_nf := \frac{\sum_{i,j=1}^n f(\boldsymbol{y}_i,\boldsymbol{z}_j)}{n^2}, \quad \widetilde{Q}_n := \frac{\sum_{1\le i\neq j\le n} f(\boldsymbol{y}_i,\boldsymbol{z}_j)}{n(n-1)},\\ &Pf := \mathrm{E}_{p_{\mathrm{yz}}}f, \quad P_nf := \frac{\sum_{i=1}^n f(\boldsymbol{y}_i,\boldsymbol{z}_i)}{n}.\end{aligned}$$

Let $\nabla g_{\boldsymbol{\theta}'}$ denote $\nabla_{\boldsymbol{\theta}} g_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = (\partial_\ell g_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}'})_\ell$ for $\boldsymbol{\theta}'\in\Theta$. Since $\widehat{\boldsymbol{\theta}}$ is the optimizer of the problem (7), we have

$$\mathbf{0}=\nabla\left(\frac{1}{2}Q_ng_{\widehat{\boldsymbol{\theta}}}^2-P_ng_{\widehat{\boldsymbol{\theta}}}\right)=\frac{1}{2}Q_n\nabla(g_{\widehat{\boldsymbol{\theta}}}^2)-P_n\nabla g_{\widehat{\boldsymbol{\theta}}}. \tag{25}$$

Therefore, as in the standard asymptotic expansion for maximum likelihood estimators (van der Vaart, 2000), we have

$$\begin{aligned}\mathbf{0}=&\frac{1}{2}Q_n\nabla(g_{\boldsymbol{\theta}^*}^2)-P_n\nabla g_{\boldsymbol{\theta}^*}+(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\left(\frac{1}{2}Q_n\nabla\nabla^\top(g_{\boldsymbol{\theta}^*}^2)-P_n\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right)\\ &+\mathcal{O}_p\left(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\right).\end{aligned}$$

This implies

$$\begin{aligned}\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*=&-\left(\frac{1}{2}Q_n\nabla\nabla^\top(g_{\boldsymbol{\theta}^*}^2)-P_n\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right)^{-1}\left(\frac{1}{2}Q_n\nabla(g_{\boldsymbol{\theta}^*}^2)-P_n\nabla g_{\boldsymbol{\theta}^*}\right)\\ &+\mathcal{O}_p\left(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\right).\end{aligned} \tag{26}$$

Since $Q\left(g_{\boldsymbol{\theta}^*}\nabla g_{\boldsymbol{\theta}^*}\right)-P\nabla g_{\boldsymbol{\theta}^*}=0$, we have

$$\frac{1}{2}Q_n\nabla(g_{\boldsymbol{\theta}^*}^2)-P_n\nabla g_{\boldsymbol{\theta}^*}=\mathcal{O}_p(n^{-1/2}),$$

and

$$\begin{aligned}
&\frac{1}{2}Q_n\nabla\nabla^\top(g_{\boldsymbol{\theta}^*}^2) - P_n\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\\
&=\frac{1}{2}Q\nabla\nabla^\top(g_{\boldsymbol{\theta}^*}^2) - P\nabla\nabla^\top g_{\boldsymbol{\theta}^*} + \mathcal{O}_p(n^{-1/2})\\
&=Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right) + Q\left(g_{\boldsymbol{\theta}^*}\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right) - P\left(\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right) + \mathcal{O}_p(n^{-1/2})\\
&=Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right) + P\left(\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right) - P\left(\nabla\nabla^\top g_{\boldsymbol{\theta}^*}\right) + \mathcal{O}_p(n^{-1/2})\\
&=Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right) + \mathcal{O}_p(n^{-1/2}). \qquad (27)
\end{aligned}$$

Thus Eq.(26) implies

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \mathcal{O}_p(n^{-1/2}) + \mathcal{O}_p(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2),$$

in particular, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \mathcal{O}_p(n^{-1/2})$. Moreover Eq.(26) becomes

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = -Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right)^{-1}\left(\frac{1}{2}Q_n\nabla(g_{\boldsymbol{\theta}^*}^2) - P_n\nabla g_{\boldsymbol{\theta}^*}\right) + \mathcal{O}_p(n^{-1}). \qquad (28)$$

Eqs.(5) and (8) give

$$\begin{aligned}
\widehat{\mathrm{SMI}}(Y,Z) - \mathrm{SMI}(Y,Z) &= -\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 + P_n g_{\widehat{\boldsymbol{\theta}}} - \left(-\frac{1}{2}Qg_{\boldsymbol{\theta}^*}^2 + Pg_{\boldsymbol{\theta}^*}\right)\\
&= -\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 + P_n g_{\widehat{\boldsymbol{\theta}}} + \frac{1}{2}Q_n g_{\boldsymbol{\theta}^*}^2 - P_n g_{\boldsymbol{\theta}^*}\\
&\quad - \left(-\frac{1}{2}Qg_{\boldsymbol{\theta}^*}^2 + Pg_{\boldsymbol{\theta}^*} + \frac{1}{2}Q_n g_{\boldsymbol{\theta}^*}^2 - P_n g_{\boldsymbol{\theta}^*}\right). \qquad (29)
\end{aligned}$$

The first four terms of the RHS can be expanded as

$$\begin{aligned}
&-\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 + P_n g_{\widehat{\boldsymbol{\theta}}} + \frac{1}{2}Q_n g_{\boldsymbol{\theta}^*}^2 - P_n g_{\boldsymbol{\theta}^*}\\
&=(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^\top\nabla\left(\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 - P_n g_{\widehat{\boldsymbol{\theta}}}\right)\\
&\quad + \frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^\top\nabla\nabla^\top\left(\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 - P_n g_{\widehat{\boldsymbol{\theta}}}\right)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) + \mathcal{O}_p(n^{-3/2})\\
&=\frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^\top\nabla\nabla^\top\left(\frac{1}{2}Q_n g_{\widehat{\boldsymbol{\theta}}}^2 - P_n g_{\widehat{\boldsymbol{\theta}}}\right)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) + \mathcal{O}_p(n^{-3/2}) \qquad (\because \text{Eq.(25)})\\
&=\frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^\top\nabla\nabla^\top\left(\frac{1}{2}Q_n g_{\boldsymbol{\theta}^*}^2 - P_n g_{\boldsymbol{\theta}^*}\right)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) + \mathcal{O}_p(n^{-3/2})\\
&=\frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^\top Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) + \mathcal{O}_p(n^{-3/2}) \qquad (\because \text{Eq.(27)})\\
&=\mathcal{O}_p(n^{-1}). \qquad (30)
\end{aligned}$$

On the other hand, by the central limit theorem,

$$-\frac{1}{2}Qg_{\boldsymbol{\theta}^*}^2 + Pg_{\boldsymbol{\theta}^*} + \frac{1}{2}Q_n g_{\boldsymbol{\theta}^*}^2 - P_n g_{\boldsymbol{\theta}^*} = \mathcal{O}_p(n^{-1/2}). \qquad (31)$$

Substituting Eqs.(30) and (31) into Eq.(29), we have the first assertion (9).

Next we prove the second assertion (10). Based on Eqs.(29) and (30), we can evaluate the expectation of $\widehat{\mathrm{SMI}}(Y, Z) - \mathrm{SMI}(Y, Z)$ as

$$\begin{aligned}
&\mathrm{E}_{D^n}[\widehat{\mathrm{SMI}}(Y, Z) - \mathrm{SMI}(Y, Z)] \\
&= \mathrm{E}_{D^n}\left[-\frac{1}{2}Q_n g^2_{\widehat{\boldsymbol{\theta}}} + P_n g_{\widehat{\boldsymbol{\theta}}} + \frac{1}{2}Q_n g^2_{\boldsymbol{\theta}^*} - P_n g_{\boldsymbol{\theta}^*}\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{A}\mathrm{E}_{D^n}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right]\right\} + o(n^{-1}),
\end{aligned}$$

where we used the fact that

$$Q\left(\nabla g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}\right) = \boldsymbol{A}.$$

Below, we will show that

$$\mathrm{E}_{D^n}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] = \frac{1}{n}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} + o(n^{-1}). \tag{32}$$

Obviously this gives Eq.(10).

Eq.(28) implies

$$\begin{aligned}
&\mathrm{E}_{D^n}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right] \\
&= \boldsymbol{A}^{-1}\mathrm{E}_{D^n}\left[\left(\frac{1}{2}Q_n\nabla(g^2_{\boldsymbol{\theta}^*}) - P_n\nabla g_{\boldsymbol{\theta}^*}\right)\left(\frac{1}{2}Q_n\nabla^\top(g^2_{\boldsymbol{\theta}^*}) - P_n\nabla^\top g_{\boldsymbol{\theta}^*}\right)\right]\boldsymbol{A}^{-1} + o(n^{-1}) \\
&= \boldsymbol{A}^{-1}\mathrm{E}_{D^n}\left[\left(Q_n(g_{\boldsymbol{\theta}^*}\nabla g_{\boldsymbol{\theta}^*}) - P_n\nabla g_{\boldsymbol{\theta}^*}\right)\left(Q_n(g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}) - P_n\nabla^\top g_{\boldsymbol{\theta}^*}\right)\right]\boldsymbol{A}^{-1} + o(n^{-1}) \\
&= \boldsymbol{A}^{-1}\mathrm{E}_{D^n}\left[\left(\widetilde{Q}_n(g_{\boldsymbol{\theta}^*}\nabla g_{\boldsymbol{\theta}^*}) - P_n\nabla g_{\boldsymbol{\theta}^*}\right)\left(\widetilde{Q}_n(g_{\boldsymbol{\theta}^*}\nabla^\top g_{\boldsymbol{\theta}^*}) - P_n\nabla^\top g_{\boldsymbol{\theta}^*}\right)\right]\boldsymbol{A}^{-1} + o(n^{-1}).
\end{aligned} \tag{33}$$

Let

$$h(\boldsymbol{w}, \boldsymbol{w}') = \frac{1}{2}\Big(g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}')\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}') + g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z})\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z}) - \nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}) - \nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z}')\Big)$$

for $\boldsymbol{w} = (\boldsymbol{y}, \boldsymbol{z})$ and $\boldsymbol{w}' = (\boldsymbol{y}', \boldsymbol{z}')$. Then h is symmetric (i.e., $h(\boldsymbol{w}, \boldsymbol{w}') = h(\boldsymbol{w}', \boldsymbol{w})$), has mean $\mathbf{0}$ (i.e., $\mathrm{E}_{p_w p_{w'}}[h(\boldsymbol{w}, \boldsymbol{w}')] = \mathbf{0}$), and satisfies

$$\widetilde{Q}_n(g_{\boldsymbol{\theta}^*}\nabla g_{\boldsymbol{\theta}^*}) - P_n\nabla g_{\boldsymbol{\theta}^*} = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} h(\boldsymbol{w}_i, \boldsymbol{w}_j) =: U_n.$$

Therefore, $\widetilde{Q}_n(g_{\boldsymbol{\theta}^*}\nabla g_{\boldsymbol{\theta}^*}) - P_n\nabla g_{\boldsymbol{\theta}^*}$ is a U-statistic with the symmetric kernel $h(\boldsymbol{w}_i, \boldsymbol{w}_j)$. It is known (Theorem 7.1 in Hoeffding, 1948) that the variance of U-statistic is given by

$$\mathrm{E}_{D^n}\left[U_n U_n^\top\right] = \frac{4}{n}\mathrm{E}[h(\boldsymbol{w}_1, \boldsymbol{w}_2)h(\boldsymbol{w}_1, \boldsymbol{w}_2')^\top] + o(n^{-1}), \tag{34}$$

where $\boldsymbol{w}_1 = (\boldsymbol{y}_1, \boldsymbol{z}_1)$, $\boldsymbol{w}_2 = (\boldsymbol{y}_2, \boldsymbol{z}_2)$, and $\boldsymbol{w}_2' = (\boldsymbol{y}_2', \boldsymbol{z}_2')$ are i.i.d. variables with probability density p_{yz}. For notational simplicity, we write

$$f_1(\boldsymbol{y}, \boldsymbol{z}) := g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}) \nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}), \quad f_2(\boldsymbol{y}, \boldsymbol{z}) := \nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}).$$

Then

$$\begin{aligned}
&4\mathrm{E}[h(\boldsymbol{w}_1, \boldsymbol{w}_2) h(\boldsymbol{w}_1, \boldsymbol{w}_2')^\top] = \mathrm{E}\Big[\\
&\underbrace{f_1(\boldsymbol{y}_1, \boldsymbol{z}_2) f_1(\boldsymbol{y}_1, \boldsymbol{z}_2')^\top}_{(A)} + \underbrace{f_1(\boldsymbol{y}_1, \boldsymbol{z}_2) f_1(\boldsymbol{y}_2', \boldsymbol{z}_1)^\top}_{(B)} - \underbrace{f_1(\boldsymbol{y}_1, \boldsymbol{z}_2) f_2(\boldsymbol{y}_1, \boldsymbol{z}_1)^\top}_{(C)} - \underbrace{f_1(\boldsymbol{y}_1, \boldsymbol{z}_2) f_2(\boldsymbol{y}_2', \boldsymbol{z}_2')^\top}_{(D)} \\
&+ \underbrace{f_1(\boldsymbol{y}_2, \boldsymbol{z}_1) f_1(\boldsymbol{y}_1, \boldsymbol{z}_2')^\top}_{(E)} + \underbrace{f_1(\boldsymbol{y}_2, \boldsymbol{z}_1) f_1(\boldsymbol{y}_2', \boldsymbol{z}_1)^\top}_{(F)} - \underbrace{f_1(\boldsymbol{y}_2, \boldsymbol{z}_1) f_2(\boldsymbol{y}_1, \boldsymbol{z}_1)^\top}_{(G)} - \underbrace{f_1(\boldsymbol{y}_2, \boldsymbol{z}_1) f_2(\boldsymbol{y}_2', \boldsymbol{z}_2')^\top}_{(H)} \\
&- \underbrace{f_2(\boldsymbol{y}_1, \boldsymbol{z}_1) f_1(\boldsymbol{y}_1, \boldsymbol{z}_2')^\top}_{(I)} - \underbrace{f_2(\boldsymbol{y}_1, \boldsymbol{z}_1) f_1(\boldsymbol{y}_2', \boldsymbol{z}_1)^\top}_{(J)} + \underbrace{f_2(\boldsymbol{y}_1, \boldsymbol{z}_1) f_2(\boldsymbol{y}_1, \boldsymbol{z}_1)^\top}_{(K)} + \underbrace{f_2(\boldsymbol{y}_1, \boldsymbol{z}_1) f_2(\boldsymbol{y}_2', \boldsymbol{z}_2')^\top}_{(L)} \\
&- \underbrace{f_2(\boldsymbol{y}_2, \boldsymbol{z}_2) f_1(\boldsymbol{y}_1, \boldsymbol{z}_2')^\top}_{(M)} - \underbrace{f_2(\boldsymbol{y}_2, \boldsymbol{z}_2) f_1(\boldsymbol{y}_2', \boldsymbol{z}_1)^\top}_{(N)} + \underbrace{f_2(\boldsymbol{y}_2, \boldsymbol{z}_2) f_2(\boldsymbol{y}_1, \boldsymbol{z}_1)^\top}_{(O)} + \underbrace{f_2(\boldsymbol{y}_2, \boldsymbol{z}_2) f_2(\boldsymbol{y}_2', \boldsymbol{z}_2')^\top}_{(P)} \Big].
\end{aligned}$$

Going through simple calculations, each term in the RHS of the above equation can be evaluated as

$$\begin{aligned}
\text{(A), (C), (I)} &= \mathrm{E}_{p_{\mathrm{z|y}} p_{\mathrm{z'|y}} p_{\mathrm{y}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z')], \\
\text{(B)} &= \mathrm{E}_{p_{\mathrm{z'|y}} p_{\mathrm{y'|z}} p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z') \nabla^\top g_{\boldsymbol{\theta}^*}(y', z)], \\
\text{(D), (H), (L), (M), (N), (O), (P)} &= \mathrm{E}_{p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z)] \mathrm{E}_{p_{\mathrm{yz}}}[\nabla^\top g_{\boldsymbol{\theta}^*}(y, z)], \\
\text{(E)} &= \mathrm{E}_{p_{\mathrm{z'|y}} p_{\mathrm{y'|z}} p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y', z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z')], \\
\text{(F), (G), (J)} &= \mathrm{E}_{p_{\mathrm{y|z}} p_{\mathrm{y'|z}} p_{\mathrm{z}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y', z)], \\
\text{(K)} &= \mathrm{E}_{p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z)].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&4\mathrm{E}[h(\boldsymbol{w}_1, \boldsymbol{w}_2) h(\boldsymbol{w}_1, \boldsymbol{w}_2')^\top] \\
&= \mathrm{E}_{p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z)] - \mathrm{E}_{p_{\mathrm{z|y}} p_{\mathrm{z'|y}} p_{\mathrm{y}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z')] \\
&\quad + \mathrm{E}_{p_{\mathrm{z'|y}} p_{\mathrm{y'|z}} p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z') \nabla^\top g_{\boldsymbol{\theta}^*}(y', z)] + \mathrm{E}_{p_{\mathrm{z'|y}} p_{\mathrm{y'|z}} p_{\mathrm{yz}}}[\nabla g_{\boldsymbol{\theta}^*}(y', z) \nabla^\top g_{\boldsymbol{\theta}^*}(y, z')] \\
&\quad - \mathrm{E}_{p_{\mathrm{y|z}} p_{\mathrm{y'|z}} p_{\mathrm{z}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z) \nabla^\top g_{\boldsymbol{\theta}^*}(y', z)] - \mathrm{E}_{p_{\mathrm{zy}}}[\nabla g_{\boldsymbol{\theta}^*}(y, z)] \mathrm{E}_{p_{\mathrm{zy}}}[\nabla^\top g_{\boldsymbol{\theta}^*}(y, z)] \\
&= \mathrm{E}_{p_{\mathrm{yz}}}\Big[\Big(\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}) - \mathrm{E}_{p_{\mathrm{z'|y}}}[\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}')] - \mathrm{E}_{p_{\mathrm{y'|z}}}[\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z})] + \mathrm{E}_{p_{\mathrm{y'z'}}}[\nabla g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z}')]\Big) \\
&\quad \times \Big(\nabla^\top g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}) - \mathrm{E}_{p_{\mathrm{z'|y}}}[\nabla^\top g_{\boldsymbol{\theta}^*}(\boldsymbol{y}, \boldsymbol{z}')] - \mathrm{E}_{p_{\mathrm{y'|z}}}[\nabla^\top g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z})] + \mathrm{E}_{p_{\mathrm{y'z'}}}[\nabla^\top g_{\boldsymbol{\theta}^*}(\boldsymbol{y}', \boldsymbol{z}')]\Big)\Big] \\
&= \boldsymbol{B}.
\end{aligned}$$

This with Eqs.(33) and (34) yields Eq.(32), and thus we have the second assertion (10). ■

C Proof of Theorem 2

Before proving Theorem 2, we show the following auxiliary lemma.

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

$$\|\widehat{g} - g^*\|_2 = \mathcal{O}_p(\lambda_n^{1/2}), \quad R(\widehat{g}) = \mathcal{O}_p(1), \tag{35}$$

where $\|\cdot\|_2$ means the $L_2(p_{\mathrm{x}}p_{\mathrm{y}})$-norm and $\mathcal{O}_p$ denotes the asymptotic order in probability.

Proof of Lemma 2. Hoeffding (1963) derived Bernstein's inequality for double sum, i.e.,

$$P\left(|(\widetilde{Q}_n - Q)f| > \frac{x}{\sqrt{n}}\right) \le 2\exp\left(-\frac{x^2}{4(\|f\|_\infty x/(3\sqrt{n}) + Qf^2)}\right)$$

for $x \ge 0$. By applying the above inequality to the proof of Theorem 2 in Birgé and Massart (1993) or Theorem 5.11 in van de Geer (2000) instead of Bernstein's inequality for i.i.d. sum, we obtain a "double sum" version of those theorems. That is, exponential decay of the tail probability of $\sqrt{n}\sup_{f\in\mathcal{F}}|(Q_n - Q)f|$, where $\mathcal{F}$ is a class of uniformly bounded measurable functions and satisfies a polynomial bracketing condition (14) as $\mathcal{G}_M$ (see Theorem 3 in the supplementary material). Later, this is used to obtain Eqs.(37) and (38).

From the definition, we obtain

$$\begin{aligned} &\frac{1}{2}Q_n\widehat{g}^2 - P_n\widehat{g} + \lambda_n R(\widehat{g})^2 \le \frac{1}{2}Q_n g^{*2} - P_n w + \lambda_n R(g^*)^2 \\ \Rightarrow\quad &\frac{1}{2}Q_n(\widehat{g} - g^*)^2 - Q_n(g^*(g^* - \widehat{g})) - P_n(\widehat{g} - g^*) + \lambda_n(R(\widehat{g})^2 - R(g^*)^2) \le 0. \end{aligned}$$

On the other hand, $Q(g^*(g^* - \widehat{g})) = P(g^* - \widehat{g})$ indicates

$$\begin{aligned} &\frac{1}{2}(Q - Q_n)(\widehat{g} - g^*)^2 - (Q - Q_n)(g^*(g^* - \widehat{g})) - (P - P_n)(\widehat{g} - g^*) - \lambda_n(R(\widehat{g})^2 - R(g^*)^2) \\ &\ge \frac{1}{2}Q(\widehat{g} - g^*)^2. \end{aligned}$$

Therefore, to bound $\|\widehat{g} - g^*\|_2$, it suffices to bound the LHS of the above inequality.

Define $\mathcal{F}_M$ and $\mathcal{F}_M^2$ as

$$\mathcal{F}_M := \{g - g^* \mid g \in \mathcal{G}_M\}, \quad \mathcal{F}_M^2 := \{f^2 \mid f \in \mathcal{F}_M\}.$$

For $g, g' \in \mathcal{G}_M$ such that $\|g - g'\|_2 \le \delta$, the $L_2(p_{\mathrm{x}}p_{\mathrm{y}})$-norm of the difference between $(g - g^*)^2$ and $(g' - g^*)^2$ is bounded by

$$\|(g - g^*)^2 - (g' - g^*)^2\|_2 = \|(g - g')(g + g' - 2g^*)\|_2 \le 2\delta(M + M_0).$$

Thus, the bracketing entropy of $\mathcal{F}_M^2$ has the following order:

$$\mathcal{H}_{[]}(\mathcal{F}_M^2, \delta, L_2(p_{\mathrm{x}}p_{\mathrm{y}})) = O\left(\left(\frac{(M + M_0)^2}{\delta}\right)^\gamma\right) \qquad \text{as } \delta \to 0.$$

Let $f := \widehat{g} - g^*$. Then, as in Lemma 5.14 and Theorem 10.6 in van de Geer (2000), we obtain

$$|(Q_n - Q)(f^2)| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f^2\|_2^{1-\frac{\gamma}{2}}(1 + R(\widehat{g})^2 + M_0^2)^{\frac{\gamma}{2}} \vee n^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right), \tag{36}$$

where $a \vee b = \max(a, b)$. Here, we have used the double sum version of Theorem 5.11 in van de Geer (2000), which is needed to obtain the same formula as Lemma 5.14 in van de Geer (2000). Since

$$\|f^2\|_2 \leq \|f\|_2\sqrt{2(1 + R(\widehat{g})^2 + M_0^2)},$$

the RHS of Eq.(36) is further bounded by

$$|(Q_n - Q)(f^2)| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\frac{\gamma}{2}}(1 + R(\widehat{g})^2 + M_0^2)^{\frac{1}{2}+\frac{\gamma}{4}} \vee n^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right). \tag{37}$$

Similarly, we can show that

$$|(Q_n - Q)(g^*(g^* - \widehat{g}))| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\frac{\gamma}{2}}(1 + R(\widehat{g}) + M_0)^{\frac{\gamma}{2}} \vee n^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g}) + M_0)\right). \tag{38}$$

Note that, for $g, g' \in \mathcal{G}$,

$$\int (g - g')^2 dp_{\mathrm{xy}} = \int (g - g')^2 w \mathrm{d}p_{\mathrm{x}}\mathrm{d}p_{\mathrm{y}} \leq M_0\|g - g'\|_2^2,$$

which implies

$$\mathcal{H}_{[]}(\mathcal{F}_M, \delta, L_2(p_{\mathrm{xy}})) = O\left(\left(\frac{M + M_0}{\delta}\right)^{\gamma}\right).$$

Thus,

$$|(P_n - P)(g^* - \widehat{g})| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\frac{\gamma}{2}}(1 + R(\widehat{g}) + M_0)^{\frac{\gamma}{2}} \vee n^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g}) + M_0)\right). \tag{39}$$

Combining Eqs.(37), (38), and (39), we can bound the $L_2(p_{\mathrm{x}}p_{\mathrm{y}})$-norm of f as

$$\begin{aligned}&\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2\\ &\leq \lambda_n R(g^*)^2 + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\frac{\gamma}{2}}(1 + R(\widehat{g})^2 + M_0^2)^{\frac{1}{2}+\frac{\gamma}{4}} \vee n^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right).\end{aligned}$$

The rest can be proved by following a similar line to Theorem 10.6 in van de Geer (2000).

We redefine $M_0 \leftarrow \max\{R(g^*), M_0\}$, and define $J = 1 + R(\widehat{g})^2 + M_0^2$. There are two possible situations, namely, **(a)** $R(\widehat{g}) \geq M_0 + 1$ or **(b)** $R(\widehat{g}) < M_0 + 1$. We show the stochastic order of $\|f\|_2$ and $R(\widehat{g})$ for the two situations separately.

(a) (i) If $\|f\|_2 \geq n^{-1/(2+\gamma)}J$, then

$$\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq \lambda_n M_0^2 + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}J^{1/2+\gamma/4}\right).$$

In this case

$$\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq 2\lambda_n M_0^2 \quad \text{or} \quad \frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}J^{1/2+\gamma/4}\right).$$

For the first event, we have

$$\|f\|_2^2 \leq 4\lambda_n M_0^2 = O(\lambda_n), \quad R(\widehat{g}) \leq \sqrt{2}M_0 = \mathcal{O}_p(1).$$

On the other hand, for the second event, we have

$$\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}J^{1/2+\gamma/4}\right) \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}R(\widehat{g})^{1+\gamma/2}\right),$$

which indicates

$$\frac{1}{2}\|f\|_2^2 \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}R(\widehat{g})^{1+\gamma/2}\right) \quad \Rightarrow \quad \|f\|_2 \leq \mathcal{O}_p\left(n^{-1/(2+\gamma)}R(\widehat{g})\right),$$

and

$$\begin{aligned} &\lambda_n R(\widehat{g})^2 \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}R(\widehat{g})^{1+\gamma/2}\right) \\ \Rightarrow \quad & R(\widehat{g}) \leq \lambda_n^{-2/(2-\gamma)}\mathcal{O}_p\left(n^{-1/(2-\gamma)}\frac{R(\widehat{g})}{n^{1/(2+\gamma)}}\right) = o_p(R(\widehat{g})) \;\; (\because \lambda_n^{-1} = o(n^{2/(2+\gamma)})). \end{aligned}$$

Thus, the probability of this event goes to 0.

(ii) If $\|f\|_2 \leq n^{-1/(2+\gamma)}J$, then

$$\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq \lambda_n M_0^2 + \mathcal{O}_p\left(n^{-2/(2+\gamma)}R(\widehat{g})^2\right).$$

In this case,

$$\lambda_n R(\widehat{g})^2 \leq 2\lambda_n M_0^2 \;\; \text{or} \;\; \lambda_n R(\widehat{g})^2 \leq \mathcal{O}_p\left(n^{-2/(2+\gamma)}R(\widehat{g})^2\right).$$

Then, similarly to the above case, we can show that the probability of the second event goes to 0. On the other hand, for the first event, $\|f\|_2 \leq n^{-1/(2+\gamma)}J \leq n^{-1/(2+\gamma)}(1+2M_0^2) = \mathcal{O}_p(\lambda_n^{1/2})$ and $R(\widehat{g}) \leq \sqrt{2}M_0 = \mathcal{O}_p(1)$.

(b) In this situation, $R(\widehat{g}) < M_0 + 1 = \mathcal{O}_p(1)$.

(i) If $\|f\|_2 \geq n^{-1/(2+\gamma)}J$, then

$$\frac{1}{2}\|f\|_2^2 + \lambda_n R(\widehat{g})^2 \leq \lambda_n M_0^2 + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}\right).$$

In this case,

$$\frac{1}{2}\|f\|_2^2 \le \mathcal{O}_p\left(\lambda_n \vee \frac{1}{\sqrt{n}}\|f\|_2^{1-\gamma/2}\right) \;\Rightarrow\; \|f\|_2 \le \mathcal{O}_p\left(\lambda_n^{1/2} \vee n^{-1/(2+\gamma)}\right) = \mathcal{O}_p(\lambda_n^{1/2}).$$

(ii) If $\|f\|_2 \le n^{-1/(2+\gamma)}J$, it is obvious that $\|f\|_2 = \mathcal{O}_p(\lambda_n^{1/2})$.

Consequently, the proof of Lemma 2 was completed. □

Based on Lemma 2, we prove Theorem 2 below.

As in the proof of Lemma 2, let $f := \widehat{g} - g^*$. Since $Q(fg^*) = Pf$, we have

$$\begin{aligned}
&\frac{1}{2}Q_n\widehat{g}^2 - P_n\widehat{g} - (\frac{1}{2}Qg^{*2} - Pg^*) \\
&= \frac{1}{2}Q_n(f+g^*)^2 - P_n(f+g^*) - \left(\frac{1}{2}Qg^{*2} - Pg^*\right) \\
&= \frac{1}{2}Qf^2 + \frac{1}{2}(Q_n - Q)f^2 + (Q_n - Q)(g^*f) - (P_n - P)f \\
&\quad + \frac{1}{2}(Q_ng^{*2} - Qg^{*2}) - (P_ng^* - Pg^*). \qquad (40)
\end{aligned}$$

Below, we show that each term of the RHS of the above equation is $\mathcal{O}_p(\lambda_n)$. By the central limit theorem, we have

$$\frac{1}{2}(Q_ng^{*2} - Q_ng^{*2}) - (P_ng^* - Pg^*) = \mathcal{O}_p(n^{-1/2}).$$

Since Lemma 2 gives $\|f\|_2^2 = \mathcal{O}_p(\lambda_n)$ and $R(\widehat{g}) = \mathcal{O}_p(1)$, Eqs.(37), (38), and (39) in the proof of Lemma 2 imply

$$\begin{aligned}
|(Q_n - Q)f^2| &= \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\|f\|_2^{1-\frac{\gamma}{2}} \vee n^{-\frac{2}{2+\gamma}}\right) \le \mathcal{O}_p(\lambda_n), \\
|(Q_n - Q)(g^*f)| &= \mathcal{O}_p(\lambda_n), \\
|(P_n - P)g^*| &= \mathcal{O}_p(\lambda_n).
\end{aligned}$$

In the above derivation, $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$ was used. Lemma 2 also implies

$$Qf^2 = \|f\|_2^2 = \mathcal{O}_p(\lambda_n).$$

Combining these inequalities with Eq.(40) implies

$$\begin{aligned}
\widehat{\mathrm{SMI}}(Y,Z) - \mathrm{SMI}(Y,Z) &= -\frac{1}{2}Q_n\widehat{g}^2 + P_n\widehat{g} - (-\frac{1}{2}Qg^{*2} + Pg^*) \\
&= \mathcal{O}_p(\lambda_n + n^{-1/2}) = \mathcal{O}_p(\max(\lambda_n, n^{-1/2})),
\end{aligned}$$

which concludes the proof of Theorem 2. ■

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.

Birgé, L., & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, *97*, 113–150.

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

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 580–598.

Bura, E., & Cook, R. D. (2001). Extending sliced inverse regression. *Journal of the American Statistical Association*, *96*, 996–1003.

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

Collins, M., & Duffy, N. (2002). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Cook, R. D. (1998a). Principal Hessian directions revisited. *Journal of the American Statistical Association*, *93*, 84–100.

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

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

Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, *100*, 410–428.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory.* Hoboken, NJ, USA: John Wiley & Sons, Inc. 2nd edition.

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.

Durand, J., & Sabatier, R. (1997). Additive splines for partial least squares regression. *Journal of the American Statistical Association*, *92*, 1546–1554.

Eberts, M., & Steinwart, I. (2011). Optimal learning rates for least squares svms using gaussian kernels. *Advances in Neural Information Processing Systems 24* (pp. 1539–1547).

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, *20*, 303–353.

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

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.

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

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory.*

Goutis, C., & Fearn, T. (1996). Partial least squares regression on smooth factors. *Journal of the American Statistical Association*, *91*, 627–632.

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.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, *19*, 293–325.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*, 13–30.

Hotelling, H. (1936). Relations between two sets of cariants. *Biometrika*, *28*, 321–377.

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

Kashima, H., & Koyanagi, T. (2002). Kernels for semi-structured data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 291–298). San Francisco, CA: Morgan Kaufmann.

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. *Proceedings of the 20th International Conference on Machine Learning*. San Francisco, CA.

Keziou, A. (2003). Dual representation of φ-divergences and applications. *C. R. Acad. Sci. Paris, Ser. I*, *336*, 857–862.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 315–322).

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

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*, 1025–1039.

Li, K. C., Lue, H. H., & Chen, C. H. (2000). Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association*, *95*, 547–560.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, *2*, 419–444.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*. to appear.

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

Reiss, P. T., & Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, *102*, 984–996.

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. *New Challenges for Feature Selection in Data Mining and Knowledge Discovery* (pp. 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. (2000). *Asymptotic statistics*. Cambridge University Press.

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

Wahba, G. (1990). *Spline model for observational data*. Philadelphia and Pennsylvania: Society for Industrial and Applied Mathematics.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis*, 391–420. New York, NY, USA: Academic Press.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, *35*, 2654–2690.

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, *101*, 630–643.