Approximating Mutual Information by Maximum Likelihood Density Ratio Estimation

Taiji Suzuki S-TAIJI@STAT.T.U-TOKYO.AC.JP
Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Jun Sese SESEJUN@IS.OCHA.AC.JP
Department of Information Science, Ochanomizu University, 2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Takafumi Kanamori KANAMORI@IS.NAGOYA-U.AC.JP
Department of Computer Science and Mathematical Informatics, Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan

Editor: Saeys et al.

Abstract

Mutual information is useful in various data processing tasks such as feature selection or independent component analysis. In this paper, we propose a new method of approximating mutual information based on maximum likelihood estimation of a *density ratio* function. Our method, called Maximum Likelihood Mutual Information (MLMI), has several attractive properties, e.g., density estimation is not involved, it is a single-shot procedure, the global optimal solution can be efficiently computed, and cross-validation is available for model selection. Numerical experiments show that MLMI compares favorably with existing methods.

1. Introduction

Detection of dependencies between random variables is highly useful in various machine learning problems such as feature selection (Guyon and Elisseeff, 2003; Torkkola, 2003) and independent component analysis (Comon, 1994). Although classical correlation analysis would be still useful in these problems, it cannot be used for discovering non-linear dependencies with no correlation. On the other hand, *mutual information* (MI), which plays an important role in information theory (Cover and Thomas, 1991), allows us to identify general nonlinear dependencies. MI is defined by

$$I(X,Y) := \iint p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y}) \log \left(\frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y})} \right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \tag{1}$$

and it vanishes if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are *independent*. For this reason, estimating MI from samples has gathered a lot of attention for many years.

A naive approach to estimating MI is to use a *kernel density estimator* (KDE) (Silverman, 1986; Fraser and Swinney, 1986), i.e., the densities $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$, $p_{\mathrm{x}}(\boldsymbol{x})$, and $p_{\mathrm{y}}(\boldsymbol{y})$ are separately estimated from samples and the estimated densities are used for computing MI. The bandwidth of the kernel functions could be optimized based on likelihood cross-validation (Härdle et al., 2004), so there remains no open tuning parameter in this approach. However, density estimation is known to be a hard problem and division by estimated densities is involved when approximating MI, which tend to magnify the estimation error. Therefore, the KDE-based method may not be reliable in practice.

Alternative methods involve estimation of the *entropies* using k-nearest neighbor (KNN) samples (Kraskov et al., 2004) or using the Edgeworth (EDGE) expansion (Hulle, 2005). The KNN-based approach was shown to perform better than KDE (Khan et al., 2007), given that the number k is chosen appropriately—a small (large) k yields an estimator with small (large) bias and large (small) variance. However, appropriately determining the value of k so that the bias-variance trade-off is optimally controlled is not straightforward in the context of MI estimation. The EDGE method works well when the target density is close to the normal distribution; otherwise it is biased and therefore not reliable.

In this paper, we propose a new MI estimator that can overcome the limitations of the existing approaches. Our method, which we call Maximum Likelihood Mutual Information (MLMI), does not involve density estimation and directly models the *density ratio*:

$$w(\boldsymbol{x}, \boldsymbol{y}) := \frac{p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y})}. \tag{2}$$

Thus it is a single-shot procedure without division by estimated quantities and therefore the estimation error is not further expanded. The density ratio is estimated by the maximum likelihood method and it is cast as a convex optimization problem. Therefore, the unique global optimal solution can be obtained efficiently. Furthermore, cross-validation (CV) is available for model selection, so the values of tuning parameters such as the kernel width can be adaptively determined in an objective manner. Our method does not assume normality of the target distribution and therefore is flexible. Numerical experiments show that MLMI compares favorably with existing methods.

A sibling of MLMI, called LSMI, is presented in Suzuki et al. (2008). In that paper, we used a least-squares method for density ratio based MI estimation and emphasized its practical usefulness in variable selection. On the other hand, the current paper employs a maximum likelihood method and focuses more on mathematical aspects of density ratio based MI estimation.

2. Approximating MI by Maximum Likelihood Density Ratio Estimation

In this section, we formulate the MI approximation problem as a density ratio estimation problem and propose a new MI estimation method.

2.1 Formulation

Let $\mathcal{D}_{\mathrm{X}}$ ($\subset \mathbb{R}^{d_{\mathrm{x}}}$) and $\mathcal{D}_{\mathrm{Y}}$ ($\subset \mathbb{R}^{d_{\mathrm{y}}}$) be data domains and suppose we are given n independent and identically distributed (i.i.d.) paired samples

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

drawn from a joint distribution with density $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$. Let us denote the marginal densities of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ by $p_{\mathrm{x}}(\boldsymbol{x})$ and $p_{\mathrm{y}}(\boldsymbol{y})$, respectively. The goal is to estimate MI defined by Eq.(1).

Our approach here is to estimate the *density ratio* $w(\boldsymbol{x}, \boldsymbol{y})$ defined by Eq.(2); then MI can be approximated using a density ratio estimator $\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$ by

$$\widehat{I}(X, Y) := \frac{1}{n} \sum_{i=1}^{n} \log \widehat{w}(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

We model the density ratio function $w(\boldsymbol{x}, \boldsymbol{y})$ by the following linear model:

$$\widehat{w}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}), \tag{3}$$

where $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$ are parameters to be learned from samples, $^\top$ denotes the transpose of a matrix or a vector, and

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) := (\varphi_1(\boldsymbol{x}, \boldsymbol{y}), \varphi_2(\boldsymbol{x}, \boldsymbol{y}), \ldots, \varphi_b(\boldsymbol{x}, \boldsymbol{y}))^\top$$

are basis functions such that

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) \geq \mathbf{0}_b \quad \text{for all } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}.$$

$\mathbf{0}_b$ denotes the b-dimensional vector with all zeros. Note that $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ could be dependent on the samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, i.e., *kernel* models are also allowed. We explain how the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ are designed in Section 2.4.

2.2 Maximum Likelihood Estimation of Density Ratio Function

Using an estimated density ratio $\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$, we may estimate the joint density $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ by

$$\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) := \widehat{w}(\boldsymbol{x}, \boldsymbol{y}) p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}).$$

Based on this, we determine the parameter $\boldsymbol{\alpha}$ in the model $\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$ so that the following log-likelihood L is maximized:

$$L(\boldsymbol{\alpha}) := \sum_{i=1}^{n} \log \widehat{p}_{\mathrm{xy}}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \sum_{i=1}^{n} \log \left(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right) + \sum_{i=1}^{n} \log p_{\mathrm{x}}(\boldsymbol{x}_i) + \sum_{i=1}^{n} \log p_{\mathrm{y}}(\boldsymbol{y}_i).$$

The second and the third terms are constants and therefore can be safely ignored. This is our objective function to be maximized with respect to the parameters $\boldsymbol{\alpha}$, which is concave. Note that this corresponds to an empirical approximation of the Kullback-Leibler divergence from $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ to $\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ up to some irrelevant constant.

$\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$ is an estimator of the density ratio $w(\boldsymbol{x}, \boldsymbol{y})$ which is non-negative by definition. Therefore, it is natural to impose $\widehat{w}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_X \times \mathcal{D}_Y$, which can be achieved by restricting $\boldsymbol{\alpha} \geq \mathbf{0}_b$. In addition to non-negativity, $\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$ should be properly normalized since $\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ is a probability density function:

$$1 = \int \widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} = \int \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \approx \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_j), \tag{4}$$

where we used the U-statistic (Serfling, 1980, p.171) for obtaining the empirical estimator.

Now our optimization criterion is summarized as follows.

$$\begin{aligned}&\underset{\boldsymbol{\alpha}\in\mathbb{R}^b}{\text{maximize}} \left[\sum_{i=1}^{n} \log\left(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right)\right] \\ &\text{subject to } \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1 \ \text{ and } \ \boldsymbol{\alpha} \ge \mathbf{0}_b. \end{aligned} \tag{5}$$

We call the above method *Maximum Likelihood Mutual Information (MLMI).*

MLMI can be characterized by *Legendre-Fenchel duality* of the convex function '$-\log$' (Rockafellar, 1970; Boyd and Vandenberghe, 2004). For $f(u) = -\log(u)$, MI is expressed as

$$I(X,Y) = \int p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y}) f\left(\frac{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})}{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

Let f^* be the Legendre-Fenchel dual function of f, which is defined and given by

$$f^*(v) := \sup_{u\in\boldsymbol{R}}\{uv - f(u)\} = -1 - \log(-v) \quad (\text{for } v < 0).$$

Then Legendre-Fenchel duality implies that $I(X,Y)$ is obtained by solving the following concave maximization problem (Nguyen et al., 2008):

$$\begin{aligned} I(X,Y) &= \sup_{\widehat{w}\ge 0}\left[\int p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})\left(-\widehat{w}(\boldsymbol{x},\boldsymbol{y})\frac{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})}{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})} - f^*(-\widehat{w}(\boldsymbol{x},\boldsymbol{y}))\right)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\right] \\ &= \sup_{\widehat{w}\ge 0}\left[\int \Big(- \widehat{w}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y}) + \log \widehat{w}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})\Big)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} + 1\right], \end{aligned}$$

where the supremum is taken over all non-negative measurable functions. If the linear model assumption (3) and the normalization constraint (4) are imposed and the expectation is approximated by the sample average, the above formulation is reduced to MLMI.

2.3 Convergence Bound

Here, we show a non-parametric convergence rate of the solution of the optimization problem (5). The set of basis functions is denoted by

$$\mathcal{F} := \{\varphi_\theta \mid \theta \in \Theta\},$$

where Θ is some parameter or index set. The set of basis functions used for estimation at n samples is characterized by a subset of the parameter set $\Theta_n \subseteq \Theta$ and denoted by

$$\mathcal{F}_n := \{\varphi_\theta \mid \theta \in \Theta_n\} \subset \mathcal{F},$$

which can behave stochastically. The set of finite linear combinations of $\mathcal{F}$ with positive coefficients and its bounded subset are denoted by

$$\mathcal{G} := \left\{\sum_l \alpha_l \varphi_{\theta_l} \,\middle|\, \alpha_l \ge 0,\ \varphi_{\theta_l} \in \mathcal{F}\right\}, \quad \mathcal{G}^M := \{g \in \mathcal{G} \mid \|g\|_\infty \le M\},$$

and their subsets used for estimation at n samples are denoted by

$$\mathcal{G}_n := \left\{ \sum_l \alpha_l \varphi_{\theta_l} \;\middle|\; \alpha_l \geq 0,\ \varphi_{\theta_l} \in \mathcal{F}_n \right\} \subset \mathcal{G}, \quad \mathcal{G}_n^M := \{g \in \mathcal{G}_n \mid \|g\|_\infty \leq M\} \subset \mathcal{G}^M.$$

Let $\widehat{\mathcal{G}}_n$ be the feasible set of MLMI:

$$\widehat{\mathcal{G}}_n := \left\{ g \in \mathcal{G}_n \;\middle|\; \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} g(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1 \right\}.$$

Then, the solution $\widehat{g}_n$ of (generalized) MLMI is given as follows:

$$\widehat{g}_n := \operatorname*{arg\,max}_{g \in \widehat{\mathcal{G}}_n} \left[\frac{1}{n} \sum_{i=1}^n \log g(\boldsymbol{x}_i, \boldsymbol{y}_i) \right].$$

For simplicity, we assume that the optimal solution can be uniquely determined. In order to derive the convergence rates of MLMI, we make the following assumptions. Let g_0 denote

$$g_0(\boldsymbol{x}, \boldsymbol{y}) := \frac{p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y})}.$$

We define the (generalized) Hellinger distance with respect to $p_{\mathrm{x}} p_{\mathrm{y}}$ as

$$h_Q(g, g') := \left(\int (\sqrt{g} - \sqrt{g'})^2 p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \right)^{1/2},$$

where g and g' are non-negative measurable functions (not necessarily probability densities).

Assumption 1

1. On the support of p_{xy}, there exists a constant $\eta_1 < \infty$ such that

$$g_0 \leq \eta_1.$$

2. All basis functions are non-negative, and there exist constants $\epsilon_0,\ \xi_0 > 0$ such that

$$\int \varphi(\boldsymbol{x}, \boldsymbol{y}) p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \geq \epsilon_0, \quad \|\varphi\|_\infty \leq \xi_0, \quad (\forall \varphi \in \mathcal{F}).$$

3. There exist constants $0 < \gamma < 2$ and K such that

$$\log N_{[]}(\epsilon, \mathcal{G}^M, h_Q) \leq K \left(\frac{\sqrt{M}}{\epsilon} \right)^\gamma, \tag{6}$$

where $N_{[]}(\epsilon, \mathcal{F}, d)$ is the ϵ-bracketing number of $\mathcal{F}$ with norm d (van de Geer, 2000)[1].

Then we obtain the following theorem and corollary (see Appendix for a sketch of proof).

1. As shown in Section 2.4, we will use Gaussian kernel models in practice. The bracketing number for Gaussian mixture models is extensively discussed, e.g., in Ghosal and van der Vaart (2001).

Theorem 1 *Let*

$$g_n^* := \max_{g \in \widehat{\mathcal{G}}_n} \int \log(g(\boldsymbol{x}, \boldsymbol{y}))p(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

In addition to Assumption 1, if there exist constants $\widetilde{c}_0, c_0$ such that g_n^* satisfies

$$P\left(\widetilde{c}_0 \leq \frac{g_0}{g_n^*} \leq c_0\right) \to 1, \tag{7}$$

then

$$h_Q(\widehat{g}_n, g_0) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + h_Q(g_n^*, g_0)),$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability.

Corollary 2 *If there exists N such that $g_0 \in \mathcal{G}_n$ $(\forall n \geq N)$, then*

$$h_Q(\widehat{g}_n, g_0) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}}).$$

In Sugiyama et al. (2008), a similar convergence result to the above theorem has been provided, where g_0 was assumed to be bounded from both below and above. On the other hand, the current proof only requires an upper bound on g_0 (see Assumption 1.1). Thus, the above theorem is more general than the result in Sugiyama et al. (2008) in terms of g_0.

More technically, we use the bracketing number with respect to the Hellinger distance for describing the complexity of a function class, while Sugiyama et al. (2008) used the bracketing number with respect to the ℓ_2 distance or the covering number. The main mathematical device used in Sugiyama et al. (2008) was to bound $h_Q(\widehat{g}_n, g_n^*)$ by the difference between the empirical mean and the expectation of $\log(2g_n^*/(\widehat{g}_n + g_n^*))$. On the other hand, the current paper employs $2\widehat{g}_n/(\widehat{g}_n + g_n^*)$ instead of $\log(2g_n^*/(\widehat{g}_n + g_n^*))$, which enables us to replace the lower bound of g_0 with the bounds of the ratio g_0/g_n^*. Then we utilize the convexity of $\widehat{\mathcal{G}}_n$ and the bracketing number condition with respect to the Hellinger distance (see Section 7 of van de Geer (2000) for more details).

2.4 Model Selection by Likelihood Cross Validation

The performance of MLMI depends on the choice of the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$. Here we show that model selection can be carried out based on a variant of CV.

First, the samples $\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$ of (approximately) the same size. Then a density ratio estimator $\widehat{w}_{\mathcal{Z}_k}(\boldsymbol{x}, \boldsymbol{y})$ is obtained using $\{\mathcal{Z}_j\}_{j \neq k}$ (i.e., without $\mathcal{Z}_k$) and the log-likelihood for the hold-out samples $\mathcal{Z}_k$ is computed as

$$L_{\mathcal{Z}_k}^{(K\text{-CV})} = \frac{1}{|\mathcal{Z}_k|} \sum_{(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_k} \log \widehat{w}_{\mathcal{Z}_k}(\boldsymbol{x}', \boldsymbol{y}'),$$

where $|\mathcal{Z}_k|$ denotes the number of sample pairs in the set $\mathcal{Z}_k$. This procedure is repeated for $k = 1, 2, \ldots, K$ and its average $L^{(K\text{-CV})}$ is outputted:

$$L^{(K\text{-CV})} = \frac{1}{K} \sum_{k=1}^K L_{\mathcal{Z}_k}^{(K\text{-CV})}.$$

For model selection, we compute $L^{(K\text{-CV})}$ for all model candidates (the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ in the current setting) and choose the one that maximizes the hold-out log-likelihood. Note that $L^{(K\text{-CV})}$ is an almost unbiased estimate of the Kullback-Leibler divergence from $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ to $\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$ up to some irrelevant constant, where the 'almost'-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting (Schölkopf and Smola, 2002).

A good model may be chosen by the above CV procedure, given that a set of promising model candidates is prepared. As model candidates, we propose using a Gaussian kernel model: for $\boldsymbol{z} = (\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top$,

$$\varphi_\ell(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{z} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{u}_\ell\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{v}_\ell\|^2}{2\sigma^2}\right),$$

where $\{\boldsymbol{c}_\ell \mid \boldsymbol{c}_\ell = (\boldsymbol{u}_\ell^\top, \boldsymbol{v}_\ell^\top)^\top\}_{\ell=1}^b$ are Gaussian centers; we choose the centers randomly from $\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i^\top, \boldsymbol{y}_i^\top)^\top\}_{i=1}^n$. Below, we fix the number of basis functions at $b = \min(200, n)$ and choose the Gaussian width σ by CV.

3. Relation to Existing Methods

In this section, we discuss the characteristics of existing and proposed approaches.

3.1 Kernel Density Estimator (KDE)

KDE is a non-parametric technique to estimate a probability density function $p(\boldsymbol{x})$ defined on $\mathbb{R}^d$ from its i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$. For the Gaussian kernel, KDE is expressed as

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n(2\pi\sigma^2)^{d/2}}\sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right).$$

The performance of KDE depends on the choice of the kernel width σ and it can be optimized by *likelihood CV* as follows (Härdle et al., 2004): First, divide the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ into K disjoint subsets $\{\mathcal{X}_k\}_{k=1}^K$ of (approximately) the same size. Then obtain a density estimate $\widehat{p}_{\mathcal{X}_k}(\boldsymbol{x})$ from $\{\mathcal{X}_j\}_{j\neq k}$ (i.e., without $\mathcal{X}_k$) and compute its hold-out log-likelihood for $\mathcal{X}_k$:

$$\frac{1}{|\mathcal{X}_k|}\sum_{\boldsymbol{x}\in\mathcal{X}_k} \log \widehat{p}_{\mathcal{X}_k}(\boldsymbol{x}).$$

This procedure is repeated for $k = 1, 2, \ldots, K$ and choose the value of σ such that the average of the hold-out log-likelihood over all k is maximized. Note that the average hold-out log-likelihood is an almost unbiased estimate of the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x})$ up to some irrelevant constant.

Based on KDE, MI can be approximated using density estimates $\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$, $\widehat{p}_{\mathrm{x}}(\boldsymbol{x})$, and $\widehat{p}_{\mathrm{y}}(\boldsymbol{y})$ (obtained from $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, $\{\boldsymbol{x}_i\}_{i=1}^n$, and $\{\boldsymbol{y}_i\}_{i=1}^n$, respectively) as

$$\widehat{I}(X, Y) := \frac{1}{n}\sum_{i=1}^n \log \frac{\widehat{p}_{\mathrm{xy}}(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\widehat{p}_{\mathrm{x}}(\boldsymbol{x}_i)\widehat{p}_{\mathrm{y}}(\boldsymbol{y}_i)}.$$

However, density estimation is known to be a hard problem and division by estimated densities may expand the estimation error. For this reason, the KDE-based approach may not be reliable in practice.

3.2 K-nearest Neighbor Method (KNN)

MI can be expressed in terms of the entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y),$$

where $H(X)$ denotes the entropy of X:

$$H(X) := -\int p_{\mathrm{x}}(\boldsymbol{x}) \log p_{\mathrm{x}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Thus MI can be approximated if the entropies $H(X)$, $H(Y)$, and $H(X,Y)$ are estimated.

Kraskov et al. (2004) developed an entropy estimator that utilizes the k-nearest neighbor distance (KNN). Let us define the norm of $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ by $\|\boldsymbol{z}\|_{\mathrm{z}} := \max\{\|\boldsymbol{x}\|, \|\boldsymbol{y}\|\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\mathcal{N}_k(i)$ be the set of k-nearest neighbor samples of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with respect to the norm $\|\cdot\|_{\mathrm{z}}$, and let

$$\epsilon_{\mathrm{x}}(i) := \max\{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \mid (\boldsymbol{x}_{i'}, \boldsymbol{y}_{i'}) \in \mathcal{N}_k(i)\}, \qquad n_{\mathrm{x}}(i) := |\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \leq \epsilon_{\mathrm{x}}(i)\}|,$$
$$\epsilon_{\mathrm{y}}(i) := \max\{\|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \mid (\boldsymbol{x}_{i'}, \boldsymbol{y}_{i'}) \in \mathcal{N}_k(i)\}, \qquad n_{\mathrm{y}}(i) := |\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \leq \epsilon_{\mathrm{y}}(i)\}|.$$

Then the KNN-based MI estimator is given by

$$\widehat{I}(X,Y) := \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}\Big[\psi(n_x(i)) + \psi(n_y(i))\Big],$$

where ψ is the *digamma* function.

An advantage of the above KNN-based method is that it does not simply replace entropies with their estimates, but it is designed to cancel the error of individual entropy estimation. A practical drawback of the KNN-based approach is that the estimation accuracy depends on the value of k and there seems no systematic strategy to choose the value of k appropriately.

3.3 Edgeworth Expansion Method (EDGE)

Hulle (2005) proposed an entropy approximation method based on the *Edgeworth expansion*, where the entropy of a distribution is approximated by that of the normal distribution and some additional higher-order correction terms. More specifically, for a d-dimensional distribution, an estimator $\widehat{H}$ of the entropy H is given by

$$\widehat{H} = H_{\mathrm{normal}} - \frac{1}{12}\sum_{i=1}^{d}\kappa_{i,i,i}^2 - \frac{1}{4}\sum_{i,j=1,i\neq j}^{d}\kappa_{i,i,j}^2 - \frac{1}{72}\sum_{i,j,k=1,i<j<k}^{d}\kappa_{i,j,k}^2,$$

where H_{normal} is the entropy of the normal distribution with covariance matrix equal to the target distribution and $\kappa_{i,j,k}$ $(1 \leq i,j,k \leq d)$ is the standardized third cumulant of the target distribution. In practice, all the cumulants are estimated from samples.

Table 1: Relation among existing and proposed MI estimators. If the order of the Edgeworth expansion is regarded a tuning parameter, model selection of EDGE should be 'Not available'.

	Division by estimated quantities	Model selection	Distribution
KDE	Involved	**Available**	**Free**
KNN	**Not involved**	Not available	**Free**
EDGE	**Not involved**	**Not necessary**	Nearly normal
MLMI	**Not involved**	**Available**	**Free**

Based on EDGE, MI can be approximated using entropy estimates $\widehat{H}(X)$, $\widehat{H}(Y)$, and $\widehat{H}(X,Y)$ as

$$\widehat{I}(X,Y) := \widehat{H}(X) + \widehat{H}(Y) - \widehat{H}(X,Y).$$

If the underlying distribution is close to the normal distribution, the above approximation is quite accurate and the EDGE method works very well. However, if the distributions are far from the normal distribution, the approximation error gets large and therefore the EDGE method is unreliable.

In principle, it is possible to include the fourth and even higher cumulants for further reducing the estimation bias. However, this in turn increases the estimation variance; the expansion up to the third cumulants seems to be reasonable.

3.4 Discussions

The characteristics of the proposed and existing MI estimators are summarized in Table 1. KDE is distribution-free and model selection is possible by CV. However, division by estimated densities is involved, which can result in magnifying the estimation error. KNN is distribution-free and does not involve division by estimated quantities. However, there is no model selection method for determining the number of nearest neighbors and therefore its practical performance is not reliable. EDGE does not involve division by estimated quantities and any tuning parameters. However, it is based on the assumption that the target distributions are close to the normal distribution and the result is not reliable if this assumption is violated. MLMI is distribution-free, it does not involve division by estimated quantities, and model selection is possible by CV. Thus MLMI overcomes the limitations of the existing approaches.

4. Numerical Experiments

In this section, we experimentally investigate the performance of the proposed and existing MI estimators using artificial datasets. We use the following four datasets (see Figure 1):

(a) Linear dependence: y has a linear dependence on x as

$$x \sim \mathcal{N}(x; 0, 0.5) \quad \text{and} \quad y|x \sim \mathcal{N}(y; 3x, 1),$$

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2.

(b) Non-linear dependence with correlation: y has a quadratic dependence on x as

$$x \sim \mathcal{N}(x; 0, 1) \quad \text{and} \quad y|x \sim \mathcal{N}(y; x^2, 1).$$

(c) Non-linear dependence without correlation: y has a lattice-structured dependence on x as

$$x \sim \mathcal{U}(x; -0.5, 0.5) \quad \text{and} \quad y|x \sim \begin{cases} \mathcal{N}(x; 0, 1/3) & \text{if } x \leq |1/6|, \\ 0.5\mathcal{N}(x; 1, 1/3) + 0.5\mathcal{N}(x; -1, 1/3) & \text{otherwise,} \end{cases}$$

where $\mathcal{U}(x; a, b)$ denotes the uniform density on (a, b).

(d) Independence: x and y are independent to each other as

$$x \sim \mathcal{U}(x; 0, 0.5) \quad \text{and} \quad y|x \sim \mathcal{N}(y; 0, 1).$$

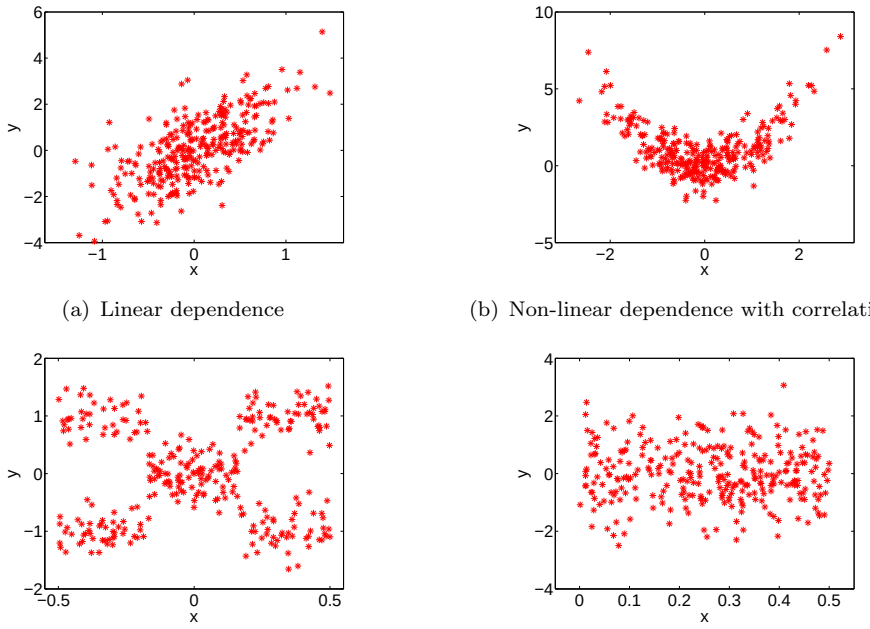
The task is to estimate MI between x and y. We compare the performance of MLMI(CV), KDE(CV), KNN(k) for $k = 1, 5, 15$, and EDGE; the approximation error of an MI estimate $\widehat{I}$ is measured by $|\widehat{I} - I|$. Figure 2 depicts the average approximation error—MLMI, KDE, KNN(5), and EDGE perform well for the dataset (a), MLMI tends to outperform the other estimators for the dataset (b), MLMI and KNN(5) show the best performance against the other methods for the dataset (c), and MLMI, EDGE, and KNN(15) perform well for the dataset (d). In the above simulation, KDE works moderately well for the datasets (a)–(c), while it performs poorly for the dataset (d). This instability would be due to devision by estimated densities, which tends to magnify the estimation error. KNN seems work well for all four datasets if the value of k is chosen optimally; the best value of k varies depending on the datasets and thus using a prefixed value of k is not appropriate—k needs to be chosen *adaptively* using the data samples. However, there is no systematic model selection strategy for KNN and therefore KNN would be unreliable in practice. EDGE works well for the datasets (a) and (b), which posses high normality[2]. However, for the datasets (c) and (d) where normality of the target distributions is low, the EDGE method performs poorly. In contrast, MLMI with CV performs reasonably well for all four datasets in a stable manner.

These experimental results show that MLMI nicely compensates for the weaknesses of the existing methods and therefore we conclude that MLMI should be regarded as a useful alternative to the existing methods of MI estimation.

5. Conclusions

In this paper, we proposed a new method of estimating mutual information. The proposed method, called MLMI, has several useful properties, e.g., it is a single-shot procedure, density estimation is not involved, it is equipped with a cross-validation procedure for model selection, and the unique global solution can be computed efficiently. We have provided a rigorous convergence analysis of the proposed algorithm as well as numerical experiments illustrating the usefulness of the proposed method.

2. Note that although the Edgeworth approximation is exact when the target distributions are completely normal, the EDGE method still suffers from some estimation error since the cumulants are estimated from samples.



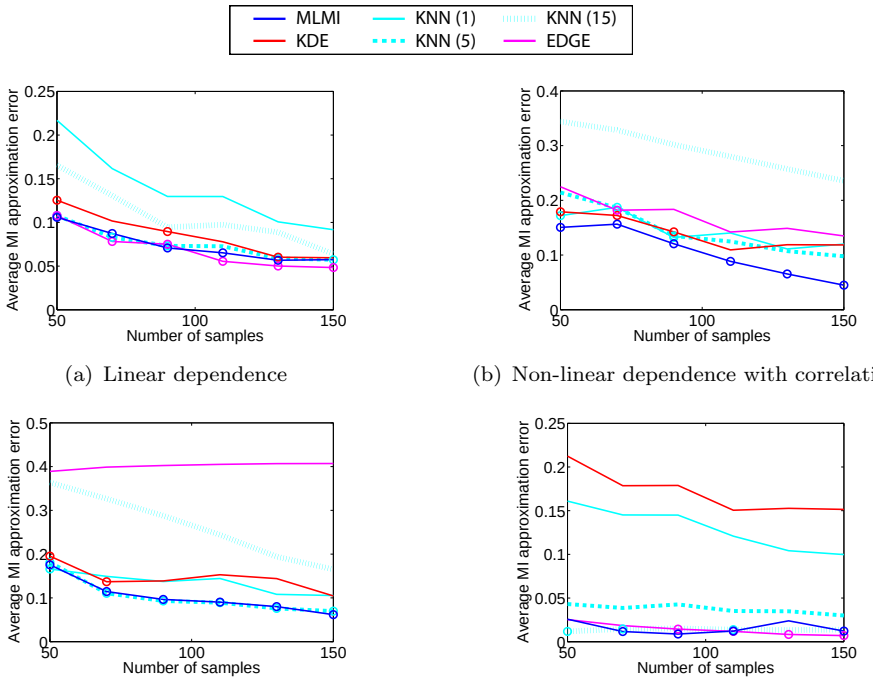
(a) Linear dependence

(b) Non-linear dependence with correlation

(c) Non-linear dependence without correlation

(d) Independence

Figure 1: Datasets used in experiments.



(a) Linear dependence

(b) Non-linear dependence with correlation

(c) Non-linear dependence without correlation

(d) Independence

Figure 2: MI approximation error measured by $|\widehat{I} - I|$ averaged over 100 trials as a function of the sample size n. The symbol '$\circ$' on a line means that the corresponding method is the best in terms of the average error or is judged to be comparable to the best method by the *t-test* at the significance level 1%.

Appendix: Sketch of Proof of Theorem 1 and Corollary 2

Proof of Theorem 1

For notational simplicity, we define the linear operators Q_n, Q, P_n, P as

$$Q_n f := \frac{\sum_{1\leq i\neq j\leq n} f(\boldsymbol{x}_i, \boldsymbol{y}_j)}{n(n-1)}, \quad Qf := \mathrm{E}_{p_{\mathrm{x}}p_{\mathrm{y}}} f, \quad P_n f := \frac{\sum_{i=1}^n f(\boldsymbol{x}_i, \boldsymbol{y}_i)}{n}, \quad Pf := \mathrm{E}_{p_{\mathrm{xy}}} f.$$

Step1: Proof of $\sup_{\varphi\in\mathcal{F}_n} |(Q_n - Q)\varphi| = \mathcal{O}_p(1/\sqrt{n})$. Hoeffding (1963) derived Bernstein's inequality for U-statistics which, in our context, is written as

$$P\left(|(Q_n - Q)f| > \frac{x}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{x}{(2x\|f\|_\infty/\sqrt{n} + 6Q(f - Q(f))^2)}\right), \tag{8}$$

for all $x > 0$ and $n \geq 2$ (see also Proposition 2.3 of Arcones and Giné, 1993). By applying the above inequality to the proof of Theorem 5.11 of van de Geer (2000) instead of Bernsten's inequality for the i.i.d. sum, we obtain a "double sum" version of Theorem 5.11 of van de Geer (2000), i.e., 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 the polynomial bracketing condition (6) as $\mathcal{G}^M$. This will be used for obtaining (9) later. By Assumption 1.2, we have $\mathcal{F}_n \subset \mathcal{G}^{\xi_0}$. Now $Q(\varphi - \varphi')^2 = Q(\sqrt{\varphi} - \sqrt{\varphi'})^2(\sqrt{\varphi} + \sqrt{\varphi'})^2 \leq 4\xi_0 h_Q(\varphi, \varphi')^2$ yields

$$N_{[]}(\epsilon, \mathcal{G}^{\xi_0}, L_2(Q)) \leq N_{[]}(\epsilon/2\sqrt{\xi_0}, \mathcal{G}^{\xi_0}, h_Q).$$

Thus the uniform convergence theorem (van de Geer, 2000, Theorem 5.11) with Assumption 1.3 implies

$$\sup_{\varphi\in\mathcal{F}_n} |(Q_n - Q)\varphi| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right). \tag{9}$$

Let $\widetilde{M} = 2\xi_0/\epsilon_0$ and $\mathcal{G}^{(\mathrm{c})}$ be

$$\mathcal{G}^{(\mathrm{c})} := \left\{ \frac{2g}{g + \bar{g}} \;\middle|\; g \in \mathcal{G}^{\widetilde{M}}, \bar{g} \in \mathcal{G}^{\widetilde{M}}, \widetilde{c}_0 \leq \frac{g_0}{\bar{g}} \leq c_0 \right\}.$$

Step2: Proof of $N_{[]}(\epsilon, \mathcal{G}^{(\mathrm{c})}, L_2(P)) \leq N_{[]}(\epsilon/8\sqrt{2c_0}, \mathcal{G}^{\widetilde{M}}, h_Q)^2$.

(9) yields

$$\inf_{\varphi\in\mathcal{F}_n} Q_n\varphi \geq \epsilon_0 - \mathcal{O}_p(1/\sqrt{n}),$$

and

$$Q(\bar{S}_n) \to 1 \text{ for } \bar{S}_n := \left\{ \inf_{\varphi\in\mathcal{F}_n} Q_n\varphi \geq \epsilon_0/2 \right\}. \tag{10}$$

On the event $\bar{S}_n$, all the elements of $\widehat{\mathcal{G}}_n$ is uniformly bounded from above:

$$1 = Q_n(\sum_l \alpha_l\varphi_l) = \sum_l \alpha_l Q_n(\varphi_l) \geq \sum_l \alpha_l\epsilon_0/2 \;\; \Rightarrow \;\; \sum_l \alpha_l \leq 2/\epsilon_0.$$

Thus on the event $\bar{S}_n$, $\widehat{\mathcal{G}}_n \subset \mathcal{G}_n^{\widetilde{M}}$ is always satisfied. We define

$$S_n := \bar{S}_n \cap \left\{ \widetilde{c}_0 < \frac{g_0}{g_n^*} < c_0 \right\}.$$

The rest of the proof goes through an analogous line to Theorem 7.6 of van de Geer (2000). Since $\widehat{\mathcal{G}}_n$ is convex, from the definition of $\widehat{g}_n$, we obtain

$$0 \leq P_n \log \frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*} \leq P_n \frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*} - 1 = (P_n - P)\frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*} - P\frac{g_n^* - \widehat{g}_n}{\widehat{g}_n + g_n^*}. \tag{11}$$

Since g_n^* maximizes $P \log g$ in the convex set $\widehat{\mathcal{G}}_n$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} P \log(\alpha g + (1-\alpha)g_n^*)\Big|_{\alpha=0} \leq 0, \quad \forall g \in \widehat{\mathcal{G}}_n.$$

This yields the inequality

$$P\frac{g - g_n^*}{g_n^*} \leq 0, \quad \forall g \in \widehat{\mathcal{G}}_n.$$

Thus the second term of the far RHS of (11) is bounded from below as

$$\begin{aligned} P\frac{g_n^* - \widehat{g}_n}{\widehat{g}_n + g_n^*} &= \frac{1}{2}P\frac{g_n^* - \widehat{g}_n}{g_n^*} + \frac{1}{2}P\frac{(g_n^* - \widehat{g}_n)^2}{g_n^*(g_n^* + \widehat{g}_n)} \geq \frac{1}{2}P\frac{(g_n^* - \widehat{g}_n)^2}{g_n^*(g_n^* + \widehat{g}_n)} \\ &= \frac{1}{2}P\frac{(\sqrt{g_n^*} - \sqrt{\widehat{g}_n})^2(\sqrt{g_n^*} + \sqrt{\widehat{g}_n})^2}{g_n^*(\widehat{g}_n + g_n^*)} \geq \frac{1}{2}Q\left((\sqrt{g_n^*} - \sqrt{\widehat{g}_n})^2 \frac{g_0}{g_n^*}\right). \end{aligned} \tag{12}$$

Combining (11) and (12), on the event $g_0/g_n^* \geq \widetilde{c}_0$, we have

$$\widetilde{c}_0 h_Q(g_n^*, \widehat{g}_n)^2 = \widetilde{c}_0 \frac{1}{2} Q(\sqrt{g_n^*} - \sqrt{\widehat{g}_n})^2 \leq (P_n - P)\frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*}. \tag{13}$$

This indicates that it suffices to bound $(P_n - P)\frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*}$. Let $\bar{g}, \widetilde{g} \in \mathcal{G}$ be $g_0/\bar{g} \leq c_0$ and $g_0/\widetilde{g} \leq c_0$, then

$$\begin{aligned} P\left(\frac{2g}{g+\bar{g}} - \frac{2g}{g+\widetilde{g}}\right)^2 &= P\frac{4g^2(\bar{g} - \widetilde{g})^2}{(g+\bar{g})^2(g+\widetilde{g})^2} = Q\frac{4g_0 g^2(\sqrt{\bar{g}} - \sqrt{\widetilde{g}})^2(\sqrt{\bar{g}} + \sqrt{\widetilde{g}})^2}{(g+\bar{g})^2(g+\widetilde{g})^2} \\ &\leq 16 c_0 Q(\sqrt{\bar{g}} - \sqrt{\widetilde{g}})^2. \end{aligned}$$

Similarly, for $g, g' \in \mathcal{G}$, on the event $g_0/g_n^* < c_0$, we have

$$P\left(\frac{2g}{g+g_n^*} - \frac{2g'}{g'+g_n^*}\right)^2 \leq 16 c_0 Q(\sqrt{g} - \sqrt{g'})^2.$$

Therefore, for $g, g' \in \mathcal{G}$ and $\bar{g}, \widetilde{g} \in \mathcal{G}$ such that $g_0/\bar{g} \leq c_0$ and $g_0/\widetilde{g} \leq c_0$, the following is satisfied:

$$\sqrt{P\left(\frac{2g}{g+\bar{g}} - \frac{2g'}{g'+\widetilde{g}}\right)^2} \leq 4\sqrt{2c_0}(h_Q(g, g') + h_Q(\bar{g}, \widetilde{g})). \tag{14}$$

Note that $P(\frac{2\widehat{g}_n}{\widehat{g}_n + g_n^*} \in \mathcal{G}^{(\mathrm{c})}) \to 1$ holds by the assumptions (7) and (10). Since $\frac{2g}{g+\bar{g}}$ is increasing with respect to g and decreasing with respect to $\bar{g}$, (14) yields that the bracketing number of $\mathcal{G}^{(\mathrm{c})}$ is bounded by

$$N_{[]}(\epsilon, \mathcal{G}^{(\mathrm{c})}, L_2(P)) \leq N_{[]}(\epsilon/8\sqrt{2c_0}, \mathcal{G}^{\widetilde{M}}, h_Q)^2. \tag{15}$$

Step3: Proof of $P(h_Q(\widehat{g}_n, g_n^*) > \delta_n) \to 0$ *where* $\delta_n = \mathcal{O}(n^{-\frac{1}{2+\gamma}})$.

Now (15) gives an upper bound of the entropy integral of $\mathcal{G}^{(\mathrm{c})}$ as follows:

$$\frac{1}{\sqrt{n}}\int_0^{\delta}\sqrt{\log(N_{[]}(\epsilon,\mathcal{G}^{(\mathrm{c})},L_2(P)))}\mathrm{d}\epsilon \leq \frac{1}{\sqrt{n}}\int_0^{\delta}\sqrt{2\log(N_{[]}(\epsilon/8\sqrt{2c_0},\mathcal{G}^{\widetilde{M}},h_Q))}\mathrm{d}\epsilon.$$

Note that on the event $g_0/g_n^* < c_0$, we have

$$\begin{aligned} P\left(\frac{2g}{g+g_n^*}-1\right)^2 &= P\left(\frac{g-g_n^*}{g+g_n^*}\right)^2 = P\left(\frac{(\sqrt{g}-\sqrt{g_n^*})^2(\sqrt{g}+\sqrt{g_n^*})^2}{(g+g_n^*)^2}\right) \\ &\leq 2c_0 Q(\sqrt{g}-\sqrt{g_n^*})^2 \leq 4c_0 h_Q(g,g_n^*)^2. \end{aligned} \tag{16}$$

Theorem 5.11 of van de Geer (2000) yields that there exist constants K', C, C', c such that for all $\delta \geq \delta_n$ where δ_n satisfies

$$\delta_n^2 \geq \frac{K'}{\sqrt{n}}\int_0^{\delta_n}\sqrt{\log(N_{[]}(\epsilon,\mathcal{G}^{(\mathrm{c})},L_2(P)))}\mathrm{d}\epsilon, \tag{17}$$

the following inequalities are satisfied:

$$\begin{aligned} &P(h_Q(\widehat{g}_n, g_n^*) > \delta) \\ &\leq P\left(\sup_{g\in\widehat{\mathcal{G}}_n : h_Q(g,g_n^*)>\delta}(P_n-P)\left(\frac{2g}{g+g_n^*}-1\right) \geq \widetilde{c}_0 h_Q(g,g_n^*)^2\right) \\ &\leq \sum_{s=0}^{\infty} P\left(\sup_{g\in\widehat{\mathcal{G}}_n : h_Q(g,g_n^*)\leq 2^{s+1}\delta}\sqrt{n}(P_n-P)\left(\frac{2g}{g+g_n^*}-1\right) \geq \widetilde{c}_0\sqrt{n}2^{2s}\delta^2 \cap S_n\right) + P(S_n^c) \\ &\leq \sum_{s=0}^{\infty} P\left(\sup_{f\in\mathcal{G}^{(\mathrm{c})} : \|f-1\|_P \leq \sqrt{c_0}2^{s+2}\delta}\sqrt{n}(P_n-P)(f-1) \geq \widetilde{c}_0\sqrt{n}2^{2s}\delta^2\right) + P(S_n^c) \\ &\leq \sum_{s=0}^{\infty} C\exp\left(-\frac{n2^{2s}\delta^2}{C'}\right) + P(S_n^c) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right) + P(S_n^c). \end{aligned} \tag{18}$$

In the above equation, S_n^c denotes the complement of S_n and the second inequality follows the application of the "peeling device" (see p.70 and Theorem 7.6 of van de Geer, 2000, for more details). Now we can take $\delta_n = \mathcal{O}(n^{-\frac{1}{2+\gamma}})$ because to ensure (17) it suffices to let δ_n satisfy

$$\delta_n^2 \geq \frac{K'}{\sqrt{n}}\int_0^{\delta_n}\sqrt{2\log(N_{[]}(\epsilon/8\sqrt{2c_0},\mathcal{G}^{\widetilde{M}},h_Q))}\mathrm{d}\epsilon = \mathcal{O}(\delta_n^{1-\gamma/2}/\sqrt{n}).$$

This yields that the far RHS of (18) converges to 0 and therefore $h_Q(\widehat{g}_n, g_n^*) = O_p(n^{-\frac{1}{2+\gamma}})$ holds. ■

Proof of Corollary 2

By the definition of g_n^*, $P\log g_n^* \geq P\log\left(\frac{g_0}{Q_n g_0}\right)$. On the other hand, since g_0 maximizes $P\log g$ in $\{g \in \mathcal{G} \mid Qg = 1\}$, we have $P\log\left(\frac{g_n^*}{Qg_n^*}\right) \leq P\log g_0$. Combining these inequalities gives

$$0 \leq P\log\left(\frac{g_0}{g_n^*/Qg_n^*}\right) = P\log\left(\frac{g_0/Q_n g_0}{g_n^*}Q_n g_0 Q g_n^*\right) \leq \log(Q_n g_0 Q g_n^*). \tag{19}$$

A similar argument to Lemma 1.3 of van de Geer (2000) yields

$$2h_Q^2(g_n^*/Qg_n^*, g_0) \leq P \log\left(\frac{g_0}{g_n^*/Qg_n^*}\right).$$

Moreover we have

$$\log(Q_n g_0 Q g_n^*) \leq Q_n g_0 Q g_n^* - 1 = (Qg_n^*)(Q_n - Q)\left(g_0 - \frac{g_n^*}{Q(g_n^*)}\right).$$

Then, by (19), we have

$$2h_Q^2(g_n^*/Qg_n^*, g_0) \leq (Qg_n^*)(Q_n - Q)\left(g_0 - \frac{g_n^*}{Q(g_n^*)}\right).$$

We can easily check that $g_n^*/Q(g_n^*)$ is contained in $\mathcal{G}^{\widetilde{M}}$. Thus

$$Q\left(g_0 - \frac{g_n^*}{Q(g_n^*)}\right)^2 = Q\left(\sqrt{g_0} - \sqrt{\frac{g_n^*}{Q(g_n^*)}}\right)^2\left(\sqrt{g_0} + \sqrt{\frac{g_n^*}{Q(g_n^*)}}\right)^2 \leq 4(\widetilde{M}+\eta_0)h_Q\left(g_0, \frac{g_n^*}{Q(g_n^*)}\right)^2.$$

Similarly, for $\widetilde{\mathcal{G}} := \{g - g_0 \mid g \in \mathcal{G}^{\widetilde{M}}\}$, we have

$$N_{[]}(\epsilon, \widetilde{\mathcal{G}}, L_2(Q)) \leq N_{[]}(\epsilon/2\sqrt{2\widetilde{M}}, \widetilde{\mathcal{G}}, h_Q).$$

Applying a similar argument to (18) with the double sum version of Theorem 5.11 of van de Geer (2000), we have

$$h_Q\left(g_0, \frac{g_n^*}{Q(g_n^*)}\right) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}}).$$

Since $Q(g_n^*) = 1 + \mathcal{O}_p(n^{-\frac{1}{2}})$ yields $h_Q\left(g_n^*, \frac{g_n^*}{Q(g_n^*)}\right) = \mathcal{O}_p(n^{-\frac{1}{2}})$, we have $h_Q(g_0, g_n^*) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}})$ by the triangle inequality. ■

References

M. A. Arcones and E. Giné. Limit theorems for U-processes. *The Annals of Probability*, 21 (3):1494–1542, 1993.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., N. Y., 1991.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986. doi: http://dx.doi.org/10.1103/PhysRevA.33.1134.

S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001.

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

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin, 2004.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

M. M. Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.

S. Khan, S. Bandyopadhyay, A. Ganguly, and S. Saigal. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76:026209, 2007.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functions and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, 2008.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4), 2008. to appear.

T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. A least-squares approach to mutual information estimation with application in variable selection. In *Proceedings of the 3rd Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery (FSDM2008)*, 2008. to appear.

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

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.