Direct Density Ratio Estimation with Dimensionality Reduction

Masashi Sugiyama[*], Satoshi Hara[†], Paul von Bünau[‡],
Taiji Suzuki[§], Takafumi Kanamori[¶], and Motoaki Kawanabe[‖]

Abstract

Methods for directly estimating the ratio of two probability density functions without going through density estimation have been actively explored recently since they can be used for various data processing tasks such as non-stationarity adaptation, outlier detection, conditional density estimation, feature selection, and independent component analysis. However, even the state-of-the-art density ratio estimation methods still perform rather poorly in high-dimensional problems. In this paper, we propose a new density ratio estimation method which incorporates dimensionality reduction into a density ratio estimation procedure. Our key idea is to identify a low-dimensional subspace in which the two densities corresponding to the denominator and the numerator in the density ratio are significantly different. Then the density ratio is estimated only within this low-dimensional subspace. Through numerical examples, we illustrate the effectiveness of the proposed method.

1 Introduction

Recently, it has been shown [28] that various data mining and machine learning tasks can be formulated in terms of the ratio of two probability density functions $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})},$$

where the subscripts 'nu' and 'de' denote 'numerator' and 'denominator', respectively. Possible usage of the density ratio includes the following tasks.

- **Importance sampling in supervised learning:** Samples in one domain (drawn from $p_{\mathrm{de}}(\boldsymbol{x})$) is utilized for learning in other domains (characterized by $p_{\mathrm{nu}}(\boldsymbol{x})$). Such data transfer is carried out by weighting the loss function according to the density ratio:

$$\begin{aligned}\mathop{\mathbb{E}}_{p_{\mathrm{nu}}(\boldsymbol{x})}[\mathrm{loss}(\boldsymbol{x})] &= \int \mathrm{loss}(\boldsymbol{x}) p_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int \mathrm{loss}(\boldsymbol{x}) r(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \mathop{\mathbb{E}}_{p_{\mathrm{de}}(\boldsymbol{x})}[\mathrm{loss}(\boldsymbol{x}) r(\boldsymbol{x})].\end{aligned}$$

  Thus, the inter-domain bias can be canceled by density-ratio weighted learning.

  Applications of importance sampling include non-stationarity adaptation [23, 43, 31, 30, 27, 29], transfer learning [26, 41], and multi-task learning [2].

- **Outlier detection:** Let us consider an outlier detection problem of finding outliers in an evaluation dataset based on another "model" dataset that only contains inliers [11, 10, 40, 24]. Defining the density ratio over the two sets of samples, one can see that the density-ratio values for inliers are close to one, while those for outliers tend to be significantly deviated from one. Thus the density-ratio value could be used as an index of the degree of outlyingness.

  The same technique can also be applied to change-point detection in time-series [17].

- **Conditional probability estimation:** Suppose we are given n i.i.d. paired samples $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{n}$ drawn from a joint distribution with density $q(\boldsymbol{x}, \boldsymbol{y})$. The goal is to estimate the conditional probability $q(\boldsymbol{y}|\boldsymbol{x})$. When the domain of $\boldsymbol{x}$ is continuous, conditional density estimation is not straightforward since a naive empirical approximation cannot be used [4]. Let us regard $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{n}$ as samples corresponding to the numerator of the density ratio and $\{\boldsymbol{x}_k\}_{k=1}^{n}$ as samples corresponding to the denominator of the density ratio, i.e., we consider the density ratio defined as

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{q(\boldsymbol{x}, \boldsymbol{y})}{q(\boldsymbol{x})} = q(\boldsymbol{y}|\boldsymbol{x}),$$

[*] Tokyo Institute of Technology and PRESTO, Japan Science and Technology Agency (JST)

[†] Tokyo Institute of Technology

[‡] Technical University of Berlin

[§] The University of Tokyo

[¶] Nagoya University

[‖] Fraunhofer FIRST.IDA

where $q(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$. Then a density-ratio estimation method directly gives an estimator of the conditional density.

The problem is conditional density estimation when $\boldsymbol{y}$ is continuous [34], while it is probabilistic classification when $\boldsymbol{y}$ is categorical.

- **Estimation of divergence functionals/mutual information:** Suppose we are given n i.i.d. paired samples $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{n}$ drawn from a joint distribution with density $q(\boldsymbol{x}, \boldsymbol{y})$. Let us denote the marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$ by $q(\boldsymbol{x})$ and $q(\boldsymbol{y})$, respectively. Then mutual information $I(X, Y)$ between random variables X and Y is defined by

$$I(X, Y) = \iint q(\boldsymbol{x}, \boldsymbol{y}) \log \frac{q(\boldsymbol{x}, \boldsymbol{y})}{q(\boldsymbol{x})q(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

which can be used for measuring independence between X and Y. Let us regard $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{n}$ as samples corresponding to the numerator of the density ratio and $\{(\boldsymbol{x}_k, \boldsymbol{y}_{k'})\}_{k,k'=1}^{n}$ as samples corresponding to the denominator of the density ratio. Then mutual information can be directly estimated using a density-ratio estimation method [20, 19, 32, 33, 38, 13, 39, 14].

Such an independence measure can be used for various purposes such as variable selection (input-output dependency) [38, 37], supervised dimensionality reduction (input-output dependency) [36], and independent component analysis (input-input dependency) [35].

Because of the wide applicability of density ratios, the problem of estimating the density ratios is attracting a great deal of attention recently and various methods have been explored [22, 6, 12, 3, 19, 32, 33, 13, 14]. A naive approach is to estimate the two densities in the ratio (corresponding to the denominator and the numerator) separately using a flexible technique such as non-parametric *kernel density estimation* [8] and then take the ratio of the estimated densities. However, this naive two-step approach is not reliable in practical situations since kernel density estimation performs poorly in high-dimensional problems; furthermore, division by an estimated density tends to magnify the estimation error.

To improve the estimation accuracy, various methods have been developed for directly estimating the density ratio without going through density estimation. The moment matching method based on reproducing kernels [1, 25] called *kernel mean matching* [12] uses the kernel trick to efficiently match the mean of two sets of samples in a reproducing kernel Hilbert space. However, model selection methods are not available for kernel mean matching. Thus, several tuning parameters such as the kernel width and the regularization parameter need to be hand-tuned using some heuristics, which is highly unreliable in practice. Furthermore, the computation of kernel mean matching is rather expensive since a quadratic programming problem has to be solved.

An alternative method based on *logistic regression* [22, 6, 3] formulates the density ratio estimation problem as the problem of separating samples from the two sets by logistic regression. An advantage of the logistic regression formulation is that standard cross-validation (CV) is available for model selection since the problem one needs to solve is a standard supervised classification problem. Thus, all the tuning parameters can be objectively determined by CV. However, it is still computationally rather demanding due to non-linearity of the optimization problem.

Maximum likelihood estimation of density ratio functions is another line of methods that allows us to avoid density estimation [20, 19, 32, 33]. An advantage of the maximum likelihood approach is that it is also equipped with CV and thus model selection is possible [32, 33]. However, this approach is also computationally rather expensive due to non-linearity of the optimization problem.

Recently, a least-squares method of density ratio estimation was proposed [13, 14]. This is also equipped with a build-in CV method, and hence all tuning parameters can be objectively determined. Furthermore, its solution can be computed *analytically* just by solving a system of linear equations. Thus it is highly advantageous in terms of computation time. The least-squares method was also shown to be numerically stable [15] under condition number analysis. Thus the least-squares method is a reliable density ratio estimator.

As described above, various methods have been developed for directly estimating the density ratios. The success of these direct density-ratio estimation methods could be intuitively understood through *Vapnik's principle* [42]:

> "When solving a problem of interest, do not solve a more general problem as an intermediate step".

The *support vector machine* would be a successful example of this principle—instead of estimating the data generation model, it directly models the decision boundary which is simpler and sufficient for pattern recognition. In the current context, estimating the two densities is more general than estimating the density ratio since knowing the two densities implies knowing

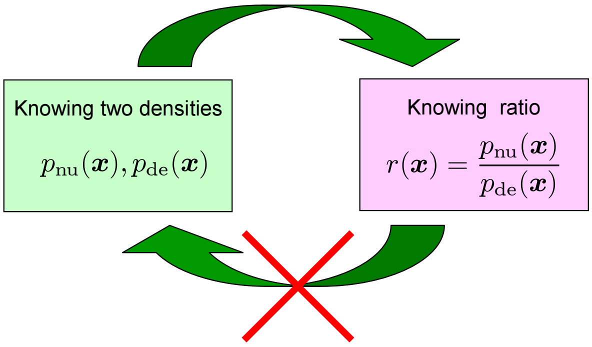

Knowing two densities
$p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})$
Knowing ratio
$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$

Figure 1: Density ratio estimation is substantially easier than density estimation. The density ratio $r(\boldsymbol{x})$ can be computed if two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are known. However, even if the density ratio is known, the two densities cannot be computed in general.

the density ratio, but not vice versa. Thus directly estimating the density ratio would be more promising than density ratio estimation via density estimation (Figure 1). Rigorous theoretical analysis along this line was carried out in the paper [16].

Although the above density ratio estimators were shown to compare favorably with naive kernel density estimation through extensive experiments, density ratio estimation in high-dimensional problems is still challenging. In this paper, we propose to incorporate dimensionality reduction in a density ratio estimation procedure. More specifically, our idea is to identify a subspace in which the two densities are significantly different (called the *hetero-distributional subspace*); Then we perform density ratio estimation only in this subspace. We derive an analytic estimator of a divergence between the two densities and this estimator is used for searching the hetero-distributional subspace. Through numerical examples, we illustrate the usefulness of the proposed method.

2 Problem Formulation

In this section, we formulate the problem of density ratio estimation with dimensionality reduction.

2.1 Density Ratio Estimation.

Let $\mathcal{D}$ $(\subset \mathbb{R}^d)$ be the data domain and suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ from a distribution with density $p_{\mathrm{de}}(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ from another distribution with density $p_{\mathrm{nu}}(\boldsymbol{x})$. We assume that the density $p_{\mathrm{de}}(\boldsymbol{x})$ is strictly positive, i.e.,

$$p_{\mathrm{de}}(\boldsymbol{x}) > 0 \text{ for all } \boldsymbol{x} \in \mathcal{D}.$$

The problem we address in this paper is to estimate the density ratio

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

from samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$.

Our basic idea is to first identify a lower-dimensional hetero-distributional subspace in which the two densities corresponding to the denominator and the numerator are significantly different, and then perform density ratio estimation only in this subspace.

2.2 Hetero-distributional Subspace.

Let $\boldsymbol{u}$ be an m-dimensional vector $(1 \leq m \leq d)$ and $\boldsymbol{v}$ is a $(d-m)$-dimensional vector defined as

$$\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{bmatrix} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\boldsymbol{x}.$$

$\boldsymbol{U}$ is an $m \times d$ matrix and $\boldsymbol{V}$ is a $(d-m)\times d$ matrix; furthermore, without loss of generality, it is possible to assume that the row vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ form an orthonormal basis, i.e., $\boldsymbol{U}$ and $\boldsymbol{V}$ correspond to "projection" matrices that are orthogonally complementary to each other (see Figure 2). Using the decomposition of $\boldsymbol{x}$ into $\boldsymbol{u}$ and $\boldsymbol{v}$, we can express the two densities $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$ as

$$\begin{aligned}p_{\mathrm{de}}(\boldsymbol{x}) &= p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{de}}(\boldsymbol{u}),\\ p_{\mathrm{nu}}(\boldsymbol{x}) &= p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u}).\end{aligned}$$

Our key theoretical assumption which forms the basis of our proposed algorithm is that the conditional densities $p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})$ and $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u})$ agree with each other, i.e., the two densities $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$ are decomposed as

$$\begin{aligned}p_{\mathrm{de}}(\boldsymbol{x}) &= p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{de}}(\boldsymbol{u}),\\ p_{\mathrm{nu}}(\boldsymbol{x}) &= p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u}),\end{aligned}$$

where $p(\boldsymbol{v}|\boldsymbol{u})$ is the common conditional density. This assumption implies that the marginal densities of $\boldsymbol{u}$ are different, but the conditional density of $\boldsymbol{v}$ given $\boldsymbol{u}$ is common. Then the density ratio is simplified as

$$r(\boldsymbol{x}) = r(\boldsymbol{u}) = \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}.$$

Thus, the density ratio does not have to be estimated in the entire d-dimensional space, but only in the m-dimensional subspace.

Let us consider the set of all subspaces such that the conditional density $p(\boldsymbol{v}|\boldsymbol{u})$ is common to $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$. We refer to the *intersection* of such subspaces

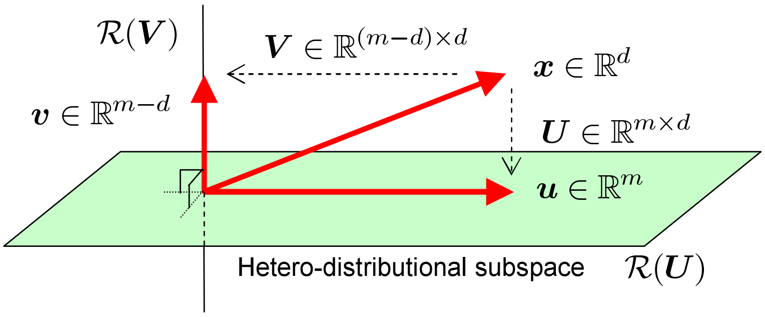

$\mathcal{R}(\boldsymbol{V})$
$\boldsymbol{V} \in \mathbb{R}^{(m-d)\times d}$
$\boldsymbol{x} \in \mathbb{R}^{d}$
$\boldsymbol{v} \in \mathbb{R}^{m-d}$
$\boldsymbol{U} \in \mathbb{R}^{m\times d}$
$\boldsymbol{u} \in \mathbb{R}^{m}$
Hetero-distributional subspace
$\mathcal{R}(\boldsymbol{U})$

Figure 2: Hetero-distributional subspace.

as the *hetero-distributional subspace*. Thus the hetero-distributional subspace is the 'smallest' subspace outside which the conditional density $p(\boldsymbol{v}|\boldsymbol{u})$ is common to $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$.

For the moment, we assume that the true dimensionality m of the hetero-distributional subspace is known. Later, we explain how m can be estimated from data in practice.

3 Direct Density Ratio Estimation with Dimensionality Reduction

In this section, we propose a new density ratio estimator which involves dimensionality reduction.

3.1 Characterizing the Hetero-distributional Subspace by the Pearson Divergence.

We use the Pearson divergence (PD) as our criterion for evaluating the discrepancy between two distributions[1]. PD from $p_{\mathrm{nu}}(\boldsymbol{u})$ to $p_{\mathrm{de}}(\boldsymbol{u})$ is defined and expressed as

$$\begin{aligned}\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] &= \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1\right)^2 p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\\ &= \int \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - 1.\end{aligned}$$

$\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ vanishes if and only if $p_{\mathrm{nu}}(\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{u})$ for all $\boldsymbol{u}$.

The following lemma characterizes the hetero-distributional subspace in terms of PD.

Lemma 1.

$$\begin{aligned}&\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] - \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]\\ &= \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (3.1)\\ &\geq 0.\end{aligned}$$

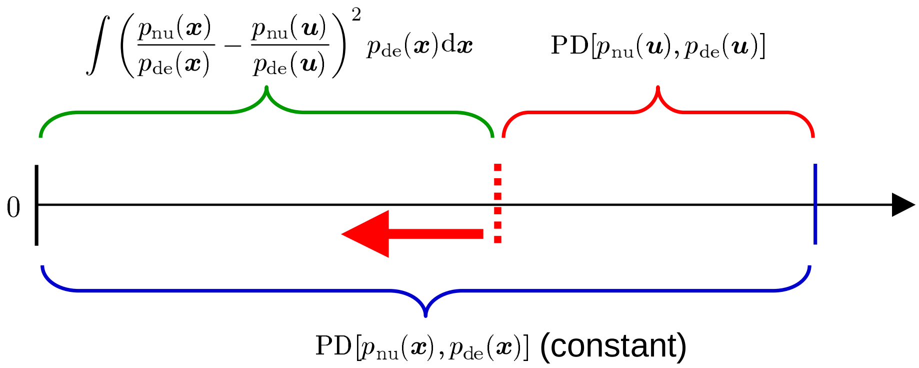
[Proof:]

$$\begin{aligned}0 &\leq \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - 1 - \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} + 1\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - 1\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\quad + \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\quad - 2\int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - 1\right)\left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1\right) p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] + \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]\\ &\quad - 2\int \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} p_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + 2\int p_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\quad + 2\int p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - 2\int p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] + \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]\\ &\quad - 2\int \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + 2\\ &= \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] - \mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})].\end{aligned}$$

■

Eq.(3.1) is non-negative and it vanishes if and only if $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})$ for all $\boldsymbol{u}, \boldsymbol{v}$. Since $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})]$ is a constant and does not depend on $\boldsymbol{U}$, maximizing $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ with respect to $\boldsymbol{U}$ leads to $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})$ for all $\boldsymbol{u}, \boldsymbol{v}$ (see Figure 3). That is, the hetero-distributional subspace can be characterized as the maximizer of $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$.

[1] It is also possible to characterize the hetero-distributional subspace by the Kullback-Leibler divergence [18]. However, as shown later, PD allows us to obtain an analytic-form estimator of the divergence which is useful in hetero-distributional subspace search.



$\int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$
$\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$
0
$\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})]$ (constant)

Figure 3: Since $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})]$ is a constant, minimizing $\int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} - \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is equivalent to maximizing $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$.

3.2 Estimation of PD.

It is not possible to directly find the maximizer of $\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ since $p_{\mathrm{de}}(\boldsymbol{u})$ and $p_{\mathrm{nu}}(\boldsymbol{u})$ are unknown. According to the *Legendre-Fenchel duality* [5], we have

$$\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] = \max_{g} J(g), \tag{3.2}$$

where

$$J(g) := -\int g(\boldsymbol{u})^2 p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + 2\int g(\boldsymbol{u}) p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - 1. \tag{3.3}$$

Let us employ a parametric model

$$g(\boldsymbol{u}) = \sum_{\ell=1}^{b} \alpha_\ell \psi_\ell(\boldsymbol{u}),$$

where $\{\psi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$ are basis functions such that $\psi_\ell(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ $(\in \mathcal{D})$ and for $\ell = 1, \ldots, b$. In our experiments, we use the Gaussian kernel model:

$$g(\boldsymbol{u}) = \sum_{\ell=1}^{n_{\mathrm{nu}}} \alpha_\ell K_\sigma(\boldsymbol{u}, \boldsymbol{u}_\ell^{\mathrm{nu}}), \tag{3.4}$$

where

$$K_\sigma(\boldsymbol{u}, \boldsymbol{u}') = \exp\left(-\frac{\|\boldsymbol{u} - \boldsymbol{u}'\|^2}{2\sigma^2}\right).$$

Let us maximize an empirical and regularized variant of $J(g)$ (see Eq.(3.3)) over the parametric model.

$$\max_{\{\alpha_\ell\}_{\ell=1}^{b}} -\sum_{\ell,\ell'=1}^{b} \alpha_\ell \alpha_{\ell'} \widehat{H}_{\ell,\ell'} + 2\sum_{\ell=1}^{b} \alpha_\ell \widehat{h}_\ell - \lambda \sum_{\ell=1}^{b} \alpha_\ell^2,$$

where λ (≥ 0) is a regularization parameter and

$$\widehat{H}_{\ell,\ell'} = \frac{1}{n_{\mathrm{de}}} \sum_{i=1}^{n_{\mathrm{de}}} \psi_\ell(\boldsymbol{u}_i^{\mathrm{de}}) \psi_{\ell'}(\boldsymbol{u}_i^{\mathrm{de}}),$$

$$\widehat{h}_\ell = \frac{1}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{nu}}} \psi_\ell(\boldsymbol{u}_j^{\mathrm{nu}}).$$

By setting the derivative of the above objective function to zero and solving it, we can obtain the maximizer analytically as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_b)^\top = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}},$$

where $^\top$ denotes the transpose of a matrix or a vector and $\boldsymbol{I}_b$ is the b-dimensional identity matrix.

Then an *analytic* estimator of the Pearson divergence $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ is given as

$$\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] = \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \widehat{h}_\ell - 1.$$

We note that the tuning parameters in the above procedure (i.e., the Gaussian width σ and the regularization parameter λ) can be determined by cross-validation (CV) over the score function $J(g)$ (see Eq.(3.3)). Using the *Sherman-Woodbury-Morrison* formula [7], we can actually compute the leave-one-out CV score analytically, which is computationally very efficient. However, we omit the details.

3.3 Hetero-distributional Subspace Search.

Given the Pearson divergence estimator $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$, our next task is to find a maximizer of $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ with respect to $\boldsymbol{U}$ and identify the hetero-distributional subspace (cf. Lemma 1).

A gradient descent approach would be a standard

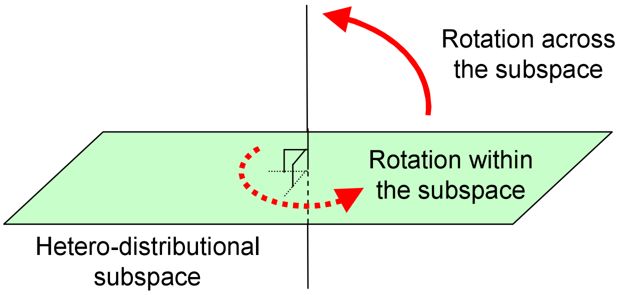


Figure 4: In the hetero-distributional subspace search, rotation which changes the subspace only matters (the solid arrow); rotation within the subspace (dotted arrow) can be ignored since this does not change the subspace. Similarly, rotation within the orthogonal complement of the hetero-distributional subspace can also be ignored (not depicted in the figure).

choice for optimization.

$$\boldsymbol{U} \longleftarrow \boldsymbol{U} + t\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{U}},$$

where t is the step size and

$$\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{U}} = -\sum_{\ell,\ell'=1}^{b}\widehat{\alpha}_{\ell}\widehat{\alpha}_{\ell'}\frac{\partial\widehat{H}_{\ell,\ell'}}{\partial\boldsymbol{U}} + 2\sum_{\ell=1}^{b}\widehat{\alpha}_{\ell}\frac{\partial\widehat{h}_{\ell}}{\partial\boldsymbol{U}},$$

$$\frac{\partial\widehat{H}_{\ell,\ell'}}{\partial\boldsymbol{U}} = \frac{1}{n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}}\left(\frac{\partial\psi_{\ell}(\boldsymbol{u}_i^{\mathrm{de}})}{\partial\boldsymbol{U}}\psi_{\ell'}(\boldsymbol{u}_i^{\mathrm{de}}) + \psi_{\ell}(\boldsymbol{u}_i^{\mathrm{de}})\frac{\partial\psi_{\ell'}(\boldsymbol{u}_i^{\mathrm{de}})}{\partial\boldsymbol{U}}\right),$$

$$\frac{\partial\widehat{h}_{\ell}}{\partial\boldsymbol{U}} = \frac{1}{n_{\mathrm{nu}}}\sum_{j=1}^{n_{\mathrm{nu}}}\frac{\partial\psi_{\ell}(\boldsymbol{u}_j^{\mathrm{nu}})}{\partial\boldsymbol{U}}.$$

For the Gaussian kernel model (3.4) which we use in the experiments, $\frac{\partial\psi_{\ell}(\boldsymbol{u})}{\partial\boldsymbol{U}}$ is given by

$$\frac{\partial\psi_{\ell}(\boldsymbol{u})}{\partial\boldsymbol{U}} = -\frac{1}{\sigma^2}(\boldsymbol{u}-\boldsymbol{c}_{\ell})(\boldsymbol{x}-\boldsymbol{x}_{\tau(\ell)}^{\mathrm{nu}})^{\top}\psi_{\ell}(\boldsymbol{u}).$$

By the gradient ascent iteration over the $m \times d$ matrix $\boldsymbol{U}$, we may find a local maximizer of $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$. On the other hand, the number of parameters to be optimized in the gradient algorithm can be actually reduced in the current setup since we are searching for a subspace—rotation *within* the subspace can be ignored (Figure 4). This idea is explained below in detail.

The matrix $\boldsymbol{U}$ can be expressed as

$$\boldsymbol{U} = \left[\boldsymbol{I}_m\ \boldsymbol{O}_{m\times(d-m)}\right]\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix},$$

where $\boldsymbol{O}_{d\times d'}$ is the $d \times d'$ matrix with all zeros. For a skew-symmetric matrix $\boldsymbol{M}$ ($\in \mathbb{R}^{d\times d}$), i.e., $\boldsymbol{M}^{\top} = -\boldsymbol{M}$, rotation of $\boldsymbol{U}$ can be expressed as

$$(3.5)\qquad \left[\boldsymbol{I}_m\ \boldsymbol{O}_{m\times(d-m)}\right]e^{\boldsymbol{M}}\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix},$$

where $e^{\boldsymbol{M}}$ is the matrix exponential of $\boldsymbol{M}$; $\boldsymbol{M} = \boldsymbol{O}_{d\times d}$ corresponds to no rotation. Our idea is not to update $\boldsymbol{U}$ directly, but through $\boldsymbol{M}$. The derivative of $\widehat{\mathrm{PD}}$ with respect to $\boldsymbol{M}$ at $\boldsymbol{M} = \boldsymbol{O}_{d\times d}$ is given by

$$\begin{aligned}\left.\frac{\mathrm{d}\widehat{\mathrm{PD}}}{\mathrm{d}\boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{O}_{d\times d}} &= \begin{bmatrix}\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{U}}\\ \frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{V}}\end{bmatrix}\left[\boldsymbol{U}^{\top}\ \boldsymbol{V}^{\top}\right] \\ &\quad - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\left[\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{U}}\right)^{\top}\ \left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{V}}\right)^{\top}\right] \\ &= \begin{bmatrix}\boldsymbol{O}_{m\times m} & \frac{\mathrm{d}\widehat{\mathrm{PD}}}{\mathrm{d}\boldsymbol{U}}\boldsymbol{V}\\ -(\frac{\mathrm{d}\widehat{\mathrm{PD}}}{\mathrm{d}\boldsymbol{U}}\boldsymbol{V})^{\top} & \boldsymbol{O}_{(d-m)\times(d-m)}\end{bmatrix},\end{aligned}$$

where we used the fact that $\frac{\mathrm{d}\widehat{\mathrm{PD}}}{\mathrm{d}\boldsymbol{V}} = \boldsymbol{O}_{(d-m)\times d}$. Then the gradient ascent update rule of $\boldsymbol{M}$ is given by

$$\boldsymbol{M} \longleftarrow t\left.\frac{\mathrm{d}\widehat{\mathrm{PD}}}{\mathrm{d}\boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{O}_{d\times d}},$$

where t is a step size. Then $\boldsymbol{U}$ (and also $\boldsymbol{V}$) are updated by Eq.(3.5). See the paper [21] for the details of geometric structures.

3.4 Estimating the Density Ratio in Hetero-distributional Subspace.

Finally, we estimate the density ratio in the hetero-distributional subspace. A notable fact of our algorithm is that the density ratio estimator in the hetero-distributional subspace has already been obtained during the hetero-distributional subspace search; thus we do not need additional computation. More specifically, the solution of the variational problem $\max_g J(g)$ (see Eq.(3.2)) is given by $\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}$ [19]. Thus, our final solution is simply given by

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b}\widehat{\alpha}_{\ell}\psi_{\ell}(\widehat{\boldsymbol{U}}\boldsymbol{x}),$$

where $\widehat{\boldsymbol{U}}$ is a projection matrix obtained by the hetero-distributional subspace search algorithm.

The above result implies that if the dimensionality is not reduced (i.e., $m = d$), the proposed method agrees with the density ratio estimator proposed in the papers [13, 14]. Thus, the proposed method could be

```
Input:  Two sets of samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ on $\mathbb{R}^d$
Output: Density ratio estimator $\widehat{r}(\boldsymbol{x})$

For each reduced dimensionality $m = 1, \ldots, d$
        Initialize embedding matrix $\boldsymbol{U}_m$ $(\in \mathbb{R}^{m \times d})$;
        Repeat until $\boldsymbol{U}_m$ converges
                Choose Gaussian width $\sigma$ and regularization parameter $\lambda$ by CV;
                Update $\boldsymbol{U}$ by the gradient method (see Section 3.3);
        end
        Obtain embedding matrix $\widehat{\boldsymbol{U}}_m$ and corresponding density ratio estimator $\widehat{r}_m(\boldsymbol{x})$;
        Compute its CV value as a function of $m$;
end
Choose the best reduced dimension $\widehat{m}$ based on the CV score;
Set $\widehat{r}(\boldsymbol{x}) = \widehat{r}_{\widehat{m}}(\boldsymbol{x})$;
```

Figure 5: Pseudo code of the proposed density-ratio estimation algorithm.

regarded as a natural extension of the existing density ratio estimator.

The dimensionality of the hetero-distributional subspace may be chosen by the CV score used for optimizing the Gaussian width σ and the regularization parameter λ. The entire procedure is summarized in Figure 5.

4 Numerical Examples

In this section, we illustrate the behavior of the proposed method through numerical examples.

4.1 Illustrative Examples.

Let us consider two-dimensional examples (i.e., $d = 2$) and suppose that the two distributions $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are different only in the one-dimensional subspace (i.e., $m = 1$) spanned by $(1, 0)^\top$:

$$\begin{aligned}\boldsymbol{x} &= (x^{(1)}, x^{(2)})^\top = (u, v)^\top,\\ p_{\mathrm{nu}}(\boldsymbol{x}) &= p(v|u)p_{\mathrm{nu}}(u),\\ p_{\mathrm{de}}(\boldsymbol{x}) &= p(v|u)p_{\mathrm{de}}(u).\end{aligned}$$

Let $n_{\mathrm{nu}} = n_{\mathrm{de}} = 1000$. We use the following two datasets.

Dataset (a) (Figure 6(a) and Figure 6(b)):

$$\begin{aligned}p(v|u) &= p(v) = N(v; 0, 1^2),\\ p_{\mathrm{nu}}(u) &= 0.5N(u; -2, 1.5^2) + 0.5N(u; 2, 1.5^2),\\ p_{\mathrm{de}}(u) &= N(u; 0, 1^2),\end{aligned}$$

where $N(u; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2 with respect to u.

Dataset (b) (Figure 7(a) and Figure 7(b)):

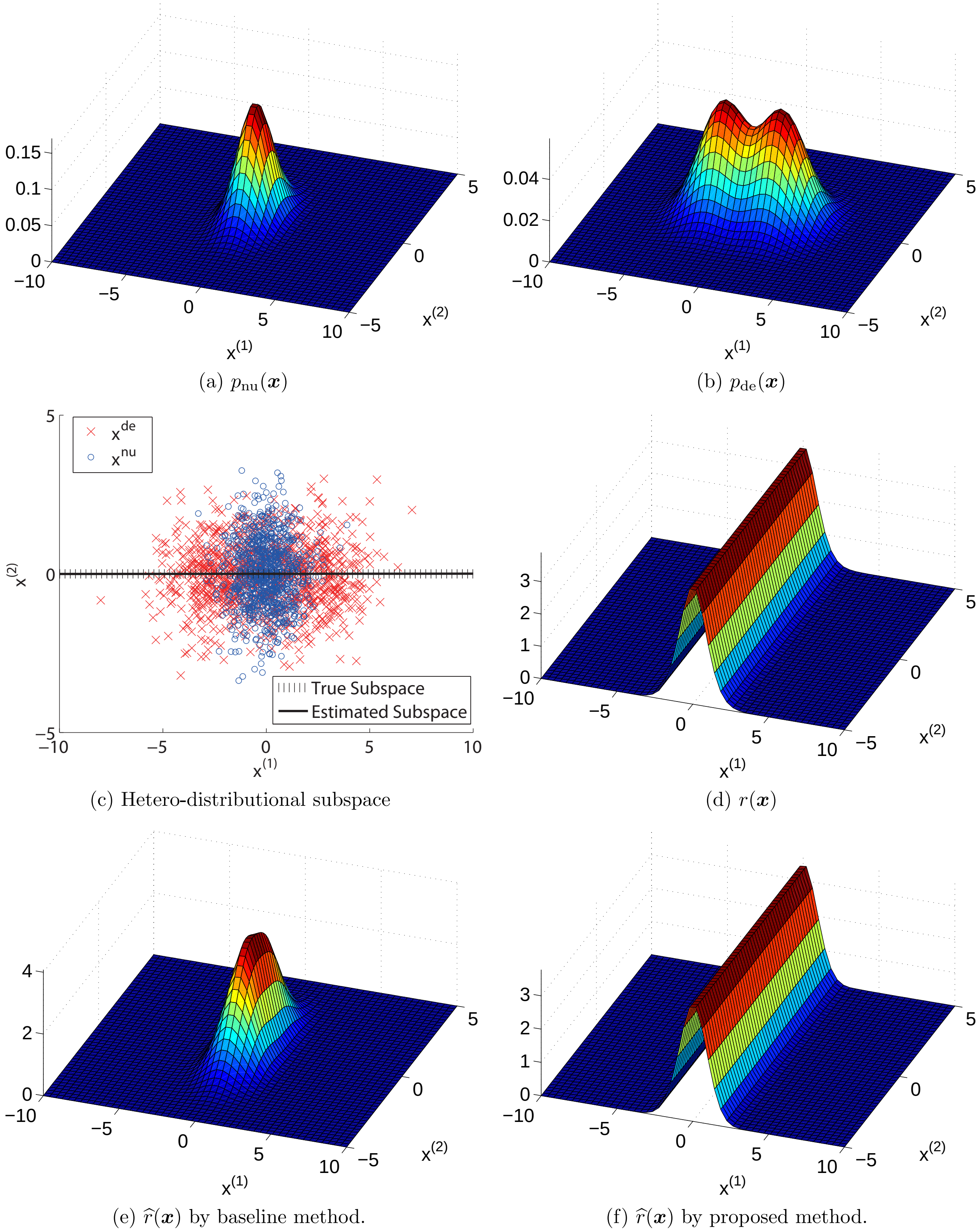
$$\begin{aligned}p(v|u) &= N(v; u, 1^2),\\ p_{\mathrm{nu}}(u) &= N(u; 0, 1^2),\\ p_{\mathrm{de}}(u) &= 0.5N(u; -2, 1.5^2) + 0.5N(u; 2, 1.5^2).\end{aligned}$$

The true and estimated hetero-distributional subspaces are depicted by the dashed and solid lines in Figure 6(c) and Figure 7(c). These plots show that the proposed method gives good estimates of the true hetero-distributional subspace. In Figure 6(e) and Figure 7(e), density-ratio functions estimated without dimensionality reduction by the baseline method proposed in the papers [13, 14] are depicted. In Figure 6(f) and Figure 7(f), density-ratio functions estimated with dimensionality reduction by the proposed method are depicted. Compared with the true density ratio functions depicted in Figure 6(d) and Figure 7(d), we can observe that the proposed method captures the redundant structure of the true density ratio functions appropriately. Consequently, the propose method gives much better estimates of the density ratio functions than the baseline method. This illustrates the usefulness of dimensionality reduction in density ratio estimation.

4.2 Performance Evaluation using Artificial Datasets.

Next, we systematically investigate the behavior of the proposed method for high-dimensional data.

For the two datasets used in the previous experiments, we increase the entire dimensionality as $d = 2, 3, \ldots, 10$ by adding dimensions consisting of standard normal noise. The dimensionality of the hetero-distributional subspace is estimated based on CV (see Section 3.4).



(a) $p_{\mathrm{nu}}(\boldsymbol{x})$

(b) $p_{\mathrm{de}}(\boldsymbol{x})$

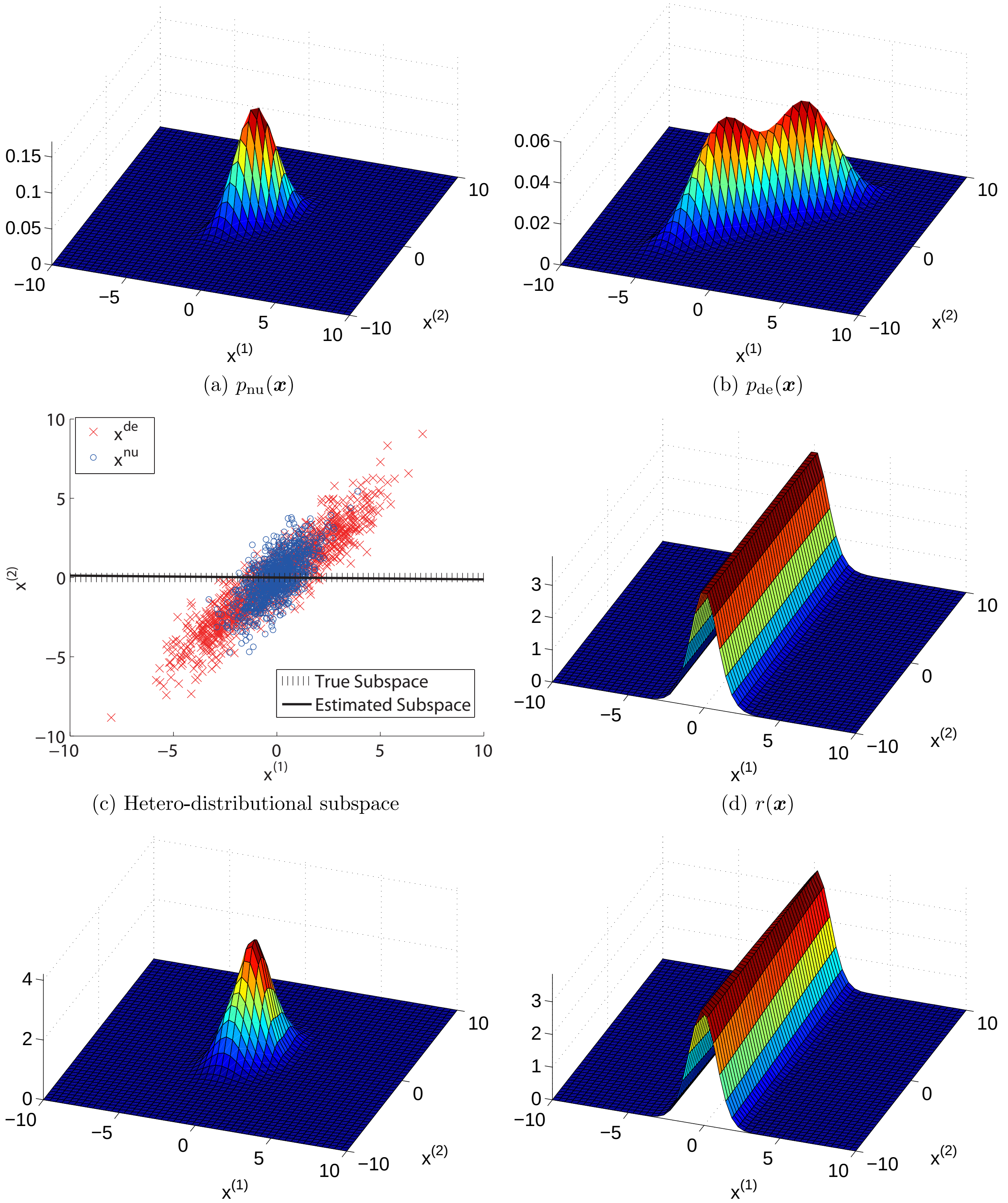
(c) Hetero-distributional subspace

(d) $r(\boldsymbol{x})$

(e) $\widehat{r}(\boldsymbol{x})$ by baseline method.

(f) $\widehat{r}(\boldsymbol{x})$ by proposed method.

Figure 6: Numerical results for dataset (a).



(a) $p_{\mathrm{nu}}(\boldsymbol{x})$

(b) $p_{\mathrm{de}}(\boldsymbol{x})$

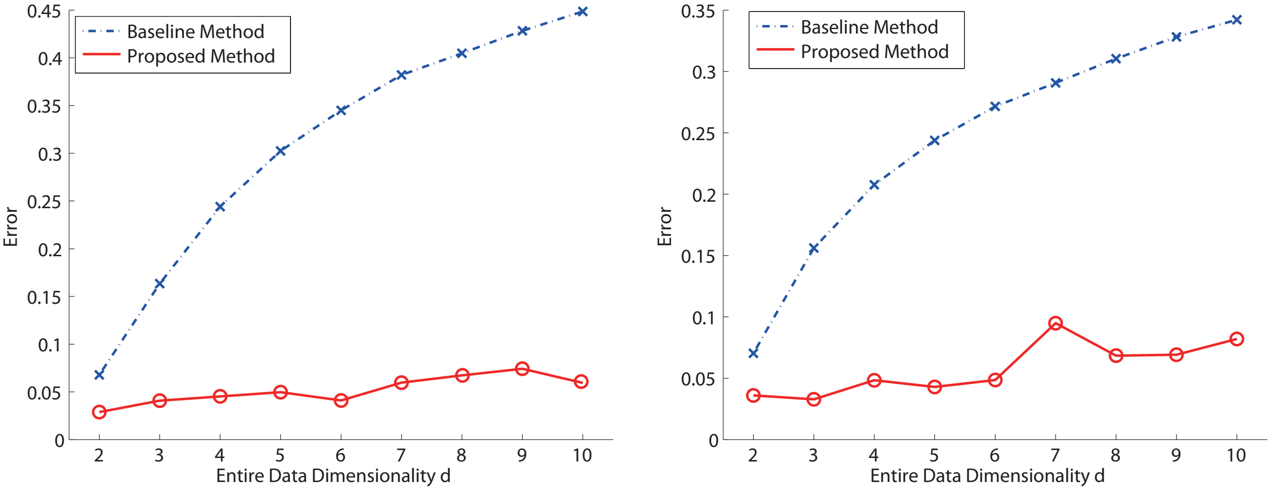
(c) Hetero-distributional subspace

(d) $r(\boldsymbol{x})$

(e) $\widehat{r}(\boldsymbol{x})$ by baseline method.

(f) $\widehat{r}(\boldsymbol{x})$ by proposed method.

Figure 7: Numerical results for dataset (b).



(a) (b)

Figure 8: Density ratio estimation error (4.6) averaged over 10 runs as a function of the entire data dimensionality d for the artificial datasets. The best method in terms of the mean error and comparable methods according to the *t-test* at the significance level 1% are specified by '$\circ$'; otherwise methods are specified by '$\times$'.

We evaluate the error of a density ratio estimator $\widehat{r}(\boldsymbol{x})$ by

$$\text{(4.6)} \qquad \text{Error} := \frac{1}{2}\int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Figure 8 shows the density ratio estimation error averaged over 10 runs as functions of the entire input dimensionality d. The best method in terms of the mean error and comparable methods according to the *t-test* [9] at the significance level 1% are specified by '$\circ$'; otherwise methods are specified by '$\times$'. This shows that, while the error of the baseline method without dimensionality reduction increases rapidly as the entire dimensionality d increases, that of the proposed method is kept moderate. Consequently, the proposed method consistently outperforms the baseline method.

5 Conclusions

The density ratio is becoming a quantity of interest in the machine learning and data mining communities since it can be used for solving various data processing tasks. In this paper, we tackled a challenging problem of estimating density ratios in high-dimensional spaces and gave a new procedure. Our key idea was to estimate the ratio only in a subspace in which two distributions (corresponding to the denominator and numerator of the density ratio) are significantly different. The proposed method was shown to be promising in experiments.

Our future work includes the application of the proposed method to various data processing tasks such as non-stationarity adaptation, outlier detection, feature selection, and independent component analysis. Improving the computational efficiency of hetero-distributional subspace search is another important issue to be further investigated.

Acknowledgments

This work has been supported by SCAT, AOARD, and the JST PRESTO program.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for HIV therapy screening. In A. McCallum and S. Roweis, editors, *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)*, pages 56–63, Helsinki, Finland, Jul. 5–9 2008. Omnipress.

[3] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 81–88, 2007.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

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

[6] K. F. Cheng and C. K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

[7] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, 1996.

[8] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models.* Springer, Berlin, 2004.

[9] R. E. Henkel. *Tests of Significance.* SAGE Publication, Beverly Hills, 1979.

[10] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems.* to appear.

[11] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Inlier-based outlier detection via direct density ratio estimation. In F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, and X. Wu, editors, *Proceedings of IEEE International Conference on Data Mining (ICDM2008)*, pages 223–232, Pisa, Italy, Dec. 15–19 2008.

[12] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 601–608. MIT Press, Cambridge, MA, 2007.

[13] T. Kanamori, S. Hido, and M Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In D. Koller, D. Schuurmans, Y. Bengio, and L. Botton, editors, *Advances in Neural Information Processing Systems 21*, pages 809–816, Cambridge, MA, 2009. MIT Press.

[14] T. Kanamori, S. Hido, and M Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, Jul. 2009.

[15] T. Kanamori, T. Suzuki, and M. Sugiyama. Condition number analysis of kernel-based density ratio estimation. Technical Report TR09-0006, Department of Computer Science, Tokyo Institute of Technology, Feb. 2009.

[16] T. Kanamori, T. Suzuki, and M. Sugiyama. Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2010. to appear.

[17] Y. Kawahara and M. Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In H. Park, S. Parthasarathy, H. Liu, and Z. Obradovic, editors, *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)*, pages 389–400, Sparks, Nevada, USA, Apr. 30–May 2 2009.

[18] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[19] X. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1089–1096. MIT Press, Cambridge, MA, 2008.

[20] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *Proceedings of IEEE International Symposium on Information Theory*, pages 2016–2020, Nice, France, 2007.

[21] M. D. Plumbley. Geometrical methods for non-negative ICA: Manifolds, Lie groups and toral subalgebras. *Neurocomputing*, 67(Aug.):161–197, 2005.

[22] J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–639, 1998.

[23] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[24] A. Smola, L. Song, and C. H. Teo. Relative novelty detection. In D. van Dyk and M. Welling, editors, *Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *JMLR Workshop and Conference Proceedings*, pages 536–543, 2009.

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

[26] A. Storkey and M. Sugiyama. Mixture regression for covariate shift. In B. Schölkopf, J. C. Platt, and T. Hoffmann, editors, *Advances in Neural Information Processing Systems 19*, pages 1337–1344, Cambridge, MA, 2007. MIT Press.

[27] M. Sugiyama, B. Blankertz, M. Krauledat, G. Dornhege, and K.-R. Müller. Importance-weighted cross-validation for covariate shift. In K. Franke, K.-R. Müller, B. Nickolay, and R. Schäfer, editors, *Pattern Recognition*, volume 4174 of *Lecture Notes in Computer Science*, pages 354–363, Berlin, 2006. Springer.

[28] M. Sugiyama, T. Kanamori, T. Suzuki, S. Hido, J. Sese, I. Takeuchi, and L. Wang. A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, 1:183–208, 2009.

[29] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

[30] M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

[31] M. Sugiyama and K.-R. Müller. Model selection under covariate shift. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, editors, *Artificial Neural Networks: Formal Models and Their Applications*, volume 3697 of *Lecture Notes in Computer Science*, pages 235–240, Berlin, 2005. Springer.

[32] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information*

Processing Systems 20, pages 1433–1440, Cambridge, MA, 2008. MIT Press.

[33] 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):699–746, 2008.

[34] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, E93-D(3), 2010. to appear.

[35] T. Suzuki and M. Sugiyama. Estimating squared-loss mutual information for independent component analysis. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation*, volume 5441 of *Lecture Notes in Computer Science*, pages 130–137, Berlin, 2009. Springer.

[36] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. Technical Report TR09-0005, Department of Computer Science, Tokyo Institute of Technology, Feb. 2009.

[37] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

[38] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. Van de Peer, editors, *JMLR Workshop and Conference Proceedings*, volume 4 of *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pages 5–20, 2008.

[39] T. Suzuki, M. Sugiyama, and T. Tanaka. Mutual information approximation via maximum likelihood estimation of density ratio. In *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)*, pages 463–467, Seoul, Korea, Jun. 28–Jul. 3 2009.

[40] M. Takimoto, M. Matsugu, and M. Sugiyama. Visual inspection of precision instruments by least-squares outlier detection. In *Proceedings of The Fourth International Workshop on Data-Mining and Statistical Science (DMSS2009)*, pages 22–26, Kyoto, Japan, Jul. 7–8 2009.

[41] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[42] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[43] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 903–910, New York, NY, 2004. ACM Press.